Chilean Journal of Statistics
Vol. 1, No. 2, September 2010, 35–49

Special Issue "In Memory of Pilar Loreto Iglesias Zuazola"
Research Paper

Linear regression with a dependent skewed Dirichlet process

Fernando A. Quintana

Departamento de Estadística, Pontificia Universidad Católica de Chile, Santiago, Chile

This article is dedicated to the memory of Pilar

(Received: 23 March 2010 · Accepted in final form: 11 August 2010)

Abstract

In the context of linear regression models, two strategies for a flexible distribution of responses are explored. The two approaches are particular instances of a skewed Dirichlet process with dependence on covariates, so that not only the mean response function but the entire distribution of responses depends on covariates. The two constructions are discussed and compared in the context of two examples.

Keywords: Bayes factor · Nonparametric Bayes · Predictive inference · Skewness.

Mathematics Subject Classification: Primary 62G08 · Secondary 62G05.

1. Introduction

Nonparametric Bayesian (NPB) models have received considerable interest over the past few years. One of the main justifications for this increased popularity is the flexibility they provide, compared to traditional parametric alternatives. A common starting point for these methods is to consider the data as generated according to a certain distribution F, assumed to belong to some infinite-dimensional class $\mathcal{F}$, and the problem is how to construct prior distributions on $\mathcal{F}$. The resulting objects are generically termed random probability measures (RPMs) and can be thought of as probability measures defined on the space of distribution functions. The best-known example of RPM is the Dirichlet process (DP), introduced by Ferguson (1973). General discussion and review of NPB methods can be found in Dey et al. (1998), Walker et al. (1999), Ghosh and Ramamoorthi (2003) and in Müller and Quintana (2004).

Iglesias et al. (2009) introduced the skewed Dirichlet process (SDP) as a flexible RPM with the property of including symmetric DPs (Dalal, 1979; Tiwari, 1988) as a special case; see also Doss (1984). The SDP includes a "skewness" parameter $\theta \in (0, 1)$ such that symmetry arises if and only if $\theta = 1/2$. This construction was specially useful when modelling the distribution of errors ϵ_i in the context of regression models. By adequately choosing the prior for θ, it is possible to derive a test of symmetry. Under a linear regression model, if ϵ_i has a symmetric distribution and if $\mathrm{E}(|\epsilon_i|) < \infty$, then it follows that $\mathrm{E}(\epsilon_i) = 0$, which facilitates the interpretation of regression coefficients.

*Corresponding author. Fernando A. Quintana. Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Avenida Vicuña Mackenna 4860, Macul, Santiago, Chile. Email: quintana@mat.puc.cl

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

This work extends models based on the SDP to include an additional dependence (or indexing) on covariates. Two different methods are used to this effect. The first method consists of stating a joint model for responses $\boldsymbol{y} = (y_1, \ldots, y_n)$, covariates $\boldsymbol{x} = (x_1, \ldots, x_n)$, and the RPM F, which then implies a conditional model for $(\boldsymbol{y}, F)$ given $\boldsymbol{x}$ that yields the desired dependence. This strategy has been used, among others, by Müller et al. (1996). The second method is adapted from the dependent models proposed by MacEachern (1999), which has been most notoriously exploited by the class of ANOVA-DDP models of De Iorio et al. (2004) and others. The two methods are compared in the context of linear regression models.

The rest of this article is organized as follows. Section 2 describes how the different components (the SDP and the dependent RPMs) are combined to form the dependent SDP (DSDP), briefly discussing the main features of the proposed models. Section 3 illustrates the model in two concrete examples. Some final remarks are stated in Section 4.

2. The Models

The proposed models have two main ingredients, which we briefly summarize in the following two subsections.

2.1 The skewed Dirichlet process

The Dirichlet process (DP) has been extensively studied and applied in an ever growing range of fields. Specially important for our context is the representation by Sethuraman (1994), which states that if $F \sim \mathcal{D}(M, F_0)$, a DP with total mass parameter $M > 0$ and baseline distribution F_0 in the appropriate space (Ferguson, 1973), then F can be represented as $F(\cdot) = \sum_{h=1}^{\infty} w_h \delta_{V_h}(\cdot)$, where $V_1, V_2, \ldots \overset{\text{i.i.d.}}{\sim} F_0$ and $w_1, w_2, \ldots$ are stochastically ordered weights that follow a stick breaking process: $w_1 = U_1$ and $w_h = U_h \times \prod_{i=1}^{h-1}(1-U_i)$ for all $h \geq 2$, with $U_1, U_2, \ldots \overset{\text{i.i.d.}}{\sim} \text{Beta}(1, M)$.

Iglesias et al. (2009) showed that if $F \sim \text{SDP}(M, F_0, \theta)$, a skewed DP with total mass parameter M, baseline measure F_0 on the positive real numbers (assumed to have density f_0) and skewness parameter θ, then F can be represented as

$$F(\cdot) = \theta \sum_{h=1}^{\infty} w_h \delta_{\theta V_h}(\cdot) + (1-\theta) \sum_{h=1}^{\infty} w_h \delta_{-(1-\theta)V_h}(\cdot), \tag{1}$$

where the $\{w_h\}_{h\geq 1}$ and $\{V_h\}_{h\geq 1}$ sequences are as before. Note that, just as the original DP, Equation (1) is an infinite mixture of point masses, drawn according to F_0. The difference is that for SDPs, the point masses are defined through a mass-splitting procedure. Concretely, each V_h is drawn from F_0 and then two point masses are added, one at θV_h, and the other at $-(1-\theta)V_h$, with respective weights θ and $(1-\theta)$. Each of these point masses is weighted by the corresponding stick-breaking weight w_h. Note that, just as the DP, this gives rise to an almost surely discrete RPM. The skewness parameter θ represents the total amount of mass assigned by F as in Equation (1) to the positive real numbers. In fact, it follows from the representation given in (1) that for a Borel set $B \subset \mathbb{R}$, the expected value of F is given by

$$\text{E}(F(B)) = \theta F_0(B/\theta) + (1-\theta) F_0(-B/(1-\theta)),$$

where $B/\theta = \{b/\theta\colon b \in B\}$ and $-B/(1-\theta) = \{-b/(1-\theta)\colon b \in B\}$. Also, when $\theta = 1/2$, Equation (1) becomes a symmetric version of the DP (Dalal, 1979; Tiwari, 1988); see further details in Iglesias et al. (2009).

It is also important to point out here the clustering structure of DPs and SDPs. The discreteness of DPs has been used many times as a mechanism to define probability distributions on partitions. Let $S_0 = \{1, \ldots, n\}$ be a collection of n indices and $\rho = \{S_1, \ldots, S_k\}$ be a partition of S_0 into k nonempty disjoint sets with $S_0 = \cup_{j=1}^{k} S_j$. Consider now $X_1, \ldots, X_n \mid F \overset{\text{i.i.d.}}{\sim} F$ and $F \sim \mathcal{D}(M, F_0)$. By the discreteness of F it follows that there are ties among the sampled values. An interpretation of this structure in terms of Pólya urns has been described in Blackwell and MacQueen (1973). Let $X_1^*, \ldots, X_k^*$ denote the unique values (also called locations) among $X_1, \ldots, X_n$ and define the cluster memberships $s_1, \ldots, s_n$ as $s_1 = 1$, and $s_i = s_j$ if and only if $X_i = X_j$, so that $X_i = X_{s_i}^*$ for all $i \in S_0$. As a convention, we assume clusters to be numbered consecutively, starting from 1. Letting ρ denote the random partition implied by the indicators just defined, it can be shown that $X_1^*, \ldots, X_k^* \overset{\text{i.i.d.}}{\sim} F_0$ and

$$p(\rho = \{S_1, \ldots, S_k\}) = \frac{M^k \prod_{j=1}^{k} (|S_j| - 1)!}{\prod_{i=1}^{n} (M + i - 1)}, \tag{2}$$

marginally over F and the locations, where $|S_j|$ is the cardinality of S_j; see, e.g., Pitman (1996). It is interesting to point out that Equation (2) also corresponds to a product partition model (PPM), as discussed in Hartigan (1990) and Barry and Hartigan (1992), with product distribution $p(\rho) \propto \prod_{j=1}^{k} c(S_j)$ and cohesion functions (which measure how tightly grouped the elements in S_j are thought to be a priori) given by $c(S_j) = M \times (|S_j| - 1)!$; see Quintana and Iglesias (2003) and Quintana (2006).

The SDP has also a similar structure of ties, but associated to the absolute values of the sample. Indeed, Iglesias et al. (2009) showed that if $X_1, \ldots, X_n \mid F \overset{\text{i.i.d.}}{\sim} F$ and $F \sim \text{SDP}(M, F_0, \theta)$ then the partition induced by the ties among $|X_1|, \ldots, |X_n|$ has the same distribution as Equation (2), $|X_1^*|, \ldots, |X_k^*| \overset{\text{i.i.d.}}{\sim} F_0$, and marginally each X_i has density

$$r(x \mid \theta) = f_0(\theta^{-1}x) I\{x \geq 0\} + f_0(-(1-\theta)^{-1}x) I\{x < 0\},$$

a continuous density over all the real numbers, provided f_0 is continuous and $\lim_{x \to 0^+} f_0(x)$ exists. Moreover, $\text{sgn}(X_1), \ldots, \text{sgn}(X_n)$ are shown to be i.i.d. with $\text{P}(\text{sgn}(X_1) = 1) = \theta = 1 - \text{P}(\text{sgn}(X_1) = -1)$, and independent of $|X_1|, \ldots, |X_n|$. In practice, this means that we can formulate SDP-based models simply in terms of regular DPs. Indeed, assuming the model $X_1, \ldots, X_n \mid F \overset{\text{i.i.d.}}{\sim} F$ and $F \sim \text{SDP}(M, F_0, \theta)$ can be equivalently represented as $X_i = Z_i \cdot |Y_i|$ where $|Y_1|, \ldots, |Y_n| \mid F \overset{\text{i.i.d.}}{\sim} F$ with $F \sim \text{DP}(M, F_0)$ and $Z_1, \ldots, Z_n$ are i.i.d. random variables, independent of $\{|Y_i|,\, 1 \leq i \leq n\}$, with $\text{P}(Z_1 = \theta) = 1 - P\,(Z_1 = -(1-\theta)) = \theta$; see further details in Iglesias et al. (2009).

2.2 Dependent nonparametric models

Of particular recent interest is the study of NPB models that can be indexed by a set of covariates. The resulting class of models has been usually termed dependent nonparametric models. Specifically, if we denote the covariate space as $\mathcal{X}$, the idea is to construct a class of RPMs $\{F_x\colon x \in \mathcal{X}\}$ such that F_x retains some interesting properties for each $x \in \mathcal{X}$. Cifarelli and Regazzini (1978) proposed a model where $F_x \sim \text{DP}(M, F_0^x)$ and F_0^x is a

distribution centered around a linear regression on x. Related models are discussed in Mira and Petrone (1996) and in Giudici et al. (2003). A much more flexible construction was proposed by MacEachern (1999), where each atom in Sethuraman (1994) representation of DPs depends on x: $F_x(B) = \sum_{h=1}^{\infty} w_h \delta_{V_h(x)}(B)$ and the $\{V_h(x)\}$ collection is such that the random variables are independent across h for given x, but dependent across x for a given h. This results on a collection of RPMs where each F_x is marginally a DP for every x. The ANOVA-DDP model discussed in De Iorio et al. (2004) considers an ANOVA-type regression for each atom. Similar approaches have been proposed for functional data analysis (Dunson and Herring, 2006), survival analysis (Jara et al., 2007; De Iorio et al., 2009), spatial data analysis (Gelfand et al., 2005; Duan et al., 2007), longitudinal data analysis (Müller et al., 2005; De la Cruz-Mesía et al., 2007) and time series (Caron et al., 2008). Other dependent extensions of the DP involve the hierarchical DP (Teh et al., 2006), the nested DP (Rodríguez et al., 2008), the kernel stick-breaking process (Dunson and Park, 2008), and the matrix stick-breaking process (Dunson et al., 2008). A recent application of nested DP models to functional data analysis is given in Rodríguez et al. (2009).

A different modeling framework consists of proposing a joint model for responses $\boldsymbol{y}$, covariates $\boldsymbol{x}$, and RPM F. By focusing on the implied conditional distribution $p(\boldsymbol{y}, F \mid \boldsymbol{x})$, a dependent model for $(\boldsymbol{y}, F)$ follows. This strategy has been applied a number of times in different contexts; see Müller et al. (1996) and Shahbaba and Neal (2009).

Finally, a different type of dependent model that benefits from the simple structure of product partition models (PPMs) (Hartigan, 1990; Barry and Hartigan, 1992) was recently proposed by Müller et al. (2010). The approach also exploits the connections between PPMs and DP-style RPMs.

2.3 The proposed models

Consider observations (y_i, x_i), $i = 1, \ldots, n$ corresponding to a response of interest y_i and a vector of associated covariates x_i of dimension $p \geq 1$. To present the discussion in a concrete framework, suppose we want to model the relationship between responses and covariates by means of a linear regression with a flexible distribution of errors. Iglesias et al. (2009) proposed a hierarchical model:

$$\begin{aligned} y_i \mid \mu_i, \boldsymbol{\beta}, \sigma^2 &\sim \mathrm{N}(\mu_i + \boldsymbol{\beta}^\top x_i, \sigma^2), \\ \mu_1, \ldots, \mu_n \mid F &\overset{\text{i.i.d.}}{\sim} F, \\ F &\sim \mathrm{SDP}(M, F_0(\tau), \theta), \end{aligned} \tag{3}$$

where $F_0(\tau)$ is the distribution of $|Z|$, with $Z \sim \mathrm{N}(0, \tau)$ and $\boldsymbol{\beta} \sim \mathrm{N}(0, S)$, $\sigma^{-2} \sim \mathrm{Gamma}(\nu_0, \nu_1)$, $\tau^{-1} \sim \mathrm{Gamma}(\lambda_0, \lambda_1)$, and a mixture prior for θ:

$$\theta \sim (1 - \pi)\,\mathrm{Beta}(a_0, b_0) + \pi \delta_{1/2}(\theta), \tag{4}$$

where $0 < \pi < 1$ represents the prior probability of symmetry. Model given in (3) implies a flexible marginal distribution of errors $y_i - \boldsymbol{\beta}^\top x_i$. Indeed, Iglesias et al. (2009) showed that the marginal distribution of errors has a density function that can be expressed as $f(x) = \int \mathrm{N}(x;\, \mu, \sigma^2)\, dF(\mu)$, a flexible mixture.

The flexibility of model given in (3) can be enlarged by incorporating covariate dependence, so that departures from the mean function specification can be captured in a better way. Two possible ways to do this are described next.

Conditional models Following Müller et al. (1996) and Shahbaba and Neal (2009), model given in (3) can be extended as follows:

$$\begin{aligned} y_i \mid x_i, \mu_i, \boldsymbol{\beta} &\sim \mathrm{N}(\mu_i + \boldsymbol{\beta}^\top x_i, \sigma^2) \\ x_i \mid \xi_i &\sim \mathrm{N}(\xi_i, \eta_i) \\ \mu_i &\sim \theta\delta_{\theta V_i}(\mu_i) + (1-\theta)\delta_{-(1-\theta)V_i}(\mu_i) \\ (V_1, \xi_1, \eta_1), \ldots, (V_n, \xi_n, \eta_n) \mid F &\sim F \\ F &\sim \mathcal{D}(M, F_0(\tau) \times Q), \end{aligned} \tag{5}$$

completed with a prior $p(\boldsymbol{\beta}, \tau, \theta) = p(\boldsymbol{\beta})p(\tau)p(\theta)$, where each of these priors are the same as model given in (3). In addition, the baseline distribution in model given in (5) has density corresponding to the triplet (V, ξ, η) assumed to have independent components with respective densities $V \mid \tau \sim f_0(\tau)$ and Q a distribution for the pair (ξ, η), with $\xi \sim \mathrm{N}(\xi_0, v_0)$ and $\eta^{-1} \sim \mathrm{Gamma}(a_\eta, b_\eta)$.

It is important to point out that, due to the equivalent formulation of SDPs in terms of a regular DP discussed in Section 2.1, model given in (5) is indeed a SDP-model for the intercept parameters and the relationship between y and x is linear at the sampling level. Interestingly, the implied clustering structure is also shared by the parameters ξ and η that define the model for covariates $\{x_i,\, 1 \leq i \leq n\}$. In fact, a model for the covariates is explicitly stated in (5), which implies a joint model for (y, x) and the RPM F. By focusing on the implied conditional model $p(y, F \mid x) = p(y, x, F)/p(x)$, it follows that the indexing of partitions (and henceforth of RPMs) is far from linear because the denominator induces nonlinearities in x.

Regression on the atoms The class of dependent models in MacEachern (1999) feature a regression at the level of the atoms in Sethuraman (1994) decomposition. An ANOVA version of such dependence was proposed in De Iorio et al. (2004). A simple way to incorporate this modeling strategy into the problem at hand is to postulate a regression at the level of the offset parameters $|\mu_1|, \ldots, |\mu_n|$. Concretely, assume

$$\begin{aligned} y_i \mid x_i, \mu_i, \boldsymbol{\beta} &\sim \mathrm{N}(\mu_i + \boldsymbol{\beta}^\top x_i, \sigma^2) \\ V_i &= \exp\left(\alpha_i + \lambda_i^\top x_i\right) \\ \mu_i &\sim \theta\, \delta_{\theta V_i}(\mu_i) + (1-\theta)\, \delta_{-(1-\theta)V_i}(\mu_i) \\ (\alpha_1, \lambda_1), \ldots, (\alpha_n, \lambda_n) \mid F &\sim F \\ F &\sim \mathcal{D}(M, F_0), \end{aligned} \tag{6}$$

with the same additional assumptions on other parameters as in (5) and where now $F_0(\alpha, \lambda) \equiv \mathrm{N}(m_0, V_0)$ is a $(p+1)$-dimensional multivariate normal distribution.

Note that under model given in (6), the $|\mu_i|$ are determined by means of a log-linear model and the coefficients of this model are cluster-specific, with a DP-style clustering structure, from which the equivalence of model given in (6) to a model formulated in terms of a dependent SDP (DSDP) follows. Finally, note that each of the models given in (5) and (6) are basically DP-style models, so that we can use the posterior simulation methods specifically proposed for such cases. The typically most involved MCMC step consists of updating the cluster-related parameters V_i (or μ_i). The algorithms discussed in MacEachern and Müller (1998) or Neal (2000) are standard approaches to deal with the clustering structure. They also have the advantage of not relying on the likelihood

and the baseline measure being conjugate, as some of the early methods do. A new set of V_i parameters implies a new set of configurations $\{s_i\}$. Following the advice in Bush and MacEachern (1996), it is customary to drop the imputed V_i^* (or μ_i^*) values and resample them from their full conditionals. These correspond to the posterior under a simple parametric model with likelihood as given in (5) or (6) and prior given by the corresponding baseline measure. Standard methods can be used for the remaining parameters in the models; see additional SDP-specific discussion in Iglesias et al. (2009).

Of special interest for models defined in (5) or (6) is the predictive inference, i.e. the density that corresponds to a new response y_{n+1} at a given value of the covariate vector x_n. This can be easily implemented on top of posterior simulation for each respective model. The key step consists of sampling a new offset parameter μ_{n+1}. The same basic structure applies to both models, so we limit the discussion to model given in (6). It follows from the Pólya urn representation of Blackwell and MacQueen (1973) that the distribution of a new $(\alpha_{n+1}, \lambda_{n+1})$ is a mixture of point-masses at previously imputed values and the baseline measure:

$$p((\alpha_{n+1}, \lambda_{n+1}) \mid \text{all else}) = \sum_{j=1}^{k} \frac{n_j}{M+n} \delta_{(\alpha_j^*, \lambda_j^*)}(\alpha_{n+1}, \lambda_{n+1}) + \frac{M}{M+n} f_0(\alpha_{n+1}, \lambda_{n+1}), \quad (7)$$

where $(\alpha_1^*, \lambda_1^*), \ldots, (\alpha_k^*, \lambda_k^*)$ are the unique values among $\{(\alpha_i, \lambda_i)\}_{i=1}^n$ (the cluster locations) and n_j is the number of (α_i, λ_i)'s equal to $(\alpha_j^*, \lambda_j^*)$ (the cluster sizes). Here, f_0 is the density function that corresponds to F_0 in model given in (6). From Equation (7) and the transformation $V_i = \exp(\alpha_i + \lambda_i^\top x_i)$ for all $i \geq 1$, the distribution of V_{n+1} easily follows as another mixture of point masses and a continuous distribution. Then, μ_{n+1} is computed as θV_{n+1} with probability θ, or $-(1-\theta)V_{n+1}$ with probability $1-\theta$, using the currently imputed value for θ. Putting all the pieces together, a predictive draw for a next response corresponding to covariate vector x_{n+1} can be generated as

(i) Draw $(\alpha_{n+1}, \lambda_{n+1})$ from Equation (7) and compute $V_{n+1} = \exp(\alpha_{n+1} + \lambda_{n+1}^\top x_{n+1})$.
(ii) Draw $\mu_{n+1} \sim \theta \delta_{\theta V_{n+1}}(\mu_{n+1}) + (1-\theta)\delta_{-(1-\theta)V_{n+1}}(\mu_{n+1})$.
(iii) Draw $y_{n+1} \sim \mathrm{N}(\mu_i + \boldsymbol{\beta}^\top x_{n+1}, \sigma^2)$,

where the currently imputed values for all parameters at a given MCMC iteration are used in the above calculations. As a more accurate alternative, one may wish to directly evaluate the predictive density of y_{n+1}, which is given by

$$\begin{aligned} p(y_{n+1} \mid x_{n+1}, \text{ all else}) &= \theta \sum_{j=1}^{k} \frac{n_j}{M+n} \mathrm{N}\left(y_{n+1};\, \theta \exp(\alpha_j^* + \lambda_j^{*\top} x_{n+1}) + \boldsymbol{\beta}^\top x_{n+1}, \sigma^2\right) \quad (8) \\ &+ (1-\theta) \sum_{j=1}^{k} \frac{n_j}{M+n} \mathrm{N}\left(y_{n+1};\, -(1-\theta) \exp(\alpha_j^* + \lambda_j^{*\top} x_{n+1}) + \boldsymbol{\beta}^\top x_{n+1}, \sigma^2\right) \\ &+ \frac{M\theta}{M+n} \int \mathrm{N}\left(y_{n+1};\, \theta \exp(\alpha_{n+1} + \lambda_{n+1}^\top x_{n+1}) + \boldsymbol{\beta}^\top x_{n+1}, \sigma^2\right) \\ &\times \mathrm{N}\left((\alpha_{n+1}, \lambda_{n+1});\, m_0, V_0\right) d\alpha_{n+1} d\lambda_{n+1} \\ &+ \frac{M(1-\theta)}{M+n} \int \mathrm{N}\left(y_{n+1};\, -(1-\theta) \exp(\alpha_{n+1} + \lambda_{n+1}^\top x_{n+1}) + \boldsymbol{\beta}^\top x_{n+1}, \sigma^2\right) \\ &\times \mathrm{N}\left((\alpha_{n+1}, \lambda_{n+1});\, m_0, V_0\right) d\alpha_{n+1} d\lambda_{n+1}, \end{aligned}$$

where $\mathrm{N}(a;\, b, c)$ is a multivariate normal density on a, with mean b and covariance matrix c. Evaluation of the integrals in Equation (8) may require numerical methods.

Table 1. Posterior summaries (means, standard deviations and 95% highest posterior density intervals) of some parameters for the stock market returns example under both models.

Parameter	Posterior Summary	Model (5)	Model (6)
β	Mean	0.665	0.554
	SD	0.200	0.195
	95% HPD	(0.259,1.055)	(0.175,0.948)
θ	Mean	0.549	0.598
	SD	0.196	0.148
	95% HPD	(0.101,0.946)	(0.349,0.909)
σ^2	Mean	0.032	0.029
	SD	0.005	0.005
	95% HPD	(0.024,0.042)	(0.022,0.039)
τ	Mean	0.448	-
	SD	0.294	-
	95% HPD	(0.138,1.187)	-

3. Data Illustrations

Two examples are considered next to illustrate the different aspects of the proposed models.

3.1 Stock market returns

Iglesias et al. (2009) considered the data on a monthly series of stock returns for the Concha y Toro wine producers. The relationship between such returns and a variable describing the market behavior (e.g. the Chilean IPSA index) is of special interest to market analysts. The capital asset pricing model (CAPM) (Elton and Gruber, 1995) is a standard tool in this context, which establishes a relation between the return on the asset in excess of the risk-free rate during the ith period y_i, and the excess return on the market portfolio of assets in that period x_i. It is common to assume this relation to be linear, namely, $y_i \sim \mathrm{N}(\mu + \beta x_i, \sigma^2)$. The slope β is called the systematic risk. It represents the sensitivity of the expected excess asset returns to the expected excess market returns, and is usually the main target of the analysis. The SDP model in Iglesias et al. (2009) provided increased flexibility and robustness to outlier-prone alternatives.

An even more flexible approach is given by the two dependent models as defined in (5) and (6). To fit these models, the following hyperparameter choices were made. For the conditional model: $M = 1$, $S = 100$, $a_0 = b_0 = 1$, $\pi = 0.2$, $\nu_0 = \lambda_0 = a_\eta = 2.01$, $\nu_1 = \lambda_1 = b_\eta = 1.01$, $\xi_0 = 0.027$ (the empirical mean of the covariate), $v_0 = 0.078$ (which amounts to 10 times de empirical variance); and for the DSDP model: $m_0^\top = (0, 0)$, $V_0 = 10 \times I_2$, and the other hyperparameters are exactly as in the previous case.

Table 1 shows posterior summaries for some of the relevant parameters. Note that τ appears only in model given in (5). Some differences exist between these estimates. The systematic risk β appears somewhat lower for the DSDP model than for the conditional model. Also, the posterior distribution for θ has some less spread in the DSDP case. All of this can be graphically seen in Figure 1, which shows the posterior densities for β, θ and σ^2 for both models, with solid lines corresponding to the conditional model given in (5). The posterior distributions for β and σ^2 mostly agree, but some discrepancies are found for that of θ. Nevertheless, the posterior of θ is skewed to the right in both cases, suggesting evidence against symmetry.

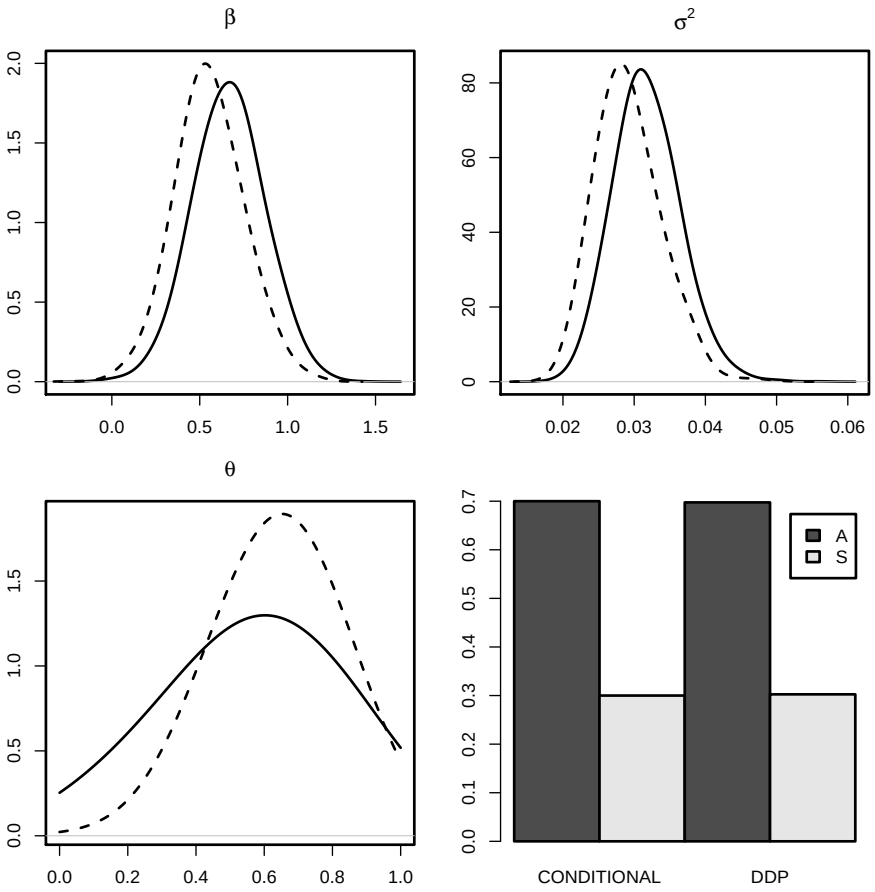

Figure 1. Posterior inference for stock market returns data under models (5) and (6). The display includes the posterior densities of β, σ^2 and θ for model (5) in solid line, and model (6) in dashed line. The remaining graph shows the posterior probabilities of symmetry ("S") and asymmetry ("A"), $\mathrm{P}(\theta = 0.5 \mid \boldsymbol{y})$ and $\mathrm{P}(\theta \neq 0.5 \mid \boldsymbol{y})$.

The bottom-right plot in Figure 1 shows the posterior distribution of symmetry, i.e. the posterior distribution of the binary indicator $I\{\theta = 1/2\}$, with values 0 (asymmetry) and 1 (symmetry). Both models have almost identical posteriors ($\mathrm{P}(\theta = 1/2 \mid \boldsymbol{y})$, which were estimated as 0.300 and 0.303 for conditional and DSDP models, respectively). In both cases there is some support for the hypothesis of symmetry. In fact, this is corroborated by the corresponding Bayes factor, computed as (Iglesias et al., 2009)

$$\mathrm{BF} = \frac{\mathrm{P}(\theta = \frac{1}{2} \mid \boldsymbol{y})(1-\pi)}{\mathrm{P}(\theta \neq \frac{1}{2} \mid \boldsymbol{y})\pi},$$

which gives 1.714 and 1.739, respectively. This also agrees with the fact that, as seen from Table 1, $\mathrm{E}(\theta \mid \boldsymbol{y}) > 1/2$ in both cases.

Another interesting aspect of the models concerns predictions. These are computed as discussed in Section 2.3. Figure 2 shows such inference for a sequence of covariate values ranging from -1 to 1 (the empirical range is from -0.282 to 0.274). Note how these densities become more dispersed as the covariate value gets away from the observed empirical range. Also, the predictions for both models are almost identical except at the extremes. Interestingly, the plotted densities for the DSDP model seem to "shrink" more towards the empirically observed range than the conditional one.

Finally, to compare how the two models fit the observed data, the CPO (conditional predictive ordinate) statistic may be used on individual observations (Gelfand et al., 1992).

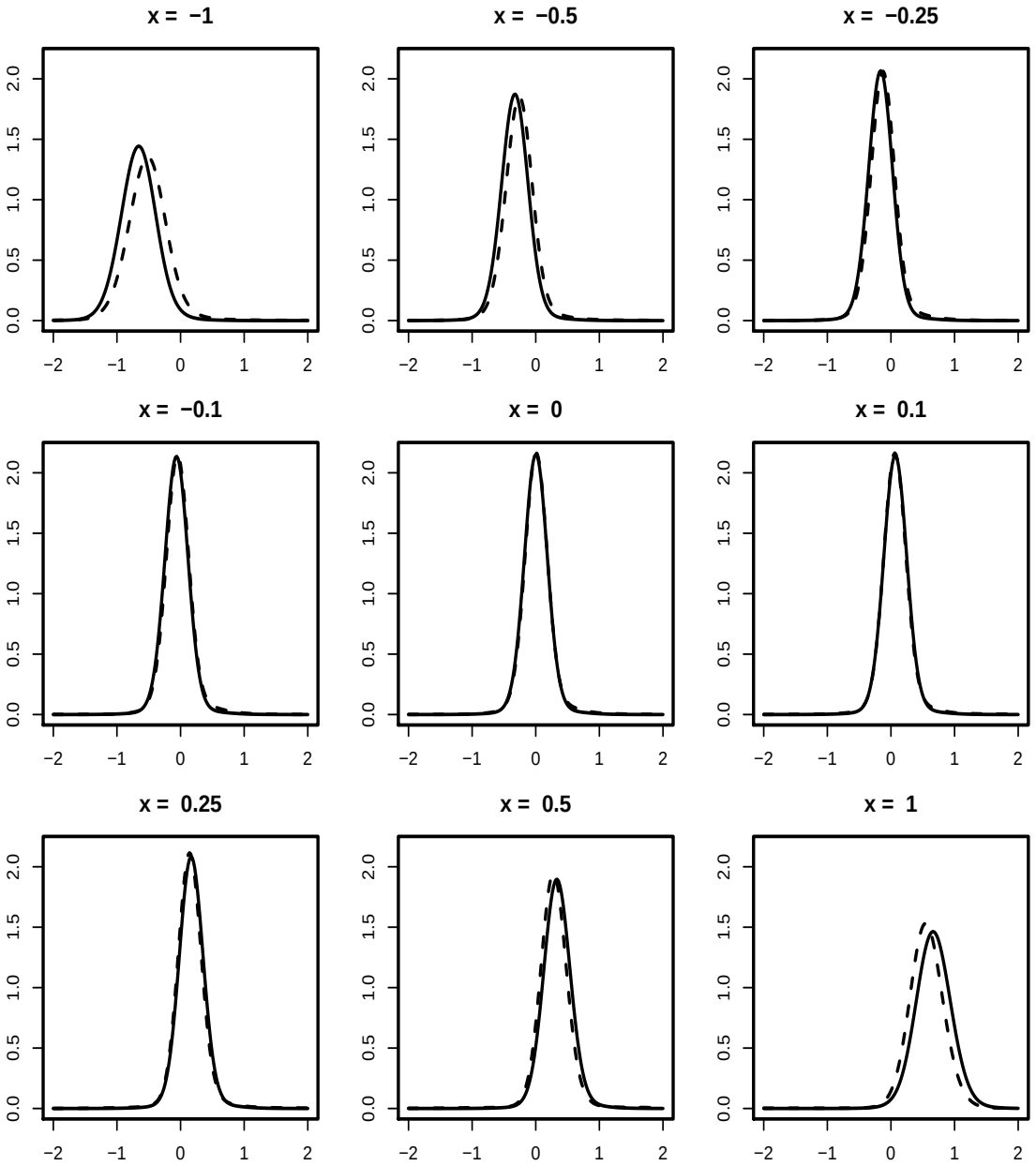

Figure 2. Predictive densities for stock market returns data under both models. The sequence of plots shows the predictive density that corresponds to the indicated covariate values. The densities for model (5) are shown in solid lines, and in dashed lines for model (6).

The idea is to assess the model fit by a cross-validation approach, comparing actual observations with how the model predicts them after deletion. All these predictions can be summarized by computing the summation of the log-CPOs over all observations, which yields the so-called log of the pseudo-marginal likelihood (LPML); see also Geisser and Eddy (1979). In general, the highest LPML value across a collection of models is regarded as providing the best fit; see further theoretical and practical details, as well as additional discussion, in Chen et al. (2000). For the conditional model the LPML was 66.646, while for the DSDP model the value was 73.856, suggesting a better fit in the later case.

Table 2. Posterior summaries (means, standard deviations and 95% highest posterior density intervals) of some parameters for the Australian athlete data under both models.

Parameter	Posterior Summary	Model (5)	Model (6)
	Mean	15.770	17.797
β_0	SD	2.377	2.198
	95% HPD	(11.010,20.360)	(13.470,22.000)
	Mean	0.856	1.272
β_1	SD	0.590	0.490
	95% HPD	(-0.324,2.048)	(0.264,2.205)
	Mean	0.953	0.399
β_2	SD	0.529	0.484
	95% HPD	(-0.066,1.989)	(-0.534,1.352)
	Mean	0.200	0.246
β_3	SD	0.108	0.095
	95% HPD	(-0.016,0.402)	(0.060,0.428)
	Mean	0.011	0.007
β_4	SD	0.004	0.004
	95% HPD	(0.003,0.020)	(-0.001,0.016)
	Mean	0.461	0.811
θ	SD	0.259	0.119
	95% HPD	(0.026,0.967)	(0.535,0.989)
	Mean	0.148	0.225
σ^2	SD	0.015	0.032
	95% HPD	(0.120,0.178)	(0.171,0.293)
	Mean	0.445	-
τ	SD	0.287	-
	95% HPD	(0.144,1.214)	-

3.2 Biomedical data

Consider now the body mass index (bmi) data for 202 Australian athletes, available online in the `sn` package in `R` and discussed in Cook and Weisberg (1994) and Azzalini and Capitanio (1999), among others. For the purpose of this illustration, bmi is taken as the response and four covariates are adopted: gender, red-cell count (rcc), white-cell count (wcc) and plasma ferritin concentration (Fe). Including an intercept term, the design vector x_i is therefore 5-dimensional.

The hyperparameter values were chosen exactly as in Section 3.1. Additionally, for model given in (5) the empirical means for the four covariates were 0.505, 4.719, 7.109 and 76.876, while the variances multiplied by a factor of 10 were 2.512, 2.097, 32.420, and 22563.677, respectively. And for model given in (6), the values $m_0^\top = (0, 0, 0, 0, 0)$ and $V_0 = 10 \times I_5$ were used.

A summary of the posterior inference on parameters can be found in Table 2 and graphical display of some of the corresponding marginal posterior densities can be seen in Figure 3.

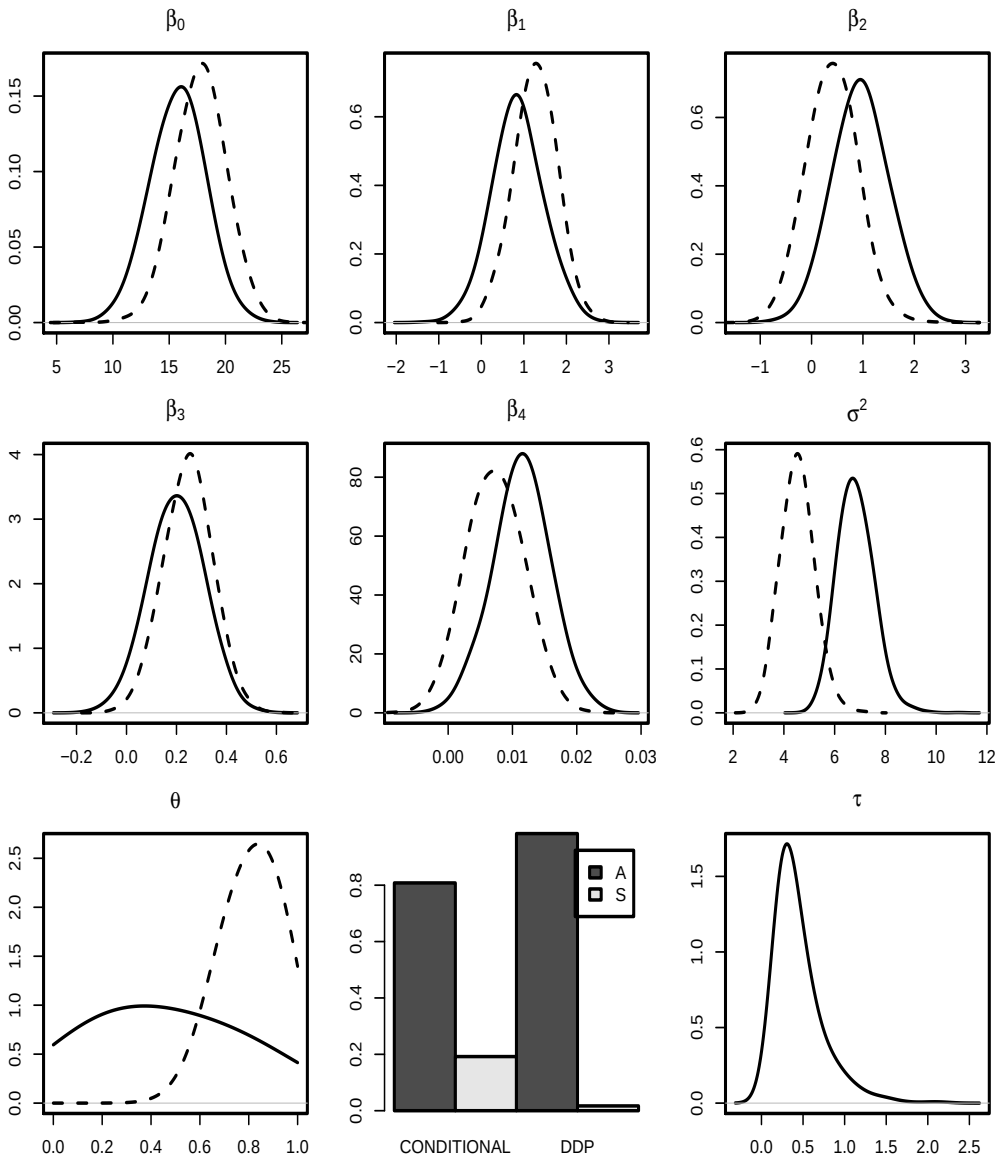

Figure 3. Posterior inference for biomedical data example under both models. The display includes the posterior densities of the regression coefficients β_0 (intercept), β_1 (gender), β_2 (rcc), β_3 (wcc), β_4 (Fe), σ^2 and θ for model (5) in solid line, and model (6) in dashed line. The display also includes the posterior probabilities of symmetry ("S") and asymmetry ("A"), $\mathrm{P}(\theta = 0.5 \mid \boldsymbol{y})$ and $\mathrm{P}(\theta \neq 0.5 \mid \boldsymbol{y})$, and the posterior density of τ for model (5) only.

Model DSDP finds strong support for skewness (the Bayes factor for symmetry is 0.0692) and the corresponding posterior distribution for θ is very skewed to the right. In contrast, for the conditional model, the Bayes factor for symmetry is 0.950 and the posterior for θ is slightly skewed to the left. It is then clear that the two models differ in the way they capture some features of the data. They both point to non-symmetric behavior of the distribution of $y_i - \beta^\top x_i$, but they differ in the extent and direction of the skewness. The differences between models are also reflected in the respective posterior distributions for regression coefficients. The significance of variables differs radically between models. By examining the HPDs in Table 2 one can conclude that only Fe is significant for the conditional model, while only gender and white-cell count have posterior distributions mostly away from zero for the DSDP model. Thus, the two models are effectively favoring different ways of expressing the mean response, which may help explaining the discrepancy in skewness and in the variance parameter σ^2.

The findings above are also reflected in the predictive densities corresponding to various combinations of the covariates. Figure 4 shows such inference for different combinations of gender and Fe. The latter values were picked over the observed Fe range. It is clear that the location of such predictions shifts to the right for the conditional model as the Fe values increase, while the predictions for the DSDP model remain mostly unchanged. In contrast, predictions for different combinations of gender and white-cell count values change for the DSDP model, but are mostly unchanged for the conditional model (data not shown).

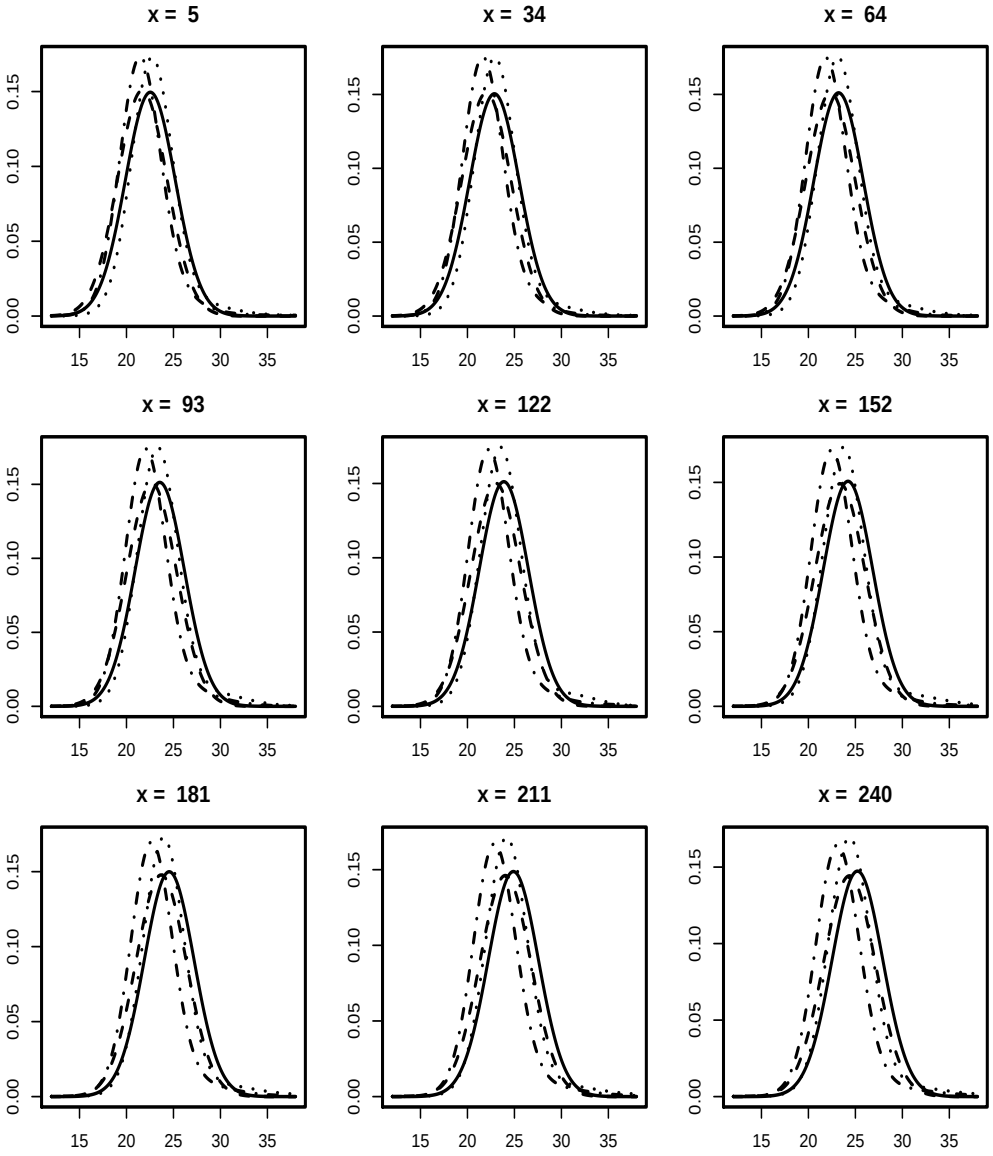

Figure 4. Predictive densities for biomedical data under both models. The sequence of plots shows the predictive density that corresponds to the combination of gender and the indicated covariate values of plasma ferritin concentration, keeping fixed rcc and wcc at their median values. The densities for model (5) are shown in solid and dashed lines (for male and female athletes, respectively), while for model (6), the densities are shown in dotted and semi-dashed lines (for male and female athletes, respectively).

As in Section 3.1, the LPMP statistic is useful to compare models. The values obtained were -482.8726 for the conditional model and -445.9674 for the DSDP model, suggesting a much better fit for the latter.

4. Discussion

The purpose of this article was to explore and compare two different ways of defining dependence for models that are based on skewed Dirichlet processes. One approach involved a joint model for responses, covariates and random measure, which would then yield dependence by conditioning on the covariates. The other approach consisted of a log-linear regression at the level of the atoms in the infinite series representation by Sethuraman (1994). Either model is capable of representing nonlinear trajectories in terms of the covariates.

In the specific applications to linear regression models discussed in Section 3, the dependence was stated for the errors distribution, also including a mean response function that is linear in terms of covariates. Both models exhibited some differences when fitted to data on stock market returns and, specially, for the biomedical features of Australian athletes. The differences simply reflect the particular way in which each model accommodates the distribution of $y_i - \beta^\top x_i$ in light of the information provided by covariates and responses. Interestingly, the LPML calculations showed that the DSDP provides a better fit in both cases for the specific implementation adopted. Nevertheless, it is possible to adopt many other definitions of dependence and it is not advisable to draw overall conclusions about which modeling strategy is best (i.e. conditional versus DSDP) from the examples discussed here.

Acknowledgments

This work was partially funded by grant FONDECYT 1100010.

References

Azzalini, A., Capitanio, A., 1999. Statistical applications of the multivariate skew normal distribution. Journal of The Royal Statistical Society Series B—Statistical Methodology, 61, 579-602.

Barry, D., Hartigan, J.A., 1992. Product partition models for change point problems. The Annals of Statistics, 20, 260-279.

Blackwell, D., MacQueen, J.B., 1973. Ferguson distributions via Pólya urn schemes. The Annals of Statistics, 1, 353-355.

Bush, C.A., MacEachern, S.N., 1996. A semiparametric Bayesian model for randomised block designs. Biometrika, 83, 275-285.

Caron, F., Davy, M., Doucet, A., Duflos, E., Vanheeghe, P., 2008. Bayesian inference for linear dynamic models with Dirichlet process mixtures. IEEE Transactions on Signal Processing, 56, 71-84.

Chen, M.-H., Shao, Q.-M., Ibrahim, J.G., 2000. Monte Carlo Methods in Bayesian Computation. Springer-Verlag, New York.

Cifarelli, D.M., Regazzini, E., 1978. Problemi statistici non parametrici in condizioni di scambiabilità parziale e impiego di medie associative: impiego di medie associative. Quaderni dell'Istituto di Matematica Finanziaria dell'Università di Torino, Serie III, Italia, pp. 1-36.

Cook, R.D., Weisberg, S., 1994. An Introduction to Regression Graphics. Wiley, New York.

Dalal, S.R., 1979. Dirichlet invariant processes and applications to nonparametric estimation of symmetric distribution functions. Stochastic Processes and their Applications, 9, 99-108.

De Iorio, M., Johnson, W.O., Müller, P., Rosner, G.L., 2009. Bayesian nonparametric nonproportional hazards survival modeling. Biometrics, 65, 762-771.

De Iorio, M., Müller, P., Rosner, G.L., MacEachern, S.N., 2004. An ANOVA model for dependent random measures. Journal of the American Statistical Association, 99, 205-215.

De la Cruz-Mesía, R., Quintana, F.A., Müller, P., 2007. Semiparametric Bayesian classification with longitudinal markers. Journal of The Royal Statistical Society Series C—Applied Statistics, 56, 119-137.

Dey, D., Müller, P., Sinha, D., (eds.) 1998. Practical Nonparametric and Semiparametric Bayesian Statistics. Springer-Verlag, New York.

Doss, H., 1984. Bayesian estimation in the symmetric location problem. Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete, 2, 127-147.

Duan, J.A., Guindani, M., Gelfand, A.E., 2007. Generalized spatial Dirichlet process models. Biometrika, 94, 809-825.

Dunson, D.B., Herring, A.H., 2006. Semiparametric Bayesian latent trajectory models. Technical report. Department of Statistical Science, Duke University, Durham.

Dunson, D.B., Xue, Y., Carin, L., 2008. The matrix stick-breaking process: flexible Bayes meta-analysis. Journal of the American Statistical Association, 103, 317-327.

Dunson, D.B., Park, J.-H., 2008. Kernel stick-breaking processes. Biometrika, 95, 307-323.

Elton, E.J., Gruber, M.J., 1995. Modern Portfolio Theory and Investment Analysis. Fifth edition. Wiley, New York.

Ferguson, T.S., 1973. A Bayesian analysis of some nonparametric problems. The Annals of Statistics, 1, 209-230.

Geisser, S., Eddy, W.F., 1979. A predictive approach to model selection. Journal of the American Statistical Association, 74, 153-160.

Gelfand, A.E., Dey, D.K., Chang, H., 1992. Model determination using predictive distributions with implementation via sampling-based methods. In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., (eds.). Bayesian Statistics 4. Oxford University Press, New York, pp. 147-167.

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

Ghosh, J.K., Ramamoorthi, R.V., 2003. Bayesian Nonparametrics. Springer, New York.

Giudici, P., Mezzetti, M., Muliere, P., 2003. Mixtures of products of Dirichlet processes for variable selection in survival analysis. Journal of Statistical Planning and Inference, 111, 101-115.

Hartigan, J.A., 1990. Partition Models. Communications in Statistics — Theory and Methods, 19, 2745-2756.

Iglesias, P.L., Orellana, Y., Quintana, F.A., 2009. Nonparametric Bayesian modelling using skewed Dirichlet processes. Journal of Statistical Planning and Inference, 139, 1203-1214.

Jara, A., García-Zattera, M.J., Lesaffre, E., 2007. A Dirichlet process mixture model for the analysis of correlated binary responses. Computational Statistics and Data Analysis, 51, 5402-5415.

MacEachern, S.N., 1999. Dependent Nonparametric Processes. American Statistical Association Proceedings of the Section on Bayesian Statistical Science, 50-55.

MacEachern, S.N., Müller, P., 1998. Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics, 7, 223-338.

Mira, A., Petrone, S., 1996. Bayesian hierarchical nonparametric inference for change-point problems. In Bernardo, J.M., Berger, J.O., Dawid, A.P., Smith, A.F.M., (eds.). Bayesian Statistics 5. Oxford University Press, New York, pp. 693-703.

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. Statistical Science, 19, 95-110.

Müller, P., Quintana, F.A., Rosner, G., 2010. A product partition model with regression on covariates. Working Paper, `http://www.mat.puc.cl/~quintana/bcwr.pdf` (submitted).

Müller, P., Rosner, G.L., De Iorio, M., MacEachern, S., 2005. A nonparametric Bayesian model for inference in related longitudinal studies. Journal of The Royal Statistical Society Series C—Applied Statistics, 54, 611-626.

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

Pitman, J., 1996. Some developments of the Blackwell-MacQueen urn scheme. In Ferguson, T.S., Shapeley, L.S., MacQueen, J.B., (eds.). Statistics, Probability and Game Theory. Papers in honor of David Blackwell. Volume 30. Institute of Mathematical Statistics, Hayward, California, pp. 245-268.

Quintana, F.A., 2006. A predictive view of Bayesian clustering. Journal of Statistical Planning and Inference, 136, 2407-2429.

Quintana, F.A., Iglesias, P.L., 2003. Bayesian clustering and product partition models. Journal of The Royal Statistical Society Series B—Statistical Methodology, 65, 557-574.

Rodríguez, A., Dunson, D.B., Gelfand, A.E., 2008. The nested Dirichlet process (with discussion). Journal of the American Statistical Association, 103, 1131-1154.

Rodríguez, A., Dunson, D.B., Gelfand, A.E., 2009. Nonparametric functional data analysis through Bayesian density estimation. Biometrika, 96, 149-162.

Sethuraman, J., 1994. A constructive definition of Dirichlet priors. Statistica Sinica, 4, 639-650.

Shahbaba, B., Neal, R., 2009. Nonlinear models using Dirichlet process mixtures. Journal of Machine Learning Research, 10, 1829-1850.

Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M., 2006. Hierarchical Dirichlet processes. Journal of the American Statistical Association, 101, 1566-1581.

Tiwari, R.C., 1988. Convergence of Dirichlet invariant measures and the limits of Bayes estimates. Communications in Statistics — Theory and Methods, 17, 375-393.

Walker, S.G., Damien, P., Laud, P.W., Smith, A.F.M., 1999. Bayesian nonparametric inference for distributions and related functions (with discussion). Journal of The Royal Statistical Society Series B—Statistical Methodology, 61, 485-527.