Bayesian Nonparametric Modelling with the Dirichlet Process Regression Smoother

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

Abstract

In this paper we discuss the problem of Bayesian fully nonparametric regression. A new construction of priors for nonparametric regression is discussed and a specific prior, the Dirichlet Process Regression Smoother, is proposed. We consider the problem of centring our process over a class of regression models and propose fully nonparametric regression models with flexible location structures. Computational methods are developed for all models described. Results are presented for simulated and actual data examples.

Keywords: Nonlinear regression; Nonparametric regression; Model centring; Stick-breaking prior

1 Introduction

Nonparametric regression has become a major area of research over the last twenty years. It offers the ability to accurately model nonlinear relationships between covariates and responses which are often observed in real data. More generally, these methods allow us to describe data whilst making a minimal amount of modelling assumptions and the conditional

*Jim Griffin is Lecturer, Department of Statistics, University of Warwick, CV4 7AL, U.K. (Email: J.E.Griffin@warwick.ac.uk) and Mark Steel is Professor, Department of Statistics, University of Warwick, Coventry, CV4 7AL, U.K. (Email: M.F.Steel@stats.warwick.ac.uk). Jim Griffin acknowledges research support from The Nuffield Foundation grant NUF-NAL/00728.

mean is only one (perhaps the most obvious) feature of the data. For example, the conditional standard deviation might be a more important feature of the data, such as occurs in certain financial time series.

A typical model assumes that a response, y, observed at a covariate value x can be expressed as

$$y = m(x) + \sigma(x)\epsilon,$$

where ϵ has expectation 0 and variance 1, whose distribution may depend on x.

In the literature, there are two main approaches to this type of problem. In Bayesian inference the main approach has assumed a parametric distribution for ϵ and constant variance $\sigma^2(x) = \sigma^2$. The mean function can be estimated using some flexible form for the mean. Possible examples include Gaussian processes (Neal 1998), or basis function regression, such as a spline basis or a wavelet basis. Gaussian process priors have a long history in Bayesian nonparametric analysis dating back to at least O'Hagan (1978) and have recently seen renewed interest in the machine learning community (*e.g.* Neal 1998). The basis function approach is described in Denison *et al.* (2002). Wahba (1978) gives a Bayesian interpretation of spline smoothing, which is further explored in Speckman and Sun (2003). Models that relax the assumption of constant $\sigma^2(x)$ are developed in Kohn *et al.* (2001). Leslie *et al.* (2007) extend these models to nonparametric inference about the distribution of ϵ (which is assumed independent of x).

A lot of non-Bayesian work has concentrated on local likelihood methods (see Tibshirani and Hastie 1987), which assumes that a local likelihood function can be constructed and maximized at any covariate value. These methods do not construct a probability model for the whole sample and extensions to Bayesian inference are difficult because we must fully specify a family of distributions for ϵ. Bayesian nonparametric methods achieve flexibility by putting priors on distribution spaces, corresponding to infinitely dimensional parameterisations. A Bayesian interpretation to local modelling links the idea to partial exchangeability, which is expressed by grouping observations into blocks. Observations are exchangeable within blocks and independent between blocks. Mallick and Walker (1997) develop this idea for nonparametric Bayesian inference with fixed blocks. In regression analysis, one popular method for defining the groups partitions the covariate space and assumes that the observations falling in an element of the partition form a group. By assuming that the partition is unknown, a prior can be defined and the posterior predictive distribution is mixed according

to the posterior distribution on the partition structure. A drawback with this approach is that many partition structures will usually seem sensible, which complicates prior selection. In a regression and classification setting, Denison *et al.* (2002) call this approach the Bayesian Partition Model. Griffin and Steel (2006) (henceforth denoted by GS) develop a nonparametric prior, called πDDP, which limits to the Bayesian Partition Model and allows much richer types of partial exchangeability to be explored.

The purpose of this paper is to extend and combine the current Bayesian nonparametric density estimation and function estimation methods in a way that captures the spirit of the classical nonparametric procedures. In particular, in functional data analysis we want to allow an unknown form for the distribution of ϵ which depends on x. In our approach, we want to specify a prior for an unknown distribution (random probability measure) which changes with the covariate value. Recent work by De Iorio *et al.* (2004), Dunson *et al.* (2007) and GS discusses such priors which extend the well-known Dirichlet process mixture distributions. The first paper is based on dependent Dirichlet processes (DDP) as defined in MacEachern (1999). Such models are flexible but make it hard to control specific features of the modelling. The methods developed in GS lead to fairly tractable prior distributions which are centred over a single distribution. One purpose of the present paper is to develop hierarchical extensions to exploit the natural idea of centring over a 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 also provides a framework for assessing the adequacy of commonly used parametric models, by centring over these models.

In this paper we will use some theoretical results derived in GS in a somewhat more general context. However, it is important to point out that the present paper makes considerable contributions to the actual modelling of data and the theoretical and practical implementation of these, very general, ideas in regression analysis.

Section 2 discusses previous approaches to regression analysis in Bayesian nonparametric analysis and introduces a new method for constructing Bayesian nonparametric regression estimators, discussing a particular class in detail: the Dirichlet Process Regression Smoother (DPRS). Section 3 introduces the idea of centring over models and uses the DPRS to develop three models which make different assumptions about how the distribution departs

from the centring model. Section 4 briefly discusses computational methods for these hierarchical models defined using the DPRS with more details of the implementation in Appendix B. Section 5 illustrates the use of these models in inference problems, and a final section concludes. Without specific mention in the text, proofs will be grouped in Appendix A.

2 Bayesian Nonparametric Regression

This section reviews existing Bayesian approaches and develops a new methodology for fully nonparametric regression modelling. By "fully" we mean models where all features of the distributions are allowed to change with the covariates, as opposed to mean regression models (such as Gaussian processes or spline regression). We restrict attention to Dirichlet process-based models since these methods dominate in the literature and our approach follows these ideas naturally. Some other methods are reviewed by Müller and Quintana (2004).

2.1 Density Regression with Stick-Breaking Priors

Much recent work on Bayesian density estimation from a sample $y_1, y_2, \ldots, y_n$ has been based around the following hierarchical model

$$y_i \sim k(\psi_i), \quad \psi_i \sim F, \quad F \sim \Pi,$$

where $k(\cdot)$ is a probability density function (pdf) parameterised by ψ and the prior Π places support across a wide range of distributions. Many choices for Π are possible including Normalized Random Measures (James *et al.* 2005) and the Dirichlet Process (Ferguson 1973). In this paper we will concentrate on the stick-breaking class of discrete distributions for which

$$F \stackrel{d}{=} \sum_{i=1}^{\infty} p_i \delta_{\theta_i}, \tag{1}$$

where δ_θ denotes a Dirac measure at θ and $p_i = V_i \prod_{j<i}(1-V_j)$. The sequence $V_1, V_2, V_3, \ldots$ is independent and $V_i \sim \text{Be}(a_i, b_i)$ while $\theta_1, \theta_2, \theta_3, \ldots$ are i.i.d. from some distribution H. Then $\text{E}[F(B)] = H(B)$ for any measurable set B, which allows us to define H as a centring distribution. The Dirichlet process occurs if $a_i = 1$ and $b_i = M$ for all i.

If we observe pairs $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$ and we wish to estimate the conditional distribution of y given x, then a natural extension of the hierarchical model above

defines

$$y_i \sim k(\psi_i), \quad \psi_i \sim F_x, \quad F_x \stackrel{d}{=} \sum_{j=1}^{\infty} p_j(x)\delta_{\theta_j(x)}.$$

Many previously defined processes fall within this class: if $p_i(x) = p_i$ then we have the single p DDP model (MacEachern, 2001) which has been applied to discrete covariates by De Iorio *et al.* (2004), spatial problems (Gelfand *et al.*, 2005) and quantile regression (Kottas and Krnjajic, 2005). Often it is assumed that $\theta_i(x)$ follows a Gaussian process. Other authors have considered a regression model for $p_i(x)$. Due to the specification of the weights p_i, values θ_i earlier in the ordering (*i.e.* with small i) will tend to have larger weights.

If $\theta_i(x)$ is a given function described by $\theta_i(x) = K(x; \phi_i)$ parameterised by ϕ_i, and $k(\cdot)$ is the pdf of a Normal$(0, \sigma^2)$ distribution, an alternative nonparametric representation is

$$y_i \sim \mathrm{N}(K(x_i; \phi_i), \sigma^2), \quad \phi_i \overset{i.i.d.}{\sim} F, \quad F \sim \Pi.$$

An example in regression analysis is the ANOVA-DDP model of De Iorio *et al.* (2004) who model replicated data from a series of experiments using combinations of treatment and block levels. A standard parametric approach would be ANOVA. The function $K(x; \phi_i)$ mimics the form of the group means in an ANOVA analysis by taking a linear combination of treatment effects, block effects and interactions.

In the present paper, we will concentrate on models where $\theta_i(x) = \theta_i$ for all x. A similar idea was implemented by GS using the πDDP prior which can capture non-linear relationships between response and regressors. An implication of this assumption is that the correlations of all moments of the distributions at two covariate levels are the same. The impact inferentially is the undersmoothing of the posterior estimates of the distribution. In particular, the posterior mean $\mathrm{E}[y|x]$ has the step form typical of piecewise constant models. This is not surprising since we have centred the process over a model that would be considered far too simple in a parametric analysis. A, perhaps more important, concern is the parsimony of the model developed above relative to a model only modelling the mean. If the mean of the response has a highly non-linear relationship to the regressor then flexible regression ideas such as Gaussian process priors (O'Hagan 1978, Rasmussen and Williams 2006) or basis regression models (Denison *et al.* 2002) could represent the relationship more accurately and parsimoniously than the piecewise constant effect achieved by the model where the mean response-regressor relationship is modelled solely through the nonparametric prior.

We therefore will concentrate on the following model

$$y_i - g(x_i) - m(x_i) \sim k(\psi_i), \quad \psi \sim F_x, \quad F_x \stackrel{d}{=} \sum_{i=1}^{\infty} p_i(x)\delta_{\theta_i},$$

where the conditional regression function consists of a parametric part $g(x)$ and a nonparametric part $m(x)$. For the latter, we will consider Gaussian process priors where $m(x_1), \ldots, m(x_n)$ are jointly normally distributed with constant mean 0 and the covariance of $m(x_i)$ and $m(x_j)$ is $\sigma_0^2\rho(x_i, x_j)$ where $\rho(x_i, x_j)$ is a proper correlation function. A popular choice of correlation function is the flexible Matèrn class (see *e.g.* Stein, 1999) for which

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

where $\mathcal{K}_{\nu}$ is the modified Bessel function of order ν. The process is q times mean squared differentiable if and only if $q < \nu$ and ζ acts as a range parameter.

In the general form of the model above, the conditional regression function is the sum of a parametric function of the covariates and a Gaussian process component. The approach is not restricted to the Gaussian process and a wavelet or spline basis could be introduced instead. Finally, the nonparametric modelling of the residual through F_x will allow for deviations in the error distribution from H.

2.2 A Bayesian Density Smoother

This section develops a process for nonparametric distributions, F_x, defined on some covariate space $x \in \mathcal{X}$ for which marginal distributions F_x follow a stick-breaking process at any x. The model is developed in the spirit of GS and shares similarities with their order-based dependent Dirichlet process (πDDP) with an "arrivals" construction. The earlier process works well in time series problem where we want to define a process of distributions changing through time. However, this construction defines a process which is not symmetric in time. If we want to smooth distributions with respect to some covariates other than time, then symmetry is often an important property. Secondly, the process changes by introducing new atoms which have higher expected weights than atoms introduced at previous time points, which may make it hard to model small changes in the unknown distribution with respect to the index. This second property is not shared by the "permutations" construction of GS, which is much closer in its properties to the prior that we propose here but has proved

difficult for developing easily implemented MCMC inference methods. The ideas here are associated with a much simpler MCMC sampling scheme, outlined in Section 4.

The πDDP of GS is based on ordering the V_is used in (1) at each covariate value. Similar orderings will induce similar distributions. In the arrivals construction of GS each atom is associated with a point in time. The ordering at a given time is defined to be all previous points ordered in reverse time (the last time point is first). If we denote the region in which points are relevant for the ordering at x by $U(x)$, then the arrivals process assumes that $U(x) \subset U(y)$ if $y > x$ and that the new atoms are placed at the start of the process. Clearly, the process is not reversible in time. If the process allows atoms to "appear" in the ordering then, to define a reversible process, we need atoms to also "disappear". A simple method that achieves this effect is to restrict $U(x)$ to be an interval. Thus, as a special case the arrivals ordering arises when the left-hand side of the interval is minus infinity for all points. If the interval is symmetric around a centre then the process will be reversible by construction.

The idea can be extended to spaces $\mathcal{X}$ that are subsets of $\mathbb{R}^p$ where we generalize the intervals to closed, convex sets of appropriate dimension, say $I_1, I_2, \ldots$. We will define the distribution F_x by applying the stick-breaking construction to $\{(V_j, \theta_j) | x \in I_j\}$. In the "arrivals" construction, these points enter according to the value of the right-hand end of the interval, largest first. This idea cannot be directly extended and our solution associates a value t_j on $\mathbb{R}^+$ with I_j. The atoms are ordered by the value of t_j with smaller values earlier. The prior introduced here is thus defined by the sequence $(V_1, \theta_1, I_1, t_1), (V_2, \theta_2, I_2, t_2), \ldots$, where $I_1, I_2, \ldots$ are i.i.d. realisations from a distribution of closed, convex sets, which are subsets of the covariate space. For example, $I_1, I_2, I_3, \ldots$ could be intervals if the covariate lies within a subset of $\mathbb{R}$, which defines an interval process.

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 a measurable set B. GS show that the correlation between the measures does not depend on B and for the case of the Dirichlet process at each covariate value is given by

$$\text{Corr}(F_s, F_v) = \frac{2}{M+2} \text{E}\left[\sum_{k \in T} \left(\frac{M}{M+2} \right)^{\#S_k} \left(\frac{M}{M+1} \right)^{\#S'_k} \right] \tag{2}$$

where T groups the relevant set indices that appear in at least one of the stick-breaking representations, *i.e.* $T = \{k | s \in I_k \text{ or } v \in I_k\}$ and $\#S_k$ is the number of shared elements (appearing in both constructions) before the k-th common element whereas $\#S'_k$ is the num-

ber of unique elements before we reach k. The expectation is taken over $\#S_k$ and $\#S'_k$.

To make the result more applicable, we now consider a situation where 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

Theorem 1 *The correlation between F_s and F_v can be expressed as*

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

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

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) 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 will tend 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 will not be shared by the ordering at s and v no matter how large M. The correlation between drawings from F_s and F_v, given by Corr$(F_s, F_v)/(M+1)$ (see GS) will, however, tend to zero as $M \to \infty$.

The model is further specified by choosing a distribution for I_k. We will restrict attention to I_k which can be expressed as a ball of radius r_k around C_k, *i.e.* $I_k = \{x | \, \|x - C_k\| < r_k\}$ where $\|\cdot\|$ is some appropriate distance measure and $(C_1, r_1, t_1), (C_2, r_1, t_1), \ldots$ follow a Poisson process on $\mathbb{R}^p \times \mathbb{R}^2_+$ with intensity $p(r)$ which is a pdf defined on $\mathbb{R}_+$. Then for a covariate value s, the set $\{i | \, \|s - C_i\| < r_i\}$ will have an infinite number of elements, which allows us to define nonparametric stick-breaking distributions, such as the Dirichlet process, for each covariate value. Properties such as covariate-reversibility (in the one-dimensional case) or, more generally, isotropy will depend on the symmetry of the sets. These properties follow from restricting the shape to be a ball with radius r, $B_r(C)$ and the reversibility or isotropy of the Poisson 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 $\{C_k | s \in I_k \text{ or } v \in I_k\} = B_{r_k}(s) \cup B_{r_k}(v)$.

Theorem 2 *For the process defined above*

$$p_{s,v} = \frac{\int \nu\left(B_{r_k}(s) \cap B_{r_k}(v)\right) p(r_k)\, dr_k}{\int \nu\left(B_{r_k}(s) \cup B_{r_k}(v)\right) p(r_k)\, dr_k},$$

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

Sofar, 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 2 leads to a simple expression.

Corollary 1 *If we consider a one-dimensional regressor then*

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

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

Throughout, we will assume the existence of a nonzero mean for r and define different correlation structures through the choice of the distribution of r. We will focus on two properties of the autocorrelation. The first property is the range, say $x^\star$, which we define as the distance at which the autocorrelation function takes the values ϵ which implies that

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

The second property is the mean square differentiability which is related to 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 for example Stein, 1999). In the case of a Gamma distributed radius, we can derive the following result.

Theorem 3 *If we assume the radius r has a Gamma distribution with shape parameter α, the process F_x is mean square differentiable of order $q = 1, 2, \ldots$ if and only if $\alpha \geq 2q - 1$.*

In this case, we will choose the shape parameter to control smoothness. The scale parameter, say β (where $\mu = \alpha/\beta$), can then be used to choose the range, $x^\star$. A closed form inverse relationship will not be available analytically in general. However, if we choose $\alpha = 1$, which gives an exponential distribution, then

$$\beta = \frac{2}{x^\star} \log\left(\frac{1+M+\epsilon}{\epsilon(M+2)}\right). \tag{3}$$

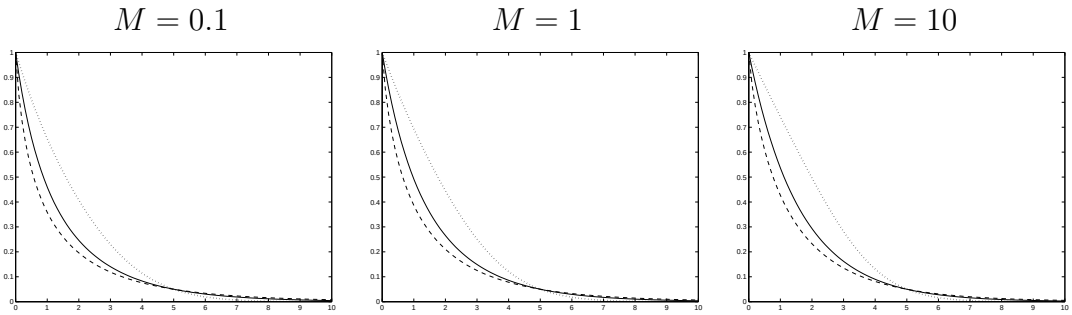

Figure 1: The autocorrelation function for a Gamma distance distribution with range 5 with shape: $\alpha = 0.1$ (dashed line), $\alpha = 1$ (solid line) and $\alpha = 10$ (dotted line)

Figure 1 shows the form of the autocorrelation for various smoothness parameters and a range fixed to 5. 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 is fixed. We will concentrate on the Gamma implementation and work with the following class

Definition 1 *Let* (t_i, C_i, r_i) *be a Poisson process with intensity* $\frac{\beta^\alpha}{\Gamma(\alpha)} r_i^{\alpha-1} \exp\{-\beta r_i\}$ *defined on* $\mathbb{R}_+ \times \mathbb{R}^p \times \mathbb{R}_+$ *with associated marks* (V_i, θ_i) *which are i.i.d. realisations of* $Be(1, M)$ *and* H. *Defining*

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

then $\{F_x | x \in \mathcal{X}\}$ *follows a* **Dirichlet Process Regression Smoother (DPRS)** *which is represented as* DPRS(M, H, α, β).

We define a prior distribution for M which can be elicited by choosing a typical value for M to be n_0 and a variance parameter η. This prior has the density function

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

and is discussed in more detail in GS. The function $1/(M + 1)$ can be useful measure of the lack of adequacy of H to represent the distribution of the data since $\text{Var}(F(B)) = H(B)(1 - H(B))/(M + 1)$.

3 Centring over models

It will be useful to define our nonparametric models to be centred over parametric models. We can then allow the nonparametric prior to model aspects of the conditional distribution that are not well captured by our parametric centring model. This approach could allow interpretation of the parameters and the use of prior information elicited for the parametric model. 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. A nonparametric model will be **centred** over a parametric model, with parameters θ, if the prior predictive distribution of the nonparametric model conditional on θ coincides with the parametric model for all covariate values.

A natural model to implement this idea assumes that the regression errors with respect to some parametric conditional regression function $g(x)$, say $g(x) = x'\gamma$, given by $\epsilon_i = y_i - g(x_i)$, can be modelled as described above, *i.e.*:

$$\epsilon_i \sim \mathrm{N}(\mu_i, a\sigma^2), \quad \mu_i \sim F_{x_i}, \quad F_x \sim \mathrm{DPRS}(M, H, \alpha, \beta), \quad H \sim \mathrm{N}(0, (1-a)\sigma^2)$$

where $0 < a < 1$. This model will be denoted here as **Model 1(a)** and is centred over $y_i \sim \mathrm{N}(g(x_i), \sigma^2)$ with the same prior on σ. Note that dependence on the covariates x enters in two different ways: it is used to define ϵ_i as deviations of y_i from $g(x_i)$ and the nonparametric distribution of the means μ_i depends on x_i through the density smoother defined in the previous section. For all values of x the model reduces to the standard Dirichlet process mixture of normals model. The parameterisation of the model is discussed by Griffin (2006), who pays particular attention to prior choice. Many distributional features, such as multimodality, are more easily controlled by a rather than M. 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.

An alternative specification, denoted by **Model 1(b)**, models the location of the errors by a Gaussian process prior as described in Subsection 2.1 and assumes

$$\epsilon_i - m(x_i) \sim \mathrm{N}(\mu_i, a{\sigma_\star}^2), \quad \mu_i \sim F_{x_i}, \quad F_x \sim \mathrm{DPRS}(M, H, \alpha, \beta), \quad H \sim \mathrm{N}(0, (1-a){\sigma_\star}^2),$$

with $\sigma^2 = {\sigma_\star}^2 + \sigma_0^2$. We define $b = \sigma_0^2/\sigma^2$, which can be interpreted as the amount 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 $y_i \sim \mathrm{N}(g(x_i) +$

$m(x_i), \sigma_\star^2)$, whereas if we integrate out both F_x and $m(x)$ with their priors we obtain $y_i \sim \mathrm{N}(g(x_i), \sigma^2)$.

Model 1 illustrates an important advantage of centring over a model: it provides a natural way to distinguish between the parametric dependence on covariates, captured by $x'\gamma$, and the nonparametric dependence, modelled through F_x and $m(x)$. Thus, by choosing $x'\gamma$ 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. It is important to 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 $g(x_i)$ and $P(\mathrm{E}[y_i|F_{x_i}, \gamma, x_i] = g(x_i)) = 0$ if y_i is a continuous random variable and H is absolutely continuous, which occur in a large proportion of potential applications. Thus, γ will not have the standard interpretation as the effects of the covariate on the mean of the response. The predictive mean, *i.e.* $\mathrm{E}[Y_i|\gamma, x_i]$ still equals $g(x_i)$, however. The prior uncertainty about this predictive mean will increase as our confidence in the centring model (as represented by M) decreases.

One solution 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 will coincide (if the mean exists). A second, wider, class of error distributions, introduced by Kottas and Gelfand (2001) to regression problems, is the class of unimodal densities with median zero (extensions to quantile regression are discussed in Kottas and Krnjajic 2005). Incorporating our Bayesian density regression smoother into this model defines **Model 2**:

$$\epsilon_i - m(x_i) \sim \mathrm{U}\left(-\sigma_\star\sqrt{u_i}, \sigma_\star\sqrt{u_i}\right), \quad u_i \sim F_{x_i}, \quad F_x \sim \mathrm{DPRS}(M, H, \alpha, \beta),$$

which leads to symmetric error distributions with $\mathrm{U}(a, b)$ denoting a Uniform distribution on (a, b). For H we choose a Gamma distribution with shape parameter $3/2$ and scale $1/2$, corresponding to a normal centring distribution. This model is centred exactly as Model 1(b).

4 Computational method

This section describes an MCMC sampler to fit to data $(x_1, y_1), \ldots, (x_n, y_n)$ the hierarchical model

$$p(y_i|\phi_i, \psi), \quad \phi_i|x_i \sim F_{x_i}, \quad F_x \sim \mathrm{DPRS}(M, H, \alpha, \beta).$$

Typically, we fix α and $x^\star$ and put appropriate priors on ψ, M and any parameters of H. Further developments for specific models described in the previous section, as well as details of the implementation are contained in Appendix B. This model can be fitted using the Retrospective Sampling methods for Dirichlet process-based hierarchical models described in Papaspiliopoulos and Roberts (2004). In fact, their ideas can be simply extended to the class of stick-breaking priors. Previous attempts to use the stick-breaking representation of the Dirichlet process in MCMC samplers have used truncation (*e.g.* Muliere and Tardella 1998, Ishwaran and James 2001 for the Dirichlet process and GS for order-based dependent Dirichlet processes). The Retrospective Sampler avoids the need to truncate and makes inference using the full model for all parameters except the distribution F_x from the correct posterior distribution. Inference about F_x must make use of a truncation method. Papaspiliopoulos and Roberts (2004) introduce latent allocation variables $s_1, \ldots, s_n$ to link the random effects to the distinct values of the stick-breaking distribution $\phi_i = \theta_{s_i}$. The algorithm exploits the fact that $K = \max\{s_1, \ldots, s_n\}$ must be finite and updating all parameters only relies on $(V_1, \theta_1), \ldots, (V_K, \theta_K)$ whereas $(\theta_{K+1}, V_{K+1}), (\theta_{K+2}, V_{K+2}), \ldots$ are not updated in the conditional posterior distribution.

The process is defined for centres, $C_1, C_2, C_3, \ldots$ in $\mathbb{R}^p$. However, we do not need to truncate the process. We define the region for which there exists an $x_i \in I$ for I drawn from the interval process. The posterior distribution of centres outside this region is not updated by the observed data and we can restrict our attention to the prior distribution of the interval I given that they fall within this region. If the intervals are balls around centres C_k with radius r_k then this conditional prior distribution can be expressed as

$$p(C_k|\text{there exists an } i \text{ for which } x_i \in I_k, r_k) = \mathrm{U}\left(\bigcup_{i=1}^{n} B_{r_k}(x_i)\right)$$

where $\mathrm{U}(A)$ represents the pdf of the uniform distribution on the set A and

$$p(r_k|\text{there exists an } i \text{ for which } x_i \in I_k) = \frac{\nu\left(\bigcup_{i=1}^{n} B_{r_k}(x_i)\right) p(r_k)}{\int \nu\left(\bigcup_{i=1}^{n} B_{r_k}(x_i)\right) p(r_k)\, dr_k}.$$

It is often hard to simulate from this conditional distribution of C_k and to calculate the normalising constant of the distribution of r_k. It will usually be simpler to use a rejection sampler from the joint distribution of (C_k, r_k) conditioned so that the interval I_k intersects a more simply-described set which is not much larger than the set of interest. In one dimension, we define $d^\star(r_k) = (x_{\min} - r_k, x_{\max} + r_k)$ where $x_{\min}$ and $x_{\max}$ are the maximum

and minimum values of $x_1, \ldots, x_n$. Then

$$p(C_k|I_k \text{ intersects } d^\star(r_k), r_k) = \mathrm{U}\left(d^\star(r_k)\right)$$

$$p(r_k|I_k \text{ intersects } d^\star(r_k)) = \frac{(x_{\max} - x_{\min} + 2r_k)\, p(r_k)}{\int (x_{\max} - x_{\min} + 2r_k)\, p(r_k)\, dr_k}$$

and we reject values of (C_k, r_k) if there is no i for which $x_i \in I_k$. If r_k follows a Gamma(α,β) distribution then

$$p(r_k|I_k \text{ intersects } d^\star(r_k)) = w f_{\mathrm{Ga}}(\alpha, \beta) + (1-w) f_{\mathrm{Ga}}(\alpha+1, \beta)$$

where $f_{\mathrm{Ga}}(\alpha, \beta)$ is the pdf of a Ga(α, β) distribution and $w = (x_{\max} - x_{\min})/(x_{\max} - x_{\min} + 2\frac{\alpha}{\beta})$.

Finally, we note the posterior distribution only depends on $t_1, t_2, t_3, \ldots$ through their ordering. We can simply extend the ideas of Papaspiliopoulos and Roberts (2004) to sample this ordering and the infinite dimensional parameters $V_1, V_2, V_3, \ldots$ and $\theta_1, \theta_2, \theta_3, \ldots$ since we can sample from the prior distribution of the intervals.

An MCMC sampler defined on the posteror distribution parameterised by r can have problems mixing. The sampler can have much better mixing properties by using the reparameterisation from r to $r^\star$ where we let $d_{il} = |x_i - C_l|$ and define $r_i^\star = r_i - \max\{d_{ij}|s_i = j\}$.

5 Examples

This section applies the methods developed on simulated data and two real datasets: the prestige data (Fox and Suschnigg, 1989) and the electricity data of Yatchew (2003). As a basic model, we take Model 1(a) with a regression function $g(x) = 0$. 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 3. Finally, we will also use Model 2, which will always have a Gaussian process specification for $m(x)$. The prior on M is as explained in Subsection 2.2 with $n_0 = 3$ and $\eta = 1$ for all examples. On the parameter a in Model 1 we adopt a Uniform prior over (0,1). The range $x^\star$ of the DPRS is such that the correlation is 0.4 at the median distance between the covariate values. Priors on σ^2 and on the parameters of the Gaussian process σ_0^2 and ζ are as in the benchmark prior of Palacios and Steel (2006). Finally, we fix the smoothness parameter ν of the Gaussian process at 1.

5.1 Example 1: Sine wave

We generated 100 observations from the following model $y_i = \sin(2\pi x_i) + \epsilon_i$ where x_i are uniformly distributed on $(0, 1)$ and the errors ϵ_i are independent and chosen to be heteroscedastic and non-normally distributed. We consider two possible formulations: *Error 1* assumes that ϵ_i follows a t-distribution with zero mean, 2.5 degrees of freedom and a conditional variance of the form $\sigma^2(x) = (x - \frac{1}{2})^2$ which equals 0 at $x = \frac{1}{2}$ and increases away from $\frac{1}{2}$. *Error 2* assumes that the error distribution has a mixture form $p(\epsilon_i|x_i) = 0.3\text{N}(0.3, 0.01) + 0.7\text{N}(-0.3 + 0.6x_i, 0.01)$. 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 can not be fitted using a mixture of uniforms. Initially, we assume *Error 1*. The results for Model

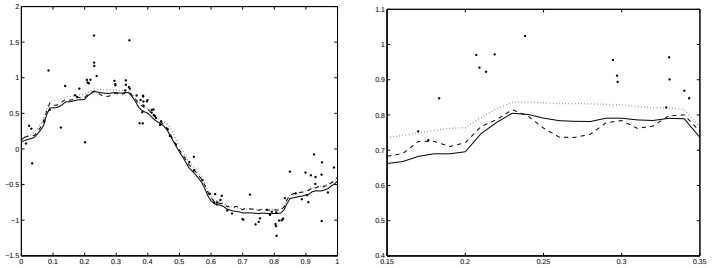

Figure 2: Example 1 with Error 1: predictive conditional mean of y given x for Model 1(a): $\alpha = 0.1$ (dashed line), $\alpha = 1$ (solid line), $\alpha = 10$ (dotted line). Data points are indicated by dots. The right panel presents a magnified section of the left panel

1(a) are illustrated in Figure 2 for three values 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 will 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 indicates 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.

We now turn our attention to the more complicated models where the conditional location is modelled through a nonparametric regression function (in this case a Gaussian process prior). Both Model 1(b) and Model 2 assume a constant prior mean for $m(x)$. Introducing

	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)

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

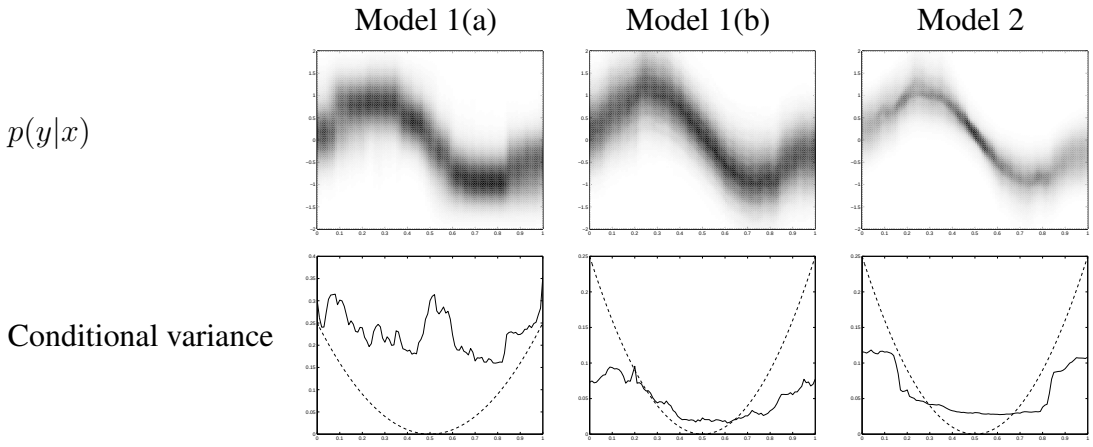

Figure 3: Example 1 with Error 1: heatmap of the posterior predictive density $p(y|x)$ and plot of the posterior conditional predictive variance $\sigma^2(x)$ (solid line) and the true value (dashed line)

the Gaussian process into Model 1 leads to smaller values of σ since some variability can now be explained by the Gaussian process prior. However, the posterior for a still favours fairly small values, reminding us that even if the conditional mean is better modelled with the Gaussian process, the tails are still highly non-Normal (see Figure 4). The estimated

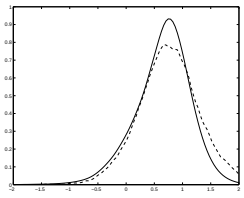

Figure 4: Example 1 with Error 1: The posterior predictive distribution of y for $x = 0.14$ using Model 1(a) (solid line) and Model 1(b) (dashed line)

posterior predictive distributions (as depicted in Figure 3) are now much smoother. Both Model 1(b) and Model 2 lead to a large estimate of σ_0 (which better fits the true variability of the mean). This leads to better estimates of the conditional predictive variance, as illustrated in Figure 3. Clearly a model of this type would struggle with estimation at the extreme values of x but the main part of the functional form is well-recovered. The parameter $\rho = 2\sqrt{\nu}/\zeta$ used in Table 1 is an alternative range parameter which is favoured by Stein (1999, p.51), and indicates that the Gaussian process dependence of $m(x)$ is similar for Model 1(a) and Model 2. The posterior median values of ρ lead to a range of the Gaussian process equal to 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) is 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 is not able to capture the bimodality by construction and only captures the changing variability. In both cases the mean is well estimated. Small values of a illustrate the difficulty in capturing 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.

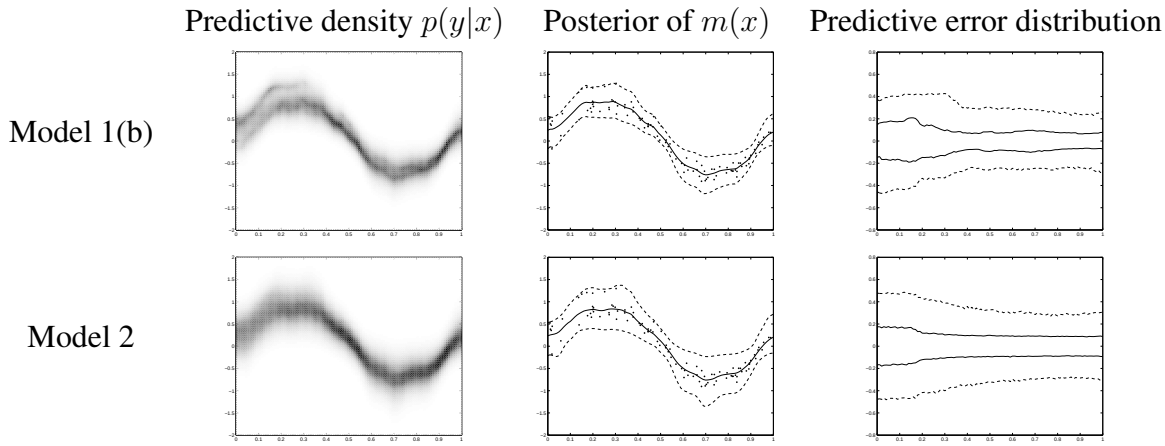

Figure 5: Example 1 with Error 2: heatmap of posterior predictive density $p(y|x)$, plot of the posterior of $m(x)$ indicating median (solid line), 95% credible interval (dashed lines) and data (dots), and the posterior predictive error distribution indicating the 2.5$^{\text{th}}$, 25$^{\text{th}}$,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

	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)

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

5.2 Prestige data

The "prestige" data was previously analysed by Fox and Suschnigg (1989) and records the income and prestige associated with 102 occupations taken from the 1971 Canadian census. The data is available to download in the R package `car`. Figure 6 shows the fitted conditional

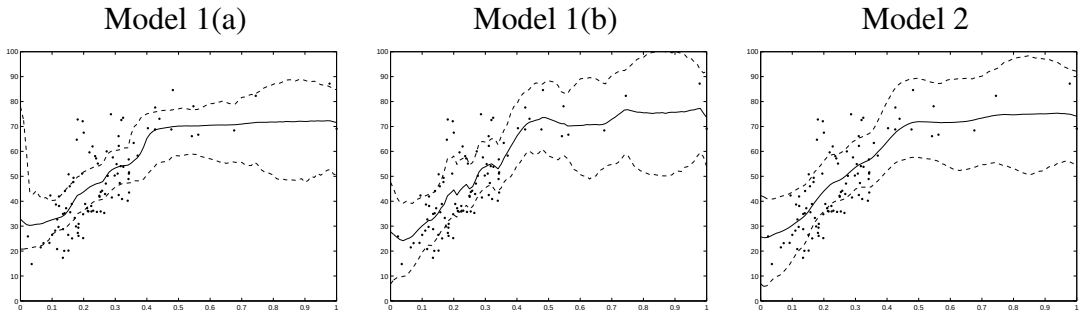

Figure 6: Prestige data: posterior distribution of the conditional mean indicating median (solid line), 95% credible interval (dashed lines) and data (dots)

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 result 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

	Model 1(a)	Model 1(b)	Model 2
σ	22.2 (14.8, 43.7)	20.0 (14.8, 30.0)	22.0 (16.2, 36.4)
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)

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

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) considers fitting a cost function for the distribution of electricity. A Cobb-Douglas model is fitted, which assumes that

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

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 will fit the DPRS model with cust as the covariate to ϵ and we will centre the model over two parametric regression models by choosing $f(\text{cust})$ as follows: Parametric 1, $\gamma_1 + \gamma_2\text{cust}$, and Parametric 2, $\gamma_1 + \gamma_2\text{cust} + \gamma_3\text{cust}^2$.

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

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 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 is much more similar for all models in Table 5. In particular, the parametric effect of customers is 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 which could be well captured by a quadratic effect, especially for Model 1(b) and 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)

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

Under both centring models, the importance of the nonparametric fitting of the error sharply decreases as a Gaussian process formulation for the regression function is used, as evidenced by the increase in a. Changing to a quadratic centring distribution leads 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.

6 Discussion

This paper shows how methods from Bayesian nonparametric density estimation and nonparametric estimation of the mean in regression models can be combined to define a range

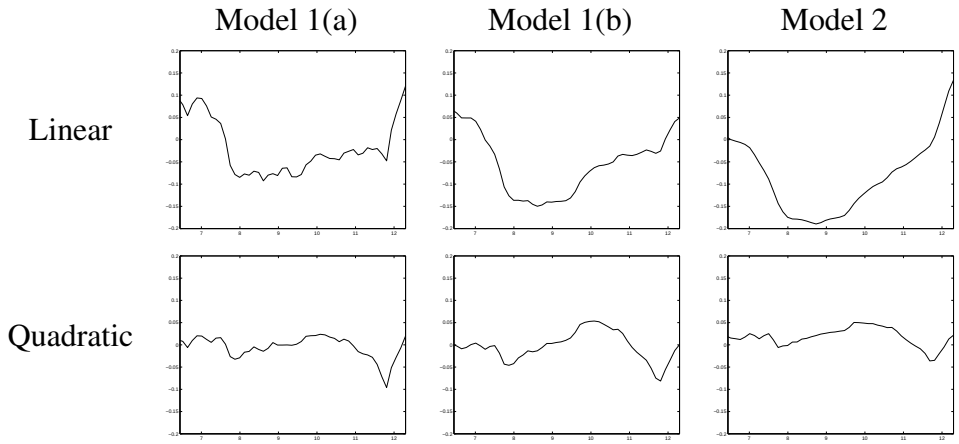

Figure 7: Electricity data: posterior mean of the nonparametric component(s) of the model

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, which 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. In this paper we have concentrated on one example which is the Dirichlet Process Regression Smoother (DPRS) prior. The latter 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.

Appendix A: Proofs

Proof of Theorem 1

The probability of observing a shared element given that the element relates to an ordering at s or v is $p_{s,v}$. Then, using the results of GS, $\#S_k = k$ and $\#S'_k$ is the number of elements appearing in only one of the orderings observed before the k-th shared element. Since these

elements follow a Poisson process, the probability that $\#S'_k = j$ is

$$\binom{j+k-1}{k-1}(1-p_{s,v})^j p_{s,v}^k$$

and so

$$\begin{aligned}
\text{Corr}(F_s, F_v) &= \frac{2}{M+2}\sum_{k=1}^{\infty}\sum_{j=0}^{\infty}\binom{j+k-1}{k-1}\left(p_{s,v}\frac{M}{M+2}\right)^k\left[(1-p_{s,v})\left(\frac{M}{M+1}\right)\right]^j \\
&= \frac{2}{M+2}\sum_{n=0}^{\infty}\sum_{i=0}^{n}\binom{n}{i}p_{s,v}\left(p_{s,v}\frac{M}{M+2}\right)^i\left[(1-p_{s,v})\left(\frac{M}{M+1}\right)\right]^{n-i} \\
&= \frac{2}{M+2}p_{s,v}\sum_{n=0}^{\infty}\left[\left(p_{s,v}\frac{M}{M+2}\right)+(1-p_{s,v})\left(\frac{M}{M+1}\right)\right]^n \\
&= \frac{2}{M+2}p_{s,v}\sum_{n=0}^{\infty}\left(\frac{M}{M+1}\right)^n\left[1-\frac{p_{s,v}}{M+2}\right]^n = 2\frac{\frac{M+1}{M+2}p_{s,v}}{1+\frac{M}{M+2}p_{s,v}}.
\end{aligned}$$

Proof of Theorem 2

Since (C, r, t) follows a Poisson process on $\mathbb{R}^p \times \mathbb{R}^2_+$ with intensity $p(r)$, $p(C_k|s \in I_k \text{ or } v \in I_k, r_k) = \mathrm{U}\left(B_{r_k}(s) \cup B_{r_k}(v)\right)$ and $p(r_k|s \in I_k \text{ or } v \in I_k) = \frac{\nu\left(B_{r_k}(s)\cup B_{r_k}(v)\right)p(r_k)}{\int \nu\left(B_{r_k}(s)\cup B_{r_k}(v)\right)p(r_k)\,dr_k}$ where $\nu(\cdot)$ is Lebesgue measure. Then

$$\begin{aligned}
p_{s,v} &= P(s, v \in I_k | s \in I_k \text{ or } v \in I_k) \\
&= \int\int_{B_{r_k}(s)\cap B_{r_k}(v)} p\left(C_k, r_k | s \in I_k \text{ or } v \in I_k\right)\, dC_k\, dr_k \\
&= \frac{\int \nu\left(B_{r_k}(s) \cap B_{r_k}(v)\right) p(r_k)\, dr_k}{\int \nu\left(B_{r_k}(s) \cup B_{r_k}(v)\right) p(r_k)\, dr_k}.
\end{aligned}$$

Proof of Theorem 3

The autocorrelation function can be expressed as $f(p_{s,s+u})$ where $f(x) = 2(\frac{M+1}{M+2})/(1+\frac{M}{M+2}x)$. Then by Faá di Bruno's formula

$$\frac{d^n}{du^n}f(p_{s,s+u}) = \sum \frac{n!}{m_1!m_2!m_3!\ldots}\frac{d^{m_1+\cdots+m_n}f}{dp_{s,s+u}^{m_1+\cdots+m_n}}\prod_{\{j|m_j\neq 0\}}\left(\frac{d^j p_{s,s+u}}{du^j}\frac{1}{j!}\right)^{m_j},$$

where $m_1 + 2m_2 + 3m_3 + \cdots + nm_n = n$ with $m_j \geq 0, j = 1, \ldots, n$, and so

$$\lim_{u\to 0}\frac{d^n}{du^n}f(p_{s,s+u}) = \sum \frac{n!}{m_1!m_2!m_3!\ldots}\lim_{u\to 0}\frac{d^{m_1+\cdots+m_n}f}{dp_{s,s+u}^{m_1+\cdots+m_n}}\prod_{\{j|m_j\neq 0\}}\left(\frac{d^j p_{s,s+u}}{du^j}\frac{1}{j!}\right)^{m_j}.$$

Since $\lim_{u\to 0}\frac{d^{m_1+\cdots+m_n}f}{dp_{s,s+u}^{m_1+\cdots+m_n}} = \lim_{p_{s,s+u}\to 1}\frac{d^{m_1+\cdots+m_n}f}{dp_{s,s+u}^{m_1+\cdots+m_n}}$ is finite and non-zero for all values of n, the degree of differentiability of the autocorrelation function is equal to the degree of differentiability of $p_{s,s+u}$. We can write $p_{s,s+u} = \left(\frac{4\mu}{a}-1\right)^{-1}$ with $a = 2\mu_2 - uI$. Now $\frac{d^k p_{s,s+u}}{a^k} = (k-1)!(4\mu-a)^{-k}$ and $\lim_{u\to 0}\frac{d^k p_{s,s+u}}{da^k} = (k-1)!(2\mu)^{-k}$ which is finite and non-zero. By application of Faá di Bruno's formula

$$\frac{d^n}{du^n}p_{s,s+u} = \sum \frac{n!}{m_1!m_2!m_3!\ldots}\frac{d^{m_1+\cdots+m_n}p_{s,s+u}}{da^{m_1+\cdots+m_n}}\prod_{\{j|m_j\neq 0\}}\left(\frac{d^j a}{du^j}\frac{1}{j!}\right)^{m_j}$$

and the degree of differentiability is determined by the degree of differentiability of a. If $p(r)\sim \text{Ga}(\alpha,\beta)$ then $\frac{d\mu_2}{du} = -\frac{1}{2}\left(\frac{u}{2}\right)^{\alpha}\exp\{-u/2\}$ and $\frac{dI}{du} = -\frac{1}{2}\left(\frac{u}{2}\right)^{\alpha-1}\exp\{-u/2\}$ and it is easy to show that $\frac{d^n a}{du^n} = C_n u^{\alpha-n+1}\exp\{-u/2\}+\zeta$ where ζ contains terms with power of x greater than $\alpha-n+1$. If $\lim_{u\to 0}u^{\alpha}\exp\{-u/2\}$ is finite then so is $\lim_{u\to 0}u^{\alpha+k}\exp\{u/2\}$ for $k>0$ and so the limit will be finite iff $\alpha-n+1\geq 0$, *i.e.* $\alpha\geq n-1$.

Appendix B: Computational details

In keeping with standard Gibbs sampler notation $s_{-i} = (s_1,\ldots,s_{i-1},s_{i+1},\ldots,s_n)$.

Updating the allocation s_i

Conditioning on $r^\star = (r_1^\star, r_2^\star,\ldots)$ rather than $r = (r_1, r_2,\ldots)$ allows each observation to be allocated to a distinct element. Initially, we condition on s_{-i} and remove s_i from the allocation. Let $K_{-i} = \max\{s_{-i}\}$ and let $r_k^{(1)} = r_j^\star + \max\{d_{jk}|s_j = k, j = 1,\ldots,i-1,i+1,\ldots,n\}$ and $r_k^{(2)} = r_j^\star + \max(\{d_{jk}|s_j = k, j = 1,\ldots,i-1,i+1,\ldots,n\}\cup\{|x_i - C_j|\})$. The proposal distribution is

$$q(s_i = k) = c^{-1}\times\begin{cases} p(y_i|\theta_k)V_k\prod_{l<k}(1-V_l)(1-V_i)^{\#\left\{m\middle|r_k^{(1)}<d_{ms_m}<r_k^{(2)},s_m>k\right\}}\dfrac{p\left(r_k^{(2)}\right)}{p\left(r_k^{(1)}\right)} & k\leq K_{-i}\\ \max_{m\leq K_{-i}}\{p(y_i|\theta_m)\}V_k\prod_{l<k}(1-V_l) & k>K_{-i}\end{cases}$$

$$\text{where } c = \sum_{l=1}^{K_{-i}} p(y_i|\theta_l)V_l\prod_{h<l}(1-V_h)(1-V_i)^{\#\left\{m\middle|r_l^{(1)}<d_{ms_m}<r_l^{(2)},s_m>l\right\}}\frac{p\left(r_l^{(2)}\right)}{p\left(r_l^{(1)}\right)}$$
$$+\max_{l\leq K_{-i}}\{p(y_i|\theta_l)\}\prod_{h\leq K_{-i}}(1-V_h).$$

If $k > K_{-i}$ we need to generate $(\theta_{K_{-i}+1}, V_{K_{-i}+1}, C_{K_{-i}+1}, d_{K_{-i}+1}), \ldots, (\theta_k, V_k, C_k, d_k)$ 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 k is such that $\sum_{l=1}^{k-1} q(s_i = l) < U \leq \sum_{l=1}^{k} q(s_i = l)$). Papaspiliopoulos and Roberts (2004) show that the acceptance probability of the proposed value is

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

Updating the centres

We update each centre $C_1, \ldots, C_K$ from its full conditional distribution Metropolis-Hastings random walk step. A new value C_i' for the i-th centre is proposed from $\mathrm{N}(C_i, \sigma_C^2)$ where σ_C^2 is chosen so that the acceptance rate is approximately 0.25. If there is no x_i such that $x_i \in (C_i' - r_i, C_i' + r_i)$ or if there is one value of j such that $s_j = i$ for which $x_i \notin (C_i' - r_i, C_i' + r_i)$ then $\alpha(C_i, C_i') = 0$. Otherwise, the acceptance probability has the form

$$\alpha(C_i, C_i') = \frac{\prod_{j=1}^{n} \prod_{h < s_j \text{ and } C_h' - r_h < x_j < C_h' + r_h} (1 - V_h)}{\prod_{j=1}^{n} \prod_{h < s_j \text{ and } C_h - r_h < x_j < C_h + r_h} (1 - V_h)}.$$

Updating the distances

The distances can be updated using a Gibbs step since the full conditional distribution of r_k has a simple piecewise form. Recall that $d_{ik} = |x_i - C_k|$ and let $\mathcal{S}_k = \{j | s_j \geq k\}$. We define $\mathcal{S}_k^{ord}$ to be a version of $\mathcal{S}_k$ where the element have been ordered to be increasing in d_{ik}, *i.e.* if $i > j$ and $i, j \in \mathcal{S}_k^{ord}$ then $d_{ik} > d_{jk}$. Finally we define $d_k^\star = \max[\{x_{\min} - C_k, C_k - x_{\max}\} \cup \{d_{ik} | s_i = k\}]$ and $m^\star$ be such that $x_i \in S_k^{ord}$ and $x_{m^\star} > d_k^\star$ and $x_{m^\star - 1} < d_k^\star$. Let l be the length of S_k^{ord}. The full conditional distribution has density

$$f(z) \propto \begin{cases} p(r_k = z) & \text{if } d_k^\star < z \leq d_{S_{m^\star}^{ord} k} \\ p(r_k = z)(1 - V_k)^{i - m^\star + 1} & \text{if } d_{S_i^{ord} k} < z \leq d_{S_{i+1}^{ord} k}, \quad i = m^\star, \ldots, l - 1 \\ p(r_k = z)(1 - V_k)^{l - m^\star + 1} & \text{if } z > d_{S_l^{ord} k} \end{cases}.$$

Swapping the positions of atoms

The ordering of the atoms should also be updated in the sampler. One of the K included atoms say $(V_i, \theta_i, C_i, r_i)$ is chosen at random to be swapped with the subsequent atom

$(V_{i+1}, \theta_{i+1}, C_{i+1}, r_{i+1})$. If $i < K$, the acceptance probability of this move is $\min\{1, (1-V_{i+1})^{n_i}/(1-V_i)^{n_{i+1}}\}$. If $i = K$, then a new point $(V_{K+1}, \theta_{K+1}, C_{K+1}, r_{K+1})$ is proposed from their prior and the swap is accepted with probability $\min\{1, (1-V_{K+1})^{n_i}\}$.

Updating θ, V and ψ

The full conditional distribution of θ_i is proportional to $h(\theta_i)\prod_{\{j|s_i\}} p(y_j|\theta_i)$, where h is the density function of H. We update V_i from a Beta distribution with parameters $1 + \sum_{j=1}^{n} \mathrm{I}(s_j = i)$ and $M + \sum_{j=1}^{n} \mathrm{I}(s_j > i, |x_j - C_i| < r_i)$. The full conditional distribution of ψ is proportional to $p(\psi)\prod_{i=1}^{n} p(y_i|\theta_{s_i})$.

Updating M

This parameter can be updated by a random walk on the log scale. Propose $M' = M\exp(\epsilon)$ where $\epsilon \sim \mathrm{N}(0, \sigma_M^2)$ with σ_M^2 a tuning parameter chosen to maintain an acceptance rate close to 0.25. The proposed value should be accepted with probability

$$\frac{M'^{K+1}\left[\prod_{i=1}^{K}(1-V_i)\right]^{M'}\beta(M')^{\alpha K}\exp\{-\beta(M')\sum_{i=1}^{K} r_i\}p(M')}{M^{K+1}\left[\prod_{i=1}^{K}(1-V_i)\right]^{M}\beta(M)^{\alpha K}\exp\{-\beta(M)\sum_{i=1}^{K} r_i\}p(M)},$$

where $\beta(M)$ is β expressed as a function of M, as in (3).

Posterior inferences on $F_{\tilde{x}}$

We are often interested in inference at some point $\tilde{x} \in \mathcal{X}$ about the distribution $F_{\tilde{x}}$. We define $(\tilde{V}_1, \tilde{\theta}_1), (\tilde{V}_2, \tilde{\theta}_2), \ldots, (\tilde{V}_J, \tilde{\theta}_J)$ to be the subset of $(V_1, \theta_1), (V_1, \theta_2) \ldots, (V_K, \theta_K)$ for which $|\tilde{x} - C_i| < r_i$. Then

$$\begin{aligned} F_{\tilde{x}} = &\sum_{i=1}^{J} \delta_{\tilde{\theta}_i} \tilde{V}_i \prod_{j<i}(1-\tilde{V}_j) \prod_{j\le i}\prod_{l=1}^{n_j}\left(1 - V_l^{(j)}\right) + \prod_{i\le J}\prod_{j=1}^{n_i}\left(1 - V_j^{(i)}\right) \sum_{l=J+1}^{\infty} \delta_{\tilde{\theta}_l} \tilde{V}_l \prod_{m<l}(1-\tilde{V}_m) \\ &+ \sum_{i=1}^{N}\sum_{j=1}^{n_i} \delta_{\theta_j^{(i)}} V_j^{(i)} \prod_{l<j}\left(1 - V_l^{(i)}\right) \prod_{l<i}(1-V_l) \prod_{m=1}^{n_l}\left(1 - V_m^{(l)}\right) \end{aligned}$$

where n_j is a geometric random variable with success probability $1 - \tilde{p}$, $\theta_j^{(i)} \sim H$, $V_j^{(i)} \sim \mathrm{Be}(1, M)$, $\tilde{\theta}_m \sim H$ and $\tilde{V}_m \sim \mathrm{Be}(1, M)$ for $m > N$. We calculate $\tilde{p}$ in the following way.

If $x_{min} < \tilde{x} < x_{max}$, define i so that $x_{(i)} < \tilde{x} < x_{(i+1)}$, where $x_{(1)}, \ldots, x_{(n)}$ is an ordered version of $x_1, \ldots, x_n$, then $\tilde{p} = \frac{\beta}{2\alpha}\tilde{q}$ where

$$\begin{aligned}\tilde{q} =&(x_{(i+1)} - x_{(i)})\mathcal{I}\left(\frac{x_{(i+1)} - x_{(i)}}{2}\right) + (x_{(i)} - \tilde{x})\mathcal{I}\left(\frac{\tilde{x} - x_{(i)}}{2}\right) - (x_{(i+1)} - \tilde{x})\mathcal{I}\left(\frac{x_{(i+1)} - \tilde{x}}{2}\right)\\ &- 2\mu^\star\left(\frac{x_{(i+1)} - x_{(i)}}{2}\right) + 2\mu^\star\left(\frac{\tilde{x} - x_{(i)}}{2}\right) + 2\mu^\star\left(\frac{x_{(i+1)} - \tilde{x}}{2}\right)\end{aligned}$$

with $\mathcal{I}(y) = \int_0^y p(r)\,dr$ and $\mu^\star(y) = \int_0^y rp(r)\,dr$. Otherwise if $\tilde{x} < x_{min}$

$$\tilde{q} = 2\mu^\star\left(\frac{x_{min} - \tilde{x}}{2}\right) + (x_{min} - \tilde{x})\left(1 - \mathcal{I}\left(\frac{x_{min} - \tilde{x}}{2}\right)\right)$$

and if $\tilde{x} > x_{max}$

$$\tilde{q} = 2\mu^\star\left(\frac{\tilde{x} - x_{max}}{2}\right) + (\tilde{x} - x_{max})\left(1 - \mathcal{I}\left(\frac{\tilde{x} - x_{max}}{2}\right)\right).$$

We use a truncated version of $F_{\tilde{x}}$ with h elements which are chosen so that $\sum_{i=1}^h p_i = 1 - \epsilon$ where ϵ is usually taken to be 0.001.

Model 2

This section is restricted to discussing the implementation when $m(x)$ follows a Gaussian process prior where we define $P_{ij} = \rho(x_i, x_j)$. We also reparametrise from u_i to $\phi_i = \sigma^2 u_i$.

Updating $u_i|s$

The full conditional distribution has the density

$$p(u_i) \propto \phi_i^{0.5(1-\sum \mathrm{I}(s_j=i, 1\le j\le n))} \exp\{-0.5\phi_i/\sigma^2\}, \qquad \phi_i > \phi_{min}$$

where $\phi_{min} = \max\left\{(y_i - m(x_i))^2 \,|s_j = i, 1 \le j \le n\right\}$. A rejection sampler for this full conditional distribution can be constructed using the envelope

$$h^\star(u) \propto \begin{cases} \phi^{0.5(1-\sum \mathrm{I}(s_j=i,1\le j\le n))} & \phi_{min} < \phi < z \\ z^{0.5(1-\sum \mathrm{I}(s_j=i,1\le j\le n))} \exp\{-0.5(\phi - z)/\sigma^2\} & \phi > z \end{cases}$$

which can be sampled using inversion sampling. The acceptance probability is

$$\alpha(u) = \begin{cases} \exp\{-0.5(\phi - \phi_{min})/\sigma^2\} & \phi_{min} < \phi < z \\ \left(\frac{\phi}{z}\right)^{0.5-0.5k} \exp\{-0.5(z - \phi_{min})/\sigma^2\} & \phi > z \end{cases}$$

and the choice $z = \sigma^2 \sum \mathrm{I}(s_j = i, 1 \le j \le n)$ maximizes the acceptance rate.

Updating σ^{-2}

Using the prior Gamma(ν_1, ν_2), the full conditional distribution of σ^{-2} is again a Gamma distribution, where we define $P = (P_{ij})$

$$\sigma^{-2} \sim \text{Ga}\left(\nu_1 + \frac{3K}{2} + \frac{n}{2}, \nu_2 + \frac{1}{2}\sum_{i=1}^{K}\phi_i + \frac{1}{2\omega}m(x)^T P^{-1} m(x)\right).$$

Updating $m(x_1), \ldots, m(x_n)$

It is possible to update $m(x_i)$ using its full conditional distribution. However this tends to lead to slowly mixing algorithms. A more useful approach uses the transformation $m(x) = C^\star z$ where $C^\star$ is the Cholesky factor of $\sigma_0^{-2}P^{-1}$, where $z \sim \text{N}(0, I)$. We then update z_j using their full conditional distribution which is a standard normal distribution truncated to the region $\cap_{i=1}^{n}(y_i - \sum_{k\neq j} C_{ik}z_k - \sqrt{\phi_i}, y_i - \sum_{k\neq j} C_{ik}z_k + \sqrt{\phi_i})$.

Updating ω

We define $\omega^2 = \sigma^2/\sigma_0^2$. If ω^2 follows a Gamma distribution with parameters a_0 and b_0 then the full conditional of σ_0^{-2} follows a Gamma distribution with parameters $a_0 + n/2$ and $b_0 + \sigma^{-2}m(x)^T P^{-1} m(x)/2$. A similar updating occurs for the Generalized inverse Gaussian prior used here.

Updating the Matèrn parameters

We update any parameters of the Matèrn correlation structure by a Metropolis-Hastings random walk. The full conditional distribution of the parameters (ζ, ν) would be proportional to

$$|P|^{-1/2}\exp\{-\sigma^{-2}\omega^{-2}m(x)^T P^{-1} m(x)\}p(\zeta, \nu).$$

References

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

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.

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

Ferguson, T. S. (1973), "A Bayesian Analysis of Some Nonparametric Problems," *The Annals of Statistics*, 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," *Canadian Journal of Sociology*, 14, 353-360.

Gelfand, A.E., Kottas, A., and MacEachern, S.N. (2005), "Bayesian Nonparametric Modelling with Dirichlet Process Mixing," *Journal of the American Statistical Association*, 100, 1021-1035.

Griffin, J.E. (2006), "On the Bayesian Analysis of Species Sampling Mixture Models for Density Estimation," Technical Report, University of Warwick.

Griffin, J.E., and Steel, M.F.J. (2006), "Order-based Dependent Dirichlet Processes," *Journal of the American Statistical Association*, 101, 179-194.

Ishwaran, H., and James, L. (2001), "Gibbs Sampling Methods for Stick-Breaking Priors," *Journal of the American Statistical Association*, 96, 161-73.

James, L. F., Lijoi, A. and Prünster, I. (2005), "Bayesian Inference via Classes of Normalized Random Measures," Technical Report, http://arxiv.org/abs/math/0503394.

Kohn, R., Smith, M., and Chan, D. (2001), "Nonparametric Regression Using Linear Combinations of Basis Functions," *Statistics and Computing*, 11, 313-322.

Kottas, A., and Gelfand, A.E. (2001), "Bayesian Semiparametric Median Regression Modeling," *Journal of the American Statistical Association*, 96, 1458-1468.

Kottas, A., and Krnjajic, M. (2005), "Bayesian Nonparametric Modeling in Quantile Regression," *Technical Report*, University of California, Santa Cruz.

Leslie, D. S., Kohn, R., and Nott, D. J. (2007), "A General Approach to Heteroscedastic Linear Regression," *Statistics and Computing*, to appear.

Mallick, B. K. and Walker, S.G. (1997), "Combining Information ¿From Several Experiments with Nonparametric Priors," *Biometrika*, 84, 697-706.

Muliere, P. and Tardella, L. (1998), "Approximating Distributions of Random Functionals of Ferguson-Dirichlet Process," *Canadian Journal of Statistics*, 26, 283-297.

Müller, P. and Quintana, F. (2004), "Nonparametric Bayesian Data Analysis," *Statistical Science*, 19, 95-110.

Neal, R. M. (1998) "Regression and Classification using Gaussian Process Priors," in *Bayesian Statistics 6* (eds: J. M. Bernardo, J. O. Berger, A. P. Dawid, A. F. M. Smith), Oxford: Oxford University Press, 475-501.

O'Hagan, A. (1978), "Curve Fitting and Optimal Design for Prediction," *Journal of the Royal Statistical Society*, B, 40, 1-42.

Palacios, M.B., and Steel, M.F.J. (2006), "Non-Gaussian Bayesian Geostatistical Modeling," *Journal of the American Statistical Association*, 101, 604-618.

Papaspiliopoulos, O., and Roberts, G. (2004), "Retrospective MCMC for Dirichlet Process Hierarchical Models," Technical Report, University of Lancaster.

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

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

Tibshirani, R. and Hastie, T. (1987), "Local Likelihood Estimation," *Journal of the American Statistical Association*, 82, 559-567.

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