Electronic Journal of Statistics
Vol. 6 (2012) 2038–2068
ISSN: 1935-7524
DOI: 10.1214/12-EJS738

Adaptive-modal Bayesian nonparametric regression

George Karabatsos

University of Illinois-Chicago
e-mail: georgek@uic.edu; gkarabatsos1@gmail.com

and

Stephen G. Walker

University of Kent, United Kingdom
e-mail: S.G.Walker@kent.ac.uk

Abstract: We introduce a novel, Bayesian nonparametric, infinite-mixture regression model. The model has unimodal kernel (component) densities, and has covariate-dependent mixture weights that are defined by an infinite ordered-category probits regression. Based on these mixture weights, the regression model predicts a probability density that becomes increasingly unimodal as the explanatory power of the covariate (vector) increases, and increasingly multimodal as this explanatory power decreases, while allowing the explanatory power to vary from one covariate (vector) value to another. The model is illustrated and compared against many other regression models in terms of predictive performance, through the analysis of many real and simulated data sets.

Keywords and phrases: Bayesian inference, nonparametric regression, unimodal distribution, binary regression.

Received April 2012.

1. Introduction

In terms of Bayesian inference, a prior is nothing more than a probability measure on distribution functions. More common in parametric inference is that one is putting all the mass of the prior onto specific shaped distribution functions, such as the normal, and hence the probability measure need only be assigned to the mean and variance parameters. In nonparametric problems, on the other hand, the need is to put a probability measure on a variety of shapes of distribution function and the prior in such cases is characterized by the random distribution functions generated from the prior. The Dirichlet process is one such random type of construction which generates distribution functions (Ferguson, 1973). The obvious drawback to these constructions is that discrete distribution functions are constructed, and hence there has been an interest in constructing random distribution functions which admit density functions. This is provided by means of the DP mixture model (Lo, 1984). Indeed, Bayesian nonparametric inference has become routinely employed in a wide range of application areas,

with the DP mixture model being the most widely-used Bayesian nonparametric model.

A considerable amount of current research has extended such nonparametric models to the regression case. The vast majority of this research is based on infinite-mixture models that take on the general form:

$$f(y|\mathbf{x}) = \int f(y|\mathbf{x},\boldsymbol{\theta})\mathrm{d}G_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{j=1}^{\infty} f(y|\mathbf{x},\boldsymbol{\theta}_j(\mathbf{x}))\omega_j(\mathbf{x}), \tag{1}$$

where the random mixing distribution,

$$G_{\mathbf{x}}(\cdot) = \sum_{j=1}^{\infty} \omega_j(\mathbf{x})\delta_{\boldsymbol{\theta}_j(\mathbf{x})}(\cdot), \tag{2}$$

depends on covariates $\mathbf{x} = (x_1, \ldots, x_p)^\intercal$, the mixture weights $\omega_j(\mathbf{x})$ sum to 1 at every $\mathbf{x} \in \mathcal{X}$, and $\delta_{\boldsymbol{\theta}}(\cdot)$ denotes a degenerate distribution with point mass at $\boldsymbol{\theta}$. The Bayesian regression model (1) is completed by the specification of a prior distribution on the weights $\{\omega_j(\mathbf{x})\}_{j=1,2,\ldots}$ and atoms $\{\boldsymbol{\theta}_j(\mathbf{x})\}_{j=1,2,\ldots}$, which are infinite collections of processes indexed by the $\mathbf{x}$-space. The Bayesian nonparametric regression model (1) is very general and encompasses many regression models; including fixed-effects and random-effects linear and generalized linear models (GLMs), finite-mixture latent-class and hierarchical mixtures-of-experts regression models, as well as infinite mixtures of Gaussian process regressions.

Nearly all of the research on the general model (1) has focused on Dependent Dirichlet process (DDP) (MacEachern, 1999; 2000; 2001). One popular approach to DDP modeling specifies a regression model for the atoms $\boldsymbol{\theta}_j(\mathbf{x})$, but assumes covariate-independent stick-breaking weights $\omega_j = \upsilon_j \prod_{k=1}^{j-1}(1-\upsilon_k)$, with beta random variates $\upsilon_j \sim_{ind} \mathrm{be}(a_j, b_j)$, as in the ordinary DP and other stick-breaking processes (Ishwaran & James, 2001). For example, the ANOVA/linear DDP model specifies the linear structure $\boldsymbol{\theta}_j(\mathbf{x}) = \mathbf{x}^\intercal\boldsymbol{\beta}_j$ with $\boldsymbol{\beta}_j \sim_{i.i.d.} G_0$ for normal kernel densities $f(y|\mathbf{x},\boldsymbol{\theta}_j(\mathbf{x})) := \mathrm{n}(y|\boldsymbol{\theta}_j(\mathbf{x}),\sigma^2)$ (e.g., De Iorio et al. 2004). Similar models take $\boldsymbol{\theta}_j(\mathbf{x}) \sim_{ind} G_{0\mathbf{x}}$, with $G_{0\mathbf{x}}$ a general Gaussian process (e.g., Gelfand et al. 2005). Another DDP modeling approach specifies covariate-dependent stick-breaking weights of the form $\omega_j(\mathbf{x}) = \upsilon_j(\mathbf{x})\prod_{k=1}^{j-1}(1-\upsilon_k(\mathbf{x}))$, with $\upsilon_j \sim Q_j$ and $\upsilon_j(\mathbf{x}) : \mathcal{X} \to [0,1]$ (e.g., Griffin & Steel, 2006; Dunson & Park, 2008; Rodríguez & Dunson, 2011). Yet another DDP modeling approach specifies a DP multivariate normal mixture model for $(\mathbf{x}, y)$, for inference of the conditional density $f(y|\mathbf{x})$ (Müller et al. 1996; Walker & Karabatsos, 2012). Other DDP-based models include the hierarchical DP (Teh et al., 2006), and the nested DP (Rodríguez, et al. 2008). An alternative, non-DDP based model of the form (1) assumes a covariate-dependent geometric distribution for the weights (Fuentes-García et al. 2010).

Several Bayesian nonparametric regression models are not mixture models of the form (1). They include regression models that are based on a continuous mixture distribution G that is assigned a mixture of Pólya trees (MPT), with $G \sim \int \mathrm{PT}(\alpha, G_0(\boldsymbol{\theta}))\mathrm{d}P(\boldsymbol{\theta})$, where $\mathrm{PT}(\alpha, G_0(\boldsymbol{\theta}))$ a finite Pólya tree prior

(Hanson, 2006), or based a dependent continuous mixing distribution $G_{\mathbf{x}}$ that is modeled by another tailfree process (Jara & Hanson, 2011). They also include models that are based on either a logistic Gaussian process for $(\mathbf{x}, y)$ (Tokdar et al., 2010), as well as product partition models (Müller & Quintana, 2010; Park & Dunson, 2010; Müller et al., 2011; Holmes et al. 2005).

While we have briefly summarized the models for Bayesian nonparametric regression, we also need to make some pertinent comments. In the parametric case, it is typical that the extent to which the covariates are influencing the outcomes is via the modeling of changes in mean and variance of a family of distributions, such as the normal. If one models normal data without covariates then one can anticipate a level of variance. If now relevant covariates are included into the model as a linear structure then one must anticipate that the variance estimator is reduced. And often quite substantially. So one notices that one is looking for a reduction in the variance of the error distribution when the covariates are explaining the outcomes.

To us at least, this idea appears lacking in the current Bayesian nonparametric regression models. The probability model for $f(y|\mathbf{x})$ is as flexible as it is for the model one would employ without covariates, i.e. $f(y)$. This is evident simply by looking at (1) and (2). So what is the equivalent feature we are looking for to reduce in the nonparametric case and which should be modeled explicitly? Since the nonparametric model without covariates has an infinite number of mixtures (and hence an arbitrary number of modes), which is clearly the dominating aspect of the model, it is the number of modes that we are seeking to reduce when modeling the covariates. Specifically we should be allowing the model to indicate the number of modes at each covariate.

Hence, with normal linear regression we aim to reduce the variance; with nonparametric regression we aim to reduce the number of modes. Here we further elaborate and explain this particular choice. If at any particular $\mathbf{x}$ it is that $f(y|\mathbf{x})$ is highly multimodal then prediction at or near this $\mathbf{x}$ is problematic. Multimodes are commonly associated with different behavior or clusters and so prediction here would be lacking conciseness. Historically, we would expect covariates to have been collected which allowed $f(y|\mathbf{x})$ to be modeled as a close to unimodal density function for each $\mathbf{x}$. To make this argument more concrete, let us consider an illustration. Suppose at each $\mathbf{x}$ the different modes of $f(y|\mathbf{x})$ correspond to an outcome based on some unobserved state of the individual. Prediction then becomes almost meaningless since one would actually be predicting by mixing over a number of possible states. It would in this case be incumbent on the experimenter to collect the necessary information for the individual, i.e. the missing information (covariate) to properly classify people and make sense of prediction. Hence, a unimodal prediction, and the necessary covariate information to effect this, are essential. This is not about collecting any covariate; just those which make and ensure meaningful prediction. A stronger point of view is that prediction with multimodal outcomes is practically meaningless. Each mode representing something different.

Current models seem not to be dealing with this issue and rather model the maximum number of modes needed to cover all covariate values. Our model

explicitly models the number of modes at each covariate. So a small or even single mode if the covariate at a particular location is explaining a lot; and many modes if the covariate at another location is explaining little.

Therefore, in this paper, we introduce a novel Bayesian nonparametric regression model, which has a random density of the form (1). Specifically, the covariate dependent mixture weights $\omega_j(\mathbf{x})$ are modeled by an ordered-category probits regression for the infinite component labels $j = 0, \pm 1, \pm 2, \ldots$, and the kernels $f(y|\boldsymbol{\theta}_j)$ are chosen to be unimodal kernel densities that are independent of $\mathbf{x}$.

Our basic modeling assumption is that the covariate $\mathbf{x}$ is explaining more precisely where the outcome y is to be found. We can assume, at one end of the scale, that precision is such that the regression density function is unimodal. So given an $\mathbf{x}$ the model will determine which $f_j(\cdot|\boldsymbol{\theta}) = f(\cdot|\boldsymbol{\theta}_j)$ the y comes from. Hence, we would model:

$$f(y|\mathbf{x}) = \sum_j \mathbb{I}(\mathbf{x} \in A_j)\, f_j(y)$$

for some set of disjoint sets $(A_j)_{j=1}^{\infty}$ which cover all possible $\mathbf{x}$, where $\mathbb{I}(\cdot)$ denotes the indicator function. This effectively gives a mixture model with weights

$$\omega_j(\mathbf{x}) = \mathbb{I}(\mathbf{x} \in A_j).$$

It is interesting to note that if we attempt to mix over the covariates, assuming they have a density function $\phi(\mathbf{x})$, then we would have

$$f(y) = \sum_j \omega_j\, f_j(y),$$

where

$$\omega_j = \int_{A_j} \phi(\mathrm{d}\mathbf{x})\,,$$

and hence returning the typical Bayesian nonparametric mixture model with no covariates. On the other hand, providing the same type of nonparametric distribution for $f(y|\mathbf{x})$ as $f(y)$ seems to make little sense and suggests that the covariate $\mathbf{x}$ is explaining little, if anything.

At the other end of the scale, if we assume the covariates are explaining nothing, then we would expect the model for y given $\mathbf{x}$ to be the same as a model one would select for y alone without covariates. The feature of our model is that we have a parameter which models this explanatory aspect of the covariates. So we can have $f(y|\mathbf{x})$ as unimodal on the one hand and $f(y|\mathbf{x})$ as an infinite-mixture model on the other hand. Clearly, in a Bayesian context, we allow a prior for this key parameter which lets the data itself determine the level of explanation.

Our model for the $\omega_j(\mathbf{x})$ is based on the Gaussian distribution function and will be presented in Section 2. Given the infinite number of $(f_j(y))$, we see that there is no reason why these densities should also depend explicitly on $\mathbf{x}$ and

perhaps the closest model to our own is provided by the product partition model of Müller and Quintana (2010).

Describing the layout of the paper: In Section 2 we fully describe our regression model, and in Appendix A we describe the Markov chain Monte Carlo (MCMC) methods for estimating its posterior distribution. In Section 3, we illustrate our model through the analysis of many real and simulated data sets. In so doing, we compare the predictive performance of two important versions of our models, against many other regression models of common usage. Section 4 concludes with a discussion.

2. The Regression Model

Using our motivation for the mixture weights given in Section 1, our Bayesian density regression model is given by:

$$f(y|\mathbf{x}) = \int f(y|\boldsymbol{\theta})\mathrm{d}G_{\mathbf{x}}(\boldsymbol{\theta}) = \sum_{j=-\infty}^{\infty} f(y|\boldsymbol{\theta}_j)\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x})), \tag{3}$$

with mixture weights

$$\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x})) = \Phi(\{j - \eta_\omega(\mathbf{x})\}/\sigma_\omega(\mathbf{x})) - \Phi(\{j - 1 - \eta_\omega(\mathbf{x})\}/\sigma_\omega(\mathbf{x})),$$

where $\Phi(\cdot)$ is the standard normal c.d.f. and the $f(y|\boldsymbol{\theta}_j)$ $(j = 0, \pm 1, \pm 2, \ldots)$ are chosen to be unimodal component (kernel) densities. Also, $\eta_\omega(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}$ and $\sigma_\omega(\mathbf{x}) : \mathbb{R}^p \to \mathbb{R}^+$ are random functions, with p the number of covariates in the vector $\mathbf{x}$. It is easy to see that the weights sum to 1 at every $\mathbf{x} \in \mathcal{X}$. The model is completed by the specification of prior densities on the $\boldsymbol{\theta}_j$ and on $(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$.

These ideas are illustrated in Figure 1, which plots the mixture weights and the corresponding density of our model, $f(y|\mathbf{x})$, for a range of $\sigma_\omega(\mathbf{x})$, given $\eta_\omega(\mathbf{x}) = .7$, and assuming normal kernel density functions $f(y_i|\boldsymbol{\theta}_j) = \mathrm{n}(y_i|\mu_j, \sigma_j^2)$, given samples of (μ_j, σ_j^2) from a normal-gamma distribution. As shown, the conditional density $f(y|\mathbf{x})$ is unimodal when $\sigma_\omega(\mathbf{x})$ is small, and $f(y|\mathbf{x})$ becomes more multimodal as $\sigma_\omega(\mathbf{x})$ increases. The parameter $\sigma_\omega(\mathbf{x})$ indicates the degree to which the covariates $\mathbf{x}$ explains the dependent variable Y. At one extreme, a small value of $\sigma_\omega(\mathbf{x})$ indicates that the covariates highly explain the dependent variable, meaning $f(y|\mathbf{x})$ is unimodal, and modeled as one of the unimodal kernel densities $f(y|\boldsymbol{\theta}_j)$. Mathematically, a small $\sigma_\omega(\mathbf{x})$ corresponds to a mixture weight $\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$ that is near 1 for $j - 1 < \eta < j$, with all the other mixture weights $\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$ being near 0 for $\eta < j - 1$ or $\eta > j$, and then $f(y|\mathbf{x}) \approx f(y|\boldsymbol{\theta}_j)$. This is because the function $\Phi(\eta/\sigma)$ is approximately 0 for $\eta < 0$, while it is approximately 1 for $\eta > 0$. At the other extreme, a large value of $\sigma_\omega(\mathbf{x})$ indicates that $\mathbf{x}$ explains little in the dependent variable Y. Specifically, as $\sigma_\omega(\mathbf{x}) \uparrow \infty$, the mixture weights become more spread out and then $f(y|\mathbf{x})$ becomes more multimodal, and then we recover typical Bayesian density regression models.

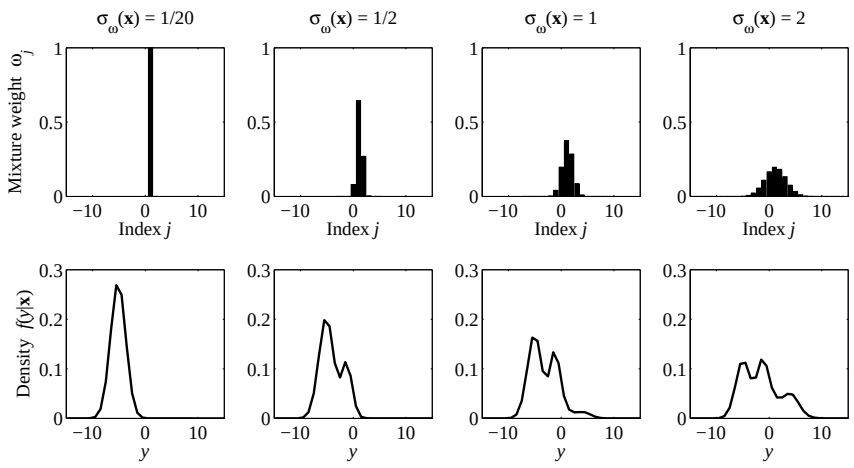

FIG 1. *Density $f(y|\mathbf{x})$ for $\sigma_\omega(\mathbf{x}) = 1/20,\ 1/2,\ 1,\ 2$, given $\eta_\omega(\mathbf{x}) = .7$ and given sampled values of (μ_j, τ_j^2) for normal (unimodal) kernel densities.*

It is possible to specify a prior on $\sigma_\omega(\mathbf{x})$, to match a prior belief on the evolution of the number of modes in the dependent response Y as a function of the covariates $\mathbf{x}$. To explain, let us suppose that for a particular $\mathbf{x}$ it is that $\eta_\omega(\mathbf{x}) = 0$, so the corresponding key j is $j = 0$. Different $\eta_\omega(\mathbf{x})$ will lead to different key j. If we want approximately $2N_\mathbf{x}+1$ modes at $\mathbf{x}$, then we would want

$$\omega_{-N_\mathbf{x}}(0, \sigma_\omega(\mathbf{x})) + \cdots + \omega_0(0, \sigma_\omega(\mathbf{x})) + \cdots + \omega_{N_\mathbf{x}}(0, \sigma_\omega(\mathbf{x})) = .9,$$

for example. This could also be .95, and so on, and in general write as $1 - \underline{\alpha}$. Then we need to choose $\sigma(\mathbf{x})$, at this $\mathbf{x}$ for which

$$\frac{1}{\sqrt{2\pi}} \int_{-N_\mathbf{x}/\sigma(\mathbf{x})}^{N_\mathbf{x}/\sigma(\mathbf{x})} e^{-0.5s^2} \mathrm{d}s = 1 - \underline{\alpha}.$$

Thus

$$\sigma(N_\mathbf{x}) = \frac{N_\mathbf{x}}{\Phi^{-1}(1 - \underline{\alpha}/2)}.$$

Hence, the prior for would have expectation $\sigma(N_\mathbf{x})$ and a variance to express the uncertainty in this choice. More generally, at a particular $\mathbf{x}$, if $\eta_\omega(\mathbf{x})$ is given, then the key j is the one closest to $\eta_\omega(\mathbf{x})$. If $2N_\mathbf{x}+1$ modes are sought then, similarly, we would now choose $\sigma_\omega(\mathbf{x})$ to have mean given by $\sigma(N_\mathbf{x})$ which satisfies

$$1 - \underline{\alpha} = \Phi\left(\frac{N_\mathbf{x} + \eta_\omega(\mathbf{x})}{\sigma(N_\mathbf{x})}\right) - \Phi\left(\frac{-N_\mathbf{x} + \eta_\omega(\mathbf{x})}{\sigma(N_\mathbf{x})}\right).$$

In the present study, we focus on linear structures for the mixture weights, namely $\eta_\omega(\mathbf{x}) = (1, \mathbf{x}^\intercal)\boldsymbol{\beta}_\omega$ and $\sigma_\omega(\mathbf{x}) = \exp[(1, \mathbf{x}^\intercal)\boldsymbol{\lambda}_\omega]^{1/2}$, along with multivariate normal prior $(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega) \sim \mathrm{n}(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega | \mathbf{m}_\omega, \mathbf{V}_\omega)$. While it is possible to

consider more flexible Gaussian process priors, the choice of linear structure for $(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$ leads to more computational tractability in the MCMC estimation of the posterior distribution of the model, especially when either the sample size (n) and the number of predictors (p) is large. Also, we argue that compared to Bayesian density regression models that assume stick-breaking mixture weights, the mixture weights $\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$ of our model are more interpretable because they follow the form of a familiar, ordered-probits regression, for infinitely many categories $j = 0, \pm 1, \pm 2, \ldots$.

As for choices of the component densities $f(y|\boldsymbol{\theta}_j)$, we may consider either of two cases. On the one hand, we may follow common practice and specify the components to be normal densities, that is $f(y|\boldsymbol{\theta}_j) := \mathrm{n}(y|\mu_j, \sigma_j^2)$ for $j = 0, \pm 1, \pm 2, \pm 3, \ldots$, with $\boldsymbol{\mu} := (\mu_j | j = 0, \pm 1, \pm 2, \ldots)$ and $\boldsymbol{\sigma} := (\sigma_j | j = 0, \pm 1, \pm 2, \ldots)$ assigned conjugate normal-gamma prior densities, with the gamma density parameterized by shape (α) and rate (β). The full prior distribution of our normal kernel regression model is:

$$\begin{aligned} \pi(\boldsymbol{\zeta}) &= \pi(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega) \\ &= \left\{ \prod_{j=-\infty}^{\infty} \mathrm{n}(\mu_j | \mu_{\mu j}, \sigma_{\mu j}^2) \mathrm{ga}(\sigma_j^{-2} | \alpha_{\sigma j}, \beta_{\sigma j}) \right\} \mathrm{n}(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega | \mathbf{m}_\omega, \mathbf{V}_\omega). \end{aligned}$$

On the other hand, we may specify each kernel $f(y|\boldsymbol{\theta}_j)$ as a more flexible, general unimodal density, modeled by a scale mixture of uniform $(\mathrm{un}(\cdot|\cdot))$ densities with mode (or mean) μ_j, defined by

$$f(y|\boldsymbol{\theta}_j) := \int \mathrm{un}(y|\mu_j - \psi, \mu_j + \psi) \mathrm{d}P_j(\psi). \tag{4}$$

The mixing distribution P_j is assigned a nonparametric, stick-breaking (SB) prior $P_j \sim \mathrm{SB}(\mathbf{a}_j = (a_{lj})_{l \geq 1}, \mathbf{b}_j = (b_{lj})_{l \geq 1}, P_{0j})$ (Ishwaran & James, 2001), in order to fully support the entire class of unimodal densities $f(y|\boldsymbol{\theta}_j)$ (Brunner, 1992). Hence, P_j and the clustering structure are allowed to change with j. This SB prior is assigned Pareto baseline distribution

$$P_{0j}(\psi) = \mathrm{Pa}(\psi | \alpha_{\psi j}, \beta_{\psi j}),$$

along with an independent prior on the mean (or mode) μ_j. In these terms, recall that each random $P_j \sim \mathrm{SB}(\mathbf{a}_j, \mathbf{b}_j, P_{0j})$ is constructed by $P_j(\cdot) = \sum_{l=1}^{\infty} \omega_{lj} \delta_{\psi_{lj}}(\cdot)$, given $\psi_{lj} \sim_{ind} P_{0j} = \mathrm{Pa}(\alpha_{\psi j}, \beta_{\psi j})$ $(l = 1, 2, \ldots)$ and stick-breaking mixture weights $\omega_{lj} = \upsilon_{lj} \prod_{k=1}^{j-1} (1 - \upsilon_{lk})$, with $\upsilon_{lj} \sim_{ind} \mathrm{be}(a_{lj}, b_{lj})$ $(l = 1, 2, \ldots)$ (Sethuraman, 1994). For example, the two parameter Poisson-Dirichlet process is given by $a_j = 1 - a$ and $b_j = b + ja$ for some $0 \leq a < 1$ and $b > -a$ (Perman, et al. 1992), with the choice $a = 0$ and $b = \alpha$ resulting in the Dirichlet process having baseline (expected) distribution P_{0j} and precision parameter α (Ferguson, 1973). Then, the general unimodal kernel in (4) can be re-written as:

$$f(y|\boldsymbol{\theta}_j) := \sum_{l=1}^{\infty} \mathrm{un}(y|\mu_j - \psi_{lj}, \mu_j + \psi_{lj}) \omega_{lj} = \sum_{l=1}^{\infty} \frac{\mathbb{I}(\mu_j - \psi_{lj} < y < \mu_j + \psi_{lj})}{2\psi_{lj}} \omega_{lj}. \tag{5}$$

Skewness can be introduced if felt necessary by making different the upper and lower limits of the uniform distribution. In all, the full prior density of our uniform-mixture (general unimodal) kernel regression model is:

$$\pi(\boldsymbol{\zeta}) = \pi(\boldsymbol{\mu}, \boldsymbol{\psi}, \boldsymbol{\upsilon}, \boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega) = \mathrm{n}(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega | \mathbf{m}_\omega, \mathbf{V}_\omega) \prod_{j=-\infty}^{\infty} \mathrm{n}(\mu_j | \mu_{\mu j}, \sigma^2_{\mu j}) \times \left\{ \prod_{l=1}^{\infty} \mathrm{be}(\upsilon_{lj} | a_{lj}, b_{lj}) \mathrm{pa}(\psi_{lj} | \alpha_{\psi j}, \beta_{\psi j}) \right\}.$$

In most applied regression settings, there is little prior information available about the parameters of the given regression model. Therefore, in such common settings where it is of interest to apply either version of our Bayesian nonparametric regression model, we may specify high prior (finite) variance for the model parameters $\boldsymbol{\zeta}$, in an attempt to specify a non-informative prior. Indeed, the specification of such a proper diffuse prior reflect a standard procedure in Bayesian analysis (e.g., see Gelman et al., 2008). So for our regression model, the following proper diffuse (high-variance) priors may be considered. We may assume a multivariate normal $\mathrm{n}(\mathbf{0}, \mathrm{diag}(10^5 \mathbf{I}_{2(p+1)}))$ prior for the parameters $(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega)$ of the mixture weights $\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$, to reflect high prior uncertainty about the number of modes in $f(y|\mathbf{x})$. For our normal kernel regression model, we may assume gamma priors $\sigma_j^{-2} \sim_{iid} \mathrm{ga}(\alpha_{\sigma j} = 1, \beta_{\sigma j} = .001)$. For our uniform-mixture kernel regression model, we may assume the Poisson-Dirichlet process as a diffuse prior for the infinite uniform scale-mixtures (i.e., $a_{lj} = .5$ and $b_{lj} = .5 + .5l$, for $\upsilon_{lj} \sim_{ind} \mathrm{be}(a_{lj}, b_{lj})$ and $l = 1, 2, \ldots$; $j = 0, \pm 1, \pm 2, \ldots$), along with Pareto prior parameters $(\alpha_{\psi j} = 1, \beta_{\psi j} = .01)$ for ψ_{lj}. Also, for the μ_j, the normal prior mean and variance parameters $(\mu_{\mu j}, \sigma^2_{\mu j})$ can be chosen according to prior knowledge about the first two moments of the dependent response Y. For example, if the observed response data have been standardized to values y_i $(i = 1, \ldots, n)$ having mean zero and variance 1, then we could choose $(\mu_{\mu j} = 0, \sigma^2_{\mu j} = 3)$. Of course, the choice of priors may vary over data sets in practice, as appropriate.

According to standard arguments of probability theory involving Bayes' theorem, given a sample set of data $\mathcal{D}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, having likelihood $\prod_{i=1}^n f(y_i|\mathbf{x}_i; \boldsymbol{\zeta})$ under our regression model, and given a proper prior density $\pi(\boldsymbol{\zeta})$ defined over the space $\Omega_{\boldsymbol{\zeta}}$ of $\boldsymbol{\zeta}$, the posterior density of $\boldsymbol{\zeta}$ is proper and given by:

$$\pi(\boldsymbol{\zeta}|\mathcal{D}_n) \propto \prod_{i=1}^{n} f(y_i|\mathbf{x}_i)\pi(\boldsymbol{\zeta}) \tag{6}$$

up to a proportionality constant. Then given any chosen $\mathbf{x}$, the posterior predictive density of Y is defined by

$$f_n(y|\mathbf{x}) = \int f(y|\mathbf{x}; \boldsymbol{\zeta})\pi(\boldsymbol{\zeta}|\mathcal{D}_n)\mathrm{d}\boldsymbol{\zeta}. \tag{7}$$

Such posterior densities can be estimated via the implementation of standard Gibbs MCMC sampling methods for infinite-dimensional models, which involve

the use of strategic latent variables (Kalli, Griffin, & Walker, 2010). For more details, see Appendix A.

It is possible to use many variations of the base regression model, after making straightforward modifications to the model's priors and MCMC algorithm. Again, see Appendix A for more details. Modeling variations can, at least, involve either: the multiple imputation of censored dependent responses y_i; the analysis of a binary or ordinal dependent variable via the modeling of latent dependent responses; the modeling of covariate-dependent kernel densities; and the specification of spike-and-slab priors for $\boldsymbol{\beta}_\omega$, for automatic Bayesian variable selection. Also, a multi-level version of our regression model, assigned a prior $\{(\boldsymbol{\beta}_{\omega q}, \boldsymbol{\lambda}_{\omega q})\}_{q=1}^{Q} \sim \pi(\boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega)$, can be used for settings where multiple observations are obtained for each one of Q populations, and the goal is to borrow information across them, while assuming exchangeability both between and within populations.

3. Illustrations

In this section, we illustrate our normal kernel regression model, and our uniform-mixture kernel regression model, through the analysis of 24 real data sets, and the analysis of 40 simulated data sets. In Section 3.1, we illustrate our models through the analysis of student literacy performance data, involving several covariates describing the student, and the student's classroom, teacher, and school. In Section 3.2, we illustrate our models through the analysis of data from a medical experiment, involving a binary (0-1) dependent response, with a small number of covariates. In Section 3.3, we describe the results of the simulation study, to investigate the predictive performance of our regression models, for a wide range of data-generation models. They include data-generation models where the dependent variable density $f(y|\mathbf{x})$ is unimodal, and include data-generation models where the number of modes in $f(y|\mathbf{x})$ depends on $\mathbf{x}$. Finally, in Section 3.4, we illustrate our models through the analyses of 22 more real data sets, which come from a range of scientific fields, and vary widely both in terms of the sample size (n), the number of covariates (p). Throughout this Section 3, we compare the predictive performance of our regression models, against the performance of many other current Bayesian and non-Bayesian regression models of common usage. Specifically, these other models are of widespread availability via publicly-contributed software packages.

For each model m, of a set of M models ($m = 1, \ldots, M$) that are fit to a common data set $\mathcal{D}_n$, we measure predictive performance using the mean-square error criterion

$$D_t(m) = t\sum_{i=1}^{n}\{y_i - \mathrm{E}_n(Y_i|\mathbf{x}_i, m)\}^2 + \sum_{i=1}^{n}\mathrm{Var}_n(Y_i|\mathbf{x}_i, m) \quad (8a)$$

$$= GF(m) + Pen(m) \quad (8b)$$

for a fixed choice $t \in [0, 1]$ (Laud & Ibrahim, 1995; Gelfand & Ghosh, 1998). Then among the M models compared, the model attaining the smallest value

of $D_t(m)$ is identified as the model with the best predictive performance. The criterion (8a) is based on predictive means and variances

$$\begin{aligned} \mathrm{E}_n(Y_i|\mathbf{x}_i, \mathcal{D}_n, m) &= \int y f_n(y|\mathbf{x}_i, m)\mathrm{d}y, \\ \mathrm{Var}_n(Y_i|\mathbf{x}_i, m) &= \int \{y - \mathrm{E}(Y_i|\mathbf{x}_i, m)\}^2 f_n(y|\mathbf{x}_i, m)\mathrm{d}y, \end{aligned}$$

respectively, given the the model's posterior predictive density, $f_n(y|\mathbf{x}_i, m)$. The first term $GF(m)$, in (8a), measures the model's goodness-of-fit to the n data samples. The second term, $Pen(m)$, is a penalty that is large when the model either over-fits or under-fits the data $\mathcal{D}_n$. For a non-Bayesian model having point estimate $\widehat{\boldsymbol{\zeta}}_n = \widehat{\boldsymbol{\zeta}}(\mathcal{D}_n)$, the criterion can be estimated via $\widehat{\mathrm{E}}_n(Y_i|\mathbf{x}_i, m) = \mathrm{E}(Y_i|\mathbf{x}_i, m, \widehat{\boldsymbol{\zeta}}_n)$ and $\widehat{\mathrm{Var}}_n(Y_i|\mathbf{x}_i, m) = \mathrm{Var}(Y_i|\mathbf{x}_i, m, \widehat{\boldsymbol{\zeta}}_n)$ $(i = 1, \ldots, n)$. Gelfand and Ghosh (1998) showed that the criterion (8a) has a Bayesian decision-theoretic justification for a general choice of $t \in [0, 1]$. See Appendix A for details on the MCMC estimation methods for $D_t(m)$. In practice, the choice $t = 1$ is usually adopted as a standard default for the criterion. For a recent example, see Gelfand and Banerjee (2010), who used the criterion to compare different spatial models. But the choice $t = 1/2$ also has a theoretical justification (Ibrahim et al., 2001). When $t = 1$, it is easy to show that:

$$D_1(m) = \sum_{i=1}^{n} \mathrm{E}_n[(y_i - y)^2|\mathbf{x}_i, m] = \sum_{i=1}^{n} \int (y - y_i)^2 f_n(y|\mathbf{x}_i, m)\mathrm{d}y.$$

So the estimate of $D_1(m)$ is obtained by generating posterior predictive sample $y_i^{\mathrm{pred}(s)}$ $(i = 1, \ldots, n)$ at each iteration $s = 1, \ldots, S$ of the MCMC chain, and then taking

$$\widehat{D}_1(m) = \frac{1}{S} \sum_{s=1}^{S} \sum_{i=1}^{n} \left(y_i - y_i^{\mathrm{pred}(s)}\right)^2.$$

For all the data sets analyzed later in subsections 3.1–3.4, Appendix B describes the software packages and priors that were used to fit all regression models that were treated as competitors to our normal-kernel model and to our uniform-mixture-kernel model. Package defaults were used. Throughout this Section 3, all the competing Bayesian regression models were assigned proper diffuse prior distributions (see Appendix B). In fact, according to user's manuals of the software packages, most of the empirical illustrations of the competing Bayesian models assumed such priors. For our normal kernel regression model and our uniform-mixture kernel regression model, we later describe the specific priors we assumed for $\boldsymbol{\mu}$, for each data set analyzed. But otherwise, throughout Section 3, we assumed proper diffuse (high variance) priors for all the other model parameters, as described at the end of Section 2. Also, to analyze the all the 64 total data sets considered in this Section, all Bayesian regression models were fit by using between 50,000 to 300,000 MCMC posterior samples. In all these cases, the estimates of the $D_t(m)$ criterion, for $t = 1/2, 3/4, 1$, stabilized

over MCMC iterations according to trace plots, and had sufficiently-small 95% Monte Carlo (MC) confidence intervals according to a consistent batch means estimator (Jones et al., 2006), after discarding a few thousand initial burn-in MCMC samples. It turned out that for all the 24 real data sets and for 39 of the 40 simulated data sets we analyzed in this manuscript, the decision of the best predictive model was the same over choices $t = 1/2, 3/4, 1$, after accounting for the 95% MC confidence intervals of the $D_t(m)$ estimates for the models compared. Thus, henceforth, we will present model comparisons in terms of the standard $D_1(m)$ criterion. Also, to enable comparisons across data sets having different sample sizes (n), we will occasionally report $D_1(m)/n$.

3.1. School Data

This subsection concerns that analysis of data from the 2006 Progress in International Reading Literacy Study (PIRLS), on 565 low-income students from 21 U.S. elementary schools. The dependent variable is student literacy score (named READSC; mean=45.2, s.d.=9.7). There are 8 covariates: a binary (0,1) indicator of male status (MALE; mean=.5), student age (AGE; mean=10.2, s.d.=.7), student's class size (SIZE; mean=22.9, s.d.=4.4), percent of English language learners in student's class (ELL; mean=7, s.d.=9.5), student's teacher years of experience (TEXP4; mean = 3.01, s.d.=2.7) and education level (TED=5 if bachelor's degree; TED=6 if master's or Ph.D.; mean=5.4, s.d.=.5), the number of students enrolled in student's school (ENROL; mean=581.4, s.d.=238.9), and the student's school safety rating (SAFE=1 is high; SAFE=3 is low; mean=1.5, s.d.=.6).

For data analysis, our normal-kernel regression model, and our uniform-mixture kernel regression model, each assumed prior $\mu_j \sim_{i.i.d.} \mathrm{n}(50, 100)$. This prior reflects previous knowledge that PIRLS literacy scores are scaled to have mean 50 and variance 100. Also, for each covariate $k = 1, \ldots, p = 8$, the data $\mathbf{x}_k = (x_{1k}, \ldots, x_{nk})^\intercal$ had been rescaled to have mean 0 and variance 1. This centering will allow us to investigate how the literacy score density changes as a function of one covariate, after controlling for (zeroing out) the effects of all the 7 other covariates.

According to the $D_1(m)$ criteria presented in Table 1, both our normal kernel regression model, and our uniform-mixture kernel regression model, attained better predictive performance, compared to 15 other regression models of common usage. (For the normal kernel model, the $D_1(m)$ estimate the half-width of the 95% MC confidence interval was 161, whereas for the uniform-mixture kernel model, the half-width was 256). For the normal kernel model, the left column of Figure 2 presents the conditional posterior predictive densities of the student literacy score, conditional on small (17), medium (25), and large (31) class sizes, after controlling for (zeroing out) the seven other covariates by fixing them to their mean z-standardized value of zero. As shown, the median literacy score decreases with increasing class size, as consistent with previous educational theory, and the inter-quartile range decreases as well. Also, each of the

TABLE 1

For the PIRLS data, a comparison of the predictive performance between models. (GCV=Generalized Cross-Validation; ML=Maximum Likelihood; REML=Restricted Maximum Likelihood; AIC=Akaike's (1973) Information Criterion. The other acronyms are described in Appendix B)

Regression Model	$D_1(m)$	$GF(m)$	$Pen(m)$
New model, normal kernel	5232	1202	4030
New model, uniform-mixture kernel	26551	6111	20440
Additive modeling via GCV (Wood 2004)	83781	41055	42726
BART (Chipman et al. 2010)	84395	39748	44647
MARS (Friedman, 1991)	84575	42213	42362
LASSO, quadratic $\mathbf{x}$ (Efron, et al. 2004)	86887	43366	43520
DP-mixed student intercepts (Ibrahim & Kleinman, 1998)	87519	40001	47518
Median regression via MPT (Hanson, 2006)	87757	52389	35368
Bayesian LASSO, quadratic $\mathbf{x}$ (Park & Casella 2008)	89452	42813	46640
HLM: Normal-mixed school intercepts (ML)	90303	46587	43716
HLM: Normal-mixed school intercepts (REML)	90741	46825	43916
Linear regression (least-squares)	92604	46220	46384
ANOVA/Linear DDP (De Iorio et al. 2004)	93855	45989	47867
Mean and variance linear regression	93899	46481	47418
Linear regression (Bayesian)	94131	46237	47894
Single-index model (Polzehl & Sperlich, 2009)	94317	47075	47242
Mixture of 13 linear regressions (ML; 13 minimized AIC)	94592	47212	47380

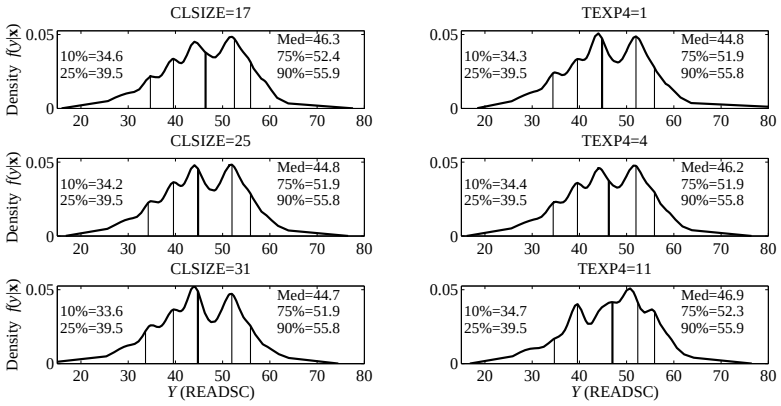

FIG 2. *For the PIRLS data, the posterior predictive density of Y given a value of one covariate, and given the average of all the 7 other covariates (zero).*

three predictive densities have about five modes of literacy scores, suggesting 5 distinct clusters of students with respect to literacy performance. The right column of Figure 2 presents the conditional posterior predictive densities of the student literacy score, conditional on 1 year, 4 years, and 11 years of teaching

experience, after controlling for (zeroing out) the seven other covariate. Here it is shown that the median literacy score increases with years of teaching experience, again consistent with previous educational theory. Also, the number and locations of the modes change along with changes in the number of years of teaching experience. Finally, the estimates of $\sigma_\omega(\mathbf{x})$, conditional on the posterior mean estimate of $\boldsymbol{\lambda}_\omega$, were found to have mean 3.8 and standard deviation of .3 over the 565 subjects. So the 8 covariates seem to have some explanatory power in the prediction of literacy, but also more covariates can be added to improve this power. This seems true, especially given the multimodal predictive densities presented in Figure 2.

3.2. AZT Data

Here we analyze a data set of $n = 338$ subjects, obtained from Agresti (1996, p. 119), where the dependent variable is a binary (0/1) indicator of the presence of AIDS symptoms (with 20.4% of subjects reporting such symptoms). The covariates include a 0-1 indicator of white race (65.1% white; 34.9% African American), a 0-1 indicator of AZT treatment receipt (50.3%), and their interaction. We analyzed these data using several regression models. They included our normal kernel regression model, and our uniform-mixture kernel regression model, each of which assumed priors $\mu_j \sim_{i.i.d.} \mathrm{n}(0, 100)$. Also, for each of our models, the kernel densities are for the latent variables underlying the binary (0-1) dependent responses. See Appendix A for more details.

Table 2 compares the predictive performance between 11 regression models. Both our normal kernel regression model, and our uniform-mixture kernel regression model, had virtually the same $D_1(m)$ predictive criterion value, and had better predictive performance than all the 9 other regression models. Also,

TABLE 2
For the AZT data, a comparison of the predictive performance between models. (PQL=Penalized Quasi-Likelihood; GLM=Generalized Linear Model; ML=Maximum-Likelihood; CV=Cross-Validation. The other acronyms are described in Appendix B)

Regression Model	$D_1(m)$	$GF(m)$	$Pen(m)$
New model, normal kernel for the link function	0.4	0.0	0.1
New model, uniform-mixture kernel for the link function	0.5	0.0	0.5
ANOVA/Linear DDP logit model (De Iorio et al. 2004)	10.6	0.5	10.0
DP model for the link function (Newton, et al. 1996)	55.7	54.0	1.7
DP-mixed intercepts logit (Mukhopadhyay & Gelfand, 1997)	59.2	16.3	42.9
MPT-mixed intercepts logit model (Hanson, 2006)	107.1	53.6	53.6
Normal-mixed intercepts logit model (PQL)	107.1	53.5	53.6
BART (Chipman et al. 2010)	107.1	53.6	53.6
Logit model (ML)	107.1	53.6	53.6
Mixture of 1 logit model (ML) (1 minimized AIC)	107.1	53.6	53.6
LASSO logit (Friedman, et al. 2010)	108.4	53.9	54.5

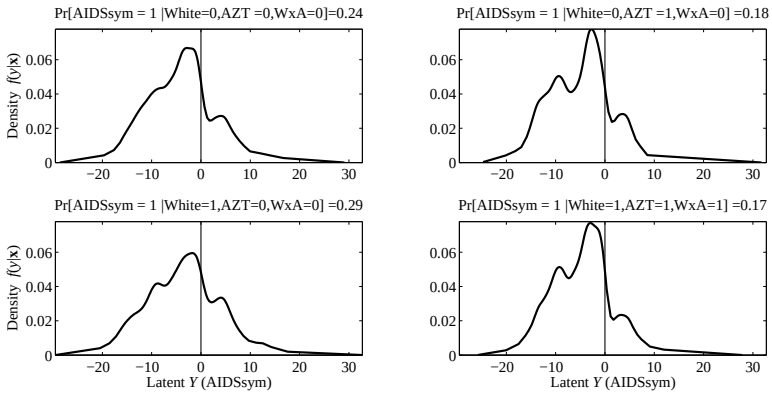

FIG 3. *Estimated predictive densities of the latent AIDS symptoms (Sym) variable, for each of the four subject groups defined by race indicator (White), AZT treatment indicator (AZT), and their interaction (W×A).*

the ANOVA-linear DDP model had much better predictive performance than 8 models. For our normal kernel regression model, Figure 3 presents the posterior predictive densities of the latent dependent response, conditional on all four possible values of the three covariates in $\mathbf{x}$. The figure shows that AZT treatment tends to lower the incidence of AIDS symptoms, for both races. Finally, with regards to the explanatory power of the 3 covariates in the regression model, the estimates of $\sigma_\omega(\mathbf{x})$, conditional on the posterior mean estimate of $\boldsymbol{\lambda}_\omega$, was found to have mean 3.1 and standard deviation .7 over the 338 subjects. Given this result, and the multimodality of the predictive densities (Figure 3), it seems that the three covariates have some, but not very much, explanatory power for prediction.

3.3. Simulation Study

Here, we consider a simulation study of many regression models, using a broad range of complex data-generating models. In terms of the $D_1(m)$ criterion, we compare the predictive performance between of our normal kernel regression model, our uniform-mixture kernel model, and 10 other models that provided the the closest competitors for the PIRLS and AZT data. The other competing models are BART, the additive model, the DP-mixed intercepts linear regression model, ANOVA/linear DDP model, the median regression model with a MPT prior for the error distribution, MARS, the LASSO estimated via the LARS algorithm, the Bayesian LASSO, and the normal normal-mixed intercepts linear regression model (either estimated by maximum-likelihood, or restricted maximum likelihood). Again, Appendix B provides details about the competing models. To analyze all the 40 simulated data sets considered in this subsection, our normal kernel model, and our uniform-mixture kernel model, each assumed priors $\mu_j \sim_{i.i.d.} \mathrm{n}(\widehat{\mu}, 100)$, with $\widehat{\mu}$ the empirical mean of the simulated Y.

First, we simulated 20 data sets according to a $5 \times 2 \times 2$ design, assuming that the true densities $f(y|\mathbf{x})$ are unimodal, normal densities. Each of the 20 cells of this simulation design are defined according to one of 5 complex mean regression functions for Y, either homoscedastic and heteroscedastic variances for Y, and either $n = 100$ or $n = 225$ samples of $(y_i, \mathbf{x}_i)$. Specifically, the 5 regression functions are given as follows.

$$\begin{aligned}
\mathrm{E}_1(Y|\mathbf{x}) &= 1.9[1.35 + \exp(x_1)\sin(13(x_1 - .6)^2)\exp(-x_2)\sin(7x_2)], && (9)\\
\mathrm{E}_2(Y|\mathbf{x}) &= (-2x_1)^{\mathbb{I}(x_1<.6)}(-1.2x_1)^{\mathbb{I}(x_1\geq.6)} + \cos(5\pi x_2)/(1 + 3x_2^2), && (10)\\
\mathrm{E}_3(Y|\mathbf{x}) &= (\mathbf{x}^\top\boldsymbol{\beta})^2\exp(\mathbf{x}^\top\boldsymbol{\beta}); \boldsymbol{\beta} = (2, 1, 1, 1)^\top/7^{1/2}, && (11)\\
\mathrm{E}_4(Y|\mathbf{x}) &= (3, 1.5, 0, 0, 2, 0, 0, 0)\mathbf{x}, && (12)\\
\mathrm{E}_5(Y|\mathbf{x}) &= 10\sin(\pi x_1 x_2) + 20(x_3 - .5)^2 + 10x_4 + 5x_5 + \sum_{k=6}^{10} 0x_k. && (13)
\end{aligned}$$

Function (9) is a complex 2-dimensional surface (Hwang, et al., 1994), function (10) is a discontinuous additive model (Hastie & Tibshirani, 1990, pp. 247-51), function (11) is a nonlinear function of 4 predictors, function (12) is a linear model with 5 of the 8 covariates irrelevant, and function (13) is a complex high-dimensional interaction with 5 of the 10 covariates irrelevant (Friedman, 1991). Respectively for the 5 mean functions, a data set of n observations (i.e., $n = 100$ or $n = 225$) was generated from the distribution with densities

$$\begin{aligned}
&\mathrm{n}(y_i|\mathrm{E}_1(Y|\mathbf{x}_i), \sigma_i^2)\mathrm{un}_2(\mathbf{x}_i|0, 1), \quad \mathrm{n}(y_i|\mathrm{E}_2(Y|\mathbf{x}_i), \sigma_i^2)\mathrm{un}_2(\mathbf{x}_i|0, 1),\\
&\mathrm{n}(y_i|\mathrm{E}_3(Y|\mathbf{x}_i), \sigma_i^2)\mathrm{un}_2(\mathbf{x}_i|0, 1), \quad \mathrm{n}(y_i|\mathrm{E}_4(Y|\mathbf{x}_i), \sigma_i^2)\mathrm{n}_8(\mathbf{x}_i|\mathbf{0}, (.5^{|j-k|})_{8\times 8}),\\
&\mathrm{n}(y_i|\mathrm{E}_5(Y|\mathbf{x}_i), \sigma_i^2)\mathrm{un}_{10}(\mathbf{x}_i|0, 1).
\end{aligned}$$

Under homoscedasticity, these 5 densities assumed $\sigma_i^2 = .0625$, $\sigma_i^2 = .5$, $\sigma_i^2 = 2$, $\sigma_i^2 = 9$, and $\sigma_i^2 = 1$, respectively. Under heteroscedasticity, the variances σ_i^2 had 5 distinct values that changed every $n/5^{\text{th}}$ sample. Specifically, for these 5 densities, the variances were $(.002\,5, .022\,5, .062\,5, .122\,5, .25)$, $(.125, .19, .25, .38, .5)$, $(2, 4, 2, 4, 2)$, $(4.49, 6.55, 9, 13.1, 17.98)$, and $(.5, .76, 1, 1.49, 1.99)$, respectively.

We simulated 20 additional data sets according to another $5 \times 2 \times 2$ design, now assuming that the $f(y|\mathbf{x})$ are multimodal, mixtures of normal densities, with the number of modes depending on $\mathbf{x}$, where $\mathbf{x}$ consists of $p = 10$ covariates with sampling density $\mathrm{un}_{10}(\mathbf{x}|0, 1)$. For each of the 5 conditions of this design, the number of modes in $f(y|\mathbf{x})$ depended on $\mathbf{x}$ in a particular way. Also, in the same manner as the unimodal density $f(y|\mathbf{x})$ simulations, each cell is defined by either homoscedastic or heteroscedastic variances for Y, and by either $n = 100$ or $n = 225$ samples of $(y_i, \mathbf{x}_i)$. Specifically, for each of these 20 simulated data sets (i.e., each cell of the simulation design), the number of modes in the true data-generating $f(y|\mathbf{x})$ depended on $\mathbf{x}$, via the function $N_{modes} = \min(\max(\mathrm{floor}(\mathrm{E}_4(Y|\mathbf{x})), 1), 4)$, where N_{modes} ranged from 1 to 4, and with $\mathrm{E}_4(Y|\mathbf{x})$ defined by equation (12). The four possible modes are defined by functions $\mathrm{E}_1(Y|\mathbf{x})$, $\mathrm{E}_2(Y|\mathbf{x})$, $\mathrm{E}_3(Y|\mathbf{x})$, and $\mathrm{E}_5(Y|\mathbf{x})$, according to equations (9), (10), (11), and (13), respectively.

TABLE 3

For 20 simulated unimodal data sets, a comparison of the predictive performance between the new nonparametric regression models, and the most-competitive alternative models. For each data set, a bold number indicates the model (or models) with the best value of a statistic, after accounting for the 95 percent Monte Carlo confidence interval. (Hom.=Homoscedastic condition; Het.=Heteroscedastic condition; kern.=kernel; unif.-mix.kern.=uniform-mixture kernel; int.=intercepts; ML=Maximum likelihood. The other acronyms are described in Appendix B)

Results of $D_1(m)/n$		n = 100		n = 225	
	Model	Hom	Het	Hom	Het
Function	New; normal kern.	**.42**	**.39**	.33	.51
$E_1(Y \mid \mathbf{x})$	New; unif.-mix. kern.	.54	.45	.81	.84
	BART	.71	.69	**.22**	.34
	ANOVA/Linear DDP	1.01	.66	.32	**.23**
$E_2(Y \mid \mathbf{x})$	New; normal kern.	**.40**	**.38**	**.27**	**.27**
	New; unif.-mix. kern.	.88	.68	.35	.29
	BART	.67	.69	.51	.52
	Additive model/GCV	.60	.72	.49	.53
$E_3(Y \mid \mathbf{x})$	New; normal kern.	**.59**	**.41**	**.76**	**1.55**
	New; unif.-mix. kern.	2.1	4.4	3.5	3.7
	BART	3.0	5.8	3.9	6.4
	Normal-mixed int. (ML)	.75	1.2	5.9	9.4
$E_4(Y \mid \mathbf{x})$	New; normal kern.	13.6	12.7	5.1	5.6
	New; unif.-mix. kern.	7.8	3.0	5.7	**5.2**
	Normal-mixed int. (ML)	**.61**	**1.2**	**2.6**	5.7
$E_5(Y \mid \mathbf{x})$	New; normal kern.	5.8	5.2	2.8	2.9
	New; unif.-mix. kern.	5.1	4.6	2.8	2.5
	BART	1.4	1.3	**.73**	**.82**
	ANOVA/Linear DDP	**.02**	**.07**	1.1	12.3

For the first of the 5 multimodal simulation conditions, as a function of N_{modes}, the modes evolved in the order of $E_1(Y|\mathbf{x})$, $E_2(Y|\mathbf{x})$, $E_3(Y|\mathbf{x})$, and $E_5(Y|\mathbf{x})$. Specifically, under this condition, each data set was generated from the distribution with densities $n(y_i|E_{z_i}(Y|\mathbf{x}_i), \sigma_i^2)$ for $i = 1, \ldots, n$, with $z_i \sim un\{1\}$ when $N_{modes} = 1$, $z_i \sim un\{1, 2\}$ when $N_{modes} = 2$, $z_i \sim un\{1, 2, 3\}$ when $N_{modes} = 3$, and $z_i \sim un\{1, 2, 3, 5\}$ when $N_{modes} = 4$. Similarly, for the 4 of the remaining 5 multimodal conditions, respectively, the modes, as a function of N_{modes}, evolved in the order of $E_5(Y|\mathbf{x})$, $E_3(Y|\mathbf{x})$, $E_2(Y|\mathbf{x})$, $E_1(Y|\mathbf{x})$; the order of $E_3(Y|\mathbf{x})$, $E_1(Y|\mathbf{x})$, $E_5(Y|\mathbf{x})$, $E_2(Y|\mathbf{x})$; the order of $E_2(Y|\mathbf{x})$, $E_5(Y|\mathbf{x})$, $E_1(Y|\mathbf{x})$, $E_3(Y|\mathbf{x})$; and the order of $E_1(Y|\mathbf{x})$, $E_5(Y|\mathbf{x})$, $E_3(Y|\mathbf{x})$, $E_2(Y|\mathbf{x})$.

Table 3 presents the results of the normalized predictive criterion $D_1(m)/n$, over the 20 data sets simulated under unimodal densities $f(y|\mathbf{x})$. For space considerations, we only present the results for the models with the best predictive performance. (For the normal-mixed intercepts regression model, the

TABLE 4

For 20 simulated multimodal data sets, a comparison of the predictive performance between the new nonparametric regression models, and the most-competitive alternative models. For each data set, a bold number indicates the model (or models) with the best value of a statistic, after accounting for the 95 percent Monte Carlo confidence interval. (Hom.=Homoscedastic condition; Het.=Heteroscedastic condition; kern.=kernel; unif.-mix.kern.=uniform-mixture kernel; int.=intercepts; ML=Maximum likelihood. The other acronyms are described in Appendix B)

Results of $D_1(m)/n$		n = 100		n = 225	
	Model	Hom	Het	Hom	Het
Multimodal Function 1	New; normal kern.	**.18**	**.12**	**1.4**	9.6
	New; unif.-mix. kern.	16.8	15.1	19.9	16.5
	DP-mixed intercepts	8.2	8.5	2.6	3.5
	ANOVA/Linear DDP	9.9	8.8	1.5	**1.4**
2	New; normal kern.	7.8	**.29**	5.8	4.8
	New; unif.-mix. kern.	15.7	3.7	9.1	7.8
	DP-mixed intercepts	34.2	30.6	1.1	1.3
	ANOVA/Linear DDP	**4.7**	6.3	**.86**	**.81**
3	New; normal kern.	9.4	5.0	5.0	4.0
	New; unif.-mix. kern.	19.9	17.7	9.0	9.5
	DP-mixed intercepts	**1.2**	2.0	1.0	.7
	ANOVA/Linear DDP	1.9	**1.7**	**.8**	**.6**
4	New; normal kern.	**1.77**	**1.4**	3.5	4.7
	New; unif.-mix. kern.	12.6	16.6	6.7	14.0
	DP-mixed intercepts	3.2	4.5	1.1	1.2
	ANOVA/Linear DDP	**1.74**	2.2	**.80**	**.87**
5	New; normal kern.	12.8	17.5	5.2	6.1
	New; unif.-mix. kern.	33.3	22.6	12.6	9.3
	DP-mixed intercepts	9.0	4.7	5.2	**1.1**
	ANOVA/Linear DDP	**2.8**	**1.9**	**1.6**	1.5

results were very similar for maximum-likelihood (ML) and for restricted maximum likelihood (REML)). As shown, our normal-kernel regression model tended to have better predictive performance than all other models, while having the smallest (best) criterion value $D_1(m)$ for half of the 20 simulated data sets. So it seems that, overall, this model tended to do best in predicting truly-unimodal densities. As consistent with our discussions in Sections 1 and 2, the DDP model tended to overfit data generated by truly-unimodal densities. Specifically, the normal kernel model tended to have superior predictive performance for conditions where the number of covariates (p) was 4 or less, and where all the covariates were not irrelevant (i.e., all covariates had truly non-zero regression coefficients). The BART, ANOVA/linear DDP and normal-mixed intercepts models tended to have superior predictive performance where there were more covariates, with a subset of these covariates being irrelevant (i.e., the covariate

subset had truly zero regression coefficients). Though, recall that the true data-generating densities $f(y|\mathbf{x})$ were normal. This partially explains why our normal kernel model tended to show better predictive performance than the uniform-mixture kernel model, and why the BART and normal-mixed intercepts model appeared competitive.

Table 4 presents the results of the normalized predictive criterion $D_1(m)/n$, over the 20 data sets simulated under multimodal densities $f(y|\mathbf{x})$. Again, only the models with the best predictive performance are shown. Here, the ANOVA/linear DDP model and our normal-kernel regression model tended to have the best predictive performance. The DDP model attained the smallest value of $D_1(m)$ for most of the 20 data sets. Our uniform-mixture kernel model did not perform as well as the other two models, because the true simulating densities $f(y|\mathbf{x})$ were multivariate normal mixtures. Finally, the apparent superiority of the DDP model is consistent with our earlier discussion (Sections 1 and 2), that DDP-based models are more-specifically designed to handle highly-complex multimodal densities $f(y|\mathbf{x})$. However, when these densities $f(y|\mathbf{x})$ are not so highly-complex and multimodal, the DDP model will tend to overfit the data, as shown by the results of the unimodal simulations.

3.4. Comparisons on 22 More Real Data Sets

Here, we compare the predictive performance between the same 11 regression models we considered in the simulation study, for a wide range of 22 real data sets that are described by Kim et al. (2007) (excluding the Bayesian LASSO). These data sets are summarized in Table 5. Also, before analyzing each of the data sets, the data for each covariate, and the dependent variable data, were separately z-standardized to have mean zero and variance 1. This facilitated model comparisons across the 22 data sets, in terms of the predictive criterion $D_1(m)/n$. Also, throughout all data analyses, our normal kernel model, and our uniform-mixture kernel model, each assumed prior $\mu_j \sim_{i.i.d.} \mathrm{n}(0, 3)$. However, we found that for 5 of the largest data sets where $np \geq 11,572$, the ANOVA/linear

TABLE 5
A summary of the 22 data sets, in terms of the name, sample size, and the number of predictors

Data Set	n	p	Data Set	n	p	Data Set	n	p
Ais	202	21	Diamond	308	15	Ozone	330	8
Attend	838	43	Edu	1400	5	Price	159	15
Baseball	263	44	Enroll	258	6	Rice	1026	20
Baskball	96	4	Fame	1076	25	Servo	167	10
Boston	506	13	Hatco	100	13	Smsa	140	13
Budget	1729	10	Laheart	200	26	Tecator	240	10
Cps	533	16	Mpg	392	9			
Diabetes	375	16	Mussels	82	4			

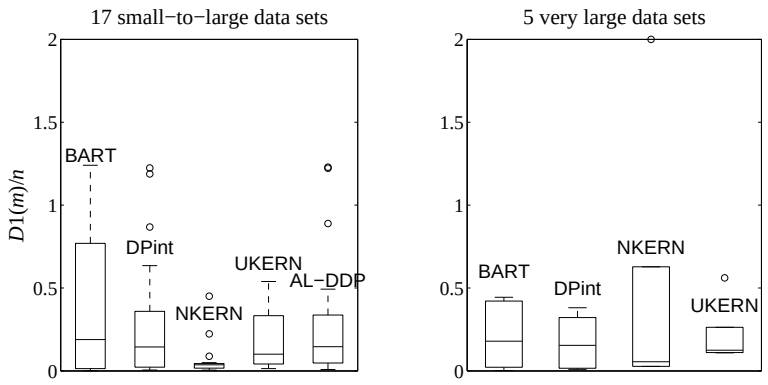

FIG 4. *Boxplots comparing the $D_1(m)$ predictive performance over the 22 data sets, for the 5 best-performing models. (NKERN=New model, normal kernel; UKERN=New model, uniform-mixture kernel; AL-DDP= ANOVA/linear DDP-model; DPint=DP-mixture of intercepts linear regression model.)*

DDP model could not be estimated, due to memory limitations of modern computational power. The posterior estimation of this model requires the computer to handle many thousands of MCMC samples, of more than $n(p+1)$ parameters.

The box-plots in Figure 4 compare the predictive performance of the models over the 22 data sets. For space considerations, this figure only presents the results for only the models having the best median $D_1(m)/n$ predictive performance, over the data sets. The left panel of Figure 4 compares models over the 17 data sets that were not too large to be estimated by the ANOVA/linear DDP model. As shown in this panel, both our normal kernel model and our uniform-mixture kernel model tended to have the best $D_1(m)/n$ predictive performance. The right panel of the figure compares models for the 5 largest data sets that could not be estimated by the ANOVA/linear DDP model. In this panel, again, both of our models had the best median $D_1(m)/n$ predictive performance. Finally, our normal kernel model had the best predictive performance for 11 of the 22 data sets, and won the most data sets, compared to all the other 10 regression models. This was followed by BART, which won 4 of the data sets.

4. Discussion

We have described a Bayesian nonparametric regression model, and demonstrated the suitability of the model through the analysis of many real and simulated data sets.

The key to our proposal, is to model weights $w_j(\mathbf{x})$ which can be close to 1 for specific regions of $\mathbf{x}$ for a particular j. This will ensure outcomes from a region of $\mathbf{x}$ values will have with high probability the observations coming from the same component. This is an appealing property. The prevailing alternative type of model is only to ensure $w_j(\mathbf{x})$ is close to $w_j(\mathbf{x}')$ for $\mathbf{x}$ close to $\mathbf{x}'$; which

has no implication for the observations coming from each covariate, other than they come from the same component with the same probability. This is not how it should work.

The stick-breaking prior is a case in point. For if $f(y|\mathbf{x})$ is a model as flexible as usual choices for $f(y)$ (e.g., the Dirichlet process mixture model), then no attempt is being made to understand how $\mathbf{x}$ is influencing the outcome of y at all. It is not possible to achieve this interpretation for the weights based on stick-breaking (DDP) models, i.e., $\omega_j(\mathbf{x}) = \upsilon_j(\mathbf{x}) \prod_{k=1}^{j-1}(1 - \upsilon_k(\mathbf{x}))$. In order to have a single $\omega_j(\mathbf{x})$ close to 1, it must be that $\upsilon_1(\mathbf{x})$ is close to 1, and so it can only be $\omega_1(\mathbf{x}) \approx 1$, and hence the lack of flexibility.

On the other hand, at one end of our model we have that there exists sets (A_j) such that $f(y|\mathbf{x}) = f_j(y)$ is a unimodal density. Mixing over the $\mathbf{x}$ then yields a standard infinite-mixture model for the y alone. The resulting model,

$$f(y|\mathbf{x}) = \sum_j f_j(y)\omega_j(\mathbf{x}),$$

where

$$\omega_j(\mathbf{x}) = \begin{cases} \approx 1 \text{ for } \mathbf{x} \in A_j \\ \approx 0 \text{ for } \mathbf{x} \notin A_j, \end{cases}$$

is a simple yet powerful model; at its heart lies the idea that $\mathbf{x}$ is explaining outcomes and hence once known, the density $f(y|\mathbf{x})$ should be a simple density, e.g., unimodal.

We achieve a more general framework by taking $\omega_j(\mathbf{x}) = \omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$. As illustrated in Figure 1, a small $\sigma_\omega(\mathbf{x})$ returns the $\omega_j(\mathbf{x}) = \mathbb{I}(\mathbf{x} \in A_j)$ with $\mathbf{x}$ at its most explanatory, while as $\sigma_\omega(\mathbf{x})$ increases the skewness and multimodality of $f(y|\mathbf{x})$ increases, and the less explanatory $\mathbf{x}$ becomes. A prior on this key parameter $\sigma_\omega(\mathbf{x})$ allows the data to determine the level of explanation.

In future research, it would be interesting to provide theoretical results that characterize the support of our Bayesian nonparametric regression model, as Barrientos et al. (2012) have done for the dependent Dirichlet process.

Appendix A: MCMC Algorithm

Our infinite-dimensional regression model can be estimated via the implementation of the MCMC sampling methods of Kalli et al. (2010). This method involves introducing strategic latent variables, to implement exact MCMC algorithms for the estimation of the model's posterior distribution. Specifically, for our normal kernel regression model (Section 2), latent variables $(u_i, z_i \in \mathbb{Z})_{i=1}^n$ are introduced, such that conditional on these variables, and given a decreasing function $\xi_j = \exp(-|j|)$, the model's data likelihood can be written as

$$\prod_{i=1}^{n} \{\mathbb{I}(0 < u_i < \xi_{|z_i|}) \xi_{|z_i|}^{-1} \mathrm{n}(y_i|\mu_{z_i}, \sigma_{z_i}^2) \omega_{z_i}(\eta_\omega(\mathbf{x}_i), \sigma_\omega(\mathbf{x}_i))\}. \tag{14}$$

Marginalizing over the each of the latent variables (u_i, z_i) in (14), for each $i = 1, \ldots, n$, returns the original likelihood,

$$\prod_{i=1}^{n} \left\{ \sum_{j=-\infty}^{\infty} \mathrm{n}(y_i|\mu_j, \sigma_j^2)\omega_j(\eta_\omega(\mathbf{x}_i), \sigma_\omega(\mathbf{x}_i)) \right\},$$

of our infinite-dimensional model. Thus, conditional on the latent variables, the infinite-dimensional model can be treated as a finite-dimensional model. This, in turn, permits the use of standard MCMC methods to sample the model's full joint posterior distribution. Given all variables, save the $(z_i)_{i=1}^n$, the choice of each z_i have minimum $-N_{\max}$ and maximum $N_{\max}$, where $N_{\max} = \max_i[\max_j \mathbb{I}(u_i < \xi_j)|j|]$.

Specifically, for our normal kernel regression model, which assumes functions $\eta(\mathbf{x}_i) = \mathbf{x}_i^\intercal \boldsymbol{\beta}_\omega$ and $\sigma_\omega(\mathbf{x}_i) = \exp((1, \mathbf{x}_i^\intercal)\boldsymbol{\lambda}_\omega)^{1/2}$, the following full conditional posterior densities are sampled at each stage s $(s = 1, \ldots, S)$ of the MCMC algorithm.

1. $\pi(\mu_j|\cdots) = \mathrm{n}\left(\mu_j \left| \frac{\mu_{\mu j}\sigma_j^2 + n_j\sigma_{\mu j}^2\overline{y}_j}{\sigma_j^2 + n_j\sigma_{\mu j}^2}, \frac{\sigma_j^2\sigma_{\mu j}^2}{\sigma_j^2 + n_j\sigma_{\mu j}^2} \right.\right)$, for $j = 0, \ldots, \pm N_{\max}$, with $n_j = \sum_{i:z_i=j} 1, \overline{y}_j = \frac{1}{n_j}\sum_{i:z_i=j} y_i$, and given draws $u_i \sim_{ind} \mathrm{un}(0, \xi_{|z_i|})$ $(i = 1, \ldots, n)$;
2. $\pi(\sigma_j^{-2}|\cdots) = \mathrm{ga}(\sigma_j^{-2}|\alpha_{\sigma j} + (n_j/2), \beta_{\sigma j} + \frac{1}{2}\sum_{i:z_i=j}(y_i - \mu_j)^2)$, for $j = 0, \ldots, \pm N_{\max}$;
3. $\Pr(z_i = j|\cdots) \propto \mathbb{I}(u_i < \xi_{|j|})\xi_{|j|}^{-1}\mathrm{n}(y_i|\mu_j, \sigma_j^2)\omega_j(\eta(\mathbf{x}_i), \sigma(\mathbf{x}_i))$, for $j = 0, \ldots, \pm N_{\max}$;
4. $\pi(\boldsymbol{\beta}_\omega|\cdots) = \mathrm{n}(\boldsymbol{\beta}_\omega|\mathbf{m}^*_{\boldsymbol{\beta}\omega}, \mathbf{V}^*_{\boldsymbol{\beta}\omega})$, given $\mathbf{V}^*_{\boldsymbol{\beta}\omega} = (\mathbf{V}^{-1}_{\boldsymbol{\beta}\omega} + \mathbf{X}^\intercal \mathrm{diag}(\sigma_\omega^{-2}(\mathbf{x}_1), \ldots, \sigma_\omega^{-2}(\mathbf{x}_n))\mathbf{X})^{-1}$, $\mathbf{m}^*_{\boldsymbol{\beta}\omega} = \mathbf{V}^*_{\boldsymbol{\beta}\omega}(\mathbf{V}^{-1}_{\boldsymbol{\beta}\omega}\mathbf{m}_{\boldsymbol{\beta}\omega} + \mathbf{X}^\intercal \mathrm{diag}(\sigma_\omega^{-2}(\mathbf{x}_1), \ldots, \sigma_\omega^{-2}(\mathbf{x}_n))\mathbf{z}^*)$, $\mathbf{X} = ((1, \mathbf{x}_i^\intercal))_{n\times(p+1)}$, and $\mathbf{z}^* = (z_1^*, \ldots, z_n^*)^\intercal$, given draws $z_i^* \sim_{ind} \pi(z_i^*|\cdots) \propto \mathrm{n}(\mathbf{x}_i^\intercal\boldsymbol{\beta}_\omega, \sigma_\omega^2(\mathbf{x}_i))\mathbb{I}(z_i^* \in (z_i - 1, z_i])$ $(i = 1, \ldots, n)$.
5. $\pi(\boldsymbol{\lambda}_\omega|\cdots) \propto \mathrm{n}(\boldsymbol{\lambda}_\omega|\mathbf{m}_{\boldsymbol{\lambda}\omega}, \mathbf{V}_{\boldsymbol{\lambda}\omega})\prod_{i=1}^n \mathrm{n}(z_i^*|\eta_\omega(\mathbf{x}_i), \exp((1, \mathbf{x}_i^\intercal)\boldsymbol{\lambda}_\omega))$.

The full conditional posterior densities in Steps 1, 2, and 4 follow from standard Bayesian normal linear models, assigned conjugate prior distributions (O'Hagan & Forster, 2004). The full conditional of $\boldsymbol{\lambda}_\omega$ in Step 5 can be sampled using an efficient random-walk Metropolis-Hastings algorithm, with multivariate normal proposal distribution having mean $\mathbf{m}^*_{\boldsymbol{\lambda}\omega} = \mathbf{V}^*_{\boldsymbol{\lambda}\omega}(\mathbf{V}^{-1}_{\boldsymbol{\lambda}\omega}\mathbf{m}_{\boldsymbol{\lambda}\omega} + \frac{1}{2}\mathbf{X}^\intercal\overline{\mathbf{z}})$ and covariance matrix $\mathbf{V}^*_{\boldsymbol{\lambda}\omega} = (\mathbf{V}^{-1}_{\boldsymbol{\lambda}\omega} + \frac{1}{2}\mathbf{X}^\intercal\mathbf{X})^{-1}$, where $\overline{z}_i = \mathbf{x}_i^\intercal\boldsymbol{\lambda}_\omega + \{(z_i - \mathbf{x}_i^\intercal\boldsymbol{\beta}_\omega)/\exp(\mathbf{x}_i^\intercal\boldsymbol{\lambda}_\omega)\} - 1$ $(i = 1, \ldots, n)$ and $\overline{\mathbf{z}} = (\overline{z}_1, \ldots, \overline{z}_n)^\intercal$ (Cepeda & Gamerman, 2001). The above 5-step sampling algorithm is repeated a large number S of times, to construct a discrete-time Harris ergodic Markov chain $\{\boldsymbol{\zeta}^{(s)} = (\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{\beta}_\omega, \boldsymbol{\lambda}_\omega)^{(s)}\}_{s=1}^S$ having a posterior distribution $\Pi(\boldsymbol{\zeta}|\mathcal{D}_n)$ as its stationary distribution, provided that the prior for $\boldsymbol{\zeta}$ is proper (Robert & Casella, Section 10.4.3 2004).

We have written MATLAB (2012, The MathWorks, Natick, MA) code that implements the MCMC sampling algorithm. Standard methods, available from

available from our own MATLAB code, can be used to check whether the MCMC algorithm has generated a sufficiently-large number of samples from the model's posterior distribution. Given a finite S number of samples $\{\boldsymbol{\zeta}^{(s)}\}_{s=1}^{S}$ generated by the MCMC algorithm, univariate trace plots can be used to evaluate the mixing of the chain (i.e., the degree to which the chain explores the support of the posterior distribution), while batch means and subsampling methods can be used to calculate standard errors of marginal posterior estimates of chosen scalar functionals of $\boldsymbol{\zeta}$ (e.g., Jones et al., 2006), after excluding burn-in samples.

Simple extensions of the above MCMC algorithm can be used to address straightforward extensions of the normal kernel model presented in Section 2, and to address other important tasks of data analysis:

- Uniform-mixture kernel densities: This choice of densities leads to our uniform-mixture kernel model (see Section 2). The MCMC sampling methods of Kalli et al. (2010) can be used to perform posterior inferences with this model, by introducing latent variables $(u_i, u_i', z_i \in \mathbb{Z}, z_i' \in \mathbb{Z}^+)_{i=1}^n$, such that conditional on these variables, the model's data likelihood is defined by

$$\prod_{i=1}^{n} \left\{ \begin{array}{c} \mathbb{I}(0 < u_i < \xi_{|z_i|})\xi_{|z_i|}^{-1}\mathbb{I}(0 < u_i' < \xi_{z_i'})\xi_{z_i'}^{-1} \\ \times \omega_{z_i}(\eta(\mathbf{x}_i), \sigma(\mathbf{x}_i))\, \omega_{z_i' z_i}\, \mathrm{un}(y_i | \mu_{z_i} - \psi_{z_i' z_i}, \mu_{z_i} + \psi_{z_i' z_i}) \end{array} \right\}. \tag{15}$$

  Marginalizing over the each of the latent variables $(u_i, u_i', z_i \in \mathbb{Z}, z_i' \in \mathbb{Z}^+)_{i=1}^n$ in (15), for each $i = 1, \ldots, n$, returns the original likelihood,

$$\prod_{i=1}^{n} \left\{ \sum_{j=-\infty}^{\infty} \left[\sum_{l=1}^{\infty} \frac{\mathbb{I}(\mu_j - \psi_{lj} < y < \mu_j + \psi_{lj})}{2\psi_{lj}} \omega_{lj} \right] \omega_j(\eta_\omega(\mathbf{x}_i), \sigma_\omega(\mathbf{x}_i)) \right\},$$

  of this infinite-dimensional, Bayesian nonparametric model. So again, conditional on the latent variables, the infinite-dimensional model can be treated as a finite-dimensional model. This, in turn, permits the use of standard MCMC methods to sample the model's full joint posterior distribution. Given all variables, save the $(z_i, z_i')_{i=1}^n$, the choice of each z_i is finitely bounded by $\pm N_{\max}$, and the choice of each z_i' has maximum finite value $N_{\max}'$, where $N_{\max} = \max_i[\max_j \mathbb{I}(u_i < \xi_j)|j|]$ and $N_{\max}' = \max_i[\max_l \mathbb{I}(u_i' < \xi_l)l]$. Specifically, for our uniform-mixture kernel regression model, the first three steps of the original 5-step MCMC algorithm are modified, to sample from the following full conditional posterior density, as follows:

1. $\pi(\mu_j | \cdots) \propto \mathrm{n}(\mu_j | \mu_{\mu j}, \sigma_{\mu j}^2)\mathbb{I}(\mu_j \in [a_{\mu j}^*, b_{\mu j}^*])$, for $j = 0, \pm 1, \ldots, \pm N_{\max}$, where $a_{\mu j}^* = \max_{i:z_i=j}\{y_i - \psi_{z_i' j}\}$ and $b_{\mu j}^* = \min_{i:z_i=j}\{y_i + \psi_{z_i' j}\}$, given draws $u_i \sim_{ind} \mathrm{un}(0, \xi_{|z_i|})$ $(i = 1, \ldots, n)$.
2. $\pi(\psi_{lj} | \cdots) = \mathrm{pa}(\psi_{lj} | \alpha_{\psi j} + \sum_{i=1}^n \mathbb{I}(z_i' = l, z_i = j), b_{\mu l j}^*)$, for $l = 1, \ldots, N_{\max}'$ and $j = 0, \pm 1, \ldots, \pm N_{\max}$, given draws $u_i' \sim_{ind} \mathrm{un}(0, \xi_{z_i'})$ $(i = 1, \ldots, n)$, where $b_{\psi lj}^* = \max\{\beta_{\psi j}, \max_{i:z_i'=l, z_i=j} |y_i - \mu_{z_i}|\}$.

3. $\Pr(z_i' = l, z_i = j|\cdots) \propto [\omega_j(\eta(\mathbf{x}_i), \sigma(\mathbf{x}_i))\upsilon_{lj} \prod_{k=1}^{l-1}(1-\upsilon_{kj})\}/(\xi_l \xi_{|j|})] \times \text{un}(y_i|\mu_j - \psi_{lj}, \mu_j + \psi_{lj})$, given draws $\upsilon_{lj} \sim_{ind} \text{be}(a_{lj} + \sum_{i=1}^{n} \mathbb{I}(z_i' = l, z_i = j),\ b_{lj} + \sum_{i=1}^{n} \mathbb{I}(z_i' > l, z_i = j))$, for $j = 0, \pm 1, \ldots, \pm N_{\max}$ and $l = 1, \ldots, N'_{\max}$.
The full conditional posterior densities in Steps 1 and 2 follow from the standard Bayesian uniform models with a conjugate, Pareto prior for the scale parameter ψ.

- Sampling the posterior predictive distribution of Y given $\mathbf{x}$: For the normal kernel regression model (uniform-mixture kernel model, respectively), a step is added to the existing MCMC algorithm, to sample from the full conditional posterior predictive distribution of Y_i given a data-observed $\mathbf{x}_i$, which is $\text{n}(y_i|\mu_{z_i}, \sigma^2_{z_i})$ (is $\text{un}(y_i|\mu_{z_i} - \psi_{z_i' z_i}, \mu_{z_i} + \psi_{z_i' z_i})$, respectively), for $i = 1, \ldots, n$. For the normal kernel model, to draw samples from the posterior predictive distribution of Y, given an unobserved value of $\mathbf{x}$ (or given an observed $\mathbf{x}$ for which y is missing from the data), another step is added to the MCMC algorithm, that draws a sample $y^{\text{pred}} \sim \text{n}(\mu_z, \sigma_z^2)$ from the full-conditional posterior predictive density, given a sample $z = j$ drawn with probability proportional to $\omega_j(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x}))$, $j = 0, \pm 1, \pm 2, \ldots, N^*_{\max}$, where $N^*_{\max}$ is a sufficiently-large value satisfying

$$\omega_{-N^*_{\max}-1}(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x})) = \omega_{N^*_{\max}+1}(\eta_\omega(\mathbf{x}), \sigma_\omega(\mathbf{x})) < \varepsilon \approx 0.$$

  For example $\varepsilon = 10^{-5}$, but it is easy to find a sufficiently-large $N^*_{\max}$ that corresponds to zero mixture weights, according to the numerical precision of a computer. For the uniform-mixture kernel model, we draw $y^{\text{pred}} \sim \text{un}(\mu_z - \psi_{z'z}, \mu_z + \psi_{z'z})$, with $(z = j, z' = l)$ sampled with probability proportional to $\omega_j(\eta(\mathbf{x}), \sigma(\mathbf{x}))\upsilon_{lj} \prod_{k=1}^{l-1}(1-\upsilon_{kj})\}$, for $j = 0, \pm 1, \pm 2, \ldots, \pm N^*_{\max}$ and $j = 0, \pm 1, \pm 2, \ldots, \pm N^{'*}_{\max}$ for sufficiently-large $(N^*_{\max}, N^{'*}_{\max})$, as before. Here, a sufficiently-large $N^{'*}_{\max}$ method can be found by using methods for approximating the DP (Ishwaran & James, 2001).
- Multiple imputation of a censored dependent response y_i: At each iteration of the MCMC algorithm, a plausible value of a dependent response y_i (e.g., a log-event time $y_i = \log(\text{time}_i)$), that is censored and known only to fall within an interval $(a_{y_i}, b_{y_i}]$, is sampled from the full conditional posterior predictive density $\pi(y|\mathbf{x}_i, \cdots) \propto \text{n}(y|\mu_{z_i}, \sigma^2_{z_i})\mathbb{I}(y \in (a_{y_i}, b_{y_i}])$, and then is imputed as the updated value of y_i. For the uniform-mixture kernel model, the sampling method proceeds similarly, after replacing $\text{n}(y|\mu_{z_i}, \sigma^2_{z_i})$ with the uniform density $\text{un}(y|\mu_{z_i} - \psi_{z_i' z_i}, \mu_{z_i} + \psi_{z_i' z_i})$.
- Discrete-valued dependent variable: Using a standard idea (Albert & Chib, 1993), our model can be extended to handle discrete-valued dependent variable responses, $y_i \in \{0, 1, \ldots, C_i^{\max} \geq 1\}$ $(i = 1, \ldots, n)$. Here, $C^{\max}$ is the maximum ordinal category, with $C^{\max} = 1$ for a binary (0-1) dependent variable. For a chosen set of unimodal kernel densities $f(y|\boldsymbol{\theta}_j)$ $(j = 0, \pm 1, \pm 2, \ldots)$, such as the normal kernels or the uniform-mixture kernels, the idea is to introduce latent responses y_i^* underlying each dis-

crete response y_i $(i = 1, \ldots, n)$, and to define the regression model by:

$$f(y_i|\mathbf{x}_i) = \int_{A_i(y_i)} \sum_{j=-\infty}^{\infty} f(y_i^*|\boldsymbol{\theta}_j)\omega_j(\eta_\omega(\mathbf{x}_i), \sigma_\omega(\mathbf{x}_i))\mathrm{d}y_i^*, i = 1, \ldots, n.$$

Here, each $A_i(\cdot)$ is a set function. Given the flexibility of our Bayesian nonparametric regression model, we may define $A(y_i) = (y_i - \infty^{\mathbb{I}(y_i=0)}, y_i\infty^{\mathbb{I}(y_i=C_i^{\max})}]$ (Kottas et al., 2005). The choice of normal kernel densities (uniform-mixture kernel densities, respectively) implies the use of a regression model with inverse-link function modeled by a scale mixture of normal c.d.f.s. (uniform c.d.f.s, respectively). For the normal kernel model (uniform-mixture kernel model, respectively), this extension to discrete-valued dependent variables is achieved by adding a step to the MCMC algorithm, to sample a latent variable y_i^* from its full conditional posterior density $\pi(y^*|\cdots) \propto \mathrm{n}(y_i^*|\mu_{z_i}, \sigma_{z_i}^2)\mathbb{I}(y^* \in A_i(y_i))$ $(\pi(y^*|\cdots) \propto \mathrm{un}(y^*|\mu_{z_i} - \psi_{z_i'z_i}, \mu_{z_i} + \psi_{z_i'z_i})\mathbb{I}(y^* \in A_i(y_i))$, respectively)), for all $i = 1, \ldots, n$. Then, Steps 1-3 of the MCMC algorithm proceed with the sampled latent variables y_i^*, instead of y_i $(i = 1, \ldots, n)$. Also, a sample of the discrete Y^{pred} from its full conditional posterior predictive distribution, given $\mathbf{x}$, is drawn by taking $y^{\mathrm{pred}} = \min(\max\{\mathrm{floor}(y^*) + 1\}, 0], C^{\max})$, given a sampled latent variable y^*. For the normal kernel model (uniform-mixture kernel model, respectively), such a latent variable is sampled by $y^* \sim \mathrm{n}(\mu_z, \sigma_z^2)$ (by $y^* \sim \mathrm{un}(\mu_z - \psi_{z'z}, \mu_z + \psi_{z'z})$, respectively).

- Covariate-dependent kernels: Consider a model with covariate-dependent kernels $\mathrm{n}(y|\mu_j + \mathbf{x}_i^\intercal\boldsymbol{\beta}, \sigma_j^2)$ $(j = 0, \pm1, \pm2, \ldots)$ and prior $\boldsymbol{\beta} \sim \mathrm{n}(\mathbf{m}_{\boldsymbol{\beta}}, \mathbf{V}_{\boldsymbol{\beta}})$. For this model, a step is added to the original 5-step MCMC algorithm, to sample from $\mathrm{n}(\boldsymbol{\beta}|\mathbf{m}_{\boldsymbol{\beta}}^*, \mathbf{V}_{\boldsymbol{\beta}}^*)$ and from

$$\pi(\mu_j|\cdots) \propto \mathrm{n}(\mu_j|\mu_{\mu j}, \sigma_{\mu j}^2) \prod_{i:z_i=j} \mathrm{n}(y_i|\mu_{z_i} + \mathbf{x}_i^\intercal\boldsymbol{\beta}, \mathrm{diag}(\sigma_1^2, \ldots, \sigma_n^2))$$

$(j = 0, \pm1, \pm2, \ldots)$, where $\mathbf{V}_{\boldsymbol{\beta}}^* = (\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}^\intercal\mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})\mathbf{X})^{-1}$, $\mathbf{m}_{\boldsymbol{\beta}}^* = \mathbf{V}_{\boldsymbol{\beta}}^*(\mathbf{V}_{\boldsymbol{\beta}}^{-1}\mathbf{m}_{\boldsymbol{\beta}} + \mathbf{X}^\intercal\mathrm{diag}(\sigma_1^{-2}, \ldots, \sigma_n^{-2})(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{z}})$, $\mathbf{X} = ((1, \mathbf{x}^\intercal))$, $\boldsymbol{\mu}_{\mathbf{z}} = (\mu_{z_1}, \ldots, \mu_{z_n})^\intercal$, and $\mathbf{y} = (y_1, \ldots, y_n)^\intercal$ (or latent $\mathbf{y} = (y_1^*, \ldots, y_n^*)^\intercal$, as applicable). Here, either $(\sigma_1^2, \ldots, \sigma_n^2) = (\sigma_{z_1}^2, \ldots, \sigma_{z_n}^2)$ as in the original normal kernel model, or $(\sigma_1^2, \ldots, \sigma_n^2) = (\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_n^2)$ as in meta-analysis where each $\widehat{\sigma}_i^2$ is the sampling variance of study effect-size y_i. Also, we may assume $(\sigma_1^2, \ldots, \sigma_n^2) = \sigma^2$ with prior $\sigma^{-2} \sim \mathrm{ga}(\alpha_\sigma, \beta_\sigma)$, in which case $\mathbf{V}_{\boldsymbol{\beta}}^* = \sigma_n^2(\mathbf{V}_{\boldsymbol{\beta}}^{-1} + \mathbf{X}^\intercal\mathbf{X})^{-1}$, and then we add an MCMC step to sample from the full conditional posterior $\mathrm{ga}(\sigma^{-2}|\alpha_\sigma + n/2, \beta_\sigma + \{(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{z}})^\intercal(\mathbf{y} - \boldsymbol{\mu}_{\mathbf{z}}) - \mathbf{m}_{\boldsymbol{\beta}}^*(\mathbf{V}_{\boldsymbol{\beta}}^*)^{-1}\mathbf{m}_{\boldsymbol{\beta}}^*\}/2)$. In all cases, posterior predictive sampling, imputation of a censored dependent response, and the sampling of latent variables for discrete dependent response proceeds as above for the original normal kernel model, after replacing $\mathrm{n}(y|\mu_z, \sigma_z^2)$ with $\mathrm{n}(y|\mu_z + \mathbf{x}^\intercal\boldsymbol{\beta}, \sigma_z^2)$ (where perhaps $z = z_i$, and/or $\sigma_z^2 = \widehat{\sigma}_i^2$, or $\sigma_z^2 = \sigma^2$, as applicable).

- Calculating $D_t(m)$: For a given Bayesian model m, the estimate of the criterion $D_t(m)$ ($t \in [0,1]$) is obtained by:

$$\begin{aligned}\widehat{D}_t(m) &= t\sum_{i=1}^{n}\{y_i - \widehat{\mathrm{E}}_n(Y_i|\mathbf{x}_i, m)\}^2 \\ &\quad + \sum_{i=1}^{n}\{\widehat{\mathrm{E}}_n(Y_i^2|\mathbf{x}_i, m) - \widehat{\mathrm{E}}_n(Y_i|\mathbf{x}_i, m)^2\},\end{aligned}$$

given $\widehat{\mathrm{E}}_n(Y_i|\mathbf{x}_i, m) = \frac{1}{S}\sum_{s=1}^{S} y_i^{\mathrm{pred}(s)}$, $\widehat{\mathrm{E}}_n(Y_i^2|\mathbf{x}_i) = \frac{1}{S}\sum_{s=1}^{S}(y_i^{\mathrm{pred}(s)})^2$, and posterior predictive samples $\{\{(y_i^{\mathrm{pred}(s)}|\mathbf{x}_i, m)\}_{i=1}^{n}\}_{s=1}^{S}$. If $t = 1$, then more simply $\widehat{D}_1(m) = \frac{1}{S}\sum_{i=1}^{n}\{y_i - y_i^{\mathrm{pred}(s)}\}^2$.

- Spike-and-slab variable selection. Spike-and-slab priors can be assigned to $\boldsymbol{\beta}_\omega$, to provide automatic variable (predictor) selection in posterior analysis, and provide added insight regarding which covariates significantly predict the dependent variable Y. Specifically, the spike-and-slab prior is defined by $\boldsymbol{\beta}_\omega|\boldsymbol{\gamma} \sim \mathrm{N}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\gamma}})$, with $\mathbf{V}_{\boldsymbol{\gamma}} = \mathrm{diag}(v_1^{\gamma_k} v_0^{1-\gamma_k}|k = 0, 1, \ldots, p)$ for some chosen large v_1 (e.g., $v_1 = 100$) and small v_0 (e.g., $v_0 = .01$), along with independent Bernoulli priors $\gamma_k \sim_{ind} \mathrm{ber}(\Pr(\gamma_k = 1))$ ($k = 0, 1, \ldots, p$) (George & McCulloch, 1997). For spike-and-slab variable selection, the following step is added to the original 5-step MCMC algorithm, to sample independently from the following full conditional posterior distributions:

$$\Pr(\gamma_k|\cdots) = \frac{\mathrm{n}(\beta_k|0, v_1^{\gamma_k} v_0^{(1-\gamma_k)})\Pr(\gamma_k = 1)^{\gamma_k}[1 - \Pr(\gamma_k = 1)]^{1-\gamma_k}}{\mathrm{n}(\beta_k|0, v_1)\Pr(\gamma_k = 1) + \mathrm{n}(\beta_k|0, v_0)[1 - \Pr(\gamma_k = 1)]},$$
$$k = 1, \ldots, p.$$

In practice, a given covariate X_k can be viewed as a significant predictor, when the marginal posterior probability $\Pr(\gamma_k = 1|\mathcal{D}_n)$ exceeds 1/2 (Barbieri & Berger, 2004).

- Multi-level modeling of mixture weights. The multi-level version of our model has prior $\{\boldsymbol{\beta}_{\omega q}\}_{q=1}^{Q}|\mathbf{V}_{\boldsymbol{\beta}\omega} \sim_{iid} \mathrm{n}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}\omega})$, along Wishart hyperprior $\mathbf{V}_{\boldsymbol{\beta}\omega}^{-1} \sim \mathrm{wi}(a_{\mathbf{V}_{\boldsymbol{\beta}\omega}}, \boldsymbol{\Sigma}_{\mathbf{V}_{\boldsymbol{\beta}\omega}})$. For this multilevel model, Step 4 of the original MCMC algorithm is modified to sample from the following full conditional posterior densities: $\pi(\boldsymbol{\beta}_{\omega q}|\cdots) \propto \mathrm{n}(\boldsymbol{\beta}_{\omega q}|\mathbf{0}, \mathbf{V}_{\boldsymbol{\beta}\omega})\prod_{i\in q}\mathrm{n}(z_i^*|(1, \mathbf{x}_i^\intercal)\boldsymbol{\beta}_{\omega q}, \exp((1, \mathbf{x}_i^\intercal)\boldsymbol{\lambda}_{\omega q}))$ ($q = 1, \ldots, Q$), $\pi(\mathbf{V}_{\boldsymbol{\beta}\omega}^{-1}|\cdots) = \mathrm{wi}(\mathbf{V}_{\boldsymbol{\beta}\omega}^{-1}|a_{\mathbf{V}_{\boldsymbol{\beta}\omega}} + Q + p, \{\boldsymbol{\Sigma}_{\mathbf{V}_{\boldsymbol{\beta}\omega}} + \sum_{q=1}^{Q}\boldsymbol{\beta}_{\omega q}\boldsymbol{\beta}_{\omega q}^\intercal\}^{-1})$, where $\pi(\boldsymbol{\beta}_{\omega q}|\cdots)$ may be sampled using a Metropolis-Hastings algorithm. Similarly, we may model $\boldsymbol{\lambda}$ as varying over the Q populations, by assigning priors $\{\boldsymbol{\lambda}_{\omega q}\}_{q=1}^{Q}|\mathbf{V}_{\boldsymbol{\lambda}\omega} \sim_{iid} \mathrm{n}(\mathbf{0}, \mathbf{V}_{\boldsymbol{\lambda}\omega})$ and $\mathbf{V}_{\boldsymbol{\lambda}\omega}^{-1} \sim \mathrm{wi}(a_{\mathbf{V}_{\boldsymbol{\lambda}\omega}}, \boldsymbol{\Sigma}_{\mathbf{V}_{\boldsymbol{\lambda}\omega}})$. The computationally-expensive matrix inversion required for the full conditional $\pi(\mathbf{V}_{\boldsymbol{\beta}\omega}^{-1}|\cdots)$ can be avoided, with perhaps only some loss of modeling flexibility, by instead modeling with the ridge priors $(\boldsymbol{\beta}_{\omega q}, \boldsymbol{\lambda}_{\omega q})|v_\omega \sim_{iid} \mathrm{n}(\mathbf{0}, v_\omega \mathbf{I}_{p+1})$ and $v_{\boldsymbol{\beta}}^{-1} \sim \mathrm{ga}(\alpha_\omega, \beta_\omega)$. Then, after treating all the coefficient vectors $\boldsymbol{\beta}_\omega = (\boldsymbol{\beta}_{\omega 1}^\intercal, \ldots, \boldsymbol{\beta}_{\omega Q}^\intercal)^\intercal$

and $\boldsymbol{\lambda}_\omega = (\boldsymbol{\lambda}_{\omega 1}^\intercal, \ldots, \boldsymbol{\lambda}_{\omega Q}^\intercal)^\intercal$ as each being $(Qp+1)$-dimensional parameters constructed from a design matrix $\mathbf{X}$ that specifies all 2-way interactions between the p covariates and the Q population (0-1) indicators, the full conditional posterior distribution of $\boldsymbol{\beta}_\omega$ is a multivariate normal according to Step 4 of the original MCMC algorithm, the full conditional of $\boldsymbol{\lambda}_\omega$ can be sampled as in Step 5 of the original algorithm. Also, a MCMC step is added to sample full conditional posterior $v_\omega^{-1} \sim \text{ga}(\alpha_\omega + Qp + 1, \beta_\omega + .5\boldsymbol{\beta}_\omega^\intercal \boldsymbol{\beta}_\omega + .5\boldsymbol{\lambda}_\omega^\intercal \boldsymbol{\lambda}_\omega)$.

Appendix B: Software and Priors for the Other Models

All the other regression models, used for comparisons in all the data analyses of Section 3, were fitted using defaults of appropriate packages of the R statistical software (R Development Core Team, 2011). They include the additive model estimated under ridge regression via generalized cross-validation (GCV) and thin-plate splines (Wood, 2004; mgcv package, Wood, 2010), the Bayesian Additive Regression Trees (BART) model using the 200 tree default (Chipman et al., 2010; BayesTree package, Chipman & McCulloch, 2010), the multivariate adaptive regression splines (MARS) model which combines the use of tensor-product splines and stepwise variable selection (Friedman, 1991; earth package, Milborrow, 2009), mean and variance regression estimated under restricted maximum likelihood (REML) (statmod package, Smyth, 2010), the single-index model under penalized estimation (Polzehl & Sperlich, 2009; EDR package, Polzehl, 2010), least-squares linear regression models, and normal random effects models with random school-level intercepts (Hierarchical Linear Models, HLM), fit either by REML or maximum likelihood (ML) estimation (nlme package, Pinheiro et al., 2010). For the PIRLS data set, the 40 simulated data sets, and the 22 additional real data sets, the LASSO model (LASSO=Least Absolute Shrinkage and Selection Operator) was estimated using the LARS algorithm, and the penalty-shrinkage parameter λ for the fixed regression coefficients was estimated by minimizing Mallows' (1973) C_p criterion (Efron, et al. 2004; using the lars package, Hastie & Efron, 2007). Both the LASSO model and the Bayesian LASSO (Park & Casella, 2008; monomvn package, Gramacy, 2010) assumed covariates $\mathbf{x}^*$ that were an order-2 polynomial ("quadratic") expansion of $\mathbf{x}$, and also each column of the $n \times p$ design matrix $((x_{i1}^*, \ldots, x_{ip}^*)^\intercal)_{n \times p}$ was standardized to have unit L^2 norm. For the AZT data, the LASSO logit binary regression model was estimated, with coefficient penalty parameter (and covariates) selected by minimizing deviance via 10-fold cross-validation (using the glmnet package; Friedman et al., 2010). For all data sets, the finite mixture of linear and logit regressions were fit by maximum likelihood (using the flexmix package; Gruen & Leisch, 2007). Here, the number of mixture components was chosen to minimize Akaike's Information Criterion (1973). Finally, nearly all of the competing Bayesian regression models were fit using the DPpackage (Jara, 2011). The Bayesian LASSO (Park & Casella, 2008) was fit using the monomvn package (Gramacy, 2010). The code that was used to fit all these other regression models can be requested from the first author.

All of the competing Bayesian regression models, used to analyze the data in Section 3, assumed proper diffuse priors. Each ANOVA/linear DDP model assigned a multivariate normal baseline distribution with density $\mathrm{n}(\boldsymbol{\beta}|\boldsymbol{\mu}, \boldsymbol{T})$ for the full vector of $p+1$ random regression coefficients $\boldsymbol{\beta}$, along with normal and inverted-Wishart hyperpriors $\boldsymbol{\mu} \sim \mathrm{n}(\mathbf{0}, 10^3\mathbf{I}_{p^*})$ and $\boldsymbol{T} \sim \mathrm{iw}(p^*+3, \mathbf{I})$, where $p^* = \dim(\boldsymbol{\beta})$. Each DP-mixed intercepts regression model, and the MPT-mixed intercepts logit regression model, assigned the same baseline distribution, and corresponding hyperpriors, for the intercept parameter, β_0. All DP-based models assumed a gamma hyper-prior $\alpha \sim \mathrm{ga}(.01, .01)$ for the DP precision parameter, α. In the Mixture of Pólya Trees (MPT) model (Hanson, 2006), a 10-level tree was specified with beta random variables having parameters αk, with k the level in the tree, with α fixed to 1. MPT models were also fit using a $\alpha \sim \mathrm{ga}(.01, .01)$ hyperprior, but this did not appear to lead to a significant advantage in terms of the $D_1(m)$ predictive criterion. All DP-based and MPT-based models assigned a gamma hyperprior density $\mathrm{ga}(.01/2, .01/2)$ to the inverse error variance σ^{-2}. This included the median regression model that assigned a MPT prior to the regression errors, with this prior having normal baseline distribution with density $\mathrm{n}(0, \sigma^2)$. Nearly all Bayesian models that were parameterized by a vector of fixed (non-mixed) regression coefficients $\boldsymbol{\beta}$ (e.g., $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\intercal$ for the DP-mixed intercepts regression model) assigned a multivariate normal $\boldsymbol{\beta} \sim \mathrm{n}(\mathbf{0}, 10^3\mathbf{I})$ prior. The Bayesian LASSO model, for the fixed coefficients, assigned priors $\beta_0 \propto 1$, independent Laplace priors $\pi(\beta_1, \ldots, \beta_p|\sigma^2) = \prod_{k=1}^{p}(\lambda_\beta/\{2\sqrt{\sigma^2}\})\exp(-\lambda_\beta|\beta_k|\}\sigma^{-1})$, $\sigma^2 \propto 1/\sigma^2$, and hyperprior $\lambda_\beta^2 \propto 1/\lambda_\beta^2$. All of these default diffuse prior specifications are very similar to the priors used for the empirical examples presented in user's manuals of the DPpackage and the monomvn package of R.

Acknowledgements

This research is supported by National Science Foundation research grant SES-1156372, from the program in Methodology, Measurement, and Statistics. The authors are grateful for the helpful suggestions of the Associate Editor and two anonymous referees, on previous versions of this manuscript.

References

Agresti, A. (1996). *An introduction to categorical data analysis*. John Wiley and Sons, New York. MR1394195

Akaike, H. (1973). Information Theory and the an Extension of the Maximum Likelihood Principle. In *Second International Symposium On Information Theory* (B. N. Petrov and F. Csaki, eds.) 267–281. Academiai Kiado, Budapest. MR0483125

Albert, J. H. and Chib, S. (1993). Bayesian Analysis of Binary and Polychotomous Response Data. *Journal of the American Statistical Association* **88** 669–679. MR1224394

Barbieri, M. and Berger, J. (2004). Optimal Predictive Model Selection. *Annals of Statistics* **32** 870–897. MR2065192

Barrientos, A. F., Jara, A. and Quintana, F. A. (2012). On the Support of MacEachern's Dependent Dirichlet Processes and Extensions. *Bayesian Analysis* **7** 277–310.

Brunner, L. J. (1992). Bayesian nonparametric methods for data from a unimodal density. *Statistics and Probability Letters* **14** 195–199. MR1173617

Cepeda, E. and Gamerman, D. (2001). Bayesian modeling of variance heterogeneity in normal regression models. *Brazilian Journal of Probability and Statistics* **14** 207–221.

Chipman, H., George, E. I. and McCulloch, R. E. (2010). BART: Bayesian Additive Regression Trees. *Annals of Applied Statistics* **4** 266–298. MR2758172

Chipman, H. and McCulloch, R. (2010). BayesTree: Bayesian Methods for Tree Based Models R package version 0.3-1.1.

DeIorio, 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. MR2054299

Dunson, D. and Park, J. H. (2008). Kernel Stick Breaking Processes. *Biometrika* **95** 307–323. MR2521586

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least Angle Regression. *Annals of Statistics* **32** 407–499. MR2060166

Ferguson, T. S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics* **1** 209–230. MR0350949

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines (With Discussion). *Annals of Statistics* **19** 1–67. MR1091842

Friedman, J. H., Hastie, T. and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* **33**.

Fuentes-García, R., Mena, R. H. and Walker, S. G. (2010). A New Bayesian Nonparametric Mixture Model. *Communications In Statistics* **39** 669–682. MR2785657

Gelfand, A. E. and Banerjee, S. (2010). Multivariate Spatial Process Models. In *Handbook of Spatial Statistics* (A. E. Gelfand, P. Diggle, P. Guttorp and M. Fuentes, eds.) 495–515. Chapman and Hall/CRC, Boca Raton. MR2730963

Gelfand, A. E. and Ghosh, J. K. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika* **85** 1–11. MR1627258

Gelfand, A. E., Kottas, A. and MacEachern, S. N. (2005). Bayesian Nonparametric Spatial Modeling With Dirichlet Processes Mixing. *Journal of the American Statistical Association* **100** 1021–1035. MR2201028

Gelman, A., Jakulin, A., Pittau, M. and Su, Y. S. (2008). A Weakly Informative Default Prior Distribution for Logistic and Other Regression Models. *The Annals of Applied Statistics* **2** 1360–1383. MR2655663

George, E. I. and McCulloch, R. E. (1997). Approaches for Bayesian Variable Selection. *Statistica Sinica* **7** 339–373.

Gramacy, R. B. (2010). Monomvn: Estimation for multivariate normal and Student-t data with monotone missingness R package version 1.8-3.

Griffin, J. E. and Steel, M. F. J. (2006). Order-Based Dependent Dirichlet Processes. *Journal of the American Statistical Association* **101** 179–194. MR2268037

Gruen, B. and Leisch, F. (2007). Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics and Data Analysis* **51** 5247–5252. MR2370869

Hanson, T. E. (2006). Inference for Mixtures of Finite Pólya Tree Models. *Journal of the American Statistical Association* **101** 1548–1565. MR2279479

Hastie, T. and Efron, B. (2007). Lars: Least Angle Regression, Lasso and Forward Stagewise R package version 0.9-7.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*. Chapman and Hall, London. MR1082147

Holmes, C. C., Denison, D. G. T., Ray, S. and Mallick, B. K. (2005). Bayesian Prediction via Partitioning. *Journal of Computational and Graphical Statistics* **14** 811–830. MR2211368

Hwang, J., Lay, S., Maechler, R., Martin, D. and Schimert, J. (1994). Regression Modelling in Back-Propagation and Projection Pursuit Learning. *IEEE Transactions of Neural Networks* **5** 342–353.

Ibrahim, J. G., Chen, M. H. and Sinha, D. (2001). Criterion-based methods for Bayesian model assessment. *Statistica Sinica* **11** 419–443. MR1844533

Ibrahim, J. G. and Kleinman, K. P. (1998). Semiparametric Bayesian Methods for Random Effects Models. In *Practical Nonparametric and Semiparametric Bayesian Statistics. Lecture Notes in Statistics 133* (D. Dey, P. Müller and D. Sinha, eds.) 89–114. Springer-Verlag, New York. MR1630077

Ishwaran, H. and James, L. F. (2001). Gibbs Sampling Methods for Stick-Breaking Priors. *Journal of the American Statistical Association* **96** 161–173. MR1952729

Jara, A. and Hanson, T. (2011). A class of mixtures of dependent tail-free processes. *Biometrika* **98** 553–566. MR2836406

Jara, A., Hanson, T. E., Quintana, F. A., Müller, P. and Rosner, G. L. (2011). DPpackage: Bayesian Semi- and Nonparametric Modeling in R. *Journal of Statistical Software* **40** 1–20.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-Width Output Analysis for Markov Chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547. MR2279478

Kalli, M., Griffin, J. and Walker, S. G. (2010). Slice Sampling Mixture Models. *Statistics and Computing* **21** 93–105. MR2746606

Kim, H., Loh, W. Y., Shih, Y. S. and Chaudhuri, P. (2007). Visualizable and interpretable regression models with good prediction power. *IEEE Transactions: Special Issue on Data Mining and Web Mining* **39** 565–579.

Kottas, A., Müller, P. and Quintana, F. (2005). Nonparametric Bayesian Modeling for Multivariate Ordinal Data. *Journal of Computational and Graphical Statistics* **14** 610–625. MR2170204

Laud, P. W. and Ibrahim, J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society, Series B* **57** 247–262. MR1325389

Lo, A. Y. (1984). On a Class of Bayesian Nonparametric Estimates. *Annals of Statistics* **12** 351–357. MR0733519

MacEachern, S. N. (1999). Dependent Nonparametric processes. *Proceedings of the Bayesian Statistical Sciences Section of the American Statistical Association* 50–55.

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

MacEachern, S. N. (2001). Decision Theoretic Aspects of Dependent Nonparametric Processes. In *Bayesian Methods with Applications to Science, Policy and Official Statistics* (E. George, ed.) 551–560. International Society for Bayesian Analysis, Creta.

Mallows, C. L. (1973). Some Comments on Cp. *Technometrics* **15** 661–675.

Milborrow, S. (2009). Earth: Multivariate Adaptive Regression Spline Models R package version 2.4-0.

Mukhopadhyay, S. and Gelfand, A. E. (1997). Dirichlet Process Mixed Generalized Linear Models. *Journal of the American Statistical Association* **92** 633–639. MR1467854

Müller, P., Erkanli, A. and West, M. (1996). Bayesian Curve Fitting Using Multivariate Normal Mixtures. *Biometrika* **83** 67–79. MR1399156

Müller, P. and Quintana, F. A. (2010). Random Partition Models with Regression on Covariates. *Journal of Statistical Planning and Inference* **140** 2801–2808. MR2651966

Müller, P., Quintana, F. A. and Rosner, G. L. (2011). A Product Partition Model with Regression on Covariates. *Journal of Computational and Graphical Statistics* **20** 260–278. MR2816548

Newton, M. A., Czado, C. and Chappell, R. (1996). Bayesian Inference for Semiparametric Binary Regression. *Journal of the American Statistical Association* **91** 142–153. MR1394068

O'Hagan, A. and Forster, J. (2004). *Kendall's Advanced Theory of Statistics: Bayesian Inference* **2B**. Arnold, London.

Park, Y. and Casella, G. (2008). The Bayesian LASSO. *Journal of the American Statistical Association* **103** 681–686. MR2524001

Park, J. H. and Dunson, D. B. (2010). Bayesian generalized product partition models. *Statistica Sinica* **20** 1203–1226. MR2730180

Perman, M., Pitman, J. and Yor, M. (1992). Size-biased sampling of Poisson point processes and excursions. *Probability Theory and Related Fields* **92** 21–39. MR1156448

Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D. and R Development Core Team (2010). Nlme: Linear and Nonlinear Mixed Effects Models R package version 3.1-97.

Polzehl, J. (2010). EDR: Estimation of the effective dimension reduction (EDR) space R package version 0.6-4.

Polzehl, J. and Sperlich, S. (2009). A note on structural adaptive dimension reduction. *Journal of Statistical Computation and Simulation* **79** 805–818. MR2751594

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods (Second Edition)*. Springer, New York. MR2080278

Rodriguez, A., Dunson, D. B. and Gelfand, A. E. (2008). The Nested Dirichlet Process. *Journal of the American Statistical Association* **103** 1131–1144. MR2528831

Rodriguez, A. and Dunson, D. B. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis* **6** 1–34. MR2781811

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Statistica Sinica* **4** 639–650. MR1309433

Smyth, G. (2010). Statmod: Statistical modeling R package version 1.4.6.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Sharing Clusters Among Related Groups: Hierarchical Dirichlet Processes. *Journal of the American Statistical Association* **101** 1566–1581. MR2279480

Tokdar, S. T., Zhu, Y. M. and Ghosh, J. K. (2010). Density regression with logistic Gaussian process priors and subspace projection. *Bayesian Analysis* **5** 316–344. MR2719655

Walker, S. G. and Karabatsos, G. (2012). Revisiting Bayesian curve fitting using multivariate normal mixtures. In *Bayesian Theory and Applications* (P. Damien, P. Dellaportas, N. Polson and D. Stephens, eds.) 297–305. Oxford University Press, New York.

Wood, S. N. (2004). Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models. *Journal of the American Statistical Association* **99** 673–686. MR2090902

Wood, S. N. (2010). GAMs with GCV/AIC/REML Smoothness Estimation and GAMMs by PQL: mgcv Package Documentation for the R Software, R Foundation for Statistical Computing, Vienna, Austria.