Statistica Sinica **20** (2010), 1507-1527

BAYESIAN NONPARAMETRIC MODELLING WITH THE DIRICHLET PROCESS REGRESSION SMOOTHER

J. E. Griffin and M. F. J. Steel

University of Kent and University of Warwick

Abstract: In this paper we discuss implementing Bayesian fully nonparametric regression by defining a process prior on distributions that depend on covariates. We consider the problem of centring our process over a class of regression models, and propose fully nonparametric regression models with flexible location structures. We also introduce an extension of a dependent finite mixture model proposed by Chung and Dunson (2011) to a dependent infinite mixture model and propose a specific prior, the Dirichlet Process Regression Smoother, which allows us to control the smoothness of the process. Computational methods are developed for the models described. Results are presented for simulated and for real data examples.

Key words and phrases: Nonlinear regression, nonparametric regression, model centring, stick-breaking prior.

1. Introduction

Standard regression techniques assume that observations $y_1, \ldots, y_n$ observed at $x_1, \ldots, x_n$, respectively, can be modelled as

$$y_i = g(x_i, \gamma) + \epsilon_i, \qquad \epsilon_i \sim \mathrm{N}(0, \sigma^2), \tag{1.1}$$

where $\mathrm{N}(\mu, \sigma^2)$ denotes a normal distribution with mean μ and variance σ^2. Misspecification of either the mean function or the error distribution may lead to biased estimates of the regression coefficients γ. For example, the error distribution may be heteroscedastic or have a changing numbers of modes. The model can be robustified by assuming that the errors $\epsilon_1, \ldots, \epsilon_n$ are modelled nonparametrically. In the Bayesian literature, an initial step in this direction was taken by Bush and MacEachern (1996) who assume that $g(x, \gamma)$ is linear in the covariates and that $\epsilon_i \sim F$ where F is modelled by a Dirichlet process mixture of normals (DPMN). The error distribution is flexibly modelled but does not depend on the covariates. Leslie, Kohn, and Nott (2007) extend this approach to allow heterogeneity in the error distribution by assuming that $\epsilon_i = \sigma^2(x_i)\eta_i$, where $\sigma^2(x_i)$ is a flexible functional form, and the η_i are drawn from a DPMN. This model allows one aspect of the error distribution to depend on the covariates, but other

aspects such as multi-modality are modelled through a single distribution. The present paper is concerned with inference in the more general model

$$y_i = g(x_i, \gamma) + m(x_i) + \epsilon_i, \quad \epsilon_i \sim k(\epsilon_i|\psi_i), \quad \psi_i \sim F_{x_i}, \quad F_{x_i} \sim \Pi(H, \xi), \quad (1.2)$$

where $m(x_i)$ has a flexible nonparametric prior, such as a Gaussian process (e.g., Rasmussen and Williams (2006)), $k(\epsilon_i|\psi_i)$ is a probability distribution with parameter ψ_i and F_{x_i} is a distribution indexed by the covariates x_i, while $\Pi(H, \xi)$ is a density regression prior centred over the distribution H with parameters ξ (so that $\mathrm{E}[F_{x_i}(B)] = H(B)$ for all B and x_i). The distribution of ϵ_i, marginalising over ψ_i, is a mixture of distributions of type k, which defines a flexible form. Recently developed density regression priors express the distribution as

$$F_x \stackrel{d}{=} \sum_{i=1}^{\infty} p_i(x)\delta_{\theta_i(x)}, \quad (1.3)$$

where $p_i(x)$ is an infinite vector of probabilities such that $\sum_{i=1}^{\infty} p_i(x) = 1$. MacEachern (2000) discusses the problem of specifying a prior for a collection of distribution $\{F_x\}_{x\in\mathcal{X}}$ for which the marginals F_x follow Dirichlet processes (Ferguson (1973)). His single-p model assumes that $p_i(x) = p_i$ for all x and the changing distribution is modelled through a process for $\theta_1, \theta_2, \ldots$, which has been applied to group experimental data (De Iorio et al. (2004)), spatial problems (Gelfand, Kottas and MacEachern (2005)) and quantile regression (Kottas and Krnjajic (2009)). Several authors have instead considered a regression model for $p_i(x)$. Dunson, Pillai and Park (2007) define dependent measures by allowing each measure to be an average of several unknown, latent distributions. Specifying weights that change with covariates allows the distributions to change. Griffin and Steel (2006) (hereafter denoted by GS), Dunson and Park (2008), and Reich and Fuentes (2007) exploit the stick-breaking construction of random measures. GS use permutations of the breaks to induce dependence whereas Dunson and Park (2008) and Reich and Fuentes (2007) introduce a kernel to allow breaks to change over x. Chung and Dunson (2011) define a specific example of this type of process, the Local Dirichlet process, which takes the kernel to be a ball of radius ϕ_i around x. In this case, it is simple to relate the correlation in $p_i(x)$ to the properties of the kernel and the choice of the prior for ϕ_i. One purpose of this paper is to extend this class of models to the nonparametric case where we have an infinite number of atoms. The methods developed in GS lead to prior distributions centred over a single distribution. This paper discusses a method of centring over a non-trivial model (in other words, allowing the centring distribution to depend on the covariates). Thus, we allow for two sources of dependence

on covariates: dependence of the random probability measure on the covariates, and dependence of the centring distribution on the same (or other) covariates. Besides extra flexibility, this provides a framework for assessing the adequacy of commonly used parametric models. We also propose a new density regression prior which allows us to control its smoothness.

The paper is organised in the following way. Section 2 introduces the idea of centring a nonparametric prior over a parametric model and develops a framework for assessing the suitability of the parametric model for observed data. Section 3 introduces a class of nonparametric priors for regression, including the Dirichlet Process Regression Smoother (DPRS). Section 4 briefly discusses computational methods for DPRS-based models, with more details of the implementation available in Appendix B of the online supplement to the paper at `http://www.stat.sinica.edu.tw/statistica`. Section 5 illustrates the use of these models, and a final section concludes. Proofs are provided in Appendix A of the online supplement.

2. Centring Dependent Nonparametric Models

Nonparametric priors for an unknown distribution, F, are usually centred over a parametric distribution, H, by setting $\mathrm{E}[F(B)] = H(B)$ for measurable sets B. It is useful to extend this idea to centre dependent nonparametric priors over parametric regression models. The nonparametric prior will model aspects of the conditional distribution that are not well captured by our parametric centring model and, by centring, we can use prior information elicited for the parametric model directly. If the parameters controlling the departures of the fitted distribution from the centring model are given priors, then we can assess how well our parametric model describes the data. The covariates used in the parametric and nonparametric parts of the model are not required to be the same, but x will generally denote the union of all covariates.

Definition 1. A nonparametric model is *centred* over a parametric model, with parameters η, if the prior predictive distribution of the nonparametric model at a covariate value x conditional on η coincides with the parametric model for each covariate value.

In this paper we concentrate on centring over the standard regression model with normally distributed errors in (1.1), where $\eta = (\gamma, \sigma^2)$. Centring then implies defining a nonparametric prior for the distribution of ϵ_i whose predictive distribution is a zero mean, normal distribution. We first consider centring the model in (1.2) when $m(x) = 0$ for all x. A suitable centring model lets $\psi_i = \mu_i$ and takes

$$k(\epsilon_i|\psi_i) = \mathrm{N}(\epsilon_i|\mu_i, a\,\sigma^2), \quad H \sim \mathrm{N}(0, (1-a)\sigma^2),$$

where $0 < a < 1$, which is denoted by **Model 1(a)**. Clearly, the prior predictive distribution of $y_i - g(x_i, \gamma)$ is a $\mathrm{N}(0, \sigma^2)$ distribution. The parameterisation of the model is discussed by Griffin (2010), who pays particular attention to prior choice. Many distributional features, such as multi-modality, are more easily controlled by a rather than ξ in (1.2). Small values of a suggest that the nonparametric modelling is crucial. A uniform prior distribution on a supports a wide range of departures from a normal distribution. The full model, denoted by **Model 1(b)**, can be centred as

$$y_i - g(x_i, \gamma) - m(x_i) \sim \mathrm{N}(\mu_i, a(1-b)\sigma^2), \quad H \sim \mathrm{N}(0, (1-a)(1-b)\sigma^2),$$

where $m(x)$ follows a Gaussian process prior with $m(x_1), \ldots, m(x_n)$ jointly normally distributed with constant mean 0, and the covariance of $m(x_i)$ and $m(x_j)$ is $b\sigma^2\rho(x_i, x_j)$ with $\rho(x_i, x_j)$ a proper correlation function. A popular choice of correlation function is the flexible Matérn class (see e.g., Stein (1999)) for which

$$\rho(x_i, x_j) = \frac{1}{2^{\tau-1}\Gamma(\tau)}(\zeta|x_i - x_j|)^\tau \mathcal{K}_\tau(\zeta|x_i - x_j|),$$

where $\mathcal{K}_\tau$ is the modified Bessel function of order τ. This process is q times mean squared differentiable if and only if $q < \tau$, and ζ acts as a range parameter. The parameter b can be interpreted as the proportion of residual variability explained by the nonparametric Gaussian process estimate of $m(x)$. If we consider the prior predictive with respect to F_x we obtain the centring model $y_i \sim \mathrm{N}(g(x_i, \gamma) + m(x_i), (1-b)\sigma^2)$, whereas if we integrate out both F_x and $m(x)$ with their priors we obtain $y_i \sim \mathrm{N}(g(x_i, \gamma), \sigma^2)$.

Dependence on the covariates x enters in three different ways: it is used in the parametric regression function $g(x, \gamma)$, it intervenes in the process $m(x)$, and the distribution of the means μ_i depends on x_i through the dependent random probability measure, Π. The distribution of y given x is marginally a standard nonparametric mixture of normals model.

Model 1(a), in combination with a density regression prior, can capture nonlinear relationships between the errors and regressors through changing weights $p_i(x)$. The following proposition shows the autocorrelation structure of the kth moment of the distribution F_x.

Proposition 1. *Suppose that* $F_x = \sum_{i=1}^\infty p_i(x)\delta_{\theta_i}$ *where* $\theta_1, \theta_2, \theta_3, \ldots \overset{i.i.d.}{\sim} H$. *If* $\mu_x^{(k)} = \sum_{i=1}^\infty p_i(x)\theta_i^k$, *then*

$$\mathrm{Corr}\left(\mu_x^{(k)}, \mu_y^{(k)}\right) = \frac{\sum_{i=1}^\infty \mathrm{E}[p_i(x)p_i(y)]}{\sum_{i=1}^\infty \mathrm{E}[p_i^2(x)]}.$$

Therefore, these priors imply that the autocorrelation structure does not change with k. This seems unsatisfactory for many applications. For example, the model would resemble a homoscedastic nonparametric regression model if the variance had a large correlation over the range of the data but we might want a different autocorrelation for the mean. The relatively crude correlation structure described in Proposition 1 can lead to undersmoothing of the posterior estimates of the distribution. In particular, the posterior mean $\mathrm{E}[y|x]$ will often have a step form typical of piecewise constant models. Introducing a Gaussian process to define Model 1(b) addresses this problem by allowing the first moment of $y_i - g(x_i, \gamma)$ to have a different correlation structure from all higher moments.

Models 1(a) and 1(b) illustrate an important advantage of centring over a model: it provides a natural way to distinguish between the parametric dependence on covariates, captured by $g(x, \gamma)$, and the nonparametric dependence, modelled through F_x and $m(x)$. Thus, by choosing $g(x, \gamma)$ appropriately, we may find that the nonparametric modelling is less critical. This will be detected by a large value of a and a small value of b, and will allow us to use the model to evaluate interesting parametric specifications. Note that the interpretation of γ is non-standard in this model, since $\mathrm{E}[y_i|F_{x_i}, \gamma, x_i]$ is merely distributed around $g(x_i, \gamma)$ and $P(\mathrm{E}[y_i|F_{x_i}, \gamma, x_i] = g(x_i, \gamma)) = 0$ if y_i is a continuous random variable and H is absolutely continuous, which occurs in a large proportion of potential applications. The predictive mean $\mathrm{E}[y_i|\gamma, x_i]$ still equals $g(x_i, \gamma)$, however. The prior uncertainty about this predictive mean increases as confidence in the centring model (usually represented by one of the parameters in ξ) decreases.

One solution to this identifiability problem is to follow Kottas and Gelfand (2001) who fix the median of ϵ_i, which is often a natural measure of centrality in nonparametric applications, to be 0. If we assume that the error distribution is symmetric and unimodal, then median and mean regression coincide (if the mean exists). An alternative, wider, class of error distributions, introduced by Kottas and Gelfand (2001) to regression problems, is the class of unimodal densities with median zero (see Kottas and Krnjajic (2009) for extensions to quantile regression). Extending the symmetric version of this model to our context yields **Model 2**:

$$k(\epsilon_i|\psi_i) \sim \mathrm{U}\left(\left[-\sigma\sqrt{(1-b)\psi_i}, \sigma\sqrt{(1-b)\psi_i}\right]\right), \quad H = \mathrm{Ga}\Big(\frac{3}{2}, \frac{1}{2}\Big),$$

where $\mathrm{U}(A)$ is the uniform distribution on the set A and $\mathrm{Ga}(3/2, 1/2)$ represents a Gamma distribution with shape $3/2$ and mean 3. This choice of H corresponds to a normal centring distribution and the model is centred exactly like Model 1(b).

3. A Bayesian Density Smoother

This section develops a measure-valued stochastic process that can be used as a prior for Bayesian nonparametric inference when we want to infer distributions, $\{F_x\}_{x\in\mathcal{X}}$, where $\mathcal{X}$ is the space of covariates. It is stationary in the sense that all marginal distributions F_x follow a Dirichlet process. We restrict attention to Dirichlet process-based models since these methods dominate in the literature and our approach follows these ideas naturally. The stick-breaking representation of the Dirichlet process (Sethuraman (1994)) is given by (1.3) without the dependence on x and with $p_i = V_i \prod_{j<i}(1-V_j)$, where $V_1, V_2, \ldots$ are i.i.d. with $V_i \sim \text{Be}(1, M)$ while $\theta_1, \theta_2, \ldots$ are i.i.d. from some distribution H. A covariate-dependent stick-breaking process can be defined by only including a subset of these V's at each x. If these subsets are similar for similar covariates, then the distribution will change in a controlled way and effective inference will be positive. We assume that the position θ_i does not depend on x. A similar idea was implemented by GS using the πDDP prior. The process is also a non-trivial generalisation of the independently proposed Local Dirichlet Process (Chung and Dunson (2011)) from finite to infinite mixture models. This is achieved by introducing extra parameters $t_1, t_2, \ldots$ that determine the order in which points enter the stick-breaking construction.

Definition 2. Let $S(\phi)$ be a shape characterized by a parameter ϕ, and (ϕ, t) be a Poisson process with intensity $f(\phi)$ with associated marks (V_j, θ_j) that are i.i.d. realisations of $\text{Be}(1, M)$ and H, respectively. We define

$$F_x = \sum_{\{i|x\in S(\phi_i)\}} \delta_{\theta_i} V_i \prod_{\{j|x\in S(\phi_j),\, t_j<t_i\}} (1-V_j),$$

and say $\{F_x|x\in\mathcal{X}\}$ follows a **Subset-based Dependent Dirichlet Process**, represented as S-DDP(M, H, f, S), where f is a non-negative function for which $\int \text{I}(x\in S(\phi))f(\phi)d\phi > 0$ for all $x\in\mathcal{X}$.

Each marginal process F_x follows a Dirichlet process. This results from the distribution of $V_1, V_2, \ldots$ and the infinite number of atoms included at each x. This can be seen as follows: the number of points included in F_x is Poisson distributed with mean $\int\int \text{I}(x\in S(\phi))f(\phi)d\phi\, dt$. This number is almost surely infinite if the condition $\int \text{I}(x\in S(\phi))f(\phi)d\phi > 0$ for all $x\in\mathcal{X}$ is met. Any atom θ_j only appears in the stick-breaking representation of F_x at points x which belong to a subset of $\mathcal{X}$, and this allows atoms to "appear" and "disappear". The construction is general and could potentially be applied to arbitrary spaces if suitable shapes are available. However, as is common with nonparametric methods, care needs to be taken in higher dimensions due to the "curse of dimensionality". Realisations of the distributions are discontinuous as are all moments. However,

conditional prior predictive distributions and moments of y given x are continuous. The dependence between distributions at different locations s and v can be easily measured using the correlation of $F_s(B)$ and $F_v(B)$ for any measurable set B.

Theorem 1. *If F follows an S-DDP(M, H, f, S) process, then*

$$Corr(F_s, F_v) = \frac{2}{M+2} E\left[\sum_{i=1}^{\infty} B_i \left(\frac{M}{M+1}\right)^{\sum_{j=1}^{i-1} A_j} \left(\frac{M+1}{M+2}\right)^{\sum_{j=1}^{i-1} B_j}\right], \quad (3.1)$$

where $A_i = I(s \in S(\phi_i)$ or $v \in S(\phi_i))$ and $B_i = I(s \in S(\phi_i)$ and $v \in S(\phi_i))$.

If $s, v \in S(\phi_i)$ for all i then the correlation is 1. If s and v do not both fall in any $S(\phi_i)$ then the correlation is 0. Suppose that the probability of observing a shared element in each subsequence is constant given two covariate values s and v and equals, say, $p_{s,v}$. Then the following holds.

Theorem 2. *The correlation between F_s and F_v is*

$$Corr(F_s, F_v) = 2\frac{[(M+1)/(M+2)]p_{s,v}}{1+[M/(M+2)]p_{s,v}} = \frac{2(M+1)p_{s,v}}{2+M(1+p_{s,v})}$$

where, for any k, $p_{s,v} = P(s, v \in S(\phi_k) | s \in S(\phi_k)$ or $v \in S(\phi_k))$.

This correlation is increasing both in $p_{s,v}$ and M, the mass parameter of the Dirichlet process, at each covariate value. As $p_{s,v}$ tends to the limits of zero and one, the correlation does the same, irrespective of M. As M tends to zero, the Sethuraman representation in (1.3) will be totally dominated by the first element, and thus the correlation tends to $p_{s,v}$. Finally, as $M \to \infty$ (the Dirichlet process tends to the centring distribution) the correlation tends to $2p_{s,v}/(1+p_{s,v})$, as other elements further down the ordering can also contribute to the correlation. Thus, the correlation is always larger than $p_{s,v}$ if the latter is smaller than one. Note that the correlation between distributions at different values of x will not tend to unity as M tends to infinity, in contrast to the πDDP constructions proposed in GS. This is a consequence of the construction: some points are not shared by the ordering at s and v no matter how large is M. The correlation between drawings from F_s and F_v, given by $\mathrm{Corr}(F_s, F_v)/(M+1)$ (see GS), however, tends to zero as $M \to \infty$. To make the result more applicable in regression, we now give a specific, simple method for choosing the subset in p-dimensional Euclidean space using a ball of radius r.

Definition 3. Let (C, r, t) be a Poisson process with intensity $f(r)$ defined on $\mathbb{R}^p \times \mathbb{R}_+^2$ with associated marks (V_j, θ_j) that are i.i.d. realisations of $\mathrm{Be}(1, M)$ and H, respectively. If

$$F_x = \sum_{\{i | x \in B_{r_i}(C_i)\}} \delta_{\theta_i} V_i \prod_{\{j | x \in B_{r_j}(C_j),\, t_j < t_i\}} (1 - V_j),$$

where $B_r(C)$ is a ball of radius r around C, we say $\{F_x|x \in \mathcal{X}\}$ follows a **Ball-based Dependent Dirichlet Process**, written B-DDP(M, H, f), where f is a non-negative function on $\mathbb{R}_+$ (the positive half-line).

The intensity f can be any positive function. However, we usually take f to be a probability density function. The following argument shows that defining f more generally does not add to the flexibility of the model. If (C, r, t) follows the Poisson process above, then writing $C_i^\star = C_i$, $r_i^\star = r_i$, and $t_i^\star = t_i/\lambda$ for $\lambda > 0$, defines a coupled Poisson process $(C^\star, r^\star, t^\star)$ with intensity $\lambda f(r)$. The ordering of t and $t^\star$ is the same for the coupled processes and the B-DDP only depends on t through its ordering. It follows that distributions $\{F_x\}_{x\in\mathcal{X}}$ defined using (C, r, t) and $(C^\star, r^\star, t^\star)$ are the same. In one dimension, we induce dependence by associating each atom with an interval, and only using that atom in the stick-breaking construction if the covariate value falls within that interval. The definition could be easily extended to ellipsoids around a central point that would allow the process to exhibit anisotropy. It is necessary to add the latent variable t for F_x to be a nonparametric prior. The set $T(x) = \{i|\,|x-C_i| < r_i\}$ indicates the atoms that appear in the stick-breaking representation of F_x. If we would instead define a Poisson process on (C, r) on $\mathbb{R}^p \times \mathbb{R}_+$ with intensity f, then $T(x)$ would be Poisson distributed with mean $2\int r\, f(r)\, dr$. This integral can be infinite but this has strong implications for the correlation structure. By including t we make $T(x)$ infinite for all choices of f and therefore define a nonparametric process. To calculate the correlation function, and relate its properties to the parameters of the distribution of r, it is helpful to consider $p_{s,v}$. This probability only depends on those centres from the set $\{C_k|s \in S_k \text{ or } v \in S_k\} = \{C_k|C_k \in B_{r_k}(s)\cup B_{r_k}(v)\}$.

Theorem 3. *If $\{F_x\}_{x\in\mathcal{X}}$ follows a B-DDP then*

$$p_{s,v} = \frac{\int \nu\left(B_r(s) \cap B_r(v)\right) f(r)\, dr}{\int \nu\left(B_r(s) \cup B_r(v)\right) f(r)\, dr},$$

where $\nu(\cdot)$ denotes Lebesgue measure in the covariate space $\mathcal{X}$.

So far, our results are valid for a covariate space of any dimension. However, in the sequel, we will focus particularly on implementations with a covariate that takes values in the real line. In this case, Theorem 3 leads to a simple expression.

Corollary 1. *If $\{F_x\}_{x\in\mathcal{X}}$ follows a B-DDP on $\mathbb{R}$ then*

$$p_{s,s+u} = \frac{2\mu_2 - |u|I}{4\mu - 2\mu_2 + |u|I},$$

where $\mu = E[r]$, $I = \int_{|u|/2}^{\infty} f(r)\, dr$, and $\mu_2 = \int_{|u|/2}^{\infty} r f(r)\, dr$, provided μ exists.

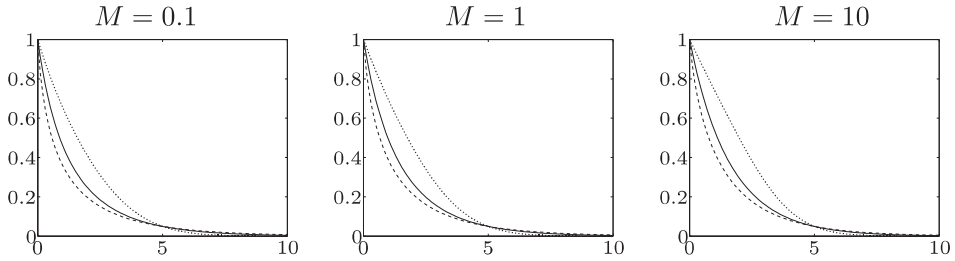

Figure 1. The autocorrelation function for a Gamma distance distribution with range 5 and shape $\alpha = 0.1$ (dashed line), $\alpha = 1$ (solid line) and $\alpha = 10$ (dotted line).

Throughout, we assume the existence of a nonzero mean for r and define different correlation structures through the choice of the distribution, $f(r)$. We focus on two properties of the autocorrelation, the first being the range, say $x^\star$, which we take to be the distance at which the autocorrelation function takes the value ε, so

$$p_{s,s+x^\star} = \frac{\varepsilon(M+2)}{M+2+M(1-\varepsilon)}.$$

The second property is the mean square differentiability or the smoothness of the process. In particular, a weakly stationary process on the real line is mean square differentiable of order q if and only if the $2q^{\text{th}}$ derivate of the autocovariance function evaluated at zero exists and is finite (see e.g., Stein (1999)). In the case of a Gamma distributed radius, we can derive the following result.

Theorem 4. *If $\{F_x\}_{x\in\mathcal{X}}$ follows a B-DDP with $f(r) = [\beta^\alpha/\Gamma(\alpha)]r^{\alpha-1}\exp\{-\beta r\}$, then F_x is mean square differentiable of order $q = 1, 2, \ldots$ if and only if $\alpha \geq 2q-1$.*

If each radius follows a Gamma distribution, then we can choose the shape parameter, α, to control smoothness and the scale parameter, β, to define the range, $x^\star$. A closed form inverse relationship is not available in general. However, if we choose $\alpha = 1$, which gives an exponential distribution, then

$$\beta = \frac{2}{x^\star}\log\left(\frac{1+M+\varepsilon}{\varepsilon(M+2)}\right). \tag{3.2}$$

Figure 1 shows the form of the autocorrelation for various smoothness parameters and a range fixed to 5 (with $\varepsilon = 0.05$). Clearly the mass parameter M, which is critical for the variability of the process, does not have much impact on the shape of the autocorrelation function once the range and smoothness are fixed. We concentrate on the Gamma implementation and work with the following class.

Definition 4. Let $\{C, r, t\}$ be a Poisson process with intensity $(\beta^\alpha/\Gamma(\alpha))r_i^{\alpha-1}$ $\exp\{-\beta r_i\}$ defined on $\mathbb{R}^p \times \mathbb{R}_+^2$, with associated marks (V_i, θ_i) that are i.i.d. realisations of $\text{Be}(1, M)$ and H. If

$$F_x = \sum_{\{i|x\in B_{r_i}(C_i)\}} V_i \prod_{\{j|x\in B_{r_j}(C_j),\, t_j<t_i\}} (1-V_j)\delta_{\theta_i},$$

then we say $\{F_x|x \in \mathcal{X}\}$ follows a **Dirichlet Process Regression Smoother (DPRS)**, represented as $\text{DPRS}(M, H, \alpha, \beta)$.

Typically, we fix α and $x^\star$ and put appropriate priors on M and any other parameters in the model. We use a prior distribution for M that can be elicited by choosing a typical value for M to be n_0 and a variance parameter η. This prior (discussed in GS) has the density function

$$p(M) = \frac{n_0^\eta \Gamma(2\eta)}{\Gamma(\eta)^2} \frac{M^{\eta-1}}{(M+n_0)^{2\eta}}.$$

4. Computational Method

This section discusses how the nonparametric hierarchical models of Section 2 with a DPRS prior can be fitted to data using a retrospective sampler. Retrospective Sampling for Dirichlet process-based hierarchical models was introduced by Papaspiliopoulos and Roberts (2008). Previous samplers based on the stick-breaking representation of the Dirichlet process used truncation (e.g., Ishwaran and James (2001)). The Retrospective Sampler avoids the need to truncate. The method produces a sample from the posterior distribution of all parameters except the unknown distribution. Inference about the unknown distribution often requires the use of some truncation method. This makes the method comparable to Pólya-urn based methods, which are reviewed by Neal (2000). Retrospective methods have been used for density regression models by Dunson and Park (2008).

We assume that data $(x_1, y_1), \ldots, (x_n, y_n)$ have been observed that are hierarchically modelled by

$$y_i \sim k(y_i|\psi_i), \quad \psi_i|x_i \sim F_{x_i}, \quad F_x \sim \text{DPRS}(M, H, \alpha, \beta).$$

The DPRS assumes that $F_x = \sum_{j=1}^\infty p_j(x)\delta_{\theta_j}$, where $\theta_1, \theta_2, \ldots \overset{i.i.d.}{\sim} H$ and p is constructed according to the definition of the DPRS. Additional parameters can be added to the sampling kernel or the centring distribution, and these are updated in the standard way for Dirichlet process mixture models. MCMC is

more easily implemented by introducing latent variables $s_1, s_2, \ldots, s_n$ and re-expressing the model as

$$y_i \sim k(y_i|\theta_{s_i}), \quad P(s_i = j) = p_j(x_i), \quad \theta_1, \theta_2, \ldots \overset{i.i.d.}{\sim} H.$$

The latent variables $s = (s_1, \ldots, s_n)$ are often called allocations, since s_i assigns the ith observation to the distinct elements of F_{x_i} (i.e., $\psi_i = \theta_{s_i}$). We define $y = (y_1, \ldots, y_n)$, $\theta = (\theta_1, \theta_2, \ldots)$, and $V = (V_1, V_2, \ldots)$. The probability $p(s_i|x_i, C, t, r, V)$ is only non-zero if $x_i \in B_r(C)$. Let (C^A, r^A, t^A) be the Poisson process (C, r, t) restricted to the set A, and let (θ^A, V^A) denote the set of associated marks. If we define the set $R = \{(C, r, t)|x \in B_r(C)\}$ with its complement R^C, the posterior distribution can be written as

$$\begin{aligned} p(\theta, s, t, C, r, V|y) &\propto p(y|\theta, s) \prod_{i=1}^{n} p(s_i|x_i, C, t, r, V) p(V) p(\theta) p(C, r, t) \\ &\propto p\left(y|\theta^R, s\right) \prod_{i=1}^{n} p\left(s_i|x_i, C^R, t^R, r^R, V^R\right) p\left(V^R\right) p\left(\theta^R\right) \\ &\quad \times p\left(C^R, r^R, t^R\right) p(V^{R^C}) \, p\left(\theta^{R^C}\right) p\left(C^{R^C}, r^{R^C}, t^{R^C}\right), \end{aligned}$$

which follows from the independence of Poisson processes on disjoint sets. Therefore we can draw inference using the restricted posterior distribution

$$\begin{aligned} p\left(\theta^R, s, t^R, C^R, r^R, V^R|y\right) &\propto p\left(y|\theta^R, s\right) \prod_{i=1}^{n} p\left(s_i|x_i, C^R, t^R, r^R, V^R\right) \\ &\quad \times p\left(V^R\right) p\left(\theta^R\right) p\left(C^R, r^R, t^R\right). \end{aligned}$$

We define a retrospective sampler for this restricted posterior distribution. A method of simulating (C^R, r^R, t^R) that is useful for our retrospective sampler is: (1) initialize $t_1 \sim \mathrm{Ga}\left(1, \int_R f(r)\, dC\, dr\right)$, and (2) $t_i = t_{i-1} + x_i$ where $x_i \sim \mathrm{Ga}\left(1, \int_R f(r)\, dC\, dr\right)$. Then $(C_1^R, r_1^R), (C_2^R, r_2^R), \ldots$, are independent of $t_1^R, t_2^R, \ldots$, and we take i.i.d. draws from the distribution

$$p\left(C_k^R|r_k^R\right) = \mathrm{U}\left(\bigcup_{i=1}^{n} B_{r_k^R}(x_i)\right), \quad p\left(r_k^R\right) = \frac{\nu\left(\bigcup_{i=1}^{n} B_{r_k^R}(x_i)\right) f(r_k^R)}{\int \nu\left(\bigcup_{i=1}^{n} B_u(x_i)\right) f(u)\, du}.$$

It is often hard to simulate from the conditional distribution of C_k^R and to calculate the normalising constant of the distribution of r_k^R. It is usually simpler to use a rejection sampler from the joint distribution of (C, R), conditioned to fall in a simpler set containing R. For example, in one dimension we take

$d^\star(r_k) = (x_{\min} - r_k, x_{\max} + r_k)$, where $x_{\min}$ and $x_{\max}$ are the minimum and maximum values of $x_1, \ldots, x_n$ and the rejection envelope is $f^\star\left(C_k^R, r_k^R\right)$, for which

$$f^\star(C_k^R | r_k^R) = \mathrm{U}\left(d^\star\left(r_k^R\right)\right), \quad f^\star(r_k^R) = \frac{(x_{\max} - x_{\min} + 2r_k)\, f(r_k^R)}{\int (x_{\max} - x_{\min} + 2u)\, f(u)\, du}.$$

Any values (C_k^R, r_k^R) are rejected if they do not fall in R. If we use a DPRS where r_k follows a Gamma(α,β) distribution, then we sample r_k^R from the rejection envelope using a mixture distribution

$$f^\star\left(r_k^R\right) = w f_{\mathrm{Ga}}(\alpha, \beta) + (1 - w) f_{\mathrm{Ga}}(\alpha + 1, \beta),$$

where $w = (x_{\max} - x_{\min})/(x_{\max} - x_{\min} + 2\frac{\alpha}{\beta})$ and $f_{\mathrm{Ga}}(\alpha, \beta)$ is the pdf of a $\mathrm{Ga}(\alpha, \beta)$ distribution. This construction generates the Poisson process underlying the B-DDP ordered in t and we use it to retrospectively sample the Poisson process in t (we can think of the DPRS as defined by a marked Poisson process where t^R follows a Poisson process and (C^R, r^R) are the marks). In fact, the posterior distribution only depends on $t_1^R, t_2^R, \ldots$ through their ordering, and we can simply extend the ideas of Papaspiliopoulos and Roberts (2008) to update the allocations s. The MCMC sampler defined on the posterior distribution parameterised by r^R can have problems mixing. The sampler can have much better mixing properties by using the reparameterisation from r^R to $r^{R\star}$, where we let $d_{il} = |x_i - C_l|$ and $r_i^{R\star} = r_i^R - \max\{d_{ij} | s_i = j\}$. Conditioning on $r^{R\star} = (r_1^{R\star}, r_2^{R\star}, \ldots)$ rather than $r^R = (r_1^R, r_2^R, \ldots)$ allows each observation to be allocated to a distinct element. Initially, we condition on $s_{-i} = (s_1, \ldots, s_{i-1}, s_{i+1}, \ldots, s_n)$ and remove s_i from the allocation. Let $K_{-i} = \max\{s_{-i}\}$, and let $r_k^{(1)} = r_j^{R\star} + \max\{d_{jk} | s_j = k, j = 1, \ldots, i-1, i+1, \ldots, n\}$ and $r_k^{(2)} = r_j^{R\star} + \max(\{d_{jk} | s_j = k, j = 1, \ldots, i-1, i+1, \ldots, n\} \cup \{|x_i - C_j|\})$. The proposal distribution is

$$q(s_i = j) = c^{-1} \times \begin{cases} k(y_i|\theta_k) V_k (1 - V_k)^{D_k} \prod_{l<k} (1 - V_l) \dfrac{f\left(r_k^{(2)}\right)}{f\left(r_k^{(1)}\right)} & j \leq K_{-i} \\ \max\limits_{m \leq K_{-i}} \{k(y_i|\theta_m)\} V_k \prod_{l<k} (1 - V_l) & j > K_{-i} \end{cases},$$

where $D_j = \#\left\{m \,\middle|\, r_j^{(1)} < d_{ms_m} < r_j^{(2)}, s_m > j\right\}$ and

$$c = \sum_{l=1}^{K_{-i}} k(y_i|\theta_l) V_l (1 - V_l)^{D_l} \prod_{h<l} (1 - V_h) \frac{f\left(r_l^{(2)}\right)}{f\left(r_l^{(1)}\right)} + \max_{l \leq K_{-i}} \{k(y_i|\theta_l)\} \prod_{h \leq K_{-i}} (1 - V_h).$$

If $j > K_{-i}$ we need to generate $(\theta_{K_{-i}+1}, V_{K_{-i}+1}, C_{K_{-i}+1}, d_{K_{-i}+1}), \ldots, (\theta_j, V_j, C_j, d_j)$ independently from their prior distribution. A value is generated from this

discrete distribution using the standard inversion method (i.e., simulate a uniform random variate U and the proposed value j is such that $\sum_{l=1}^{j-1} q(s_i = l) < U \leq \sum_{l=1}^{j} q(s_i = l)$). Papaspiliopoulos and Roberts (2008) show that the acceptance probability of the proposed value is

$$\alpha = \begin{cases} 1 & \text{if } j \leq K_{-i} \\ \min\left\{1, \frac{k(y_i|\theta_j)}{\max_{1\leq l\leq K_{-i}} k(y_i|\theta_l)}\right\} & \text{if } j > K_{-i} \end{cases}.$$

The other full conditional distributions of the Gibbs sampler are given in Appendix B of the online supplement.

5. Examples

This section applies the models developed in Section 2, in combination with the DPRS of Section 3 to simulated data and two real data sets: the prestige data (Fox and Suschnigg (1989)) and the electricity data of Yatchew (2003). As a basic model, we took Model 1(a) with a regression function $f(x) = 0$. This model tries to capture the dependence on x exclusively through the Dirichlet process smoother. Model 1(b) is a more sophisticated version of Model 1, where $m(x)$ is modelled through a Gaussian process, as explained in Section 2. Finally, we also used Model 2, which always has a Gaussian process specification for $m(x)$. The prior on M is as explained in Section 3, with $n_0 = 3$ and $\eta = 1$ for all examples. For the parameters a and b we adopted a Uniform prior over (0,1). The range $x^\star$ of the DPRS was such that the correlation was 0.4 at the median distance between covariate values. Priors on σ^2 and on the parameter of the Gaussian process ζ were as in the benchmark prior of Palacios and Steel (2006), and we fixed the smoothness parameter τ of the Gaussian process at 1.

5.1. Example 1: Sine wave

We generated 100 observations from the model $y_i = \sin(2\pi x_i) + \epsilon_i$ where x_i were uniformly distributed on $(0, 1)$ and the errors ϵ_i were independent and chosen to be heteroscedastic and non-normally distributed. We considered two possible formulations: *Error 1* assumed that ϵ_i follows a t-distribution with zero mean, 2.5 degrees of freedom and a conditional variance of the form $\sigma^2(x) = (x - 1/2)^2$, thus 0 at $x = 1/2$ and increasing away from $1/2$. *Error* 2 assumed that the error distribution was a mixture of normals $p(\epsilon_i|x_i) = 0.3\mathrm{N}(0.3, 0.01) + 0.7\mathrm{N}(-0.3 + 0.6x_i, 0.01)$. This error distribution is bimodal at $x_i = 0$ and unimodal (and normal) at $x_i = 1$. The first error distribution can be represented using both a mixture of normals and a scale mixture of uniforms whereas the second error distribution cannot be fitted using a mixture of uniforms. Initially, we assumed *Error* 1. The results for Model 1(a) are illustrated in Figure 2 for three values

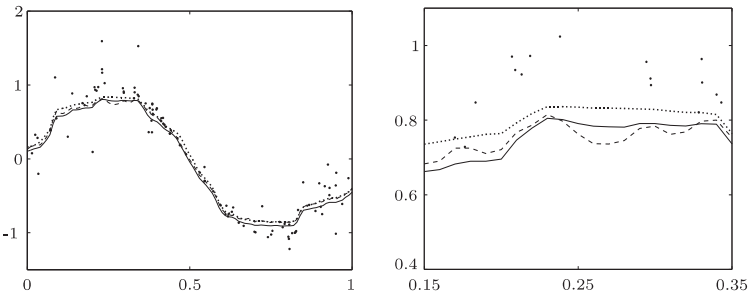

Figure 2. Example 1 with Error 1: predictive conditional mean of y given x for Model 1(a): $\alpha = 0.1$ (dashed line), $\alpha = 1$ (solid line), $\alpha = 10$ (dotted line). Data points are indicated by dots. The right panel presents a magnified section of the left panel.

Table 1. Example 1 with Error 1: posterior median and 95% credible interval (in parentheses) for selected parameters.

	Model 1(a)	Model 1(b)	Model 2
σ	0.71 (0.48, 1.13)	0.64 (0.46, 0.96)	0.68 (0.49, 1.08)
a	0.09 (0.02, 0.24)	0.05 (0.01, 0.33)	
b		0.75 (0.54, 0.88)	0.76 (0.53, 0.90)
ρ		0.53 (0.31, 0.96)	0.62 (0.31, 1.19)
M	0.38 (0.14, 0.95)	1.84 (0.61, 5.27)	1.57 (0.46, 3.64)

of α. Smaller values of α lead to rougher processes and the effect of its choice on inference is clearly illustrated. In the sequel, we only present results where $\alpha = 1$.

Under Model 1(a), we infer a rough version of the underlying true distribution function as illustrated by the predictive density in Figure 3. The small values of a in Table 1 indicate a lack of normality. The results are similar to those of GS who find that the estimated conditional mean is often "blocky", which reflects the underlying piecewise constant approximation to the changing distributional form.

In the more complicated models the conditional location was modelled through a nonparametric regression function (in this case a Gaussian process prior). Both Model 1(b) and Model 2 assumed a constant prior mean for $m(x)$. Introducing the Gaussian process into Model 1 led to smaller values of σ, since some variability can now be explained by the Gaussian process prior. However, the posterior for a still favoured fairly small values, reminding us that even if the conditional mean was better modelled with the Gaussian process, the tails were still highly non-Normal (see Figure 4). The estimated posterior predictive distributions (as depicted in Figure 3) were now much smoother. Both Model 1(b) and Model 2 led to large values for b and thus the variance of the Gaussian process (which

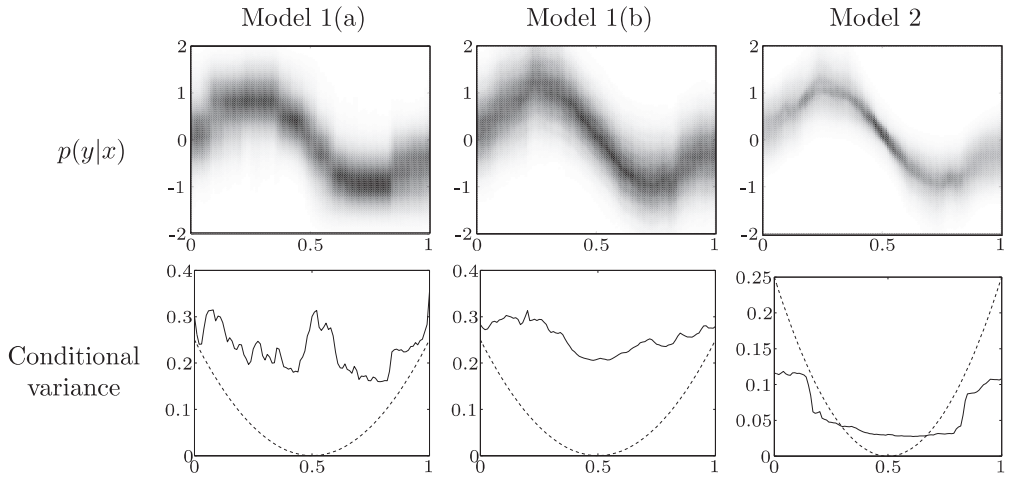

Figure 3. Example 1 with Error 1: heatmap of the posterior predictive density $p(y|x)$ and plot of the posterior conditional predictive variance $\sigma^2(x)$ (solid line) and the true value (dashed line).

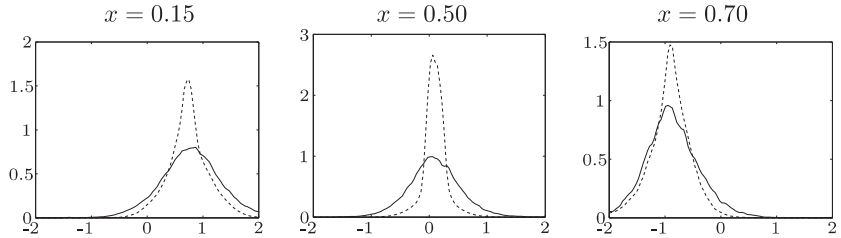

Figure 4. Example 1 with Error 1: The posterior predictive distribution of y given x using Model 1(a) (solid line) and Model 1(b) (dashed line)

better fits the true variability of the mean). This led to better estimates of the conditional predictive variance, as illustrated in Figure 3. Clearly a model of this type struggles with estimation at the extreme values of x, but the main part of the functional form is well-recovered. The parameter $\rho = 2\sqrt{\tau}/\zeta$ used in Table 1 is an alternative range parameter favoured by Stein (1999, p.51), and indicates that the Gaussian process dependence of $m(x)$ was similar for Model 1(a) and Model 2. The posterior median values of ρ led to a range of the Gaussian process of 1.89 and 1.61 for Models 1(b) and 2, respectively.

Results for data generated with the second error structure are shown in Figure 5 and Table 2 (for selected parameters). Model 1(b) was able to infer the bimodal distribution for small values of x and the single mode for large x as well as the changing variance. Model 2 was not able to capture the bimodality by construction and only captured the changing variability. In both cases the mean was well estimated. Small values of a illustrate the difficulty in capturing

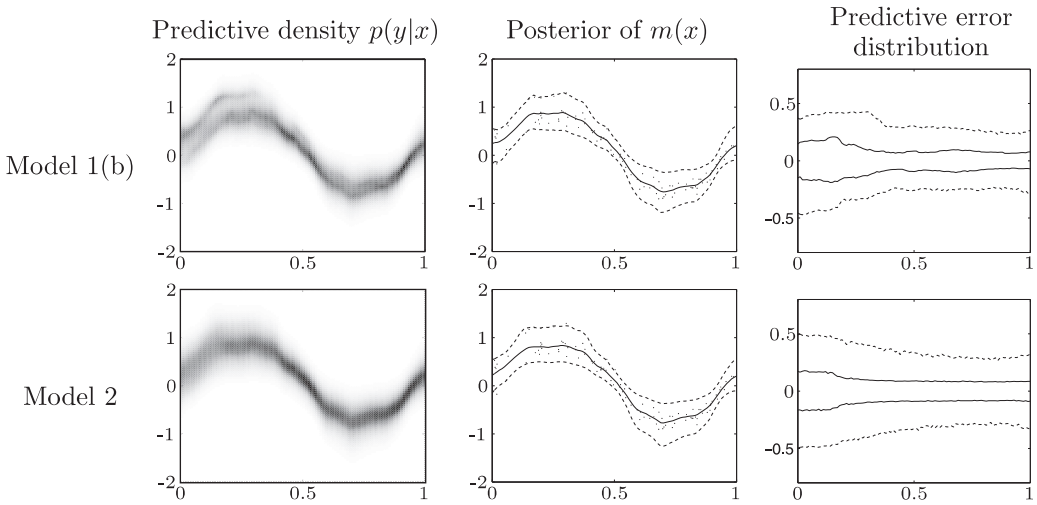

Figure 5. Example 1 with Error 2: heatmap of posterior predictive density $p(y|x)$, plot of the posterior of $m(x)$ indicating median (solid line), 95% credible interval (dashed lines), and data (dots), and the posterior predictive error distribution indicating the 2.5^{th}, 25^{th}, 75^{th}, and 97.5^{th} percentiles

Table 2. Example 1 with Error 2: posterior median and 95% credible interval (in parentheses) for selected parameters.

	Model 1(a)	Model 1(b)	Model 2
σ	0.70 (0.41, 1.66)	0.47 (0.34, 0.71)	0.54 (0.38, 0.98)
a	0.12 (0.02, 0.31)	0.13 (0.02, 0.38)	
b		0.84 (0.66, 0.92)	0.84 (0.65, 0.94)

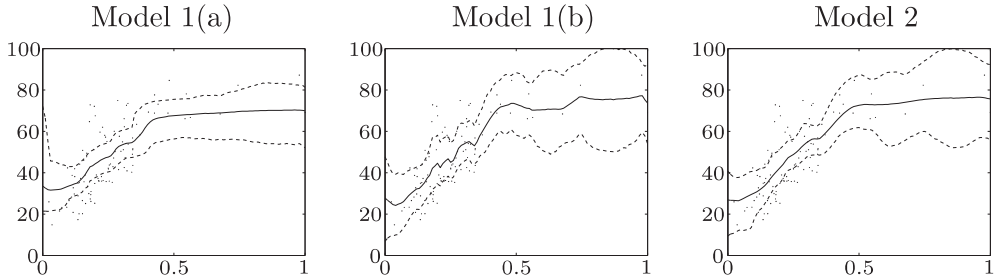

Figure 6. Prestige data: posterior distribution of the conditional mean indicating median (solid line), 95% credible interval (dashed lines) and data (dots).

the error structure. The large values of b indicate that the centring model (a constant model with mean zero) does a poor job in capturing the mean.

Table 3. Prestige data: posterior median and 95% credible interval (in parentheses) for selected parameters

	Model 1(a)	Model 1(b)	Model 2
σ	22.20 (14.80, 43.70)	20.00 (14.80, 30.00)	22.00 (16.20, 36.40)
a	0.12 (0.03, 0.31)	0.28 (0.08, 0.69)	
b		0.66 (0.38, 0.85)	0.69 (0.40, 0.88)

5.2. Prestige data

Fox (1997) considered the relationship between income and prestige of 102 occupations using the 1971 Canadian census. The prestige of the jobs was measured through a social survey. We treat income as the response and the prestige measure as a covariate. The data is available to download in the R package `car`. Figure 6 shows the fitted conditional mean. In all cases the relationship between income and prestige show an increasing trend for smaller income before prestige flattens out for larger incomes. The results are very similar to those described in Fox (1997). The inference for selected individual parameters is presented in Table 3. As in the previous example, the Gaussian process structure on $m(x)$ accounts for quite a bit of variability, rendering the error distribution not too far from normal in Model 1(b), as indicated by the fairly large values of a.

5.3. Scale economies in electricity distribution

Yatchew (2003) considered a cost function for the distribution of electricity. A Cobb-Douglas model was fitted,

$$\text{tc} = f(\text{cust}) + \beta_1\,\text{wage} + \beta_2\,\text{pcap} + \beta_3\,\text{PUC} + \beta_4\,\text{kwh} + \beta_5\,\text{life} + \beta_6\,\text{lf} + \beta_7\,\text{kmwire} + \epsilon,$$

where tc is the log of total cost per customer, cust is the log of the number of customers, wage is the log wage rate, pcap is the log price of capital, PUC is a dummy variable for public utility commissions, life is the log of the remaining life of distribution assets, lf is the log of the load factor, and kmwire is the log of kilometres of distribution wire per customer. The data consist of 81 municipal distributors in Ontario, Canada during 1993. We fitted the DPRS model with cust as the covariate to ϵ and we centred the model over two parametric regression models by choosing $f(\text{cust})$ as Parametric 1, $\gamma_1 + \gamma_2\,\text{cust}$, and Parametric 2, $\gamma_1 + \gamma_2\,\text{cust} + \gamma_3\,\text{cust}^2$.

The results of Yatchew (2003) suggest that a linear $f(\text{cust})$ is not sufficient to explain the effect of number of customers. The results for selected parameters are shown in Tables 4 and 5 when centring over Parametric 1 and over Parametric 2, respectively. When fitting both parametric models we see differences in the estimates of the effects of some other covariates. The parameters β_1 and β_6 have

Table 4. Electricity data: posterior median and 95% credible interval (in parentheses) for selected parameters of Parametric model 1 (linear) and the nonparametric models centred over Parametric model 1.

	Parametric 1	Model 1(a)	Model 1(b)	Model 2
γ_1	0.42(-4.14, 5.10)	-0.70(-4.88, 3.20)	-0.90(-4.98, 3.09)	-0.67(-4.79, 4.30)
γ_2	-0.07(-0.13, -0.02)	-0.07(-0.14, -0.01)	-0.10(-0.20, 0.02)	-0.09(-0.20, 0.00)
β_1	0.48(-0.25, 1.16)	0.67(0.05, 1.20)	0.71(0.07, 1.32)	0.70(0.00, 1.53)
β_4	0.12(-0.06, 0.31)	0.07(-0.10, 0.25)	0.04(-0.14, 0.22)	0.06(-0.14, 0.23)
β_6	0.97(0.03, 1.92)	1.11(0.29, 2.00)	1.24(0.40, 2.10)	1.19(0.14, 2.04)
σ	0.17(0.15, 0.21)	0.20(0.14, 0.36)	0.23(0.17, 0.39)	0.27(0.19, 0.48)
a		0.19(0.05, 0.45)	0.75(0.25, 0.99)	
b			0.41(0.11, 0.77)	0.55(0.21, 0.84)

Table 5. Electricity data: posterior median and 95% credible interval (in parentheses) for selected parameters of Parametric model 2 (quadratic) and the nonparametric models centred over Parametric model 2

	Parametric 2	Model 1(a)	Model 1(b)	Model 2
γ_1	2.77(-1.53, 6.96)	2.78(-1.83, 6.88)	2.52(-2.44, 7.56)	2.77(-4.20, 7.79)
γ_2	-0.83(-1.19, -0.48)	-0.92(-1.42, -0.41)	-0.91(-1.69, -0.23)	-0.96(-1.57, -0.32)
γ_3	0.04(0.02, 0.06)	0.05(0.02, 0.07)	0.04(0.01, 0.09)	0.05(0.01, 0.08)
β_1	0.83(0.20, 1.48)	0.79(0.16, 1.38)	0.80(0.14, 1.43)	0.78(-0.03, 1.41)
β_4	-0.02(-0.20, 0.15)	-0.02(-0.22, 0.17)	0.00(-0.18, 0.18)	0.00(-0.18, 0.19)
β_6	1.25(0.38, 2.09)	1.31(0.52, 2.18)	1.32(0.47, 2.15)	1.31(0.48, 2.59)
σ	0.16(0.13, 0.19)	0.17(0.14, 0.23)	0.21(0.16, 0.34)	0.22(0.16, 0.38)
a		0.13(0.02, 0.40)	0.77(0.24, 0.99)	
b			0.30(0.08, 0.75)	0.37(0.15, 0.77)

larger posterior medians under Parametric 2, while β_4 has a smaller estimate. If we centre our nonparametric models over the linear parametric model then we see the same changes for β_1 and β_6 and a smaller change for β_4. Posterior inference on regression coefficients was much more similar for all models in Table 5. In particular, the parametric effect of customers was very similar for Parametric 2 and for all the nonparametric models centred over it. The estimated correction to the parametric fit for the effect of customers is shown in Figure 7. For models centred over the linear model, it shows a difference that could be well captured by a quadratic effect, especially for Model 1(b) and Model 2.

Under both centring models, the importance of the nonparametric fitting of the error sharply decreased as a Gaussian process formulation for the regression function was used, as evidenced by the increase in a. Changing to a quadratic centring distribution led to decreased estimates of b, indicating a more appropriate fit of the parametric part. This is corroborated by the nonparametric correction to this fit as displayed in Figure 7.

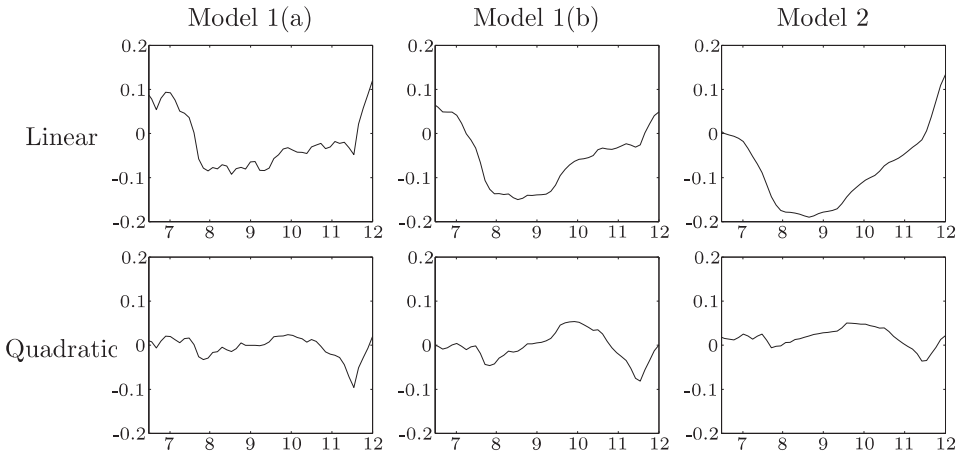

Figure 7. Electricity data: posterior mean of the nonparametric component(s) of the model

6. Discussion

This paper shows how ideas from Bayesian nonparametric density estimation and nonparametric estimation of the mean in regression models can be combined to define a range of useful models. We introduce novel approaches to nonparametric modelling by centring over appropriately chosen parametric models. This allows for a more structured approach to Bayesian nonparametrics, and can greatly assist in identifying the specific inadequacies of commonly used parametric models. An important aspect of the methodology is separate modelling of various components, such as important quantiles, like the median, or moments, like the mean, that allows the nonparametric smoothing model to "do less work". These ideas can be used in combination with any nonparametric prior that allows distributions to change with covariates. Here we have concentrated on one example, the Dirichlet Process Regression Smoother (DPRS) prior, introduced in this paper. We have concentrated on univariate regression problems but the methods could be extended to higher dimensions. However, we imagine that computation will become harder with increasing dimension. The DPRS is related to the πDDP methods of GS, but allows simpler computation (and without truncation) through retrospective methods. The parameters of the DPRS can be chosen to control the smoothness and the scale of the process.

Acknowledgements

The authors would like to acknowledge the helpful comments of an associate editor and two anonymous referees. Jim Griffin acknowledges research support from The Nuffield Foundation grant NUF-NAL/00728.

References

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

Chung, Y. and Dunson, D. B. (2011). The local Dirichlet process. *Ann. Inst. Statist. Math.*, forthcoming.

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

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

Dunson, D. B., Pillai, N. and Park, J. H. (2007). Bayesian density regression. *J. Roy. Statist. Soc. Ser. B* **69**, 163-183.

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

Fox, J. (1997). *Applied Regression Analysis, Linear Models and Related Methods*. Sage, California.

Fox, J. and Suschnigg, C. (1989). A note on gender and the prestige of occupations. *Canad. J. Sociology* **14**, 353-360.

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian nonparametric modelling with Dirichlet process mixing. *J. Amer. Statist. Assoc.* **100**, 1021-1035.

Griffin, J. E. (2010). Default priors for density estimation with mixture model. *Bayesian Analysis* **5**, 45-64.

Griffin, J. E. and Steel, M. F. J. (2006). Order-based dependent Dirichlet processes. *J. Amer. Statist. Assoc.*, 101, 179-194.

Ishwaran, H. and James, L. (2001). Gibbs sampling methods for stick-breaking priors. *J. Amer. Statist. Assoc.* **96**, 161-73.

Kottas, A. and Gelfand, A. E. (2001). Bayesian semiparametric median regression modeling. *J. Amer. Statist. Assoc.* **96**, 1458-1468.

Kottas, A. and Krnjajic, M. (2009). Bayesian semiparametric modeling in quantile regression. *Scand. J. Statist.* **36**, 297-319.

Leslie, D. S., Kohn, R. and Nott, D. J. (2007). A general approach to heteroscedastic linear regression. *Statist. Comput.* **17**, 131-146.

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

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

Palacios, M. B. and Steel, M. F. J. (2006). Non-Gaussian Bayesian geostatistical modeling. *J. Amer. Statist. Assoc.* **101**, 604-618.

Papaspiliopoulos, O. and Roberts, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, **95**, 169-186.

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. MIT Press, Boston.

Reich, B. and Fuentes, M. (2007). A multivariate semiparametric Bayesian spatial modeling framework for hurricane surface wind fields. *Ann. Appl. Statist.* **1**, 249-264.

Sethuraman, J. (1994). A constructive definition of the Dirichlet process. *Statist. Sinica* **4**, 639-650.

Stein, M. (1999). *Interpolation of Spatial Data*. Springer, New York.

Yatchew, A. (2003). *Semiparametric Regression for the Applied Econometrician*. Cambridge University Press, Cambridge.

School of Mathematics, Statistics and Actuarial Science, University of Kent, CT2 7NF, U.K.

E-mail: J.E.Griffin-28@kent.ac.uk

Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K.

E-mail: M.F.Steel@stats.warwick.ac.uk

(Received May 2008; accepted July 2009)