# Modeling Covariates with Nonparametric Bayesian Methods [1]

Alejandro Cruz-Marcelo [2 6] Gary L. Rosner [3]

Peter Müller [4] Clinton F. Stewart [5]

February 2010

**Abstract**

A research problem that has received increased attention in recent years is extending Bayesian nonparametric methods to include dependence on covariates. Limited attention, however, has been directed to the following two aspects. First, analyzing how the performance of such extensions differs, and second, understanding which features are worthwhile in order to produce better results. This article proposes answers to those questions focusing on predictive inference and continuous covariates. Specifically, we show that 1) nonparametric models using

[1]This research was supported by the National Cancer Institute grant R01CA075981; The Brown Foundation Fellowship, Center for Computational Finance and Economic Systems; and the NSF VIGRE grant DSM-0739420.

[2]Department of Statistics, Rice University, Houston, TX 77030.

[3]Division of Oncology Biostatistics & Bioinformatics, Sidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD 21205.

[4]Department of Biostatistics, M. D. Anderson Cancer Center, University of Texas, Houston, TX 77030.

[5]Pharmaceutical Sciences, St. Jude Children's Research Hospital, Memphis, TN 38105.

[6]Corresponding author: ac6@rice.edu

different strategies for modeling continuous covariates can show noteworthy differences when they are being used for prediction, even though they produce otherwise similar posterior inference results, and 2) when the predictive density is a mixture, it is convenient to make the weights depend on the covariates in order to produce sensible estimators. Such claims are supported by comparing the Linear DDP (an extension of the Sethuraman representation) and the Conditional DP (which augments the nonparametric distribution to include the covariates). Unlike the Conditional DP, the weights in the predictive mixture density of the Linear DDP are not covariate-dependent. This results in poor estimators of the predictive density. Specifically, in a simulation example, the Linear DDP wrongly introduces an additional mode into the predictive density, while in an application to a pharmacokinetic study, it produces unrealistic concentration-time curves.

**KEY WORDS:** Dirichlet process mixture; Hierarchical model;Nonparametric Bayes; Covariates modeling; Dependent Dirichlet process.

# 1 Introduction

A modeling framework that has become increasingly popular is the use of Bayesian nonparametric methods (Müller et al. 2004). In particular, the Dirichlet process (DP) (Ferguson 1973) is the most popular prior model for an unknown random measure. The popularity of the DP is due to its elegance, simplicity and the existence of computationally efficient Markov chain Monte Carlo (MCMC) posterior simulation algorithms (MacEachern and Müller 1998). In addition, hierarchical models based on random effects distributions with DP priors are able to accommodate outliers, multimodality, and dependence

in multivariate, longitudinal, and functional data (Dunson 2009). Examples of nonparametric methods based on the DP include applications in pharmacokinetics (Rosner and Müller 1997), econometrics (Griffin and Steel 2004), spatial modelling (Gelfand et al. 2005), meta-analysis (Burr and Doss 2005), variable selection (Kim et al. 2006), genetics (Xing et al. 2007), density estimation (Dunson et al. 2007; Rodriguez and ter Horst 2008), and survival analysis (De Iorio et al. 2009).

The DP as described by Ferguson (1973) does not incorporate covariates. Hence, an active area of research is extending nonparametric models to allow the unknown distribution to depend on covariates. A popular approach uses as a starting point the Sethuraman (1994) representation of a DP. This representation states that a random measure $G$ that follows a DP with total mass parameter $M$ and base measure $G_0$, denoted as $DP(M, G_0)$, can be represented as

$$G = \sum_{h=1}^{\infty} w_h \delta(\theta_h), \quad w_h = v_h \prod_{i=1}^{h-1}(1 - v_i), \quad v_h \sim Beta(1, M), \quad \theta_h \sim G_0, \quad (1)$$

where $\delta(\theta)$ is a point mass at $\theta$. Generalizations of (1) achieve the desired dependence of $G$ on covariates by making the weights, $w_h$, and/or the locations, $\theta_h$, vary with the covariates according to a stochastic process (MacEachern 1999). Such extensions of the Sethuraman representation are probability models on a collection of dependent random probability measures $\{G_x, x \in X\}$, where $X$ is the corresponding covariate space. Generalizations of the Sethuraman representation that introduce covariates via the locations have been applied to the analysis of variance (De Iorio et al. 2004), spatial modeling (Gelfand et al. 2005), time series (Caron et al. 2006), and regression (De Io-

rio et al. 2009). On the other hand, there are also generalizations based on making the weights covariate-dependent. Griffin and Steel (2006) proposed an order-based dependent Dirichlet process that makes the order of $v_h$ in the stick-breaking construction be a function of the covariates, Dunson and Park (2008) express $v_h$ as a covariate-dependent kernel multiplied by beta weights, and Fuentes-García et al. (2009) models $w_h$ with a nonparametric mixture model that depends on covariates.

Approaches for modeling covariates not based on the Sethuraman representation exist in the literature as well. If the covariates only take a finite number of values, then the product of Dirichlet processes described in Cifarelli and Regazzini (1978) can be used to introduce dependence on covariates. Specifically, a covariate-dependent regression model is used as the base measure of independent Dirichlet processes at each level of the covariates (Carota and Parmigiani 2002; Griffin and Steel 2004). Another approach for modeling dependence is forming convex combinations of independent DP (Dunson et al. 2007; Müller et al. 2004). Finally, Müller et al. (1996) and Müller and Rosner (1998) include covariates $x_i$ in an augmented response vector $(y_i, x_i)$ and obtain the desired dependence by focusing on the conditional distribution given $x_i$. We refer to this method as Conditional DP.

Although modeling dependence on covariates has been a very active area of research, limited research has examined the relative performance of such methods or improved understanding of which features are suitable in order to produce better results. This article considers such a comparison, focusing on predictive inference and continuous covariates. We show that different approaches for modeling dependence on continuous covariates can lead to very similar posterior fits and yet produce very different results when used for pre-

diction. In addition, when the predictive density is a mixture, this paper shows that making the weights depend on the covariates plays a major role in determining the quality of the predictions. Such findings are illustrated by comparing the Linear DDP (De Iorio et al. 2009) to the Conditional-DP (Müller et al. 1996); we apply those methods to a simulated data set and to data from a pharmacokinetic meta-analysis. Section 2 describes and compares both methods. Implementation and empirical results are reported in Section 3. Finally, conclusion and discussion appear in Section 4.

# 2 Modeling Approaches

This section describes the Linear DDP and the Conditional DP, points out their differences regarding the form of the predictive density, and explains how such differences affect the performance of each method.

Both methods introduce continuous covariates into the typical DP mixture (DPM) model given by

$$\boldsymbol{y}_i \overset{iid}{\sim} \int f\left(\boldsymbol{y}|\boldsymbol{\mu}\right) dG(\boldsymbol{\mu}), \qquad G \sim DP(M, G_0), \tag{2}$$

where $f$ is a probability density. The DPM model (2) is a mixture model with a DP prior on the mixing measure $G$. In many practical applications, the kernel $f\left(\boldsymbol{y}|\boldsymbol{\mu}\right)$ is set to be a normal multivariate density with mean $\boldsymbol{\mu}$ and common covariance matrix $\mathbf{S}$.

## 2.1 Linear DDP

The Linear DDP introduced in De Iorio et al. (2009) models the relationship between continuous covariates and the unknown distribution by replacing the random probability measure $G$ in the DPM model (2) with a collection of random probability measures indexed by $\boldsymbol{x}$, $\{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\}$, where $\boldsymbol{x} = (x_1, \ldots, x_d)$ denotes a $d$-dimensional vector of continuous covariates and $X$ is the corresponding covariate space.

De Iorio et al. (2009) set a prior on $\{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\}$ using as a starting point the Dependent DP (DDP), as defined in MacEachern (1999). The DDP specifies

$$G_{\boldsymbol{x}} = \sum_{h=1}^{\infty} w_h \delta(\boldsymbol{\theta}_{\boldsymbol{x}h}), \qquad \text{for any } \boldsymbol{x} \ . \tag{3}$$

The point masses $\boldsymbol{\theta}_{\boldsymbol{x}h}$ satisfy the condition that $\theta_h = \{\boldsymbol{\theta}_{\boldsymbol{x}h}, \boldsymbol{x} \in X\}$ are iid realizations of a stochastic process in $\boldsymbol{x}$. The weights, $w_h$, follow a stick-breaking prior as in Sethuraman (1994), that is, $w_h = v_h \prod_{i=1}^{h-1}(1 - v_i)$, $v_h \sim Beta(1, M)$. The DDP model (3) implies that, for each $x$, $G_x$ follows a DP. Specifically, $G_{\boldsymbol{x}} \sim DP(M, G_{0\boldsymbol{x}})$, where the base measure $G_{0\boldsymbol{x}}$ is the marginal distribution at $\boldsymbol{x}$ of the stochastic process on the point masses $\boldsymbol{\theta}_{\boldsymbol{x}h}$.

In general, the DDP model induces dependence of the random measures $G_{\boldsymbol{x}}$ by assuming that the sample paths $\theta_h$ are dependent across $\boldsymbol{x}$. De Iorio et al. (2009), in particular, impose a linear model on $\boldsymbol{\theta}_{\boldsymbol{x}h}$ given by

$$\boldsymbol{\theta}_{\boldsymbol{x}h} = \boldsymbol{m}_h + \sum_{i=1}^{d} \boldsymbol{\beta}_{ih} x_i, \tag{4}$$

where $d$ is the number of continuous covariates, $\boldsymbol{m}_h \overset{iid}{\sim} p^0_{\boldsymbol{m}}$ and $\boldsymbol{\beta}_{ih} \overset{iid}{\sim} p^0_{\boldsymbol{\beta}_i}$, $i = 1 \ldots d$. It follows that the base measure, $G_{0\boldsymbol{x}}$, is given by the convolution

of $p^0_{\boldsymbol{m}}$ and $p^0_{\boldsymbol{\beta}_i}$, $i = 1 \dots d$. The distributional assumptions on $\boldsymbol{\theta}_{\boldsymbol{x}h}$ proposed by De Iorio et al. (2009) imply that the random measures $G_{0\boldsymbol{x}}$ share the common main effect given by $\boldsymbol{m}_h$. In addition, for each $i$, $\boldsymbol{\beta}_i$ represents a slope coefficient as in a standard linear model.

The prior given in (3) with the linear model on the locations as in (4) is the Linear DDP (De Iorio et al. 2009). When such a prior is used to introduce continuous covariates into the DPM model (2) with a normal kernel, it leads to models given by

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i = \boldsymbol{x}) &\overset{iid}{\sim} \int N(\boldsymbol{y}; \boldsymbol{\mu}, \mathbf{S})\, dG_{\boldsymbol{x}}(\boldsymbol{\mu}), \\ \{G_{\boldsymbol{x}}, \boldsymbol{x} \in X\} &\sim \text{Linear DDP}(M, G_{0X}), \end{aligned} \tag{5}$$

where $G_{0X}$ denotes the set of distributions for the main effect and slope coefficients in (4). The model is completed with appropriate hyperpriors for $S$, $M$, and $G_{0X}$.

It is possible to rewrite model (5) in terms of a mixture of linear models. Specifically, we define the random matrix $\tilde{\boldsymbol{\Gamma}}_h = \left[\boldsymbol{m}_h, \boldsymbol{\beta}_{1h}, \dots, \boldsymbol{\beta}_{ph}\right]$ and design vectors $\boldsymbol{d}_i = (1, \boldsymbol{x}_i)$, $i = 1, \dots, n$, such that $\boldsymbol{\theta}_{\boldsymbol{x}_i h} = \tilde{\boldsymbol{\Gamma}}_h \boldsymbol{d}_i$ , where $\boldsymbol{x}_i$ denote the continuous covariate vector of subject $i$. Model (5) can then be rewritten as

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i) &\overset{iid}{\sim} \int N\left(\boldsymbol{y}; \tilde{\boldsymbol{\Gamma}}\boldsymbol{d}_i, \mathbf{S}\right) dG(\tilde{\boldsymbol{\Gamma}}), \\ G &\sim DP(M, G_0), \end{aligned} \tag{6}$$

with base measure $G_0 = \left(p^0_{\boldsymbol{m}}, p^0_{\boldsymbol{\beta}_1}, \dots, p^0_{\boldsymbol{\beta}_p}\right)$. The reformulation (6) is convenient because it has the form of a DPM model (2). Hence, it is possible to carry out posterior inference in the Linear DDP model by using well-known MCMC algorithms designed for DPM models, such as those described in MacEachern

and Müller (1998).

For later reference we replace the mixture in (6) by an additional level in the hierarchical model

$$\begin{aligned} (\boldsymbol{y}_i|\boldsymbol{x}_i, \boldsymbol{\Gamma}_i, S) &\overset{iid}{\sim} N\left(\boldsymbol{y}; \boldsymbol{\Gamma}_i \boldsymbol{d}_i, \mathbf{S}\right), \\ \boldsymbol{\Gamma}_i \sim G, \quad \text{and} \quad G &\sim DP(M, G_0). \end{aligned} \tag{6a}$$

Since a DP is almost surely discrete, there is a positive probability for ties among the $\boldsymbol{\Gamma}_i$. Let $\{\boldsymbol{\Gamma}_1^\star \ldots \boldsymbol{\Gamma}_k^\star\}$, $k \leq n$, be the set of unique values in $\{\boldsymbol{\Gamma}_1 \ldots \boldsymbol{\Gamma}_n\}$, and let $n_j$ denote the number of $\boldsymbol{\Gamma}_i$ equal to $\boldsymbol{\Gamma}_j^\star$.

We conclude the description of the Linear DDP by deriving the predictive density of model (5). Let $\Theta$ denote the set of all model parameters and let $\Theta^{(t)}$, $t = 1, \ldots, n$ denote a posterior Monte Carlo sample. We will generically use the superscript $(t)$ to indicate elements of $\Theta^{(t)}$. For a new subject with vector of covariates $\boldsymbol{x}_{n+1}$, the predictive density can be approximated as follows:

$$\begin{aligned} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y\right) &= E\left[p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y, \Theta\right)|Y\right] \\ &\approx \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y, \Theta^{(t)}\right) \\ &= \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right), \end{aligned}$$

where $Y$ denotes the current data. The specific expressions for the probabilities in the last average are easily obtained from (6a).

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) = \sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)}\, p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) \tag{7}$$

with

$$\alpha_j^{(t)} \propto \begin{cases} n_j^{(t)} & j = 1 \dots k^{(t)}, \\ M^{(t)} & j = k^{(t)} + 1, \end{cases}$$

where $\sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)} = 1$, and

$$p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right) = \begin{cases} N\left(\boldsymbol{y}_{n+1}; \boldsymbol{\Gamma}_j^{\star(t)}\boldsymbol{d}_{n+1}, \boldsymbol{S}^{(t)}\right) & j = 1 \dots k^{(t)}, \\ \int N\left(\boldsymbol{y}_{n+1}; \boldsymbol{\Gamma}\boldsymbol{d}_{n+1}, \boldsymbol{S}^{(t)}\right) dG_0(\boldsymbol{\Gamma}) & j = k^{(t)} + 1. \end{cases}$$

## 2.2 Conditional DP

The Conditional DP approach introduces regression into the DPM model (2) by including the continuous covariates in an augmented response vector $\tilde{\boldsymbol{y}} = (\boldsymbol{y}, \boldsymbol{x})$ in the nonparametric model. Specifically, Müller and Rosner (1998) make the unknown distribution $p(\boldsymbol{y})$ depend on covariates $\boldsymbol{x}$ by defining a DP mixture model for the joint model $p(\boldsymbol{y}, \boldsymbol{x})$. Considering a normal kernel, the Conditional DP modifies the DPM model as follows:

$$(\boldsymbol{y}_i, \boldsymbol{x}_i) \overset{iid}{\sim} \int N\left((\boldsymbol{y}, \boldsymbol{x}); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu}), \qquad G \sim DP(M, G_0), \tag{8}$$

which is equivalent to the hierarchical model

$$\begin{aligned} (\boldsymbol{y}_i, \boldsymbol{x}_i) &\overset{iid}{\sim} N\left(\boldsymbol{y}; \boldsymbol{\mu}_i, \mathbf{S}\right), \\ \boldsymbol{\mu}_i \sim G, \quad \text{and} \quad & G \sim DP(M, G_0). \end{aligned} \tag{8a}$$

We refer to this modeling approach as Conditional DP, because the implied conditional distribution $p(\boldsymbol{y}|\boldsymbol{x})$ formalizes the desired regression on $\boldsymbol{x}$. Particularly, as explained in Müller and Rosner (1998), the mixture of normals $\int N\left((\boldsymbol{y}, \boldsymbol{x}); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu})$ implies a locally weighted mixture of normal linear

regressions for $E\left[\boldsymbol{y}_i|\boldsymbol{x}_i\right]$. Implementation of posterior inference for the Conditional DP model (8) is straightforward, because it has the form of a DPM model. Hence, the MCMC algorithms described in MacEachern and Müller (1998) can be used.

The predictive density for a new $(n+1)$-th subject, conditional on covariates $x_{n+1}$, can be estimated using

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, Y\right) \approx \frac{1}{N}\sum_{t=1}^{N} p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1}, \Theta^{(t)}\right). \tag{9}$$

The conditional density in the last average can be derived from the density $p\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ corresponding to the DPM model (8a) as follows. Let $\{\boldsymbol{\mu}_1^\star \ldots \boldsymbol{\mu}_k^\star\}$, $k \leq n$, be the set of distinct vectors from the set $\{\boldsymbol{\mu}_1 \ldots \boldsymbol{\mu}_n\}$, with $n_j$ the number of $\boldsymbol{\mu}_i$ equal to $\boldsymbol{\mu}_j^\star$. The predictive density for $\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}\right)$ is

$$p\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) = \sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)}\, p_j\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) \tag{10}$$

with

$$\alpha_j^{(t)} \propto \begin{cases} n_j^{(t)} & j = 1 \ldots k^{(t)}, \\ M^{(t)} & j = k^{(t)} + 1, \end{cases}$$

where $\sum_{j=1}^{k^{(t)}+1} \alpha_j^{(t)} = 1$, and

$$p_j\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}|\Theta^{(t)}\right) = \begin{cases} N\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}; \boldsymbol{\mu}_j^{\star(t)}, \boldsymbol{S}^{(t)}\right) & j = 1 \ldots k^{(t)}, \\ \int N\left(\boldsymbol{y}_{n+1}, \boldsymbol{x}_{n+1}; \boldsymbol{\mu}, \boldsymbol{S}^{(t)}\right) dG_0(\boldsymbol{\mu}) & j = k^{(t)} + 1. \end{cases}$$

It follows that the conditional density needed in (9) is given by

$$p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1},\Theta^{(t)}\right)=\sum_{j=1}^{k^{(t)}+1}\left(\frac{\alpha_j^{(t)}\;p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)}{\sum_{\ell=1}^{k^{(t)}+1}\alpha_\ell^{(t)}\;p_\ell\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)}\right)p_j\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1},\Theta^{(t)}\right), \tag{11}$$

where $p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ is obtained by integrating out $\boldsymbol{y}_{n+1}$ from the joint distribution $p_j\left(\boldsymbol{y}_{n+1},\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$ in the mixture (10).

## 2.3 Comparing Approaches

Both extensions of the DPM model, the Linear DDP and the Conditional DP, estimate the predictive density by averaging the mixture distribution $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1},\Theta^{(t)}\right)$ with respect to a Monte Carlo sample from the posterior distribution. We claim, however, that the specific factors determining the weights in such mixtures can affect the resulting predictive inference.

As shown in (7), the weights in the mixture density $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1},\Theta^{(t)}\right)$ corresponding to the Linear DDP are solely determined by $n_j^{(t)}$ and $M^{(t)}$. That is, the relative importance of each component in the mixture is mainly determined by the size of the "clusters" induced by the unique values of $\Gamma_i^{(t)}$. This feature is an important limitation of the Linear DDP. It implies that, for each $t$, the weights of the mixture remain the same for any new subject rather than change as a function of the specific covariates $x_{n+1}$. In other words, there is no built-in mechanism in the predictive density to favor those components in the mixture that are more likely to provide a better fit to the specific characteristics of the new subject.

Unlike the formulas in the Linear DDP, the weights in the mixture $p\left(\boldsymbol{y}_{n+1}|\boldsymbol{x}_{n+1},\Theta^{(t)}\right)$ corresponding to the Conditional DP are a function of the covariates via the marginal density $p_j\left(\boldsymbol{x}_{n+1}|\Theta^{(t)}\right)$, as shown in (11). It follows that components

in the mixture with a high marginal density on $\boldsymbol{x}_{n+1}$ will tend to have higher weights. Hence, the mixture is adjusted to reflect the specific characteristics of a given new subject. The inclusion of covariates in the weights increases the chance of using a regression structure suitable for the given subject.

In summary, the differences mentioned above suggest that, in terms of predictive inference, the Conditional DP should outperform the Linear DDP. We present empirical evidence corroborating such a claim in the next Section.

Finally, we comment on a weakness of the Conditional DP that somewhat offsets the discussed features. The sampling model in (7) can be factored as $p(y_i|x_i) * p(x_i)$, highlighting the fact that the likelihood includes an additional factor for the covariates $x_i$. This is technically inappropriate when the covariates are chosen and fixed by design.

# 3 Empirical Implementation

This section provides evidence to illustrate the superior performance of the Conditional DP over the Linear DDP for predictive inference. Two data sets are considered; the first one is a simulated data set, while the second derives from a population pharmacokinetic study.

## 3.1 Simulation Example

The simulated data set corresponds to a multiple regression model with two dependent variables and a single predictor variable. Specifically, the data are generated from a model similar to (6) as follows. Let $\boldsymbol{\Gamma} = [\boldsymbol{m}, \boldsymbol{\beta}]$ be a $2 \times 2$ matrix where $\boldsymbol{m}$ is a vector of constants and $\boldsymbol{\beta}$ is a vector of slope coefficients. Our simulated data set includes a sample of bivariate observations $\boldsymbol{y}_i = (y_{i1}, y_{i2})^{\mathrm{T}}$,

$i = 1, \ldots, 100$, generated from the bivariate distribution $N(\boldsymbol{\Gamma}_i \boldsymbol{d}_i s\boldsymbol{I})$, where $\boldsymbol{I}$ is the $2 \times 2$ identity matrix, $s = 0.10$, and $\boldsymbol{d}_i = (1, x_i)^{\mathrm{T}}$ is a design vector including the subject-specific covariate $x_i$. We introduce two underlying regression structures into the simulated data set by randomly setting $\boldsymbol{\Gamma}_i$ equal to one of the following:

$$\begin{cases} \boldsymbol{\Lambda}_1 = \begin{bmatrix} 0 & 3 \\ 1 & 0 \end{bmatrix}, & \text{w.p. } 1/2 \\ \\ \boldsymbol{\Lambda}_2 = \begin{bmatrix} 0 & 0 \\ 1 & 3 \end{bmatrix}, & \text{w.p. } 1/2 \end{cases} \tag{12}$$

Finally, each $x_i$ is generated from a uniform distribution. If $\Gamma_i = \Lambda_1$ then $x_i \sim \mathrm{U}(0, 1)$, otherwise $x_i \sim \mathrm{U}(-1, 0)$. Hence, the value of the covariate $x_i$ provides information about the specific underlying regression structure. As shown later, the ability to incorporate such information is a crucial difference between the Linear DDP and the Conditional DP.

The resulting simulated data set reflects two regression structures given by the product $\boldsymbol{\Gamma}_i \boldsymbol{d}_i$; one describes the horizontal line segment $(0, 1)^{\mathrm{T}} + x(3, 0)^{\mathrm{T}}$, $x \in (0, 1)$, while the other follows the vertical line segment $(0, 1)^{\mathrm{T}} + x(0, 3)^{\mathrm{T}}$, $x \in (-1, 0)$ (see Figure 1). The simulated data are generated from a model that resembles (6). Thus, the data set matches the modeling assumptions of the Linear DDP. Such a feature was chosen in order to rule out the characteristics of the simulated data as an explanation for the poor performance of the Linear DDP.

We used posterior MCMC simulation to generate posterior Monte Carlo

samples under Linear DDP (6) and the Conditional DP (8), respectively. In both cases the uncertainty on the common matrix $\boldsymbol{S}$ is modeled by adopting the conjugate inverse Wishart prior $\boldsymbol{S}^{-1} \sim Wishart(r, (r\boldsymbol{R})^{-1})$ with $r$ degrees of freedom and mean $r(r\boldsymbol{R})^{-1} = \boldsymbol{R}^{-1}$. In addition, the base measure is assumed to be multivariate normal, $G_0 \sim N(\boldsymbol{b}, \boldsymbol{B})$, and M is given a gamma distribution, $M \sim Ga(a_m, b_m)$. In both cases, we considered 10,000 iterations of the MCMC algorithm; convergence of such algorithms was reached after 2000 iterations. In general, the posterior fit under the Linear DDP is slightly better than the one obtained with the Conditional DP. For example, we estimated the first, second, and third quartiles, (Q1,Q2,Q3), of the distribution of the Euclidean distance between the posterior mean of $\boldsymbol{y}_i$ and its sample value. Those statistics were equal to $(0.07, 0.12, 0.18)$ for the Linear DDP, while the Conditional DP showed somewhat greater values, $(0.09, 0.16, 0.32)$.

Substituting the posterior Monte Carlo samples in (7) and (11), we evaluated the posterior predictive density corresponding to a new subject with covariate $x \in (-1, 1)\backslash\{0\}$. The results for $x = 0.5$ and $x = -0.5$ highlight the shortcoming of the Linear DDP. It wrongly assigns positive probability to regions in the plane without any sample points (see Figure 1). Such results can be explained as follows. Since the Linear DDP correctly recognizes the two regression structures, the mixture in (7) is dominated by two components, each one of them corresponding to one of the regression structures. It follows that for any given covariate $x$, the mixture in (7) gives positive probability to two regions, one for each regression structure. In contrast, the Conditional DP incorporates the values of the covariates into the weights of the mixture (11). Therefore, the Conditional DP is able to identify those regression structures that are more likely, given the value of $x$.

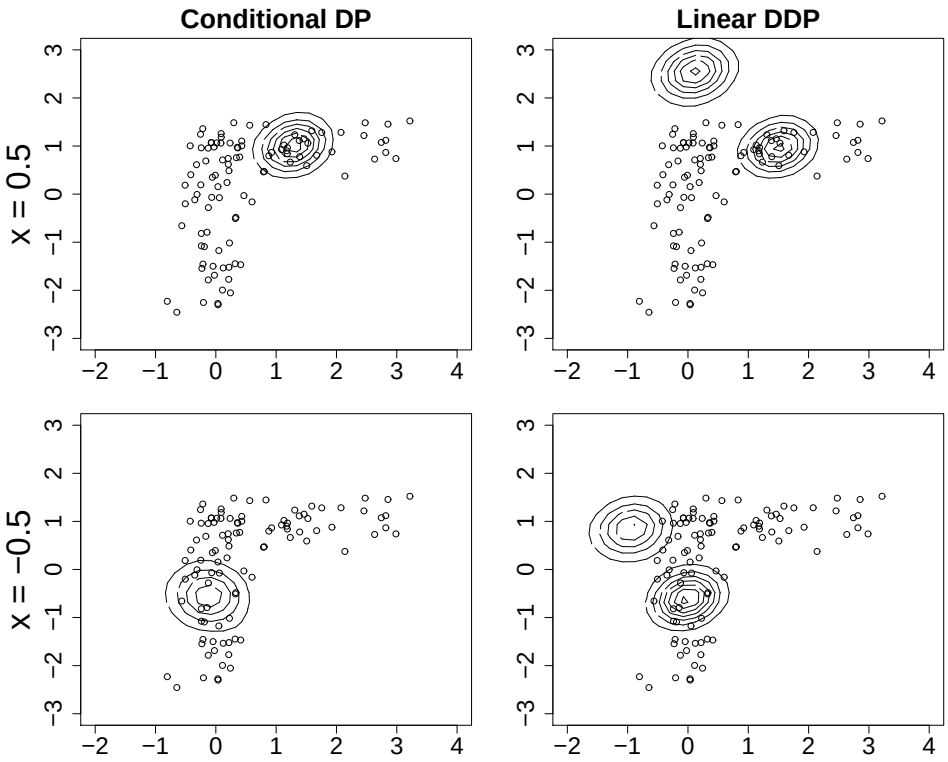



Figure 1: Contours of the predictive density. The elements of the simulated data set are marked with dots. The plots are organized by modeling approach (columns) and by covariate level used to perform predictive inference (rows). The estimators produced by the Linear DDP are not acceptable because they wrongly introduce an extra mode to the predictive density. Such results reflect the fact that the Linear DDP selects a regression structure based solely on the cluster size, which by design is approximately half for each one of the two underlying regression structures, rather than using the covariate level of the new subject.

## 3.2 Pharmacokinetics Example

We consider a data set that includes clinical covariates and longitudinal data consisting of drug concentration measurements. The measurements came from patients receiving the anti-cancer drug topotecan in several different studies. We fit a Bayesian population pharmacokinetic (PK) model to estimate and predict plasma concentration-time curves. This section compares the performance of the Linear DDP and Conditional DP when introducing dependence on patient covariates into the PK model. The inclusion of covariates enable us to predict individual concentration-time curves appropriate for an individual patient instead of the group average. Such estimated curves provide valuable information that can be used, for example, to design dose individualization schemes for future patients who are starting topotecan treatment.

Before describing the Bayesian PK model, here are some details about the the data set. The population consisted of 138 children enrolled in seven clinical studies (see Table 1). These data have previously been used by Schaiquevich et al. (2007) to characterize the population pharmacokinetics of topotecan lactone in children with cancer and to identify covariates related to topotecan disposition. The data include concentration-time measurements corresponding to several treatment occasions for each patient. We use only those corresponding to the first treatment in our sample. This is because we are interested in making inference (and prediction) for patients with no previous topotecan treatment. In all cases, topotecan was administered via IVAC Controller (IVAC corp.) with duration of infusion of 30 minutes. See Schaiquevich et al. (2007) and references therein for more details regarding the eligibility criteria for the clinical studies, drugs administration, blood collection, and patient demographics.

Table 1: Characteristics of the clinical trials from which PK data was obtained

| Trial no. | No. of patients | Trial type | Topotecan dosage |
|---|---|---|---|
| 1 | 15 | Phase I recurrent solid tumors | Target AUC = 120-180 ng · h/mL |
| 2 | 21 | Phase I recurrent acute leukemia | Fixed dosage = 2.4 mg/m$^2$ |
| 3 | 28 | Phase I recurrent solid tumors | 0.8 and 1.1 mb/m$^2$ |
| 4 | 10 | Phase II newly diagnosed medulloblastoma | Target AUC = 120-160 ng · h/mL |
| 5 | 22 | Phase I recurrent solid tumors | 1.4, 1.7, 2.0, and 2.4 mg/m$^2$ |
| 6 | 30 | Phase II newly diagnosed high-risk medulloblastoma | Target AUC = 80-120 ng · h/mL |
| 7 | 12 | Phase II recurrent Wilms tumor | Target AUC = 70-90 ng · h/mL |

The structure of a Bayesian population PK model is as follows. Let $y_{ij}$ denote the $j$-th measurement for the $i$th patient, and $\boldsymbol{\theta}_i$ the vector of random effects of patient $i$. The vector $\boldsymbol{x}_i$ represents the patient-specific covariates. The probability model for the PK data is given by

$$p(y_{ij}|\boldsymbol{\theta}_i), \qquad p(\boldsymbol{\theta}_i|\phi), \qquad p(\phi). \tag{13}$$

Here, $p(y_{ij}|\boldsymbol{\theta}_i)$ is a parametric, and typically non-linear, regression model for the concentration-time curve, $p(\boldsymbol{\theta}_i|\phi)$ is the prior for $\boldsymbol{\theta}_i$, and, finally, $p(\phi)$ denotes the probability model of the hyperparameters. Bayesian models similar to (13) have been considered in Zeger and Karim (1991) for generalized linear mixed models and in Wakefield (1994) using a multivariate normal population distribution as population model.

Both methods, the Linear DDP and the Conditional DP, can be used to

introduce covariates into the PK model (13) via the prior distribution on $\boldsymbol{\theta}_i$. The resulting prior has the form (5) or (8), respectively, where $\boldsymbol{y}_i$ is replaced with $\boldsymbol{\theta}_i$. That is, using the Linear DDP model, the prior on $(\boldsymbol{\theta}_i|\boldsymbol{x}_i)$ is a mixture of normals with a mixing measure given by a family of random measures indexed by $\boldsymbol{x}$. If the Conditional DP is used, the vector of parameters is augmented to include the covariates. Hence, the prior $p(\boldsymbol{\theta}_i, \boldsymbol{x}_i)$ is a mixture of normals

$$p\left(\boldsymbol{\theta}_i, \boldsymbol{x}_i\right) = \int N\left(\left(\boldsymbol{\theta}_i, \boldsymbol{x}_i\right); \boldsymbol{\mu}, \mathbf{S}\right) dG(\boldsymbol{\mu}),$$

with a DP prior on the mixing measure $G$. The model achieves the desired nonlinear, semi-parametric regression of the parameters on the covariates via the conditional distribution

$$p(\boldsymbol{\theta}_i|\boldsymbol{x}_i) = \frac{p(\boldsymbol{\theta}_i, \boldsymbol{x}_i)}{\int p(\boldsymbol{\theta}_i, \boldsymbol{x}_i) d\boldsymbol{\theta}_i} \propto p(\boldsymbol{\theta}_i, \boldsymbol{x}_i).$$

Both modeling approaches assign a flexible non-parametric prior distribution to the population parameters. Hence, they both are able to accommodate heterogeneity in the patient population, such as outliers, over-dispersion, and multimodality.

Since we want to emphasize the differences between the Linear DDP and the Conditional DP, similar criteria are used to specify $p(y_{ij}|\boldsymbol{\theta}_i)$ and $p(\phi)$, regardless of the prior on $\boldsymbol{\theta}_i$ being considered.

The model for the concentration-time curve, $p(y_{ij}|\boldsymbol{\theta}_i)$, is determined by the nonlinear regression

$$\log(y_{ij}) = \log(f(\boldsymbol{\theta}_i, \tau_{ij})) + \epsilon_{ij},$$

where $y_{ij}$ is the $j$th concentration measurement for the $i$th patient at time $\tau_{ij}$, $\epsilon_{ij} \sim N(0, \lambda^{-1})$ is the noise term with precision $\lambda$, and the function $f$ is a two-compartment model with constant rate intravenous infusion (Wagner 1968) that describes the concentration at time $\tau$ as

$$f(\boldsymbol{\theta}, \tau) = \frac{D/\gamma}{V_1 K_2} \left\{ 1 - \left[ \left( \frac{K_2 - \alpha}{\beta - \alpha} \right) e^{-\beta\tau + \beta(\tau-\gamma)\mathbf{1}_{(\tau \geq \gamma)}} - \left( \frac{K_2 - \beta}{\beta - \alpha} \right) e^{-\alpha\tau + \alpha(\tau-\gamma)\mathbf{1}_{(\tau \geq \gamma)}} \right] \right\} \times \left[ \left( \frac{K_2 - \alpha}{\beta - \alpha} \right) e^{-\beta(\tau-\gamma)} - \left( \frac{K_2 - \beta}{\beta - \alpha} \right) e^{-\alpha(\tau-\gamma)} \right]^{\mathbf{1}_{(\tau \geq \gamma)}} \tag{14}$$

with

$$\begin{aligned} \alpha &= \frac{1}{2} \left[ (K_1 + K_2 + K_{-1}) + \sqrt{(K_1 + K_2 + K_{-1})^2 - 4K_{-1}K_2} \right], \\ \beta &= \frac{1}{2} \left[ (K_1 + K_2 + K_{-1}) - \sqrt{(K_1 + K_2 + K_{-1})^2 - 4K_{-1}K_2} \right]. \end{aligned}$$

Here, the four parameters $\{V_1, K_2, K_1, K_{-1}\}$ are all positive, $\gamma$ is the duration of infusion, $D$ is the dose, and $\mathbf{1}_{(.)}$ denotes an indicator function. Finally, we parameterize (14) with

$$\boldsymbol{\theta} = (log(V_1), log(K_2), log(K_1), log(K_{-1}))^{\mathrm{T}}.$$

Such a parameterization guarantees that the subject-specific parameters can take any value, and thus the use of a mixture of normals as a prior is appropriate.

Finally, we introduce distributional assumptions for the hyperparameters, $\phi$, that allow the implementation of a MCMC algorithm that is computationally efficient. In the discussion that follows, it is assumed that the Linear

DDP is written in the equivalent form (6). Regarding the parameters of the DP, the total mass parameter $M$ is given a gamma prior $Ga(a_m, b_m)$, while the base measure $G_0$ follows a multivariate normal distribution $N(\boldsymbol{b}, \boldsymbol{B})$. The moments of $G_0$ are assumed random with hyperpriors $\boldsymbol{b} \sim N(\boldsymbol{b_0}, \boldsymbol{B_0})$ and $\boldsymbol{B}^{-1} \sim Wishart\,(w, (w\boldsymbol{W})^{-1})$, where $w$ is the degrees of freedom and $\boldsymbol{W}^{-1}$ is the mean of $\boldsymbol{B}^{-1}$. Finally, a Wishart prior is also used for $\boldsymbol{S}$, $\boldsymbol{S}^{-1} \sim Wishart(r, (r\boldsymbol{R})^{-1})$, and $\lambda$ is given a gamma prior $G(a_\lambda, b_\lambda)$.

Unlike $p(y_{ij}|\boldsymbol{\theta}_i)$, the characteristics of $p(\phi)$ change in accordance with the prior being used for $\boldsymbol{\theta}$. Specifically, the dimension of $G_0$ changes as follows. Let $p$ denote the number of model parameters; that is, $p$ equals the dimension of $\boldsymbol{\theta}$. Let $d$ be the number of covariates being modeled. For the Conditional DP, $G_0$ is a distribution on a vector with dimension $p + d$, while under the Linear DDP, $G_0$ is a distribution on the columns of the matrix $\boldsymbol{\Gamma}$. In the latter case, it is useful to think of $G_0$ as the distribution on the vector with dimension $p\,(d+1)$ which results from stacking the columns of $\boldsymbol{\Gamma}$ one on top of the other.

Putting together the assumptions described above for each component in (13), it follows that two population PK models have been completely specified, each one implementing a different prior on $\boldsymbol{\theta}$. For brevity, we will refer to those PK models by using only the name of the modeling approach used for the prior on $\boldsymbol{\theta}$, that is, Conditional DP or Linear DDP. The analysis that follows includes the covariates age, body surface area (BSA), and glomerular filtration rate (GFR). Such covariates were chosen because, as shown in Schaiquevich et al. (2007), they are significant for explaining topotecan disposition. The observed measurements for each covariate were centered at zero.

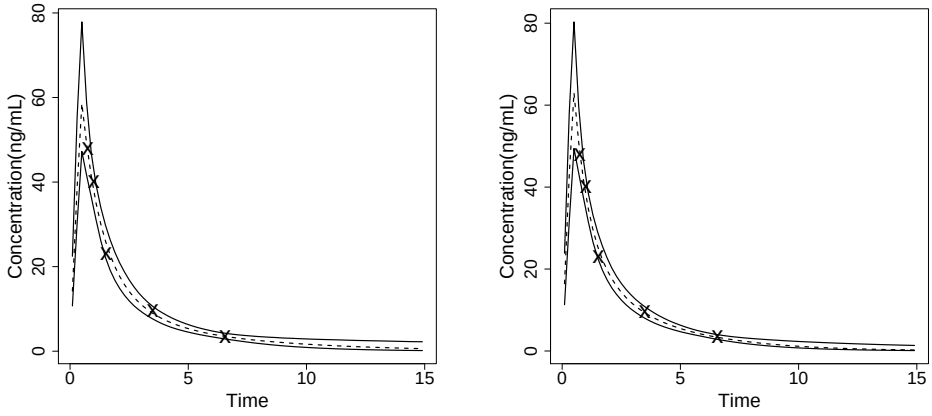
Implementation of both PK models requires a MCMC scheme to sample from the corresponding posterior distribution. Such sampling schemes can

be efficiently implemented, since both the kernel of the mixture and the base measure $G_0$ are normally distributed (MacEachern and Müller 1998). Since the full conditional of all the parameters, with the exception of $\boldsymbol{\theta}$, have a closed form, they can be updated via Gibbs sampling. For $\boldsymbol{\theta}$, the non-linearity in (14) implies that its full conditional is not a known distribution. Therefore, $\boldsymbol{\theta}$ is updated using the adaptive Metropolis Hasting algorithm introduced by Haario et al. (2001). For each model, the corresponding MCMC sampling scheme is used to draw a Monte Carlo sample of size $30,000$ from the posterior distribution, after a burn-in period of $20,000$ iterations.

Despite the differences between the Linear DDP and the Conditional DP, they lead to very similar posterior fits on the concentration-time curves per child. Specifically, posterior mean estimates of each curve, along with pointwise 95% probability intervals, are practically indistinguishable (see Figure 2). Although the curves correspond to a single patient in the data set, similar results are found for the 138 children in the study population.

It is in terms of prediction, however, that the Linear DDP and the Conditional DP show different performance. In particular, we compare the predictive distribution for the concentration-time curves of the 15 patients belonging to study 1. We obtained a Monte Carlo sample from those distributions by first fitting each PK model to the other patients in the data set (123 children in studies 2 - 7), then generating a Monte Carlo sample from the predictive distribution of $\boldsymbol{\theta}$ for each patient belonging to study 1 (using formulas (7) and (11), along with the patient-specific covariates), and, finally, using each of those samples to evaluate (14).

Comparison of the results by method shows that only the Conditional DP is able to produce sensible estimators. As seen in Figure 3, the Linear



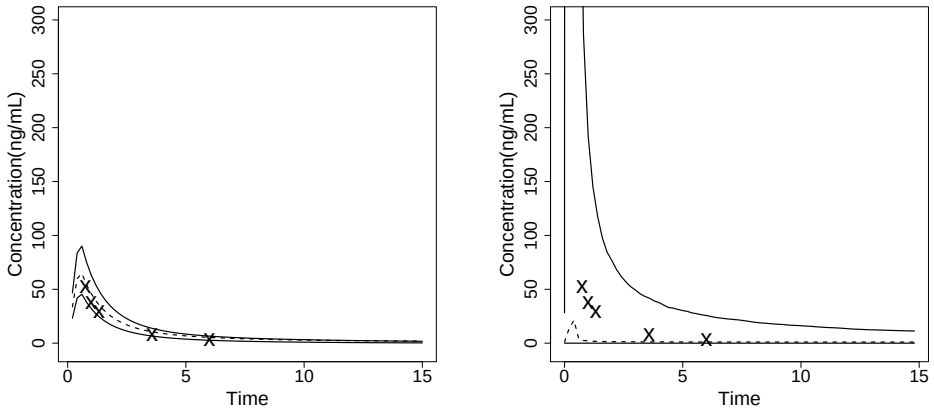

(a) Conditional DP (b) Linear DDP

Figure 2: Posterior concentration-time curves by method. The dashed lines denote the posterior mean concentration curves, the solid lines refer to the 95% pointwise probability intervals, and the "x"s denote the observed concentration-time combinations. The estimated curves show that the posterior fits produced by both methods are practically the same.

DDP has greater predictive uncertainty than the Conditional DP and led to unrealistic predicted concentration-time profiles. When using the Linear DDP, the estimated predictive distribution provides no information about the real concentration-time curves. Although the curves shown correspond to a single patient, similar results were found for all the patients in study 1.

# 4 Discussion

In this paper we have focused on extensions of nonparametric Bayesian models that introduce dependence on covariates. We have shown that good posterior fits with those extensions do not necessarily translate to good prediction. In addition, we have shown that, when the predictive density of such extensions is estimated by averaging mixture distributions, better predictions are produced when the weights in that mixture depend on the covariates. The arguments




(a) Conditional DP (b) Linear DDP

Figure 3: Predictive concentration-time curves by method. The curves denote the pointwise quartiles Q1 $\leq$ Q2 $\leq$ Q3, and the "x"s mark the observed concentration-time combinations.

above have been illustrated by comparing the Linear DDP (De Iorio et al. 2009) and the Conditional DP (Müller and Rosner 1998).

When using the Linear DDP, the weights in the predictive density are solely determined by the size of the clusters induced by the DP, with each cluster representing a different regression structure. Such a feature is a drawback, because it leads to the Linear DDP being unable to identify those regression structures that are more likely for a new subject, given that subject's and the other observed covariates. Combining the covariates with inadequate regression structures leads to unrealistic estimators. Evidence of such behavior has been shown in this paper. In the simulation example, the Linear DDP wrongly introduced an additional mode to the estimated predictive density, while in the pharmacokinetic example, it predicted unrealistic concentration curves.

In contrast, the predictive density of the Conditional DP is a mixture with weights that depend on the covariates of the new subject. This results in an agreement between the covariates and the regression structures. Such

an agreement leads to considerable improvement in the predictive inference when compared to the Linear DDP. In spite of the clear differences in terms of prediction, both methods produce identical results when used for posterior inference.

Although we have focused on continuous covariates, it is also desirable when modeling categorical and discrete covariates to be able to identify the components in the predictive mixture that produce the best fit. Hence, in those cases it is also advantageous to have predictive densities with covariate-dependent weights. However, under some modeling approaches for introducing dependence on finite covariates (numeric or categorical), such as the ANOVA DDP (De Iorio et al. 2004) or the Spatial DDP (Gelfand et al. 2005), the limitations from not having covariate-dependent weights may not be as noticeable as those shown by the Linear DDP. This happens because, in those models, the mechanism through which the covariates determine the form of the components in the mixture can vary, if needed, depending on the specific values of the covariates; this is not achieved by the Linear DDP because, in each component, the corresponding regression structure remains the same regardless of the specific values of the continuous covariates. In particular, the ANOVA DDP models dependence on categorical covariates in an ANOVA fashion. Assuming, for simplicity, that there is only one categorical covariate, the ANOVA DDP implies that the mean of each component in the predictive mixture distribution is equal to a mean effect plus an offset vector. Since such an offset vector varies according to the value of the covariate, the means of the mixture components automatically account for component-specific associations with the discrete covariate level. Hence, it is possible to generate reasonable inferences, even though the weights in the predictive density are

only based on the size of the clusters induced by the DP.

Finally, the results in this paper provide insight regarding which strategies for introducing dependence on covariates can lead to better posterior inferences. Specifically, it provides evidence in favor of extensions of the Sethuraman representation that make the weights vary with the covariates, because such extensions result in nonparametric Bayesian models with a prediction rule based on a mixture with covariate-dependent weights. Example of those extensions include the order-based DDP (Griffin and Steel 2006) and the kernel stick-breaking process (Dunson and Park 2008).

# References


Burr, D. and Doss, H. (2005), "A Bayesian Semiparametric Model for Random-Effects Meta-Analysis," *Journal of the American Statistical Association*, 100, 242–251.

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2006), "Bayesian Inference for Dynamic Models with Dirichlet Process Mixtures," *IEEE Transactions on Signal Processing*, 56, 71–84.

Carota, C. and Parmigiani, G. (2002), "Semiparametric Regression for Count Data," *Biometrika*, 89, 265–285.

Cifarelli, D. and Regazzini, E. (1978), "Nonparametric Statistical Problems under Partial Exchangeability. The use of Associative Means," *Annali del Instituto di Matematica Finianziara dell Universita di Torino*, 12, 1–36.

De Iorio, M., Johnson, W., Muller, P., and Rosner, G. (2009), "Bayesian

Nonparametric Nonproportional Hazards Survival Modeling," *Biometrics*, 65, 762–771.

De Iorio, M., Müller, P., Rosner, G., and MacEachern, S. (2004), "An ANOVA Model for Dependent Random Measures," *Journal of the American Statistical Association*, 99, 205–215.

Dunson, D. (2009), "Bayesian nonparametric hierarchical modeling," *Biometrical Journal*, 51, 273–284.

Dunson, D. and Park, J. (2008), "Kernel Stick-Breaking Processes," *Biometrika*, 95, 307–323.

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

Ferguson, T. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Fuentes-García, R., Mena, R., and Walker, S. (2009), "A Nonparametric Dependent Process for Bayesian Regression," *Statistics and Probability Letters*, 79, 1112–1119.

Gelfand, A., Kottas, A., and MacEachern, S. (2005), "Bayesian Nonparametric Spatial Modeling with Dirichlet Process Mixing." *Journal of the American Statistical Association*, 100, 1021–1036.

Griffin, J. and Steel, M. (2004), "Semiparametric Bayesian Inference for Stochastic Frontier Models," *Journal of Econometrics*, 123, 121–152.

— (2006), "Order-Based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179–194.

Haario, H., Saksman, E., and Tamminen, J. (2001), "An Adaptive Metropolis Algorithm," *Bernoulli*, 7, 223–242.

Kim, S., Tadesse, M., and Vannucci, M. (2006), "Variable Selection in Clustering via Dirichlet Process Mixture Models," *Biometrika*, 93, 877–893.

MacEachern, S. (1999), "Dependent Nonparametric Processes," in *ASA Proceedings of the Section on Bayesian Statistical Science*, Alexandria, VA: American Statistical Association.

MacEachern, S. and Müller, P. (1998), "Estimating Mixture of Dirichlet Process Models," *Journal of Computational and Graphical Statistics*, 7, 223–238.

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

Müller, P., Quintana, F., and Rosner, G. (2004), "A Method for Combining Inference across Related Nonparametric Bayesian Models," *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 66, 735–749.

Müller, P. and Rosner, G. (1998), "Semi-Parametric PK/PD Models," in *Practical Nonparametric and Semiparametric Bayesian Statistics*, eds. Dey, D., Müller, P., and Sinha, D., New York: Springer-Verlag, pp. 323–337.

Rodriguez, A. and ter Horst, E. (2008), "Bayesian Dynamic Density Estimation," *Bayesian Analysis*, 3, 339–366.

Rosner, G. and Müller, P. (1997), "Bayesian Population Pharmacokinetic and Pharmacodynamic Analyses using Mixture Models," *Journal of Pharmacokinetics and Pharmacodynamics*, 25, 209–233.

Schaiquevich, P., Panetta, J. C., Iacono, L. C., Freeman, B. B., Santana, V. M., Gajjar, A., and Stewart, C. F. (2007), "Population Pharmacokinetic Analysis of Topotecan in Pediatric Cancer Patients," *Clinical Cancer Research*, 13, 6703–6711.

Sethuraman, J. (1994), "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Wagner, J. (1968), "Kinetics of Pharmacologic Response. I. Proposed Relationships Between Response and Drug Concentration in the Intact Animal and Man," *Journal of Theoretical Biology*, 20, 173–201.

Wakefield, J. (1994), "An Expected Loss Approach to the Design of Dosage Regimens via Sampling-Based Methods," *The Statistician*, 43, 13–29.

Xing, E., Jordan, M., and Sharan, R. (2007), "Bayesian Haplotype Inference via the Dirichlet Process," *Journal of Computational Biology*, 14, 267–284.

Zeger, S. and Karim, M. (1991), "Generalized Linear Models With Random Effects; A Gibbs Sampling Approach," *Journal of the American Statistical Association*, 86, 79–86.