Bayesian Analysis (2013) **8**, Number 1, pp. 63–88

A Simple Class of Bayesian Nonparametric Autoregression Models

Maria Anna Di Lucca [*], Alessandra Guglielmi [†], Peter Müller [‡]
and Fernando A. Quintana [§]

Abstract. We introduce a model for a time series of continuous outcomes, that can be expressed as fully nonparametric regression or density regression on lagged terms. The model is based on a dependent Dirichlet process prior on a family of random probability measures indexed by the lagged covariates. The approach is also extended to sequences of binary responses. We discuss implementation and applications of the models to a sequence of waiting times between eruptions of the Old Faithful Geyser, and to a dataset consisting of sequences of recurrence indicators for tumors in the bladder of several patients.

Keywords: binary data, dependent Dirichlet process, hierarchical Bayesian model, latent variables, longitudinal data

1 Introduction

Consider a sequence of continuous random variables $\{Y_t : t \geqslant 1\}$. A very popular class of models for such time series data is autoregressive models that relate Y_t with a number of lagged terms $Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p}$. In the simplest scenario, it is assumed that $p = 1$, and that conditional on Y_{t-1}, $Y_t = \beta + \alpha Y_{t-1} + \epsilon_t$, for $t \geqslant 2$, where $\{\epsilon_t\}$ is a conveniently chosen sequence of residuals.

The assumptions made on $\{\epsilon_t\}$ are crucial for the specification and statistical analysis of AR(1) models. Consider, for instance, a white noise process $\epsilon_t \mid \sigma^2 \overset{\text{iid}}{\sim} N(0, \sigma^2)$. It then follows that, conditionally on σ^2, all random variables Y_t are normally distributed. While convenient, such assumptions may be too restrictive in many practical cases.

We present here a general framework for nonparametric autoregressive modeling, that can be easily modified to accommodate the special cases of binary and ordinal outcomes. The main idea is to provide an extension of the usual normal dynamic models. We focus on a joint model for $\{Y_t\}$, which can be equivalently done by considering the sequence of increasing conditionals $Y_t \mid Y_{t-1}, \ldots, Y_1$. To fix ideas, consider again the order-one dependence case, and assume that the conditional distribution $Y_t \mid Y_{t-1}, \ldots, Y_1$ depends only on Y_{t-1} for $t \geqslant 2$, and denote $Y_t \mid Y_{t-1} = y \sim F_y$ for any $t \geqslant 2$. We also assume homogeneity, in the sense that the distribution of $Y_t \mid Y_{t-1} = y$ does not change with t. We define a prior probability model for $\mathcal{F} = \{F_y : y \in \mathcal{Y}\}$.

[*] Karolinska Institutet, Stockholm, SWEDEN maria.di.lucca@ki.se
[†] Politecnico di Milano, Milano, ITALY alessandra.guglielmi@polimi.it
[‡] The University of Texas, Austin (TX), USA pmueller@utexas.edu
[§] Pontificia Universidad Católica de Chile, Santiago, CHILE quintana@mat.puc.cl

DOI:10.1214/13-BA803

We construct the desired family of random probability measures (RPMs) using the Dirichlet Process (DP) introduced by Ferguson (1973). It is well known that the DP is almost surely discrete, and that if $G \sim DP(M, G_0)$, a DP with total mass parameter $M > 0$ and baseline distribution G_0, then G can be represented as (Sethuraman 1994)

$$G(\cdot) = \sum_{h \geqslant 1} w_h \delta_{\theta_h}(\cdot), \tag{1}$$

where $\delta_\theta(\cdot)$ is a point-mass at θ, the weights follow a stick-breaking process, $w_h = \prod_{i<h}(1 - V_i)V_h$, with $V_h \overset{\text{iid}}{\sim} \text{Beta}(1, M)$, and the atoms $\{\theta_h\}_{h \geqslant 1}$ are such that $\theta_h \overset{\text{iid}}{\sim} G_0$.

MacEachern (1999, 2000) extended (1) by introducing the dependent Dirichlet process (DDP) as a collection of RPMs of the form $G_y = \sum_{h \geqslant 1} w_h(y)\delta_{\theta_h(y)}$, $y \in \mathcal{Y}$, such that each RPM G_y is marginally distributed according to a certain DP, *and* with the property that G_y varies smoothly with y. In particular, this implies that G_y and $G_{y'}$ should be correlated for $y \neq y'$ and that $G_y \to G_{y'}$ in some sense as $y \to y'$. A primary application of dependent models is to the case where y represents some (possibly vector-valued) covariate. De Iorio et al. (2004) explored an ANOVA formulation for categorical covariates, with weights that do not vary with y. Many other variations of DDPs have been proposed for various settings. De Iorio et al. (2009) applied the DDP to survival analysis, and Caron et al. (2008) and Rodríguez and ter Horst (2008) considered a time-dependent version of DDPs. Griffin and Steel (2006) proposed the order-based DDP, where weights are sorted according to the values of covariates. Other approaches that explicitly introduce covariate dependence in the weights include the kernel-stick breaking of Dunson and Park (2008), and the probit-stick breaking of Chung and Dunson (2011). See additional references in Hjort et al. (2010). An early development of dependent Dirichlet models appears in Cifarelli and Regazzini (1978), where the dependence on the covariates is introduced as a regression in the base measure of marginally Dirichlet process distributed random probability measures. Cruz-Marcelo et al. (2010) review and compare some covariate-dependent models. For an approach via parametric mixtures of autoregressive models with a common but unknown lag see Wood et al. (2011).

In practice, the discreteness associated with DPs and the DDP extension is inappropriate for the modeling of continuous data. A common way of addressing this limitation is by introducing an additional convolution with a continuous kernel, so that the resulting model can be expressed as a countable mixture of absolutely continuous distributions.

Our model uses the DDP. We propose modeling a sequence of continuous outcomes by means of a DDP with an additional normal kernel as a prior for the regression on lagged terms in an autoregression. In the general case, denoting $\boldsymbol{y} = (y_{t-1}, \ldots, y_{t-p})$, with possible values on $\mathcal{Y}$, we assume that $Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim F_{\boldsymbol{y}}$ where $F_{\boldsymbol{y}}$ is a location mixture of normals, with a DDP prior on the mixing measures. We further assume that the weights $\{w_h(\boldsymbol{y})\}$ and atoms $\{\theta_h(\boldsymbol{y})\}$ are defined in terms of two independent sequences of stochastic processes defined on $\mathcal{Y}$, as discussed in MacEachern (1999, 2000) and in Barrientos et al. (2012).

Of course, there are similarities between the proposed approach and those in the soaring literature on Bayesian nonparametric dependent models. For instance, the common framework between our models and those in Caron et al. (2008), Rodríguez et al. (2010), and Rodríguez and Dunson (2011) consists of the adoption of mixture models with (dependent) stick-breaking random probabilities as mixing measures, where the dependence is on spatial or temporal covariates, and the type of dependence for the observations (or latent processes, or hidden states) is Markovian.

In particular, Rodríguez and Dunson (2011) propose mixture models where the stick-breaking mixing measure has constant atoms, but covariate-varying weights. In contrast, we consider varying atoms. In Rodríguez and Dunson (2011) the weights are defined via a probit transformation of a Gaussian latent process that determines the stick-breaking ratios. Specifications of this model include autoregressive dependence in the Gaussian latent process and random effects models for different population distributions, and the latent variables defining probit weights share information across populations (but apparently there is no autoregression). In contrast, we propose a model where dependence on previous states or observations *and* on covariates defining population subgroups is considered in the state of the latent process at time t. Rodríguez et al. (2010) assume the stick-breaking framework for the mixing measure of the mixture model; however marginally their stick-breaking covariate-dependent processes share the same dependent distribution. The latent process there drives the selection of the atoms at each location/covariate. In contrast, all the stick-breaking processes we propose have marginal distributions depending on time and covariates. Caron et al. (2008) is one of the first papers where Dirichlet processes were used in the context of dynamic linear models; more recent works include Fox et al. (2011). In particular, they model the hidden/latent state at time t, which is a known linear combination of the hidden state at time $t-1$, plus an error which is distributed as a Dirichlet process mixture of Gaussian distributions. This approach is similar to one of our models, where we adopt a Gaussian latent process as well. In both cases, conditioning on the label which identifies the component in the mixture, the latent process at time t is Gaussian distributed, but the temporal assumptions on the means differ. Specifically, in Caron et al. (2008) the mean at time t is the sum of a linear transformation of the previous latent state and a linear transformation of the mean of the Gaussian component of the error. On the other hand, in our case, the mean is a linear transformation of the mean of the Gaussian component of the error, but the transformation matrix depends on the previous latent state.

Our proposed approach differs also from previous models for hidden Markov models (see Zucchini and MacDonald 2009, for instance) in that the dependence is directly either on previous outcomes or latent variables, rather than on a sequence of likelihood parameters. The type of dependence that we consider includes also that implied by the model in Lau and So (2008), who considered the case where the atoms are defined as linear combinations of lagged terms. As we will discuss later, posterior simulation for the proposed model can be carried out using standard techniques for DP mixtures.

The rest of this paper is organized as follows. Section 2 introduces the model, discussing some of its main features. Extensions to ordinal outcomes by means of a latent autoregressive process are also considered. Section 3 illustrates the model in two examples. In Section 3.1 the Old Faithful geyser dataset (Härdle 1991) is analyzed, while Section 3.2 addresses the bladder cancer example of Quintana and Müller (2004) and Giardina et al. (2011). In Section 3.1 we consider more complex alternative models for the Old Faithful geyser dataset, but conclude that no substantial differences are found when comparing with the previous results. We conclude with a discussion in Section 4.

2 The Model

2.1 Setup

The class of models that we consider is based on dependent Dirichlet processes (DDP). Given the vector $\boldsymbol{y} \in \mathcal{Y}$ of p lagged responses at times $t-1, \ldots, t-p$, we consider a model for the conditional distribution of Y_t given $\boldsymbol{y}$, i.e., we assume $p[Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y}] = F_{\boldsymbol{y}}$. We define a prior distribution on the collection of random probability measures $\mathcal{F} = \{F_{\boldsymbol{y}} : \boldsymbol{y} \in \mathcal{Y}\}$. To do so, we consider two sequences of independent stochastic processes, $\{V_h(\boldsymbol{y}) : y \in \mathcal{Y}, h \geqslant 1\}$ and $\{\theta_h(\boldsymbol{y}) : y \in \mathcal{Y}, h \geqslant 1\}$ on $\mathcal{Y}$ such that marginally for every $\boldsymbol{y} \in \mathcal{Y}$ and $h \geqslant 1$ we have $V_h(\boldsymbol{y}) \sim \text{Beta}(1, M)$ and $\theta_h(\boldsymbol{y}) \sim G_{0,\boldsymbol{y}}$. We also require continuity of trajectories for all these processes, which is satisfied when they are constructed using suitable families of copulas, as described in Barrientos et al. (2012). Setting $w_h(\boldsymbol{y}) = \prod_{j<h} V_j(\boldsymbol{y})(1 - V_h(\boldsymbol{y}))$ for $h \geqslant 1$, we define

$$G_{\boldsymbol{y}}(\cdot) = \sum_{h=1}^{\infty} w_h(\boldsymbol{y}) \delta_{\theta_h(\boldsymbol{y})}(\cdot). \tag{2}$$

The above choices guarantee that $G_{\boldsymbol{y}} \sim DP(M, G_{0,\boldsymbol{y}})$ for every $\boldsymbol{y} \in \mathcal{Y}$. The proposed model can then be expressed in the general case as

$$Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y}, m_t \sim N(m_t, \sigma^2), \qquad m_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim G_{\boldsymbol{y}}, \tag{3}$$

where $N(m, S^2)$ indicates a normal distribution with mean m and variance S^2.

We will discuss M and $G_{0,\boldsymbol{y}}$ for specific applications later. Note that the resulting conditional distribution for Y_t, given $(Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y}$ is a location mixture of normals, where the mixing measure G comes from the DP. By the discussion around (1), this implies that the model for $Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y}$ can be equivalently represented as

$$Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim \sum_{h \geqslant 1} w_h(\boldsymbol{y}) N(Y_t \mid \theta_h(\boldsymbol{y}), \sigma^2). \tag{4}$$

Assuming common weights, i.e. $w_h(\boldsymbol{y}) \equiv w_h$, model (3) can be further simplified to a countable mixture of autoregressive models, where each mixture component has a mean defined by its own stochastic process $\theta_h(\boldsymbol{y})$, $h \geqslant 1$. Moreover, the exact nature of the dependence on lagged terms $\boldsymbol{y}$ encoded in the random probability measure (2) is very general.

Despite the great generality of the proposed construction, it is in practice useful to resort to simple and manageable specifications. The main motivations for the simplification are easier implementation and parsimony. As we will later demonstrate, inference in the more general model is practically equivalent to inference under the proposed simplification. We just mentioned the simplification with a common-weights DDP (Barrientos et al. 2012), where it is assumed that $w_h(\boldsymbol{y}) = w_h$ for all $h \geqslant 1$, i.e., we assume the stochastic processes $\{V_h : h \geqslant 1\}$ to have constant trajectories on $\mathcal{Y}$. In addition, we may assume specific forms for the stochastic processes defining the atoms. For instance, we may assume that each $\theta_h(\boldsymbol{y})$ is a Gaussian process with mean and covariance functions expressed in terms of $\boldsymbol{y}$. Later in Section 3 we will consider the special case where each $\theta_h(\boldsymbol{y})$ corresponds to a polynomial function of the first lagged response, or to a linear combination of the p elements of $\boldsymbol{y}$. A related mixture model approach, with weights depending on previous responses was proposed in Müller et al. (1997). Lau and So (2008) considered similar types of models, where each atom included a formulation involving infinite mixtures of order-p linear autoregressions.

Observe that, when $\{G_{\boldsymbol{y}} : y \in \mathcal{Y}\}$ is a common-weights DDP, i.e., $w_h(\boldsymbol{y}) = w_h$, model (4) can be alternatively expressed as a DP mixture (DPM) model as follows. This is best seen in the marginal model for Y_t. Marginally, for each t,

$$Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim \int N(Y_t \mid \theta(\boldsymbol{y}), \sigma^2)\, dG(\theta), \qquad G \sim DP(M, G_{0,\boldsymbol{y}}), \tag{5}$$

where the above integration is interpreted as a marginalization over the stochastic process θ, which does not eliminate the dependence on $\boldsymbol{y}$ in (5).

As is usual in DPM models, computation is simplified by introducing latent variables and breaking the mixture (5). Since details are model-specific, we consider here as an illustration, the case where $p = 1$, $w_h(y) = w_h$, and $\theta_h(y) = \beta_h + \alpha_h y$, i.e., the common-weights DDP where the atoms correspond to linear trajectories of the first lagged response. We call this the AR(1)-DDP model. The model can alternatively be written as

$$\begin{aligned} Y_t \mid Y_{t-1} = y, (\beta_t, \alpha_t), \sigma^2 &\sim N(Y_t \mid \beta_t + \alpha_t y, \sigma^2), \\ (\beta_t, \alpha_t) \mid G \overset{\text{iid}}{\sim} G, &\qquad G \sim DP(M, G_0), \end{aligned} \tag{6}$$

Representation (6) provides a hierarchical definition and also highlights the fact that the dependence is introduced at the level of responses, and not in terms of the latent parameters $\{(\beta_t, \alpha_t)\}$. The Bayesian model specification would then be completed by assigning a prior distribution to σ^2 and a distribution for Y_1. Specific prior choices for relevant parameters will be later discussed in Section 3.

A simplified version of models (4) or (5) can be sometimes convenient from a computational viewpoint. This is achieved by truncating the infinite mixture implied by the DP to a finite mixture of a sufficiently large number of components, say H. This simplified model also implies a stick-breaking definition of the mixture weights $\{w_h(\boldsymbol{y}) : h = 1, \ldots, H\}$, with $w_h(\boldsymbol{y}) = \prod_{i<h}(1 - V_i(\boldsymbol{y}))V_h(\boldsymbol{y})$, for $h = 1, \ldots, H$, where each $V_h(\boldsymbol{y})$ has marginally a Beta$(1, M)$ distribution for $h < H$, and $V_H(\boldsymbol{y}) = 1$ for all

$\boldsymbol{y} \in \mathcal{Y}$, which guarantees $P(\sum_{h=1}^{H} w_h(\boldsymbol{y}) = 1) = 1$ for all $\boldsymbol{y} \in \mathcal{Y}$ (Ishwaran and James 2001). In the particular case of (6) and $w_h(\boldsymbol{y}) = w_h$ for all $h, \boldsymbol{y}$, and introducing latent mixture component indicators $\{r_t\}$, with $P(r_t = h) = w_h$, $1 \leqslant h \leqslant H$, the model becomes

$$\begin{aligned} Y_t \mid Y_{t-1} = y, r_t = h, \{(\beta_j, \alpha_j)\}, \sigma^2 &\sim N(\beta_h + \alpha_h y, \sigma^2), \\ P(r_t = h) = w_h, \qquad (\beta_h, \alpha_h) &\overset{\text{iid}}{\sim} G_0, \qquad h = 1, \ldots, H. \end{aligned} \tag{7}$$

Finally, it is worth pointing out some properties of the proposed model. There are no constraints to stationarity of the time series. In fact, the prior puts zero probability on stationarity, which would only arise as a special case of the auto-regression. Only the regressions $p(y_t \mid y_{t-1})$ are assumed to be constant through time, which contrasts with the method of Mena and Walker (2005) for constructing strictly stationary AR(1)-type models via nonparametric Bayes. The model inherits regularity properties of density estimation with a DP mixture of normals.

2.2 Binary Outcomes

The previous construction can easily be extended to the binary case, using ideas from Albert and Chib (1993), and model (5), or (4), on a latent scale. Assume Y_t is binary for all t, and introduce latent scores Z_t so that Y_t is defined by means of

$$Y_t = 1 \qquad \text{if and only if} \qquad Z_t > 0, \tag{8}$$

and, as consequence, $Y_t = 0$ iff $Z_t \leqslant 0$. The extension can now be stated as

$$Z_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim \int N(Z_t \mid \theta(\boldsymbol{y}), \sigma^2)\, dG(\theta), \qquad G \sim DP(M, G_{0,\boldsymbol{y}}). \tag{9}$$

In other words, the proposed continuous nonparametric autoregressive model is used to define the distribution of the latent score Z_t. Of course, given $\{Z_t\}$, the observations $\{Y_t\}$ are deterministic, which means that the desired distribution for the observed binary sequence is completely specified. Moreover, in terms of the observables, this model has the following probit-type structure:

$$Y_t \mid (Y_{t-1}, \ldots, Y_{t-p}) = \boldsymbol{y} \sim Be(p_t(\boldsymbol{y}, \sigma)), \quad p_t(\boldsymbol{y}, \sigma) = \int \Phi\left(\frac{\theta(\boldsymbol{y})}{\sigma}\right) dG(\theta). \tag{10}$$

Here Be(p) indicates a Bernoulli distributed (binary) random variable with success probability p. Note that a truncated version of (9) can be also considered, exactly as in the discussion leading to (7).

An alternative model specification for the binary case considers the sequence of conditional distributions entirely in the latent scale. By this we mean a nonparametric autoregressive model directly in terms of the latent sequence of scores $\{Z_t\}$. Parallelling (5) we consider

$$Z_t \mid (Z_{t-1}, \ldots, Z_{t-p}) = \boldsymbol{z} \sim \int N(Z_t \mid \theta(\boldsymbol{z}), \sigma^2)\, dG(\theta), \qquad G \sim DP(M, G_{0,\boldsymbol{z}}). \tag{11}$$

In this case, the model representation in terms of the observables is not as simple as before; in fact, conditioning on the parameters, the joint distribution of Y_t, $(Y_{t-1}, \dots, Y_{t-p})$, can be expressed as

$$\begin{aligned}
&\mathcal{L}(y_t, \boldsymbol{y}|par)\\
&\quad = \int \mathcal{L}(y_t, \boldsymbol{y}|Z_t = z_t, (Z_{t-1}, \dots, Z_{t-p}) = \boldsymbol{z}, par) f_{Z_t,(Z_{t-1},\dots,Z_{t-p})}(z_t, \boldsymbol{z}) dz_t d\boldsymbol{z}\\
&\quad = \int \left(\mathbb{I}_{(0,+\infty)}(z_t)\right)^{y_t} \left(\mathbb{I}_{(-\infty,0])}(z_t)\right)^{1-y_t} \mathcal{L}(\boldsymbol{y}|(Z_{t-1}, \dots, Z_{t-p}) = \boldsymbol{z}, par)\\
&\qquad\qquad\qquad \times f_{Z_t|(Z_{t-1},\dots,Z_{t-p})}(z_t|\boldsymbol{z}) f_{(Z_{t-1},\dots,Z_{t-p})}(\boldsymbol{z}) dz_t d\boldsymbol{z}.
\end{aligned} \tag{12}$$

From (10) and (12), it is clear that the two models are different. In particular, the former defines a Markovian process of order p on $\{Y_t\}$, unlike the latter. A formulation in terms of latent variables has an advantage though: it can be readily extended to ordinal outcomes. Indeed, assume Y_t is ordinal, with support $\{0, 1, \dots, \kappa - 1\}$ for some integer $\kappa \geqslant 2$. The binary case follows when $\kappa = 2$. Let $-\infty = \gamma_0 < \gamma_1 < \cdots < \gamma_{\kappa-1} < \gamma_\kappa = \infty$ be ordered cutoffs. We then assume the $\{Y_t\}$ to be defined through a latent sequence $\{Z_t\}$ by means of

$$Y_t = j \qquad \text{if and only if} \qquad \gamma_j < Z_t \leqslant \gamma_{j+1}, \qquad j = 0, 1, \dots, \kappa - 1. \tag{13}$$

Kottas et al. (2005) argue that the cutoffs can be fixed without loss of generality. In particular, when $\kappa = 2$ we take $\gamma_1 = 0$, and we have $Y_t = 1$ if and only if $Z_t > 0$, just as before. Conditional on the latent variables Z_t, the observations are deterministic, and it is therefore natural to consider exactly the same nonparametric autoregressive model (3), or (4), on this latent scale.

The model can easily be extended to multiple subjects. In particular, in the following section, we will fit model (8) with (9) or (11) using a common-weights model with $p = 1$ and atoms defined as simple linear trajectories, to binary data representing the recurrences of a disease in patients at different times. In this case, since the AR(1)-dependence is on the latent scores, we guess that this model could produce a good fit when the data consist of several short sequences, a situation that prevents us from using a higher-order dependence specification. On the other hand, we expect that model (3) will fit one single longer sequence of data well, since it assumes the order p Markovian property directly on the continuous responses.

3 Applications

In this section, we illustrate the class of models with applications to two datasets, the Old Faithful geyser (Sect. 3.1) and the bladder cancer (Sect. 3.2) dataset. First we summarize some implementation details for the following examples. In the Old Faithful geyser example, inference was implemented in R as Markov chain Monte Carlo (MCMC) posterior simulation, using the first 100,000 iterations as burn-in, and saving every 20-th iteration after burn-in. On the other hand, all inference for the latter example was coded in JAGS with the same burn-in, but with a larger thinning interval (100 iterations). In all cases, a posterior Monte Carlo sample of size $5,000$ was saved. Standard convergence diagnostics criteria such as those available in the R package CODA (Plummer et al. 2006) were applied to all parameters, indicating that convergence had been achieved.

3.1 Old Faithful Geyser

Inference under the AR(1)-DDP Model

We illustrate the proposed AR(1)-DDP model (6), or its simplified version (7), using the easily accessible Old Faithful geyser data set. For an extensive description of the data, see Härdle (1991) and Azzalini and Bowman (1990). Old Faithful is a geyser in the Yellowstone National Park in Wyoming, USA. The data consist of 299 pairs of measurements, referring to the time interval between the starts of successive eruptions, and the duration of the subsequent eruption. Here we only use the 272 data points that are readily available in the R dataset (Härdle 1991).

We focus on the waiting times $\{y_t,\, t = 1, \ldots, 272\}$ between the eruptions (y_t is the waiting time before the t-th eruption of the geyser), and fit model (7) to the dataset. Figure 1 plots y_t versus y_{t-1}. As the lagged data point y_{t-1} varies across the x-axis one can clearly recognize how the autoregressive model $p(Y_t \mid Y_{t-1} = y_{t-1})$ changes from a unimodal distribution around $y_{t-1} = 50$ to a bimodal distribution around $y_{t-1} = 80$. For later comparison, three pairs of vertical lines pick out three groups of data, with lagged waiting times y_{t-1} in the interval 50 ± 5, 65 ± 5 and 80 ± 5, respectively. We fitted model (7) using $H = 20$ and a total mass parameter of $M = 1$. The point masses in the DDP are assumed to be simple linear functions, $\theta_h(y) = \beta_h + \alpha_h y$. The base measure G_0 is a (bivariate) Gaussian distribution with independent components, with mean $(0, 0)^T$, and variances of the β_h and α_h components equal to 400 and 5, respectively. We consider fixed kernel variance $\sigma^2 = 25$, and alternatively an inverse gamma prior, $p(1/\sigma^2) = \mathrm{Ga}(2, 2)$, i.e., $E(1/\sigma^2) = 1$, and $\mathrm{Var}(1/\sigma^2) = 0.5$.

Figure 2 shows the posterior mean of the autoregressive model $F_y(\cdot)$ in (5) for y corresponding to a first-lag response of $\tilde{y}_1 = 50$, $\tilde{y}_2 = 65$, and $\tilde{y}_3 = 80$, respectively. Let $\bar{F}_y(\cdot)$ denote the posterior expectation

$$\bar{F}_y = E(F_y \mid \text{data}),$$

and let $\bar{f}_y(\cdot)$ denote the corresponding probability density function. The three panels of Figure 2 show $\bar{f}_{\tilde{y}_j}$, $j = 1, \ldots, 3$. For comparison the figures also show a kernel density estimator (dashed line) using a subset of the data with y_{t-1} within $\tilde{y}_j \pm 5$. Figure 3 shows the posterior mean of $f_{y_{t-1}}(\cdot)$ for $y_{t-1} = 80$ under the $1/\sigma^2 \sim \mathrm{Ga}(2, 2)$ prior, $M = 1$ and $H = 50$, together with 95% point-wise posterior credible bands.

Finally, we carried out sensitivity analysis to investigate variations in the prior assumptions. For example, we found that substantially increasing the value of fixed σ^2 beyond $\sigma^2 = 50$ lead to poorly mixing MCMC. On the other hand, increasing the prior mean for $1/\sigma^2$ by assuming $1/\sigma^2 \sim \mathrm{Ga}(2, 10)$ leads to only negligible changes in the

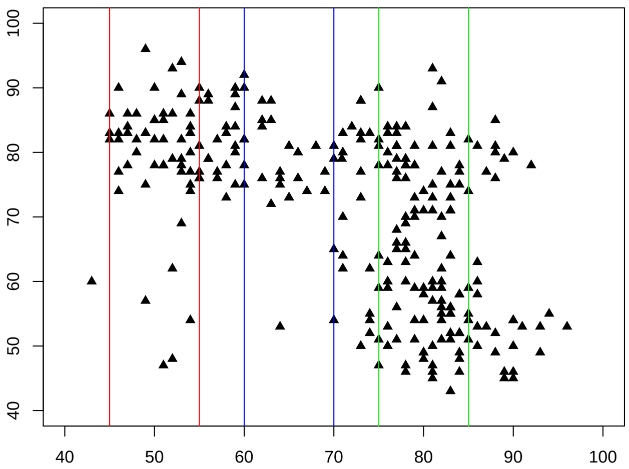

Figure 1: Old Faithful geyser: y_t versus y_{t-1}. The pairs of vertical lines pick out groups of data with y_{t-1} around 50 ± 5, 65 ± 5 and 80 ± 5, respectively. Notice the different form of the empirical distributions of y_t within each of the groups.

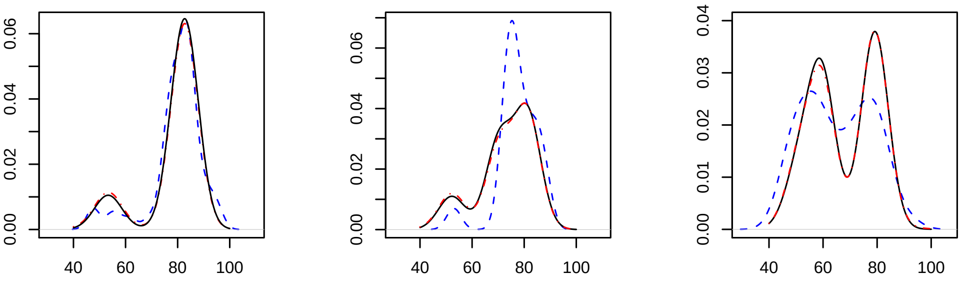

Figure 2: Old Faithful geyser data. Posterior means $\bar{f}_{y_{t-1}}(y_t)$ for $y_{t-1} = 50$ (left panel), 65 (central) and 80 (right). The continuous (black) line shows inference under the prior $1/\sigma^2 \sim \text{Ga}(2, 2)$, the (red) dash-dotted line shows inference under $\sigma^2 = 25$ (practically indistinguishable from the solid line), and the dashed (blue) shows a kernel density estimate.

inference. We investigated robustness with respect to prior parameters of the finite DP prior. Increasing H to 50 and M to 10, we observed little change in the estimated autoregressive models $\bar{f}_{y_{t-1}}(y_t)$. Figure 4 shows the estimates of $f_{y_{t-1}}(\cdot)$ for $y_{t-1} = 80$ under a variety of choices for M and H. The different curves are almost indistinguishable.

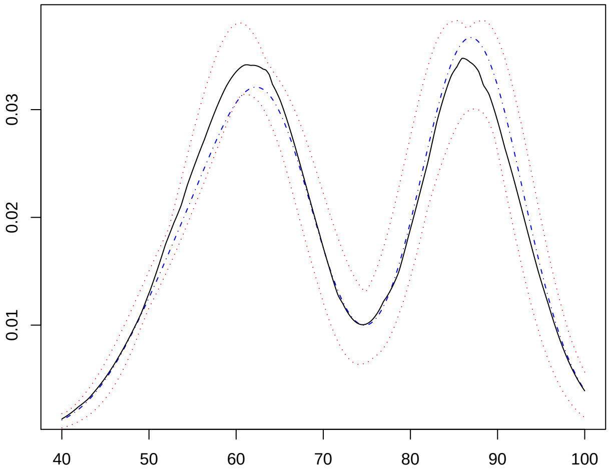

Figure 3: Old Faithful geyser data. Posterior mean $\bar{f}_{y_{t-1}}(\cdot)$ for $y_{t-1} = 80$ (blue semi-dashed line), together with pointwise 95% credible bands (red dotted lines) and median (solid black line).

Model Variations

In the construction of the proposed AR(1)-DDP model (6) we made a sequence of simplifying assumptions. The question arises whether a more general model without some of these simplifications could lead to a practically meaningful extension, trading parsimony with more flexibility. The answer, of course, is always dependent on the particular application. In a sequence of alternative implementations we investigate this question for the particular example of the Old Faithful Geyser example.

We first considered the truncation to the finite DDP. We implemented an alternative model as in (6), without approximating the DP random measure G to finitely many, H, point masses. Figure 5 shows the resulting posterior means of $f_{y_{t-1}}(\cdot)$, for $y_{t-1} = 50, 65$ and 80. Compared with Figure 2 we find virtually the same inference. Another major simplification was the use of simple polynomials for the trajectories $\theta_h(\boldsymbol{y})$ to replace more flexible alternatives, such as a Gaussian process (GP) prior for $\theta_h(\boldsymbol{y})$. In the special case of lag $p = 1$ regression the more general GP model is easy to implement. In particular, we considered an Ornstein-Uhlenbeck (OU) process, a GP with covariance function $\text{cov}[\theta(s), \theta(t)] = \tau^2\rho^{|s-t|}$, for $0 < \rho \leqslant 1$. The attraction of the OU process is the Markovian nature of the process that greatly simplifies posterior computation. We thus implemented (4) with common weights w_h and point masses $\theta_h(y) = b + a_h y + \text{OU}(\rho, \tau^2)$, where $\text{OU}(\rho, \tau^2)$ denotes the OU process with parameters

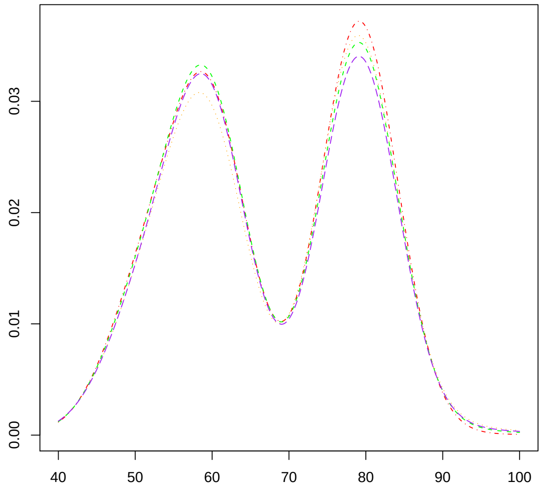

Figure 4: Old Faithful geyser data. Posterior means of $f_{y_{t-1}}(\cdot)$, for $y_{t-1} = 80$. The (red) semi-dashed line is the estimate under $M = 1$, $H = 20$, the (orange) dotted line for $M = 10$, $H = 20$, the (green) dashed line is for $M = 1$, $H = 50$ and the (blue) long dashed line is for $M = 10$, $H = 50$. The estimates are almost indistinguishable.

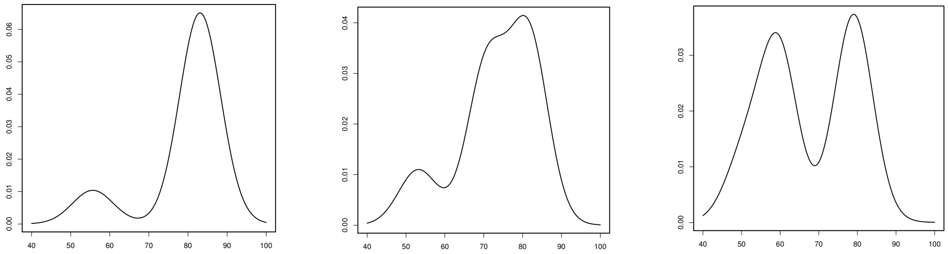

Figure 5: Old Faithful geyser data. Posterior means $\bar{f}_{y_{t-1}}(\cdot)$ under the AR(1)-DDP model (6) with $H = \infty$, i.e., without truncation, for $y_{t-1} = 50$ (left), $y_{t-1} = 65$ (center) and $y_{t-1} = 80$ (right).

(ρ, τ^2) in the covariance function. A priori, ρ was assumed uniform on (0,1), while $1/\tau$ was given a Ga(0.1, 0.1) prior, while $b \sim N(110, 1)$ and $a_h \mid a_0 \sim N(a_0, \sigma_a^2)$, with $a_0 \sim N(-0.5, 1)$, $\sigma_a^{-2} \sim \text{Ga}(0.1, 0.1)$. For a fair comparison we used the same setup as above, now saving $10,000$ iterations for the inference. The estimated distributions $\bar{f}_y(\cdot)$ for $y = 50, 65$ and 80 (not shown) after the same number of iterations look very different from Figure 2, including a unimodal distribution $\bar{f}_{50}(\cdot)$ and $\bar{f}_{65}(\cdot)$ lacking the secondary mode around $y = 50$. We conclude a serious lack of convergence with the same number of iterations, which may be related to the fact that the GP model is over-parametrized. This leads us to prefer the parsimonious implementation of the AR(1)-DDP.

Finally we considered a variation with varying weights $w_h(\boldsymbol{y})$. Similar to what was proposed in Rodríguez and Dunson (2011), we used a logit model to replace the beta distributed fractions V_h in (2), with $\text{logit}(V_h(y)) = \eta_{h1} + \eta_{h2} y$. We continue to use a finite truncation with $H = 20$. The resulting estimates $\bar{f}_y(\cdot)$ are shown in Figure 6. We see no practically meaningful differences in the inference. We therefore recommend the more parsimonious model (6).

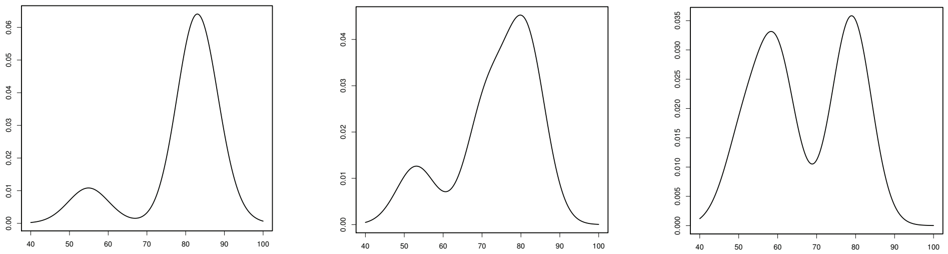

Figure 6: Old Faithful geyser data. Posterior means $\bar{f}_{y_{t-1}}(\cdot)$ under the AR(1)-DDP model (6) with $H = \infty$, but with varying weights $w_h(\boldsymbol{y})$ for $y_{t-1} = 50$ (left), $y_{t-1} = 65$ (center) and $y_{t-1} = 80$ (right).

3.2 Bladder Cancer Data

To illustrate the nonparametric autoregressive approach for latent scores, as described in the previous section, we consider many short sequences of binary variables. The dataset is part of a bladder cancer study conducted in the USA by the Veterans Administration Cooperative Urological Research Group (VACURG). The purpose of the study was to compare the effectiveness of three treatments (placebo, pyridoxine, and topical thiotepa) in preventing recurrence of Stage I bladder cancer (Byar et al. 1977).

Many authors, including Quintana and Müller (2004), have analyzed this dataset. The study conducted by VACURG enrolled $m = 81$ patients with up to a maximum of $n_i = 12$ observations taken every three months for each patient. We restrict ourselves to only patients grouped into treatment (thiotepa) and placebo: group T (36 subjects) and group P (45 subjects). See Davis and Wei (1988) for the original dataset. Each observation records an indicator of recurrence of bladder cancer tumors, *i.e.* $y_{it} = 1$ if an increased number of tumors was detected at time t for patient i, and $y_{it} = 0$ otherwise, where $i = 1, \ldots, m$ denotes individuals and $t = 1, \ldots, n_i$ denotes the measurement time for each individual i. We record treatment information as a binary covariate. Denote $x_i = 0$ if patient i belongs to the P group, and $x_i = 1$ otherwise, for $i = 1, \ldots, m$. The binary r.v.'s Y_{it} are modeled as

$$Y_{it} = 1 \qquad \text{if and only if} \qquad Z_{it} > 0, \quad i = 1, \ldots, m, \; t = 1, \ldots, n_i. \tag{14}$$

We compare two different classes of models for the latent variables, one as described in (11), and the other as in (9), using the AR(1)-DDP specification, which defines a Markovian process on each $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{in_i})'$. The covariate x_i will be included in the autoregression, together with the past value of the latent score, in both models. We mention that these models can also be considered as nonparametric generalizations of earlier parametric work in Giardina et al. (2011), where more details on data construction can be found. However, the description of the models here is self-contained.

AR(1)-latent model

We consider the following AR(1)-DDP model on the latent variables $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$, where $\mathbf{Z}_i = (Z_{i1}, \ldots, Z_{in_i})'$:

$$\begin{aligned} Z_{it}|Z_{i\,t-1} = z_{i\,t-1}, x_i, \beta_0, \beta_1 &\sim \int_{\mathbb{R}^2} N(\beta_0 + \beta_1 x_i + \alpha_1 z_{i\,t-1} + \alpha_2 x_i z_{i\,t-1}, \sigma^2) dG(\alpha_1, \alpha_2), \\ G &\sim DP(M, G_0), \end{aligned} \tag{15}$$

for $i = 1, \ldots, m$, $t = 2, \ldots, n_i$, where G_0 is a bivariate distribution. To complete the model definition, a prior distribution for the initial latent variables $\{Z_{i1}, i = 1, \ldots, m\}$ must be given. We further assume that, conditionally on the latent variables $\mathbf{Z}_1, \ldots, \mathbf{Z}_m$, the vectors $\mathbf{Y}_1, \ldots, \mathbf{Y}_m$ are independent, with binary components as in (14), where each Z_{it} follows (15). This is not a probit-type model. Analogously to (12), the joint distribution of all observables given parameters and covariates is

$$\begin{aligned} &\mathcal{L}(\boldsymbol{y}_1, \ldots, \boldsymbol{y}_m | par) \\ &= \int \prod_{\{y_{it}=1\}} \mathbb{I}_{[0,+\infty)}(z_{it}) \prod_{\{y_{it}=0\}} \mathbb{I}_{(-\infty,0)}(z_{it}) \prod_{i=1}^{m} \left(f\left(z_{i1}\right) \prod_{j=2}^{n_i} f\left(z_{ij}|z_{it-1}\right) dz_{i1} \cdots dz_{in_i} \right). \end{aligned}$$

Moreover, observe that β_0 in (15) is the intercept of the regression model, and β_1 represents the treatment effect on the response variable. A finite approximation of equation (15) is

$$\begin{aligned} Z_{it} \mid Z_{it-1} = z_{it-1}, r_{it} = h, par &\sim N(\beta_0 + \beta_1 x_i + \alpha_{1h} z_{it-1} + \alpha_{2h} x_i z_{it-1}, \sigma^2), \\ \{r_{it}\}_i &\overset{\text{iid}}{\sim} \text{according to } P(r_{it} = h) = w_h, \\ (\alpha_{1h}, \alpha_{2h}) &\overset{\text{iid}}{\sim} G_0, \quad h = 1, \ldots, H, \end{aligned} \tag{16}$$

i.e. the distribution of $Z_{it} - (\beta_0 + \beta_1 x_i)$, given $Z_{i\,t-1} = z_{i\,t-1}$, is a location-mixture of Gaussian distributions with fixed variance, where the mixing distribution is a truncated single-p (constant weights as defined in Section 2.1) DDP. Since the Z_{it}'s are latent variables representing the observations according to (14), $\boldsymbol{Z}_i$ and $C\boldsymbol{Z}_i$ yield the same distribution of $\boldsymbol{Y}_i$, whatever positive constant C we choose. Identification may be achieved by fixing σ^2; the interested reader could refer to Giardina et al. (2011), Section 3.1, for a discussion about identifiability issues.

Finally note that model (15) is a slight generalization of (11). The dependence in the random mixing distribution includes both the previous latent variable z_{it-1} and the covariate x_i. Also, although the treatment variable x is the only covariate in this application, the model could easily be adapted for inference with more covariates if desired.

Regarding σ^2, β, G_0 and M, we assume that $\sigma^2 = 0.25$, $(\beta_0, \beta_1)' \sim \mathcal{N}_2((\beta_{00}, \beta_{01})', V_\beta)$ with $\beta_{00} = \beta_{01} = 0$ and $V_\beta = I_2$, the bivariate identity matrix, and $G_0(\alpha_{1h}, \alpha_{2h})$ is determined by

$$\begin{aligned} (\alpha_{1h}, \alpha_{2h})' \mid (\alpha_{01}, \alpha_{02}) &\overset{iid}{\sim} \mathcal{N}_2((\alpha_{01}, \alpha_{02})', V_\alpha),\ h = 1, 2, \dots \\ (\alpha_{01}, \alpha_{02})' &\sim \mathcal{N}_2((\alpha_{001}, \alpha_{002})', V), \end{aligned} \tag{17}$$

where $\alpha_{001} = \alpha_{002} = 0$, $V = 10I_2$, $V_\alpha = I_2$, and $(\beta_0, \beta_1)'$ and $\{(\alpha_{1h}, \alpha_{2h})'\}$ are independent. The model is completed by assuming two different prior distributions for Z_{i1} for T and P patients as follows:

$$\begin{aligned} Z_{i1}|x_i, \mu_{x_i} &\sim \mathcal{N}\left(\mu_{x_i}, \sigma_1^2\right), \quad i = 1, \dots, m,\ x_i = 0, 1, \\ \mu_1 &\sim \text{ logistic-beta}(a, b) \\ \mu_0 &= \mu_1 + D,\ \text{ where } D \sim \text{ log-normal}(\mu_D, \sigma_D), \\ \mu_1, D &\text{ independent.} \end{aligned} \tag{18}$$

Prior (18) was proposed for the first latent variables so as to ensure that $\mu_0 \geqslant \mu_1$ almost surely, since we assume that the patients under treatment will have a lower probability of recurrence. The logistic-beta(a, b) is assumed for μ_1, i.e. $(1 + \exp(-\mu_1))^{-1}$ is distributed according to a Beta(a, b), where the specific choices of hyperparameter values will be discussed later. Here a more standard assumption for μ_1 would be $\Phi(\mu_1/\sqrt{\sigma_1})$ distributed as a Beta(a, b). In practice, however, this requires evaluating Φ^{-1}, which is notorious for being a numerically unstable operation, unlike the case implied by the logistic assumption. Here the parametrization of the log-normal distribution is such that $\mathbb{E}(D) = \mathrm{e}^{\mu_D + \sigma_D^2/2}$ and $\mathrm{Var}(D) = (\mathrm{e}^{\sigma_D^2} - 1)\mathrm{e}^{2\mu_D + \sigma_D^2}$. We have fixed $a = b = 3$, $\mu_D = -1$, $\sigma_D = 1$.

A simpler alternative to (15) is to assume

$$\begin{aligned} Z_{it}|Z_{i\,t-1} = z_{i\,t-1}, x_i, \beta_0, \beta_1 &\sim \int_{\mathbb{R}^2} N(\beta_0 + \beta_1 x_i + \alpha_1 z_{i\,t-1}, \sigma^2) dG(\alpha_1), \\ G &\sim DP(M, G_0), \end{aligned} \tag{19}$$

for $i = 1, \dots, m$, $t = 2, \dots, n_i$, i.e. the "slope" α_1 is constant over the two groups of patients (P and T). Note that G_0 denotes a univariate distribution for (19), fixed here as the corresponding marginal of that in (17). We will refer to this latter model as AR(1)-latent 3P, since the regression parameters included here are only 3, while the former model with 4 regression parameters will be obviously referred to as AR(1)-latent 4P. Summary posterior inferences and posterior distributions about regression parameters for both models can be found in Table 1 and in Figure 7. Unless otherwise stated, these estimates and those in the following tables were computed with $H = 30$ and $M = 1$. Observe that the marginal posterior distributions of β_0, β_1 and μ_1 are concentrated on the negative numbers. This means that the baseline probability of tumor recurrence is less than 0.5 for both groups, and that treated patients have lower baseline probability than the ones in the placebo group. The posterior of D confirms that there is a difference between the two treatments.

	$M = 1$				$M \sim U(0.5, 10)$		$M \sim$ trunc-$\mathcal{IG}(2, 2)$	
	3P		4P		4P		4P	
	mean	sd	mean	sd	mean	sd	mean	sd
β_0	-0.2171	0.0410	-0.2221	0.0439	-0.2206	0.0433	-0.2207	0.0429
β_1	-0.1348	0.0749	-0.1547	0.1299	-0.1301	0.1038	-0.1286	0.0995
α_{01}	0.0798	3.1894	0.3576	0.9326	0.4703	0.9552	0.4128	0.9386
α_{02}	-	-	-0.2642	0.9937	-0.1596	0.9635	-0.1969	0.9562
μ_1	-0.4275	0.0890	-0.4240	0.0876	-0.4252	0.0883	-0.4249	0.0882
D	0.1475	0.0811	0.1483	0.0816	0.1482	0.0815	0.1465	0.0809
K	4.0524	1.5484	4.2164	1.6007	3.7666	1.6754	4.2758	1.6719
M	-	-	-	-	0.8411	0.3331	1.1115	0.2748

Table 1: Posterior means and standard deviations of the parameters of the AR(1)-latent models 3P and 4P.

AR(1)-latent-Y model

As a second model, we assume a finite approximation of (9):

$$\begin{aligned} Z_{it} \mid Y_{it-1} = y_{it-1}, r_{it} = h, par &\sim N(\beta_0 + \beta_1 x_i + \alpha_{1h} y_{it-1} + \alpha_{2h} x_i y_{it-1}, \sigma^2), \\ \{r_{it}\}_i &\overset{\text{iid}}{\sim} \text{according to } P(r_{it} = h) = w_h, \\ (\alpha_{1h}, \alpha_{2h})' &\overset{\text{iid}}{\sim} G_0, \quad h = 1, \ldots, H, \end{aligned} \tag{20}$$

i.e. the distribution of $Z_{it} - (\beta_0 + \beta_1 x_i)$, given $Y_{i\,t-1} = y_{i\,t-1}$ is a location-mixture of Gaussian distributions with fixed variance, where the mixing distribution is a truncated single-p (constant weights) DDP. The prior for the "regression" parameters and the initial latent variables Z_{i1}'s is as in (17)-(18). Of course, the meaning of β_0, β_1, α_{1h} and α_{2h} is completely different. But we can meaningfully compare the resulting predictive recurrence probabilities of the two classes of models. Summary posterior inferences and posterior distributions for the regression parameters for model AR(1)-latent-Y as specified in (20) are reported in Table 2 and Figure 8. Note that, as was the case for the AR(1)-latent models, the marginal posterior distributions for β_0, β_1 and μ_1 are all concentrated on the negative numbers.

Comparison between models

For comparison purposes, we report estimates of the predictive probabilities for the two models considered, corresponding to additional measurement for already observed patients (Table 3), and for new patients (see Figure 9). Both are reported separately for patients under treatment groups P and T.

Figure 9 displays predicted recurrence probabilities for a new placebo (upper set of lines) and a new treated patient (lower set of lines). We observe no significant differences

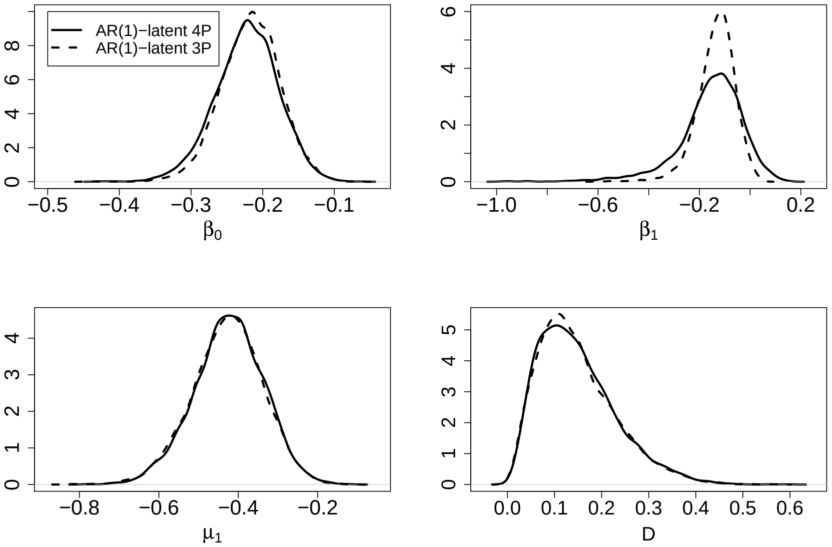

Figure 7: Posterior marginal distribution of AR(1)-latent models parameters when $H = 30$ and $M = 1$ for models 4P (continuous) and 3P (dashed).

in these predictions between the three considered models, for both types of patients.

We have also examined robustness of these results to choices of H and M. When increasing H to 50, we found no substantial differences on the predictives for new patients or on the posterior distributions of K, the number of components in the mixtures (see Figure 10, first row). On the other hand, the predictive probabilities of additional measurements for already observed patients are very robust, and for this reason we have not reported them.

As far as the total mass parameter M is concerned, we have assumed it random, i.e. M has a Uniform prior on the interval $(0.5, 10)$, or inverse-gamma with parameter $(2, 2)$ with support $(0.5, +\infty)$ ($M = 0.5 + X$, with $1/X \sim \text{gamma}(2, 2)$). The total mass parameter was assumed bounded away from zero due to numerical instability of the posterior simulation algorithms, as implemented in JAGS. In any case, these two choices imply quite different prior assumptions for M. Table 1 reports the regression parameter estimates for model 4P-AR(1)-latent, while Figure 11 displays some of these posterior distributions. Even if the posteriors of M, under the two priors, are different (see Figure 12), the posteriors of the number of clusters K in Figure 10 (c)-(d) are quite similar.

Figures 13 and 14 display the posterior distributions of the regression parameters, and of the number K of components in the mixture, respectively, for AR(1)-latent-Y model when M is random (as before). We note that posterior distribution of M is

	$M = 1$		$M \sim U(0.5, 10)$		$M \sim$ trunc-$\mathcal{IG}(2, 2)$	
	mean	sd	mean	sd	mean	sd
β_0	-1.0797	0.0881	-1.0818	0.0891	-1.0816	0.0891
β_1	-0.4039	0.1483	-0.4009	0.1532	-0.4007	0.1497
α_{01}	0.8921	0.9371	0.8870	0.9370	0.8851	0.9219
α_{02}	0.2114	0.9766	0.2234	0.9521	0.2136	0.9411
μ_1	-0.7454	0.1656	-0.7479	0.1675	-0.7465	0.1667
D	0.2143	0.1361	0.2173	0.1376	0.2157	0.1373
K	4.3454	1.6996	3.9334	1.8607	4.8270	2.0100
M	-	-	0.8615	0.3582	1.1450	0.3103

Table 2: Posterior means and standard deviations of the parameters of the AR(1)-latent-Y model, when $H = 30$, $M = 1$ and σ^2=1.

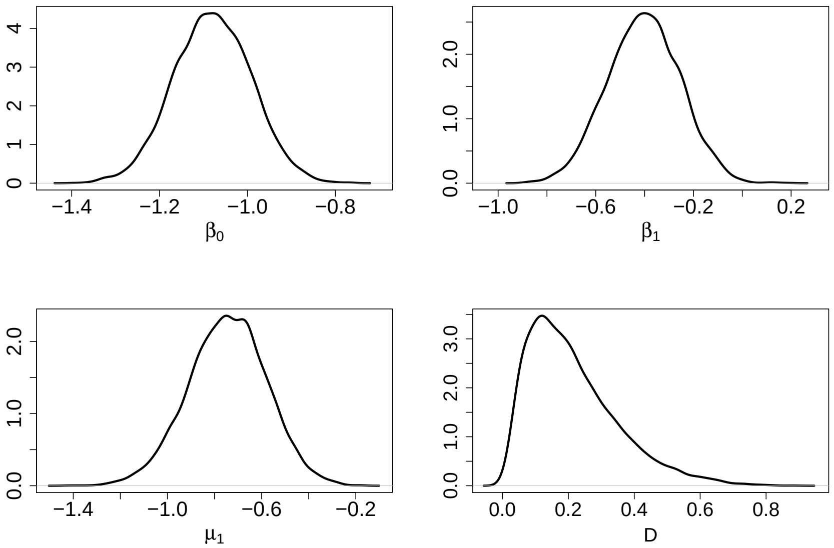

Figure 8: Posterior marginal distributions of AR(1)-latent-Y model parameters when $H = 30$ and $M = 1$, for $\sigma^2 = 1$.

similar to Figure 12, and therefore, not shown here. This suggests that inference on M is not affected by the specific choice of autoregressive dependence, i.e. latent variables or responses. It is also worth mentioning that under these priors, the predictive probabilities for "new" placebo and treated patients shown in Figure 15 are very robust to all the model choices explored here. Finally, we incidentally remark that, for the AR(1)-latent models only, the MCMC algorithm may fail to converge if choosing a value of σ^2 larger than 0.25, as traceplots corresponding to latent variables for some patients with too many zero responses diverged to $-\infty$. This suggests that, in this case, the latent variables, due to the lack of identifiability problem mentioned earlier, need to be tightly controlled.

	AR(1)-latent				AR(1)-latent-Y	
	3P		4P		4P	
	Prob.	MCse	Prob.	MCse	Prob.	MCse
$Y_{9,9}$	0.5412	0.0070	0.5226	0.0071	0.4550	0.0070
$Y_{16,10}$	0.0934	0.0041	0.1036	0.0043	0.1354	0.0048
$Y_{23,11}$	0.5698	0.0070	0.5532	0.0070	0.4530	0.0070
$Y_{33,13}$	0.1038	0.0043	0.0992	0.0042	0.1392	0.0049
$Y_{60,10}$	0.0688	0.0036	0.0680	0.0036	0.0744	0.0037
$Y_{71,9}$	0.0590	0.0033	0.0498	0.0031	0.0724	0.0037
$Y_{74,12}$	0.0532	0.0032	0.0526	0.0032	0.0684	0.0036

Table 3: Estimates of the predictive probabilities of a new measurements for subjects 9, 16, 23, 33 (PLACEBO) and 60, 71, 74 (TREATMENT), including Monte Carlo standard errors.

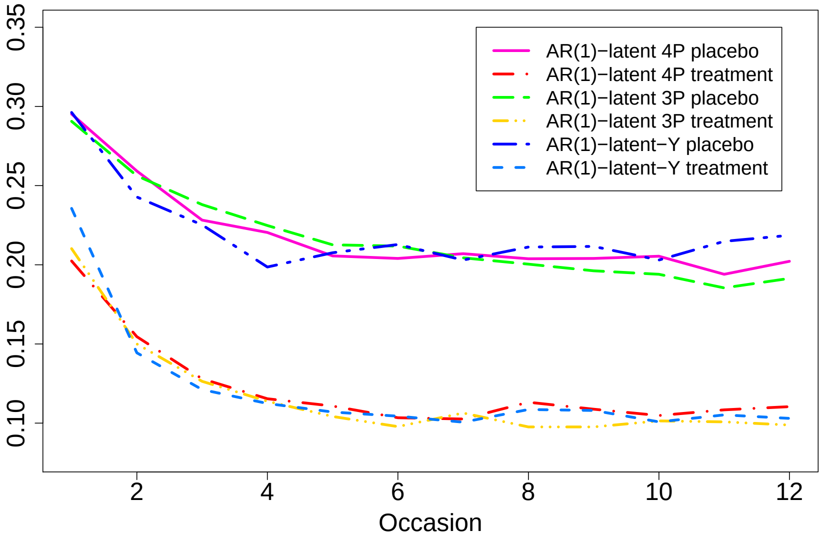

Figure 9: Predicted recurrence probabilities for a new placebo and a new treated patient under different models.

To summarize, though the predictions for this particular dataset are quite robust to the proposed models, we point out that the two approaches are actually very different. As we mentioned before, the AR(1)-latent model is not Markovian while the AR(1)-latent-Y model is. Moreover, the former presents similarities with the nonparametric linear dynamic model by Caron et al. (2008), where they assume that the hidden state at time t is a known linear combination of the hidden state at time $t-1$ plus an error which is distributed as a Dirichlet process mixture of Gaussian distributions. The AR(1)-latent-Y model does not seem to fit such structure. Of course, from a computational point of view, the latter model yields a better mixing of the MCMC algorithm. Setting

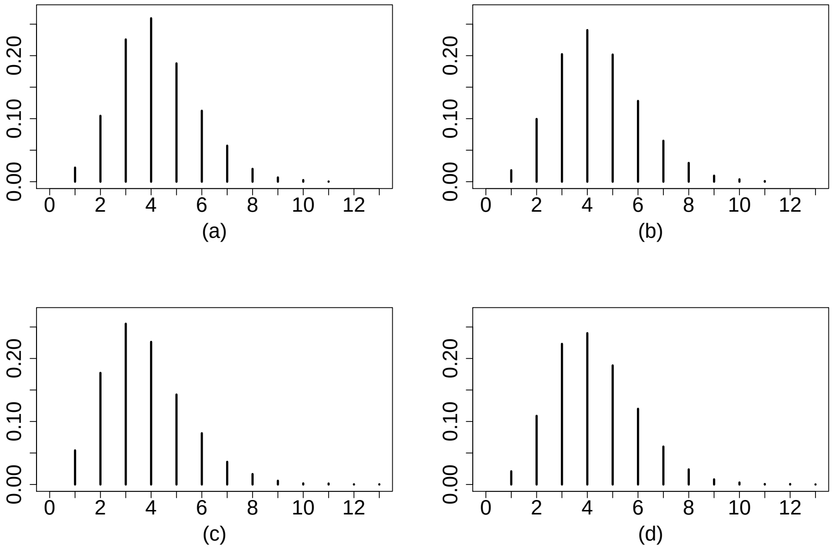

Figure 10: Posterior distributions of the number of components K in the mixture in the AR(1)-latent 4P model when $M = 1$ and $H = 30$ (a) or $H = 50$ (b), and when $H = 30$ and M is $U(0.5, 10)$- (c) or $trunc - \mathcal{IG}(2, 2)$-distributed (d).

aside the problem of inference on the random measure G, which was not of interest for these data, this seems to be the only practical difference between the two models.

4 Discussion

We have presented a framework for nonparametric modeling of either one or multiple time series of observations. The model is based on dependent Dirichlet processes (DDPs), where the dependence is on lagged responses. The proposal can be characterized as using non-parametric Bayesian density regression, i.e., fully nonparametric regression, to define the regression on lagged data in an autoregressive model. For the sake of clarity, we have limited the presentation to simple implementations of the nonparametric regression. A simplification of the models to a finite number of mixture components was also discussed. The framework can be also applied to binary or ordinal responses, where the key is to apply the model to sequences of latent variables defining the observations. Applications to both types of data were considered.

We characterized and introduced the model as a DDP. However, it is worth reiterating that the model can alternatively be written as a simple DP mixture. We showed this representation in (5). Recognizing this representation greatly simplifies computation. We still prefer to think of the model as a special case of the DDP be-

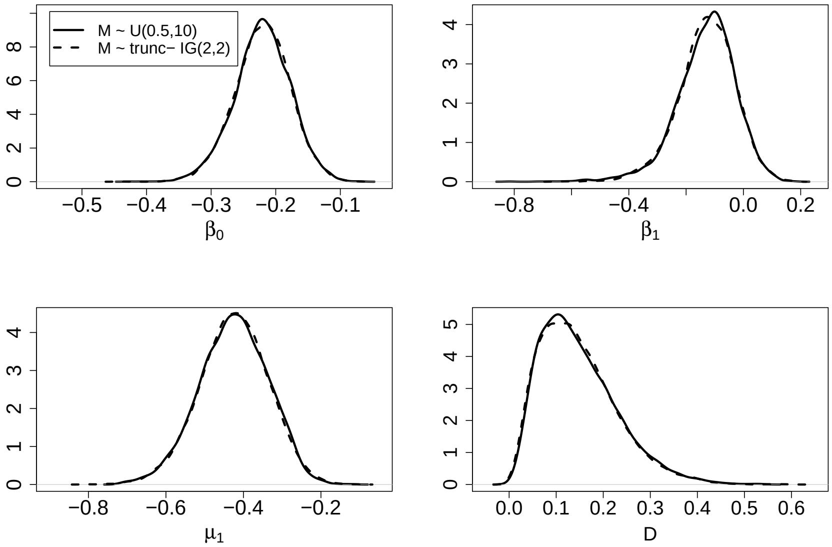

Figure 11: Posterior marginal distribution of 4P-AR(1)-latent model when $H = 30$ and M is $U(0.5, 10)$- (continuous) or $trunc - \mathcal{IG}(2, 2)$- distributed (dashed).

cause this highlights the nature of the problem as inference about a family of random probability measures $\mathcal{F} = \{F_y\}$.

The class of models considered here can adopt many different forms. The linear dependence discussed in Section 2.1 is just one example. Higher order polynomials or other nonlinear functions of lagged terms can be accommodated under the general framework, for instance b-splines (Eilers and Marx 1996). Another option consists of including dependence on $p \geqslant 2$ lagged terms, i.e., a nonparametric AR(p) model. Although computational convenience is achieved by linearity assumptions on the autoregression coefficients, the model for the point masses $\theta_h(\boldsymbol{y})$ in the DDP can be arbitrarily specified. In practice, however, one would like to retain some interpretability of the mixture components, which poses some practical restriction on the way lagged terms enter the model.

In the discussion and the examples we did not focus on inference for the random mixing measure G in (6), as this is usually not an inference target. However, if such inference were desired it could easily be obtained as part of the MCMC. Even without the constraint to the finite DP with finite H, one could report inference on G by means of the slice sampler proposed in Walker (2007) and Kalli et al. (2011).

Finally, extensions to the current approach include further comparison between different ways of specifying the dependence on lagged terms, assessing the number of lagged terms to include in the autoregression, and multivariate formulations of the autoregres-

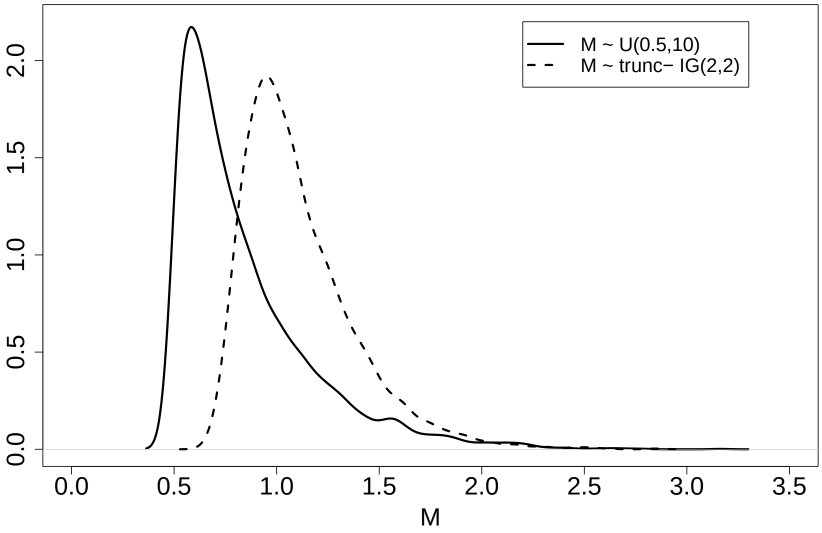

Figure 12: Posterior marginal distribution of the total mass parameter M in the AR(1)-latent 4P model when $H = 30$ and M is $U(0.5, 10)$- (continuous) or $trunc - \mathcal{IG}(2, 2)$-distributed (dashed).

sive models. These and other topics are the subject of current research.

References

Albert, J. H. and Chib, S. (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88: 669–679. 68

Azzalini, A. and Bowman, A. W. (1990). "A look at some data on the Old Faithful Geyser." *Journal of the Royal Statistical Society, Series C-Applied Statistics*, 39: 357–365. 70

Barrientos, A. F., Jara, A., and Quintana, F. A. (2012). "On the support of MacEachern's dependent Dirichlet processes." *Bayesian Analysis*, 7: 277–310. 64, 66, 67

Byar, D. P., Blackard, C., and the Veterans Administration Cooperative Urological Research Group (1977). "Comparisons of placebo, pyridoxine, and topical thiotepa in preventing recurrence of Stage I bladder cancer." *Urology*, 10: 556–561. 74

Caron, F., Davy, M., Doucet, A., Duflos, E., and Vanheeghe, P. (2008). "Bayesian inference for linear dynamic models with Dirichlet process mixtures." *IEEE Transactions on Signal Processing*, 56: 71–84. 64, 65, 80

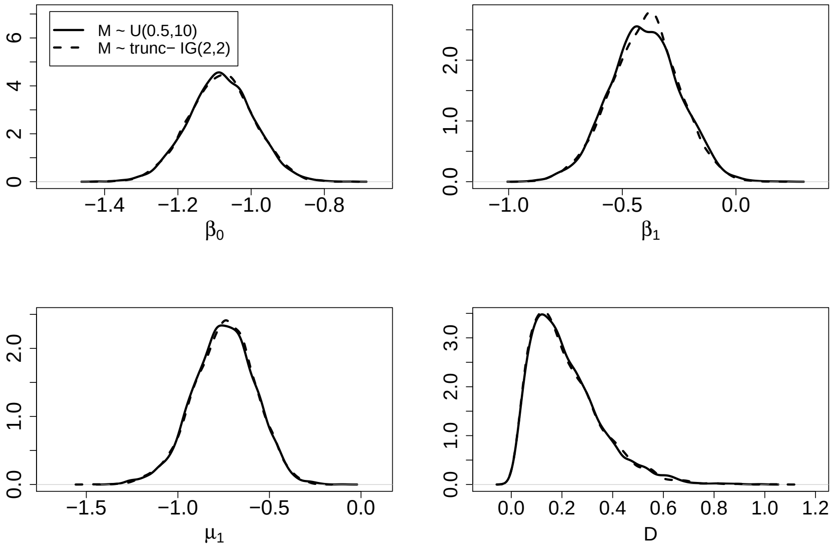

Figure 13: Posterior marginal distribution of the regression parameters in the AR(1)-latent-Y model when $H = 30$ and M is $U(0.5, 10)$- (continuous) or $trunc - \mathcal{IG}(2, 2)$-distributed (dashed).

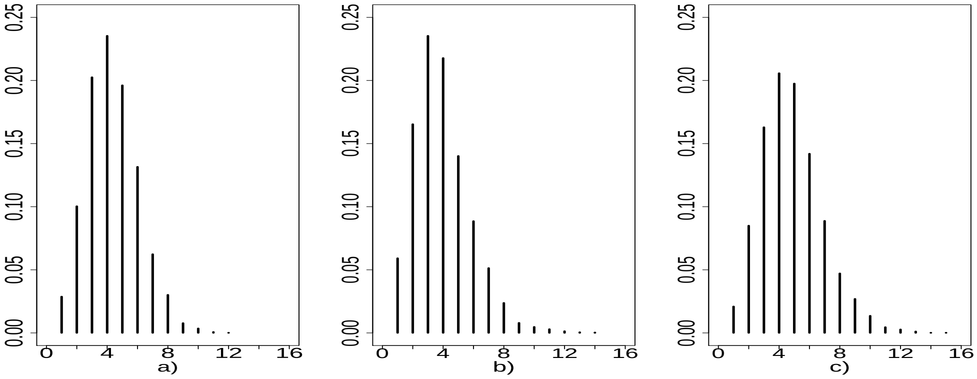

Figure 14: Posterior distributions of the number of components K in the mixture in the AR(1)-latent-Y model when M is equal to 1 (left), or $U(0.5, 10)$- (central) or $trunc - \mathcal{IG}(2, 2)$- distributed (right panel).

Chung, Y. and Dunson, D. B. (2011). "The local Dirichlet process." *Annals of the Institute of Statistical Mathematics*, 63: 59–80. 64

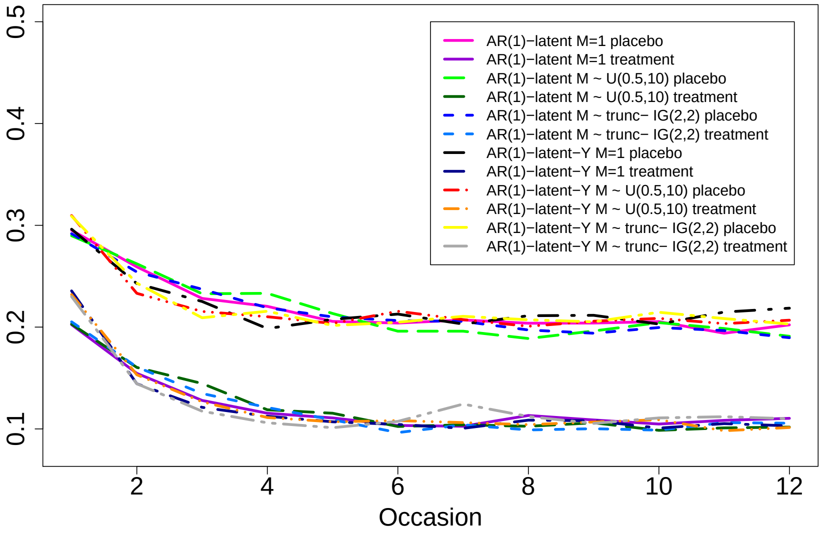

Figure 15: Predicted recurrence probabilities for a new placebo (continuous line) and a new treated patient (dashed line), under different models with M random.

Cifarelli, D. M. and Regazzini, E. (1978). “Problemi statistici non parametrici in condizioni di scambiabilità parziale: impiego di medie associative.” Technical Report Quaderni Istituto di Matematica Finanziaria, Serie III, n.12, Universitá di Torino. 64

Cruz-Marcelo, A., Rosner, G. R., Müller, P., and Stewart, C. (2010). “Modeling Covariates with Nonparametric Bayesian Methods.” Technical Report Available at SSRN: http://ssrn.com/abstract=1576665. 64

Davis, C. S. and Wei, L. J. (1988). “Nonparametric Methods for Analyzing Incomplete Nondecreasing Repeated Measurements.” *Statistics in Medicine*, 44: 1005–1018. 74

De Iorio, M., Johnson, W. O., Müller, P., and Rosner, G. L. (2009). “Bayesian nonparametric nonproportional hazards survival modeling.” *Biometrics*, 65: 762–771. 64

De Iorio, M., Müller, P., Rosner, G. L., and MacEachern, S. N. (2004). “An ANOVA model for dependent random measures.” *Journal of the American Statistical Association*, 99: 205–215. 64

Dunson, D. B. and Park, J. H. (2008). “Kernel stick-breaking processes.” *Biometrika*, 95: 307–323. 64

Eilers, P. H. C. and Marx, B. D. (1996). “Flexible Smoothing with B-splines and penalties.” *Statistical Science*, 11: 89–121. 82

Ferguson, T. S. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1: 209–230. 64

Fox, E., Sudderth, E. B., Jordan, M. I., and Willsky, A. S. (2011). "Bayesian nonparametric inference for switching dynamic linear models." *IEEE Transactions on Signal Processing*, 59: 1569–1585. 65

Giardina, F., Guglielmi, A., Quintana, F. A., and Ruggeri, F. (2011). "Bayesian first order auto-regressive latent variable models for multiple binary sequences." *Statistical Modelling*, 11: 471–488. 66, 75

Griffin, J. E. and Steel, M. (2006). "Order-based dependent Dirichlet processes." *Journal of the American Statistical Association*, 101: 179–194. 64

Härdle, W. (1991). *Smoothing Techniques: With Implementation in S*. New York: Springer. 66, 70

Hjort, N., Holmes, C., Müller, P., and Walker, S. G. (eds.) (2010). *Bayesian Nonparametrics*. Cambridge, UK: Cambridge University Press. 64

Ishwaran, H. and James, L. F. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96: 161–173. 68

Kalli, M., Griffin, J. E., and Walker, S. G. (2011). "Slice Sampling Mixture Models." *Statistics and Computing*, 21: 93–105. 82

Kottas, A., Müller, P., and Quintana, F. A. (2005). "Nonparametric Bayesian Modeling for Multivariate Ordinal Data." *Journal of Computational and Graphical Statistics*, 14: 610–625. 69

Lau, J. W. and So, M. K. P. (2008). "Bayesian mixture of autoregressive models." *Computational Statistics and Data Analysis*, 53: 38–60. 65, 67

MacEachern, S. N. (1999). "Dependent nonparametric processes." In *ASA Proceedings of the Section on Bayesian Statistical Science*. Alexandria, VA: American Statistical Association. 64

— (2000). "Dependent Dirichlet processes." Technical report, Department of Statistics, The Ohio State University. 64

Mena, R. H. and Walker, S. G. (2005). "Stationary autoregressive models via a Bayesian nonparametric approach." *Journal of Time Series Analysis*, 26: 789–805. 68

Müller, P., West, M., and MacEachern, S. N. (1997). "Bayesian models for non-linear autoregressions." *Journal of Time Series Analysis*, 18: 593–614. 67

Plummer, M., Best, N., Cowles, K., and Vines, K. (2006). "CODA: Convergence Diagnosis and Output Analysis for MCMC." *R News*, 6: 7–11. 70

Quintana, F. A. and Müller, P. (2004). "Optimal Sampling for Repeated Binary Measurements." *Canadian Journal of Statistics*, 32: 73–84. 66, 74

Rodríguez, A. and Dunson, D. B. (2011). "Nonparametric Bayesian models through probit stick-breaking processes." *Bayesian Analysis*, 6: 145–178. 65, 74

Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2010). "Latent stick-breaking processes." *Journal of the American Statistical Association*, 105: 647–659. 65

Rodríguez, A. and ter Horst, E. (2008). "Bayesian dynamic density estimation." *Bayesian Analysis*, 3: 339–366. 64

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. 64

Walker, S. G. (2007). "Sampling the Dirichlet mixture model with slices." *Communications in Statistics: Simulation and Computation*, 36: 45–54. 82

Wood, S., Rosen, O., and Kohn, R. (2011). "Bayesian mixtures of autoregressive models." *Journal of Computational and Graphical Statistics*, 20: 174–195. 64

Zucchini, W. and MacDonald, I. L. (2009). *Hidden Markov Models for Time Series*. London: Chapman & Hall. 65

Acknowledgments

We thank Annalisa Cadonna for running the JAGS code of the bladder cancer application. Alessandra Guglielmi was partially funded by MIUR, grant 2008MK3AFZ, and she would like to thank people at Departamento de Estadística at PUC, Chile, for their kind hospitality. Fernando Quintana was partially funded by grant FONDECYT 1100010. Peter Müller was supported in part by NIH/NCI R01CA075981.