Bayesian Analysis (2011) **6**, Number 4, pp. 793–828

A Computational Bayesian Method for Estimating the Number of Knots In Regression Splines

Minjung Kyung*

Abstract. To determine the size of the drug-involved offender population that could be served effectively and efficiently by partnerships between courts and treatment in the United States, a synthetic dataset is created by Bhati and Roman (2009). Because of hidden structure and aggregation necessary to create variables, there exists latent variance that can not be explained fully by a normal random effect model. Semiparametric regression is a well-known and frequently used tool to capture the functional dependence between variables with fixed effect parametric and nonlinear regression. A new Gibbs sampler is developed here for the number and positions of knots in regression splines by expressing semiparametric regression as a linear mixed model with a random effect term for the basis functions. Our Gibbs sampler exploits the properties of the multinomial-Dirichlet distribution, and is shown to be an improvement, in terms of operator norm and efficiency, over add/delete one MCMC algorithms. We find that the Dirichlet distribution with small values of the parameters results in a smaller number of knots and, in general, good fit to the data. This approach is shown to reveal previously unseen structures in the synthetic dataset of Bhati and Roman.

Keywords: Regression Splines, Multinomial-Dirichlet distribution, Bayesian Semiparametric Regression

1 Introduction

Is substance abuse treatment an effective method for reducing drug-involved offenses? What treatment options are the most effective and cost-beneficial? These are important policy questions for reducing crime and improving the lives of addicted individuals. A key measurement objective in this literature is estimating the size of the drug-involved offender population that could be served effectively and efficiently by partnerships between the courts and treatment regimes. Unfortunately, obtaining individual-level data

*Department of Statistics, Duksung Women's University Seoul, Korea, mkyung@duksung.ac.kr

 DOI:10.1214/11-BA629

that can be usefully aggregated is a much more difficult challenge than expected since there exists considerable under-reporting, privacy restrictions, and medical issues in such settings. Bhati and Roman (2009) recently approached this dilemma in a creative way by constructing micro-level data from three nationally representative sources to construct a 40,320 case *synthetic dataset*, which uses population profiles rather than sampled observations.

Researchers create synthetic datasets (Rubin 1993) when the actual data in its raw form are either unavailable or partially restricted due to privacy concerns. However, it is often the case that in the aggregation process that creates synthetic data hidden structures and latent information are formed in unexpected ways. What this suggests is that the set of parametric models in our standard toolkit, linear and generalized linear models, will be inadequate for explaining the key underlying relationships in such data. Additionally, the data creation process used by Bhati and Roman (2009) relies heavily on simulation models to estimate substance abuse treatment effect. The work here develops a new Bayesian semiparametric regression model that balances user-defined restrictions against pure data information, and is sufficiently flexible that it has the ability to capture unusual or hidden features of the data that would ordinarily be missed by conventional approaches. This approach requires a new variant of the Gibbs sampler to provide posterior estimates.

Semiparametric regression is concerned with the flexible incorporation of nonlinear functional relationships in regression analysis. Consider the typical setup of an n-length outcome variable vector $\mathbf{Y}$ associated with the explanatory matrix $\mathbf{R}$, which is partitioned into two components, $n \times p$ matrix $\boldsymbol{X}$ and $n \times q$ matrix $\mathbf{W}$, such that for the ith case, $i = 1, \ldots, n$, Y_i is modeled as

$$\mathrm{E}\left[\mathrm{Y_i}|\mathbf{R_i}\right] = \mathrm{h}\left(\mathrm{X_i}\boldsymbol{\beta} + \mathrm{g}(\mathbf{W_i})\right), \tag{1}$$

where $g(\cdot)$ is some unspecified "smooth" function to be estimated. This specification, with known link function $h(\cdot)$, is therefore a form of semiparametric regression due to the partitioned treatment of the covariates. Details about semiparametric regression models are found in Ruppert *et al.* (2003) and Härdle *et al.* (2004). Recent applications of these models are contained in Zhang and Davidian (2001), Zeng and Lin (2007), Yin *et al.* (2008), and Maity *et al.* (2009). A likelihood-based approach is developed by Ke and Wang (2001) with a semiparametric nonlinear mixed effects model (SNMM) that extends the nonlinear mixed effects models where self-modeling nonlinear regression models are used to fit repeated measures data. Fahrmeir and Lang (2001) proposed an approach for Bayesian inference via Markov chain Monte Carlo (MCMC) methods in generalized

additive and semiparametric mixed models. Our approach is a modernization of this idea whereby the smoothing parameters of $g(\cdot)$ are updated on each cycle of the Gibb sampler.

The creation and use of synthetic data has dramatically increased over the last decade due to heightened attention to privacy issues, more powerful computation, demand for larger datasets, a huge upswing in data mining work, and improved estimation algorithms. When information is combined from multiple data sources in this process, profiles are created, which are combinations of attributes that are considered in the same way as actual individual observations. Thus, cases are created which are not actual individuals but characterizations of individuals that are in the aggregate unbiased. In the case of Bhati and Roman (2009) the synthetic data contains a profile for every client permutation, therefore allowing estimation of the effect of drug treatment on every combination of client attributes and characteristics. Additionally, each profile is tested against four differing treatment regimes to understand their efficacy in reducing substance abuse. The application of the Bayesian semiparametric regression model developed in this work shows that Bhati and Roman *miss key features* present in their synthetic data, and that these features substantially change the policy interpretation of the results. Further evidence is supplied here to show that the adaptive semiparametric algorithm reveals latent structure in benchmarking nonparametric datasets suggesting that this approach is useful in very general contexts as well.

1.1 Relationship to Mixed Models

A semiparametric model can be expressed as a set of penalized regression splines, and, more generally, as a linear mixed model. To allow the flexibility in the estimation of a unknown function, constraints are widely used, typically with a roughness penalty, which leads to parametric statistical models. In a classic work, Wahba (1977) derived theoretical details showing that a nonparametric regression model can be rewritten in the form of a linear combination of the fixed effect (trends) and the random effects (the small-scale variations). Silverman (1985) then showed that spline smoothing provides a natural and flexible approach to curve estimation and the smoothing parameter determines the degree to which the data are smoothed to produce the estimate. Stone (1985) also showed that the closed additive approximation to a nonparametric regression model g with explanatory variables is $h^*(\cdot) = \mu + \sum_{j=1}^{J} f^*(\cdot)$, which has been chosen subject to the constraint that $Df_j^* = 0$ for $j = 1, \ldots, J$, minimizes mean squared error. French *et al.* (2001) argued that good low-rank approximations (reduced knot, $K < n$) exist

with the mixed model representation of smoothing splines by showing the equivalence of the BLUP coefficient estimator of low-rank smoothing splines and to the exact smoothing splines. These results allow for mixed model software solutions to perform the entire fitting algorithm and for inference within the mixed model framework.

For the ease of notation, we consider again the following linear semiparametric regression model

$$Y_i|\mathbf{R}_i = \boldsymbol{X}_i\boldsymbol{\beta} + g(W_i) + \epsilon_i, \tag{2}$$

where $\epsilon_i \sim$ iid $N\left(0, \sigma_\epsilon^2\right)$ for $i = 1, \ldots, n$. Here, g is a unknown function and needs to be estimated. The model (2) can be treated as a semiparametric regression when the covariate is measured with error. We can consider smoothing splines to estimate the unknown function g, but smoothing splines become less practical when n is large, because they can use up to n knots. Thus, an alternative approach to spline fitting is penalized splines, which are given by

$$g(W_i) = \sum_{k=1}^{K} \gamma_k \mathbf{B}_k\left(W_i\right) \tag{3}$$

where $K < n$ is the number of knots with the degree of the B-spline or the degree of polynomial, and $\mathbf{B}_k$s denote basis functions. Here we define the order of positions of K knots as $\boldsymbol{\kappa}_K = (\kappa_1, \ldots, \kappa_K)'$. Thus, penalized spline fitting can be written generally as

$$\min_{\boldsymbol{\gamma}} ||y - \boldsymbol{X}\boldsymbol{\beta} - \mathbf{S}\boldsymbol{\gamma}||^2 + \lambda\boldsymbol{\gamma}'\mathbf{D}\boldsymbol{\gamma} \tag{4}$$

where $\lambda > 0$ is the penalty parameter, $\mathbf{S}$ is an $n \times K$ smoother matrix of $\mathbf{B}_k$ with coefficient vector $\boldsymbol{\gamma}$, and $\mathbf{D}$ is a symmetric positive semidefinite matrix.

For the semiparametric regression in exponential families with suitable link function of $h(\cdot)$ in (1), we need to investigate a new parameterization of the hierarchical model to derive a Gibbs sampler that more fully exploits the structure of the model. Also, for the multivariate semiparametric regression, we can consider generalized additive models (GAM) by an additive model predictor, interaction models with tensor products of spline bases and bivariate radial basis functions, and the bivariate smoothing models with thin plate spline family of smoothers. These issues are left to further research in this area.

There are various basis functions that have been used in this context. The truncated power functions of degree p, natural cubic splines (Green and Silverman 1994), B-spline basis function (Eilers and Marx 1996), and the radial basis functions (Ripley 1996) are the most popular choices. Eilers and Marx (1996) showed that B-spline basis functions provide more stable numerical properties than truncated power functions. They have

local support, thus with different positions of knots with the same number of knots, B-spline basis functions will have flexible forms.

For the degree of a spline p, Ruppert *et al.* (2003) discussed that if one is using a linear spline with enough knots so that increasing the number of knots has no appreciable effect on the penalized fit, then increasing the degree of the spline is also unlikely to have a noticeable effect. In this paper, we do not restrict or prefix the degree of a spline. Also, we do not restrict to a specific set of basis functions, as the following methods can be applied to all basis functions.

Kauermann *et al.* (2009) discussed asymptotic properties of generalized penalized spline smoothing if the spline basis increases with the sample size. They argue that the equivalence of penalized spline fitting and generalized linear mixed models is asymptotically justified only if the Laplace approximation holds. Also, they make use of a fully Bayesian viewpoint by imposing prior distributions on all parameters and coefficients, and show that a fully Bayesian formulation of the model yields approximately the same results as a Laplace approximation even for growing dimensions of the spline basis. This means that even though we express the semiparametric model as a form of mixed model, these models share the same theoretical properties.

Furthermore, one of the merits of semiparametric models is that the Gauss-Markov theorem generally holds, as in the linear regression model. There are various versions of Gauss-Markov theorems for linear random effect models (Harville 1976) and for a heteroscedastic linear model (Carroll 1982). Also, Pfeffermann (1984) discussed extensions of the Gauss-Markov theorem to the case of stochastic regression coefficients and Robinson (1988) discussed a $\sqrt{N}$ consistent version of semiparametric regression. More relevantly, for semiparametric models, Huang and Lu (2001) and French *et al.* (2001) discuss Gauss-Markov theorems for fixed effect coefficient estimators by re-expressing the nonparametric mixed effects model as a linear combination of the fixed effects and the random effects by describing the space of fixed effects and the space of random effects via subspaces of certain Reproducing kernel Hilbert spaces (RKHS) (Aronszajn 1950).

For better estimation and fitting of models to a given dataset with basis functions, we need to carefully consider the number of knots K and the order of positions $\boldsymbol{\kappa}_K = (\kappa_1, \ldots, \kappa_K)'$. Early discussions about the penalized linear spline models with smaller number of basis functions than sample size ($K \leq n$) can be found in Parker and Rice (1985), O'Sullivan (1986), Kelly and Rice (1990), Gray (1994), Hastie (1996), Eilers and Marx (1996) and Ruppert and Carroll (2000). More discussions in recent years are

presented in French *et al.* (2001), Ruppert (2002), Wand (2003), Ruppert *et al.* (2003) and Claeskens *et al.* (2009).

1.2 Choosing Knots

In a regular penalized linear model (4), K, $\boldsymbol{\kappa}_K$ and λ will control the smoothness of the model, and how much variational information can be captured in random effects terms. The penalty parameter λ is a parameter that controls the bias and variance of the random effects. Also, in practice, the number of knots K and their position in $\boldsymbol{\kappa}_K$ are unknown, so we need to estimate them.

There has been much written on the choice of K and $\boldsymbol{\kappa}_K$ based on generalized cross validation (GCV) using various smoothing methods (Friedman and Silverman 1989; Hastie and Tibshirani 1990; Ruppert 2002; Ruppert *et al.* 2003; Woods 2006). Also, there are knot selection methods based on stepwise selection (Stone *et al.* 1997) and on a componentwise boosting algorithm with radial basis functions (Leitenstorfer and Tutz 2007). For Bayesian methods, reversible jump MCMC has been widely used (Denison *et al.* 1998; Biller 2000; DiMatteo *et al.* 2001; Holmes and Mallick 2001).

Recently, Claeskens *et al.* (2009) provided a theoretical justification that, depending on the number of knots, sample size and penalty, the theoretical properties of penalized regression spline estimators are either similar to those of regression splines or to those of smoothing splines, with a clear breakpoint distinguishing the cases. They prove that a smaller number of knots leads to a smaller averaged mean squared error. Also, they showed that using truncated polynomial basis functions leads to an optimal rate of convergence independent of the assumption made on the number of knots.

Stochastic methods that move simultaneously in model space and parameter space allow us a limitless range of possibilities for the choice of K and $\boldsymbol{\kappa}_K$ (Denison *et al.* 1998; Biller 2000; DiMatteo *et al.* 2001; Holmes and Mallick 2001, 2003). Especially for the Bayesian methods, since the dimensionality of the parameter space generally changes with the model, reversible jump MCMC (RJMCMC) proposed by Green (1995), is sometimes used. However, the RJMCMC algorithm usually results in a slow mixing chain because empty components arise and the sampler retains these for extended periods of time.

In this paper, we develop an overtly Bayesian method to find the number of knots and the position of knots in a Gibbs sampling scheme. For this we need to consider

various candidates for the position of knots. Typically, the observed data points have been used as candidates for knots. If we consider a different point of view, the position of knots can be thought of as a changepoint problem in the data sense (Moreno *et al.* 2005; Girón *et al.* 2007). This means that before and after a knot, the curve changes directions or amount of curvature, or it drops or goes up suddenly. Thus, we propose a method for the choice of K and $\boldsymbol{\kappa}_K$ based on the standard changepoint problem by using Bayes Factors (BF) for the update instead of the likelihood function. In this process, we apply a Dirichlet prior on the process of choosing the number of knots, K and the order of positions, $\boldsymbol{\kappa}_K$, and instead of λ, a Dirichlet prior on a changepoint setup that controls the bias and variance of the random effects. Thus, in our model, we don't need to consider the penalty part with $\mathbf{D}$ and λ in a regular penalized linear model (4). More details about the MCMC sampling schemes for the regression splines are provided in the following sections.

2 A Bayesian Approach to Knot Selection

We now outline a Gibbs sampler for the semiparametric model, which is represented as a linear mixed model where, at each iteration, we generate a length n vector $\boldsymbol{\kappa}_K$ and recover the knot size through marginalization for K and $\boldsymbol{\kappa}_K$.

To make the sampling scheme easier, let $W_1 \leq W_2 \leq \cdots \leq W_n$. Given the number of knots K and their positions $\boldsymbol{\kappa}_K$, the linear mixed model representation of the semiparametric model with basis functions can be written as

$$\mathbf{y} = \boldsymbol{X}\boldsymbol{\beta} + \mathbf{S}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\mathbf{S}$ is an $n \times (K+1)$ smoother matrix with K knots in the positions of $\boldsymbol{\kappa}_K = (\kappa_1, \ldots, \kappa_K)'$. Thus, $\mathbf{S}_{ik}$, the element in the ith row and kth column of $\mathbf{S}$, is $\mathbf{B}_k(W_i)$ in (3), which are basis functions only related to K knots. Also,

$$\boldsymbol{\epsilon} \sim N_n\left(\mathbf{0}, \sigma^2\mathbf{I}\right) \text{ and } \quad \boldsymbol{\gamma} \sim N_{K+1}\left(\mathbf{0}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}\right).$$

Here, for $\mathbf{W} \in [a, b]$ with $a = \kappa_0 < \kappa_1 < \cdots < \kappa_K < \kappa_{K+1} = b$, we define

$$n_k = \text{ number of observations in } (\kappa_k, \kappa_{k+1}], \quad k = 0, \ldots, K,$$

$n_0 + n_1 + \cdots + n_K = n$, and we use

$$\boldsymbol{\Sigma}_{\boldsymbol{\gamma}} = \text{diag}\{n_0, n_1, \ldots, n_K\}.$$

The given structure can also be considered as a subclustering of W's with subcluster sizes $n_0, n_1, \ldots, n_K$. This means that the position of knots will determine the subgroups and the subgroup size. Because the W's are assumed to be in increasing order, it is easy to see that $\kappa_1 = W_{n_0}$, $\kappa_2 = W_{n_0+n_1}$, ..., and $\kappa_K = W_{\sum_{k=0}^{K-1} n_k}$. Thus, we are marginalizing n knots into K knots in a manner similar to that of Kyung *et al.* (2010).

2.1 Sampling Scheme

The joint likelihood for regression parameters $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}_K$ with normal assumptions can be written as:

$$\begin{aligned} & L_K\left(\boldsymbol{\beta}, \boldsymbol{\gamma}, \sigma^2, \boldsymbol{\kappa}_K | \mathbf{y}\right) \qquad (5) \\ & = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} e^{-\frac{1}{2\sigma^2}|\mathbf{y}-\boldsymbol{X}\boldsymbol{\beta}-\mathbf{S}\boldsymbol{\gamma}|^2} \left(\frac{1}{2\pi\sigma^2}\right)^{(K+1)/2} |\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2} e^{-\frac{1}{2\sigma^2}\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\gamma}}. \end{aligned}$$

With a flat prior on $\boldsymbol{\kappa}_K$ and K, such as $\pi(K) \propto c$ and $\pi(\boldsymbol{\kappa}_K|K) \propto c$, and with priors on the regression parameter $\pi(\boldsymbol{\theta})$, we get the joint posterior distribution as:

$$\pi(\boldsymbol{\theta}, \boldsymbol{\kappa}_K \mid \mathbf{y}) = \frac{L_K(\boldsymbol{\theta}, \boldsymbol{\kappa}_K|\mathbf{y})\pi(\boldsymbol{\theta})}{\int_\Theta \sum_{J=0}^{n-1} \sum_{\boldsymbol{\kappa}_J \in \mathscr{K}_J} L_J(\boldsymbol{\theta}, \boldsymbol{\kappa}_J|\mathbf{y})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}},$$

where

$$\begin{aligned} \mathscr{K}_K &= \{\text{all possible } K \text{ knots} : (\kappa_1, \ldots, \kappa_K)\} \\ &= \{(n_0, \ldots, n_K) : \text{all possible } n_k\text{'s such that } n_0 + n_1 + \cdots + n_K = n\}. \end{aligned}$$

Thus, the full conditional posteriors of $\boldsymbol{\theta}$ and $\boldsymbol{\kappa}_K$ are

$$\begin{aligned} \pi(\boldsymbol{\theta} \mid \boldsymbol{\kappa}_K, \mathbf{y}) &= \frac{L_K(\boldsymbol{\theta}, \boldsymbol{\kappa}_K|\mathbf{y})\pi(\boldsymbol{\theta})}{\int_\Theta L_K(\boldsymbol{\theta}, \boldsymbol{\kappa}_K|\mathbf{y})\pi(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}} \\ \pi(\boldsymbol{\kappa}_K \mid \boldsymbol{\theta}, \mathbf{y}) &= \frac{L_K(\boldsymbol{\theta}, \boldsymbol{\kappa}_K|\mathbf{y})}{\sum_{J=0}^{n-1} \sum_{\boldsymbol{\kappa}_J \in \mathscr{K}_J} L_J(\boldsymbol{\theta}, \boldsymbol{\kappa}_J|\mathbf{y})}. \end{aligned}$$

We add the following uniform, normal, and inverted gamma (IG) priors:

$$\begin{aligned} \mu &\sim \pi(\mu) \propto c, \quad -\infty < \mu < \infty \\ \beta|\mu, \sigma^2 &\sim N\left(\mathbf{b}, d\sigma^2\mathbf{I}\right) \\ \sigma^2 &\sim \mathrm{IG}(a_1, b_1), \end{aligned}$$

where $\mathbf{b} = (\mu, 0, \ldots, 0)'$, $a_1 > 0$ is the shape parameter, $b_1 > 0$ is the scale parameter, and $c, d > 0$ are constants. Here, $\boldsymbol{\theta} = \left(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\gamma}, \sigma^2\right)'$. For the priors on K and $\boldsymbol{\kappa}_K$, we add:

$$\pi(K) = \frac{1}{n} \text{ and } \pi(\boldsymbol{\kappa}_K | K) \propto w^{K+1} \prod_{j=0}^{n-1} \Gamma(n_j),$$

where w is a weight. Alternatively for the prior on K, we can consider a truncated Poisson distribution with hierarchical parameter ν that can be specified with a small number for higher probability on smaller value of K. As stated in DiMatteo *et al.* (2001), posteriors appear not to be sensitive to the precise specification of the prior on K.

We now outline a Gibbs sampler that will generate from the conditionals by generating a length n vector $\boldsymbol{\kappa}_K$ and recovering the knot size through marginalization.

For $t = 1, \ldots T$, at iteration t

1. Starting from $\left(\boldsymbol{\theta}^{(t)}, \boldsymbol{\kappa}_K^{(t)}\right)$, with $\boldsymbol{\kappa}_K^{(t)} = (n_0^{(t)}, \ldots, n_K^{(t)})$

$$\text{draw}: \quad \boldsymbol{\theta}^{(t+1)} \sim \pi\left(\boldsymbol{\theta} \mid \boldsymbol{\kappa}_K^{(t)}, \mathbf{y}\right), \tag{6}$$

2. Given $\boldsymbol{\theta}^{(t+1)}$,

$$\begin{aligned} \text{draw}: \quad \boldsymbol{q}^{(t+1)} &= (q_0^{(t+1)}, \ldots, q_{n-1}^{(t+1)}) \\ &\sim \text{Dirichlet}\left(\alpha_0^{(t)}, \ldots, \alpha_K^{(t)}, \alpha_{K+1}, \ldots, \alpha_{n-1}\right) \\ \text{where} \quad \boldsymbol{\kappa}_{K'}^{(t+1)} &\propto w^{K'+1} m(\mathbf{y}|\boldsymbol{\kappa}'_{K'}) \binom{n}{n'_0 \ \cdots \ n'_{n-1}} \prod_{j=0}^{n-1} [q_j^{(t+1)}]^{n'_j}, \end{aligned} \tag{7}$$

and $n'_0 + \cdots + n'_{n-1} = n$ with $K'+1$ of the $n'_j > 0$, and $m(\cdot)$ is the marginal distribution. We note that the length of $\kappa^{(t+1)}$ is $K' = K^{(t+1)}$.

Sampling of the model parameters $\boldsymbol{\theta}$ in (6) is straightforward (Appendix 6), so we will concentrate on the sampling of $\boldsymbol{\kappa}_K$ and $\boldsymbol{q}$, a vector of probabilities that decides the number of knots and the order of positions.

The marginal distribution of K knots in positions $\boldsymbol{\kappa}_K$ will have the form of

$$\begin{aligned} m(\mathbf{y}|M_K) \quad \propto \quad & \left| \boldsymbol{X}' \left\{ \mathbf{I} - \mathbf{S} \left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \right)^{-1} \mathbf{S}' \right\} \boldsymbol{X} \right|^{-1/2} \left| \left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \right) \right|^{-1/2} \\ & \times \left[b_1 + \frac{1}{2} \mathbf{y}' \left\{ \boldsymbol{\Sigma}_*^{-1} - \boldsymbol{\Sigma}_*^{-1} \boldsymbol{X} \left(\boldsymbol{X}' \boldsymbol{\Sigma}_*^{-1} \boldsymbol{X} \right)^{-1} \boldsymbol{X}' \boldsymbol{\Sigma}_*^{-1} \right\} \mathbf{y} \right]^{-\left(\frac{n-p}{2} + a_1\right)} \end{aligned} \tag{8}$$

where M_K is a model with K knots in positions $\boldsymbol{\kappa}_K$ and $\boldsymbol{\Sigma}_*^{-1} = \mathbf{I} - \mathbf{S}\left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right)^{-1}\mathbf{S}'$. Thus, the Bayes Factor of a model with K knots in positions $\boldsymbol{\kappa}_K$ and a model with no knots can be written as

$$\begin{aligned} B_{K0}(\mathbf{y}) &= \frac{m(\mathbf{y}|M_K)}{m(\mathbf{y}|M_0)} \\ &\propto \frac{\left|\boldsymbol{X}'\left\{\mathbf{I} - \mathbf{S}\left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right)^{-1}\mathbf{S}'\right\}\boldsymbol{X}\right|^{-1/2}\left|\left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right)\right|^{-1/2}}{|\boldsymbol{X}'\boldsymbol{X}|^{-1/2}} \\ &\quad \times \left[\frac{b_1 + \frac{1}{2}\mathbf{y}'\left\{\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\right\}\mathbf{y}}{b_1 + \frac{1}{2}\mathbf{y}'\left\{\mathbf{I} - \boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\right\}\mathbf{y}}\right]^{-\left(\frac{n-p}{2}+a_1\right)} \end{aligned} \tag{9}$$

Therefore, the posterior distribution of K and $\boldsymbol{\kappa}_K$ is

$$\begin{aligned} \pi(\boldsymbol{\kappa}_K|\mathbf{y}, K) &= \frac{\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}B_{K0}(\mathbf{y})}{\sum_{\boldsymbol{\kappa}_K \in \mathscr{K}_K}\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}B_{K0}(\mathbf{y})} \\ &= \frac{\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}m(\mathbf{y}|M_K)}{\sum_{\boldsymbol{\kappa}_K \in \mathscr{K}_K}\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}m(\mathbf{y}|M_K)} \end{aligned} \tag{10}$$

and

$$\pi(K|\mathbf{y}) = \frac{w^{K+1}\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}m(\mathbf{y}|M_K)}{\sum_{K=0}^{n-1}w^{K+1}\sum_{\boldsymbol{\kappa}_K \in \mathscr{K}_K}\prod_{j=0}^{K}\Gamma(n_j)|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2}m(\mathbf{y}|M_K)}. \tag{11}$$

Here, $|\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1} = \prod_{k=0}^{K}\left(\frac{1}{n_k}\right)$.

The transition kernel of Markov chain in (7) is

$$k\left((\boldsymbol{\theta}, \boldsymbol{\kappa}_K), (\boldsymbol{\theta}', \boldsymbol{\kappa}'_{K'})\right) = \pi(\boldsymbol{\theta}' \mid \boldsymbol{\kappa}_K, \mathbf{y})\int_Q P(\boldsymbol{\kappa}'_{K'} \mid \mathbf{q}, \boldsymbol{\theta}')f(\mathbf{q} \mid \boldsymbol{\kappa}_K)\mathrm{d}\mathbf{q}, \tag{12}$$

with

$$\begin{aligned} P(\boldsymbol{\kappa}_K|q, \theta) &= \frac{w^{K+1}B_{K0}(\mathbf{y})\binom{n}{n_0 \cdots n_{n-1}}\prod_{j=0}^{n-1}q_j^{n_j}}{\sum_{K=0}^{n-1}w^{K+1}\sum_{\boldsymbol{\kappa}_K \in \mathscr{K}_K}B_{K0}(\mathbf{y})\binom{n}{n_1 \cdots n_{n-1}}\prod_{j=0}^{n-1}q_j^{n_j}} \\ &\text{and} \\ f(\boldsymbol{q}|\boldsymbol{\kappa}_K) &= \frac{\Gamma(\sum_{j=0}^{n-1}\alpha_j)}{\prod_{j=0}^{n-1}\Gamma(\alpha_j)}\prod_{j=0}^{n-1}q_j^{\alpha_j - 1}. \end{aligned}$$

If we take $\alpha_j = n_j + 1$ for all $j = 0, \ldots, n-1$ with $n_{K+1} = \cdots = n_{n-1} = 0$,

$$\frac{\Gamma(\sum_{j=0}^{n-1}\alpha_j)}{\prod_{j=0}^{n-1}\Gamma(\alpha_j)}\prod_{j=0}^{n-1}q_j^{\alpha_j - 1} = \frac{\Gamma(2n)}{n!}\binom{n}{n_0 \cdots n_K}\prod_{j=0}^{K}q_j^{n_j}.$$

Thus, with this choice, the transition kernel has $\pi(\boldsymbol{\theta}, \boldsymbol{\kappa}_K|\mathbf{y})$ as its stationary distribution. The verification of the stationary distribution is the same as in Kyung *et al.* (2010) with BF instead of the likelihood function.

For the estimation of w, we consider a gamma prior. Escobar and West (1995) use a gamma prior on w and update Gibbs with an auxiliary variable method for a Bayesian density estimation. Also, Blei and Jordan (2006) discussed that in the stick-breaking representation, the gamma distribution is conjugate to the stick lengths. In the stick-breaking representation, the mixing proportions are given by successively breaking a unit length "stick" into an infinite number of pieces. The size of each successive piece (stick length), proportional to the rest of the stick, is given by an independent draw from a Beta$(1, w)$ distribution with our notation. This means that the gamma distribution is a good candidate for a prior on w. For all cases, w is a parameter that controls the number of mixtures or the smoothness of the model in Dirichlet. However, in our approach, the number of mixtures or the smoothness is controlled by $\boldsymbol{\kappa}_K$. Thus w is then insensitive to choosing the number of knots and their positions.

3 Generating the Positions of Knots

In this section we show how to generate the positions of knots according to (7). Then we examine convergence rates, and establish that our sampler is an improvement, in terms of operator norm and efficiency, over commonly used one knot add/delete algorithms.

3.1 Dirichlet Distribution

To generate the positions of knots based on (7), we use the Dirichlet distribution with parameters $\boldsymbol{\alpha}$. The Dirichlet distribution is the multivariate generalization of the beta distribution with density given by

$$f(\mathbf{q}) = \frac{\Gamma(\sum_{j=0}^{n-1} \alpha_j)}{\prod_{j=0}^{n-1} \Gamma(\alpha_j)} \prod_{j=0}^{n-1} q_j^{\alpha_j - 1} \qquad \text{with } \alpha_i > 0 \text{ for all } j, \text{ and } \sum_{j=0}^{n-1} q_j = 1.$$

First consider the simple case, $(p, 1-p)$, with $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)'$. Then, $(p, 1-p)$ has a beta distribution with $\alpha_1 > 0$ and $\alpha_2 > 0$. For the beta distribution with $\alpha_1 < 1$ and $\alpha_2 < 1$, if $p \to 0$, then $1 - p \to 1$ with high probability and vice versa. However, $p \approx 1 - p$ will occur with small probability. This means that one of $(p, 1-p)$ will be much larger than the other. For the case with $\alpha_1 \geq 1$ and $\alpha_2 \geq 1$, the probability of $p \approx 1 - p$ is high.

Thus, p and $1-p$ will tend to be close to each other.

Now we consider the more general case with $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_m)'$. Due to the restriction that $\sum_{j=1}^{m} q_j = 1$, if $\alpha_j < 1$ for all $j = 1, \ldots, m$, some q_j's will have larger values than the others. This means that in the Gibbs sampler (7), if $\alpha_j < 1$ for all $j = 1, \ldots, n$, the number of knots tends to be smaller. However if we consider $\alpha_j \geq 1$, the number of knots tends to be close to $n/2$.

The general goal of the penalized linear spline models is good prediction with a smaller number of basis function and knots than sample size, $K < n$. If we believe that the original number of knots should be small, instead of $K \approx n/2$, then we update the Gibbs sampler for the number of knots and the position of these knots with

$$\alpha_j^{(t)} = \begin{cases} \frac{n_j^{(t)}+1}{n} & \text{for } j = 0, \ldots, K \\ \frac{1}{n} & \text{for } j = K+1, \ldots, n-1. \end{cases} \tag{13}$$

In the following section, we discuss how the positions of knots with the number of knots are generated based on a Dirichlet prior.

3.2 Generating $\boldsymbol{\kappa}_K$

We follow the methods described in Kyung *et al.* (2010) for the number and the positions of knots, updating the positions of knots with the marginalization of a multinomial-Dirichlet distribution using a Metropolis-Hastings algorithm. In other words, based on the value of the q_j in (7), we generate a length n vector of indices of subclusters from the multinomial, grouping the objects with the same indices, then remove the subclusters with no elements. For example, let a candidate vector of indices for subclusters from the multinomial be $(1, 3, 1, 4, 5, 4, 1, 5, 5, 3)'$ with $n = 10$ and W in increasing order. Thus we sort the vector in increasing order, $(1, 1, 1, 3, 3, 4, 4, 5, 5, 5)$. Here we remove zero counts of indices, $2, 6, \ldots, 10$, and sum over the frequencies of indices. Then, $n_0 = 3$, $n_1 = 2$, $n_2 = 2$, $n_3 = 3$ and $n_0 + \cdots + n_3 = 10$. Therefore, the updated $\boldsymbol{\kappa}'_K$ is (W_3, W_5, W_7) with 3 knots ($K' = 3$). The details about the marginalization of multinomial-Dirichlet distributions are provided in the Appendix of Kyung *et al.* (2010).

3.3 Convergence Properties

Given $\boldsymbol{\kappa}_K$, the sampling of the model parameters from $\pi(\theta|\boldsymbol{\kappa}_K\, \mathbf{y})$ is straightforward. Thus, in investigating convergence we only need to be concerned with the convergence

of the Markov chain on the number and position of knots.

If we ignore the model parameters for now, then we are concerned only with convergence of the chain to the stationary distribution, that is

$$\pi(\boldsymbol{\kappa}_K) = \pi(n_0, \ldots, n_K) = \frac{\Gamma(w)}{\Gamma(n+w)} w^{K+1} \prod_{j=0}^{K} \Gamma(n_j),$$

with the expected value of K,

$$E(K) = \sum_{i=1}^{n} \frac{w}{w+i-1}. \tag{14}$$

The full conditionals, ignoring the model parameters, are given by

$$P(a = j | n_0, \ldots, n_K) = \begin{cases} \frac{n_j}{n-1+w} & \text{for } j = 0, \ldots, K \\ \\ \frac{w}{n-1+w} & \text{for } j = K+1. \end{cases}, \tag{15}$$

where a is an index. With a similar argument, the full conditionals from the chain in (7) are

$$P(a = j | n_0, \ldots, n_K) \propto \begin{cases} \frac{n_j}{n-1+w} \frac{q_{j+1}}{\alpha_{j+1}} & \text{for } j = 0, \ldots, K \\ \\ \frac{w}{n-1+w} \frac{q_{j+1}}{\alpha_{j+1}} & \text{for } j = K+1. \end{cases} \tag{16}$$

Notice that for $q_j = \alpha_j$ $(j = 0, \ldots, n-1)$ (the normalization is not important), we see that the one-knot add Gibbs sampler (15) is the same as (16). Based on Hobert and Marchev (2008), it can be shown that for $\alpha_j = n_j + 1$ for $j = 0, \ldots, K$ and $\alpha_j = 1$ for $j = K+1, \ldots, n-1$, the kernel of (16) dominates the kernel of (15) with smaller variance for any square-integrable function with any $m > 0$. Details about the derivation of full conditionals and the dominated convergency are in Kyung *et al.* (2010).

However, if α_j's have the form in (13), we are not able to compare the transition kernel of (16) to the kernel of (15), because the transition kernel of our chain has a different stationary distribution. If we consider more details about the full conditionals in (16), we know from the properties of the Dirichlet distribution that

$$E(q_j) = \frac{\alpha_j}{\sum_{j=0}^{n-1} \alpha_j} \qquad \text{for } j = 0, \ldots, n-1.$$

Also, we know that q_j's determine K and $\boldsymbol{\kappa}_K$. Thus, the α_j's have an important role in the Markov chain to update the number of knots, K and the positions of knots, $\boldsymbol{\kappa}_K$, similar to m in (14).

Halpern (1973) proved that when the locations of all possible knots are assumed to be known but the subset of these which are the actual knots is unknown, the posterior distribution of the Bayesian spline regression is proper for natural conjugate and vague priors. The subset of all possible knots, that is, the set of actual knots, forms a model. Thus, with conjugate priors of the form of a marginal distribution on the index of the model, or with non-informative priors which are a function of number of elements in each subset of knots, the posterior distribution of the Bayesian spline is proper. For the location of the knot, a prior probability based on a discrete distribution is assigned to each subset of the possible locations, and the posterior probability has been calculated to choose the model with highest posterior probability. Also, the optimal predictor based on the marginal distribution for the model and the mean vector for the coefficients is derived for a loss function which is a generalization of that appearing in Lindley (1968). This loss function is squared error loss plus the amount of shrinkage which controls the number of knots.

Our model setting fits Halpern's Bayesian spline setting with conjugate priors except it has one more step to generate the number and positions of knots. A uniform distribution is assigned as a prior and our chain is updated to a new number and position of knots with higher posterior probability. Specifically when we generate the number and positions of knots, we start from n knots then marginalize over the empty sets; this implies conditionally consistent multivariate proper normal priors on coefficients $\boldsymbol{\gamma}$. We can then combine the parametric part $\boldsymbol{X\beta}$ to a component of the nonparametric part. Therefore, from Lemma 2 of Halpern (1973), any sample from the posterior distribution of our proposed model is proper and the set of multivariate normal posteriors on coefficients will be conditionally consistent with the chosen number (K) and positions ($\boldsymbol{\kappa}_K$) of knots. Also, with non-empty intervals between knots in our setting, from Theorem 1 of Halpern (1973), the posterior mean of coefficients $\boldsymbol{\gamma}$ in our model is the optimal predictor for the expected loss function which is an addition of the expected squared error loss and the risk of wrong number (K) and position ($\boldsymbol{\kappa}_K$) of knots.

4 Simulations and Data Analysis

To illustrate the proposed nonparametric approach and required Gibbs sampler, we present a simulation study and two data analyses. The parameter w is known to be insensitive to the choice of data, thus we fix $w = 1$ because the number of knots and the position of knots will be controlled by $\boldsymbol{\alpha}$ in the Dirichlet distribution. We compare small

and large parameter values over $\boldsymbol{\alpha}$ in the Dirichlet distribution. In these applications, a Poisson distribution with mean ν as a prior on the number of knots K is tested. As discussed in Section 2.1, the posteriors appear insensitive to the value of λ, thus we fix $\lambda = 3$ for convenience.

We also compare our proposed Gibbs sampler to recently proposed nonstationary methodologies that couple stationary Gaussian processes (GP) with treed partitioning done in Gramacy and Lee (2008). GP regressions accommodate prior knowledge in the form of covariance functions. The covariance term is considered with the correlation function of a neighbor. Tree GP regression uses treed partitioning through RJMCMC and in each partition (branch of the tree), independent local GP regressions are applied. The `tgp` package for `R` is developed for Bayesian nonstationary, semiparametric nonlinear regression and design by treed Gaussian processes with jumps to the limiting linear model by Gramacy (2007). GP models are well known for effectively fitting arbitrary functions or surfaces. The GP sampling proposed by Gramacy and Lee (2008) commences with RJMCMC which allows a simultaneous fit of the tree and the GPs at its leafs. In this paper, we consider GP regression and tree GP regression models to compare to our proposed semiparametric models.

4.1 Simulations

We begin with the simple smooth test functions:

1. $f(x) = x + 2\exp\left(-16x^2\right), \quad x \in [-2, 2]$

   and

2. $f(x) = \sin(x) + 2\exp\left(-30x^2\right), \quad x \in [-2, 2].$

Simulated data are drawn from the rescaled function with support in the unit interval. The first function comes from Denison *et al.* (1998). In the original setting, it is evaluated at 200 regularly spaced points with normally distributed noise having mean 0 and standard deviation $\sigma = 0.4$ but in our simulation, we evaluated at 200 regularly spaced points with $\sigma = 0.3$. The second function comes from DiMatteo *et al.* (2001). Originally, it has been evaluated at 101 regularly spaced points with $\sigma = 0.3$ in DiMatteo *et al.* (2001) and we evaluate at 101 regularly spaced points with $\sigma = 0.3$.

Our implementation focuses on Gibbs sampling with the marginal distribution, $m(\mathbf{y}|M_K)$. We computed the mean squared error for our Bayes method and GP models

	Function 1.		Function 2.	
	AMSE	Number of knots	AMSE	Number of knots
Small valued	0.008 (0.002)	13.33 (0.916)	0.019 (0.007)	8.57 (0.365)
Large valued	0.026 (0.007)	83.70 (3.092)	0.025 (0.008)	39.82 (2.103)
GP regression	0.008 (0.003)	–	0.017 (0.005)	–
tree GP	0.008 (0.003)	2	0.017 (0.006)	2

Table 1: Average mean squared errors with standard errors on 50 samples. Small and large valued parameters of Dirichlet distribution in MCMC, and GP and tree GP regressions are compared. (Number of knots for tree GP regressions is the number of change points to fit independent GP models.)

as:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}\left\{\hat{f}(x_i) - f(x_i)\right\}^2,$$

where $\hat{f}(x_i)$ is an estimate and $f(x_i)$ is the true function. The Bayesian estimates of $E[f(x)|y]$ are found from our Markov chain Monte Carlo process with runs of 5000 following burn-ins of 5000.

Here, with MCMC, we compare:

- a large valued parameter in the Dirichlet distribution

$$\alpha_j = \begin{cases} n_j + 1 & \text{for } j = 0, \ldots, K \\ 1 & \text{for } j = K+1, \ldots, n-1 \end{cases}$$

- versus a small valued parameter in the Dirichlet distribution

$$\alpha_j = \begin{cases} \frac{n_j+1}{n} & \text{for } j = 0, \ldots, K \\ \frac{1}{n} & \text{for } j = K+1, \ldots, n-1. \end{cases}$$

The average mean squared error (AMSE) with standard errors based on 50 samples of data, and the average number of knots are reported in Table 1. We observe that the semiparametric model as a linear mixed model fits well for all examples shown by small mean squared error (MSE). For generated data from Function 1, the MSE and number of knots from a small valued parameter in the Dirichlet distribution are smaller than these from a large valued parameter. For this function, the number of knots has a big

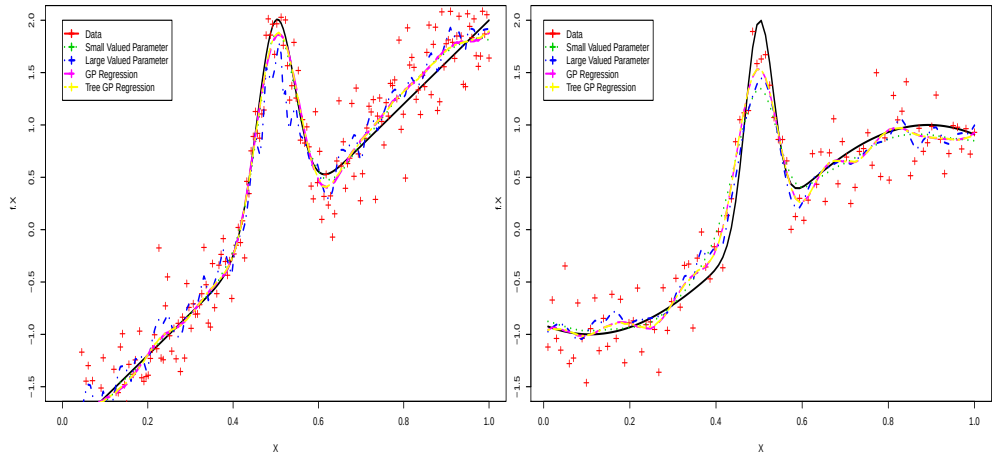

Figure 1: Left panel is the rescaled smooth Function 1 and right panel is the rescaled smooth Function 2.

effect on the fit with a preference on a small number of knots. However, for Function 2, the number of knots does not have a big effect on the fit.

Compared to GP and tree GP regressions, our semiparametric model with a small valued parameter in the Dirichlet distribution tends to recover precision on GP regressions. AMSEs are not different numerically for Function 1 and Function 2. Tree GP has been fitted with three independent local GP regressions. From Figure 1, we clearly observe three partitions if we consider local regressions for Function 1 and Function 2. Graphically, in Figure 1, we also observe that Function 1 has a small number of change points, but Function 2 shows a large number of change points. Also, the estimated curve with small valued parameter and the estimated curves of GP and tree GP regressions are close to the true function , but the estimated function with large valued parameter is close to the data points.

We now consider one-dimensional simulated data which is partly a mixture of sines and cosines, and partly linear that is considered in Gramacy (2007):

$$f(x) = \begin{cases} \sin\left(\frac{\pi x}{5}\right) + \frac{1}{5}\cos\left(\frac{4\pi x}{5}\right) & x < 9.6 \\ \frac{x}{10} - 1 & \text{otherwise.} \end{cases}$$

From the true model itself, we expect that the independent local regression in each partition, tree GP regression, will show superiority among others. However, in Figure

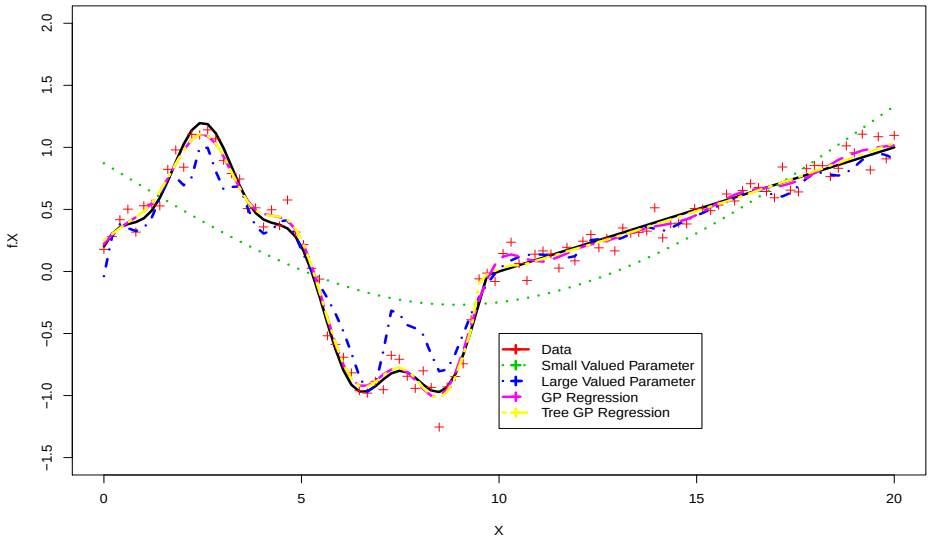

Figure 2: Mixture of sines and cosines and partly linear function.

	Small valued	Large valued	GP regression	tree GP
AMSE	0.054 (0.007)	0.014 (0.004)	0.002 (0.001)	0.002 (0.001)
Number of knots	7.42 (3.761)	38.65 (2.305)	–	1

Table 2: Average mean squared errors with standard errors on 50 samples. Small and large valued parameters of Dirichlet distribution in MCMC, and GP and tree GP regressions are compared. (Number of knots for tree GP regressions is the number of change points to fit independent GP models.)

2, we observe that our semiparametric regression with a large valued parameter in the Dirichlet distribution also tends to capture wiggly curves well. The average mean squared error (AMSE) with standard errors based on 50 samples of data, and the average number of knots are reported in Table 2. The AMSEs of GP regressions are smaller than AMSEs of our models. Tree GP regressions fit two independent GP regressions on each partition and the partition point is the given degenerate point $x < 9.6$. In our model, the number and position of knots depend on the subsample sizes instead of the correlation with neighbors. Thus, our method might not recover the true model as much as tree GP regression, but without considering all possible partitions, the semiparametric regression tends to fit the appropriate smoothed curve on data. In other words, the proposed semiparametric model is simpler and faster, while capturing

the main features.

4.2 Data Analysis: Nitrogen Oxides in Engine Exhaust

Originally, Brinkman (1981) conducted an experiment of a single-cylinder engine with ethanol or indolene to see how the nitrogen oxides (NO_x) concentration in the exhaust depended on the compression ratio (C) and the equivalence ratio (E). There were 88 runs with ethanol; for these runs, E varied from .535 to 1.232, C took one of five values ranging from 7.5 to 18, and the values of E and C were nearly uncorrelated. There were 22 runs with indolene; for these runs, C took just one value, 7.5, and E ranged from 0.665 to 1.224. In this example, we only consider the data with ethanol. This data has been analyzed with various methods, but Cleveland and Devlin (1988), Breiman (1991) and Gu (2002) used smoothing methods. Cleveland and Devlin (1988) applied the locally quadratic smoother based on an adaptation of Mallows's C_p, the M plot and argued that an additive fit of E and C is inappropriate because of a substantial interaction. Breiman (1991) used a product method of multivariate functions for estimating an underlying smooth function of noisy data by a sum of products of the univariate functions. For the NO_x data, it has been argued that 2 knots are enough to estimate the function of E and there is a non-removable interaction in these data. In these analysis, an $NO_x^{1/3}$ transformation has been used, but Gu (2002) made a log-transformation because the NO_x concentrations are positive with some near-zero readings. As it has been argued, the effect of equivalence ratio was dominant, but the compression ratio had little impact. Gu (2002) used a cubic spline fit with estimating smoothing parameter λ estimated after seeing a rough cross-validation fit.

We use a log-transformation on NO_x and we compare our methods to the cubic spline fit of Gu (2002), GP regression and tree GP regression based on the residual sum of squares (RSS)

$$\text{RSS} = \frac{1}{n}\sum_{i=1}^{n}\{\hat{y}_i - y_i\}^2,$$

where $\hat{y}_i$ is an estimate from the cubic spline and y_i is the observation. Two models have been applied to this data previously:

Model 1. $\log_{10}(NO_x) = \beta_0 + \beta_1 E + g(E) + \epsilon$

Model 2. $\log_{10}(NO_x) = \beta_0 + \beta_1 E + \beta_2 C + \beta_3 E \times C + g(E) + \epsilon.$

The compression ratio, C, has five distinct values; it could have been treated as an

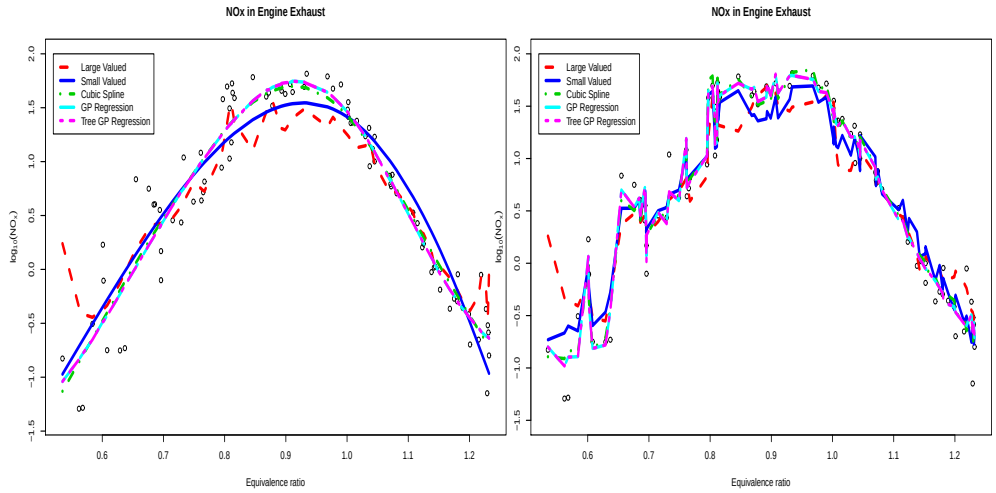

Figure 3: Left panel is a scatterplot of NO_x with estimated curves of Model 1 and right panel is with estimated curves of Model 2.

ordinal discrete variable and interaction has been considered. The numerical results are displayed in Table 3 and the graphical functional fits are in Figure 3.

From Table 3, we observe that the number of knots which has been detected by the proposed methods with small valued parameters in Dirichlet distribution is the same as Breiman (1991). With C and interaction between E and C (Model 2), the RSS is smaller than the RSS of Model 1, because as it has been argued, there exists a non-removable interaction in these data. The curve estimation with cubic splines tries to smooth out the dataset by reducing a function of RSS (cross-validation); thus the cubic spline curve fit has smaller RSSs. GP regression behaves similar to cubic spline models, and tree GP has the smallest RSS among the techniques. From Figure 3, we observe that an estimated curve based on small valued parameter in the Dirichlet distribution is showing smaller estimated valued in the highest peaks of data, but bigger estimated valued in the tails of the data. From the left panel in Figure 3, for Model 1, we observe that compared to the cubic spline curve, GP regression and tree GP regression, the small valued parameter curve has a lower peak with thick tails. It might be the reason why RSS of the small valued parameter curve is larger than the RSS of the cubic spline curve. As discussed with the simulations, the large valued curve is more sensitive to the data points, but less smooth. The right panel in Figure 3 is based on the Model 2. C is an ordinal discrete variable, thus the estimated curves are not as smooth as in the left panel. However, if we consider C and interaction in the model, the estimated curves are

	Model 1		Model 2	
	RSS	Number of knots	RSS	Number of knots
Cubic Spline	0.067	–	0.027	–
GP regression	0.066	–	0.023	–
tree GP	0.061	3	0.013	3
Small valued	0.091	2	0.050	2
Large valued	0.119	34	0.088	31

Table 3: Residual sum of squares and number of knots from NO_x concentration data. For Model 1 and Model 2, small and large valued parameters of the Dirichlet distribution in MCMC are compared. Also, cubic spline fit of Gu (2002), GP regression and tree GP regression are compared. (Number of knots for tree GP regressions is the number of change points to fit independent GP models.)

more sensitive to the data points compared to the model with E only. In this data set, there are more data points in the right hand side tail; thus a small valued parameter curve tries to give more weight onto the right tail with somewhat less weight on higher peaks. The cubic spline curve is attempting to smooth out the data points based on the point value and GP regressions are trying to smooth out the data points considering the correlation with neighbor points. However, the proposed semiparametric regression is also attempting to smooth out the data points with the basis functions which are determined based on the subsample sizes $n_0, n_1, \ldots, n_K$. This technique is a somewhat different setup for positions of knots compared to others and it might be the reason for the above results.

5 Crime Reducing Benefits In Treating Drug Involved Offenders

Returning to our primary data interest, we look at the synthetic dataset problem described in Section 1. Empirical investigation of "Going to Scale" in drug interventions in the United States, 1990, 2003 (Bhaty and Roman, 2009) has been done to determine the size of the drug-involved offender population that could be served effectively and efficiently by partnerships between the courts and treatment regimes. From micro-level data of three nationally representative sources, a dataset of 40,320 cases which is de-

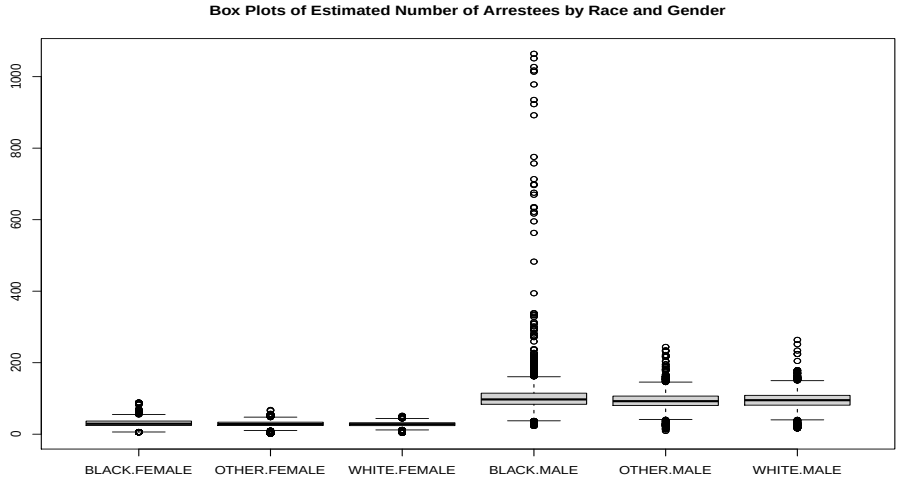

Figure 4: Box Plots of Estimated Number of Arrestees by Race and Gender

fined using population profiles was constructed. The principal investigators combined information from the National Survey on Drug Use and Health, 2003 (USDHHS 2006) and the Arrestee Drug Abuse Monitoring (ADAM) Program in the United States, 2003 (USDJNIJ 2004) to estimate the likelihood of drug addiction or dependence problems and develop nationally representative prevalence estimates. They used information in the Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States] (USDHHS 2010) to compute expected crime reducing benefits of treating various types of drug involved offenders under four different treatment modalities.

In this dataset, age, race, gender, offense, history of violence, history of treatment, co-occurring alcohol problem, criminal justice system status, geographic location, arrest history, and a total of 134 prevalence and treatment effect estimates and variances are all included. Moreover, the principal investigators obtained estimates of crime reducing benefits for all crimes as well as select sub-types. The four different treatment modalities considered from DATOS were Long-Term Residential Treatment (Modality 1), Short-Term Inpatient Treatment (Modality 2), Outpatient Methadone Treatment (Modality 3) and Outpatient Drug-Free Treatment (Modality 4).

Among these 145 variables, in this example, the estimated number of arrestees (`ESIZE`) by weighting on the prevalence in the population of interest is considered as the dependent variable Y. In the original data, ADAM provides information on the number

of times individuals were arrested in the year prior to the current booking. Using the empirical similarity between synthetic profiles and ADAM sample members, Bhaty and Roman computed the expected number of arrests for particular profiles. Thus, by rescaling based on the prevalence in the population of interest, `ESIZE` has been created. For independent variables, we consider age (`AGE`), current offense status (`OFFENSE`) - Violent, Drug, Property, or Other, history of substance abuse or dependent treatment (`THIST`) - Yes or No, geographical location (`GEO`) - Rural, Urban and Suburban areas, number of prior arrests (`AHIST`) - 0, 2, 5, 10, or 20 , and eight abuse/dependence variables which are estimated based on the four treatment modalities;

$$\texttt{TEC}_{ij}, \text{ for } i = 1, \ldots, 4 \text{ and } j = 0, 1.$$

Here subscript i is an index of Modality, and $j = 0$ is for ABUSE and $j = 1$ is for DEPENDENCE.

This is a large dataset ($n = 40,320$), so all variables are trivially significant at $\alpha = 0.05$ with a standard GLM. Figure 4 shows box plots of `ESIZE` by race and gender. For the BLACK MALE case, it is showing skewness to the right with large values of `ESIZE`. Compared to distributions of other race and gender combinations, the distribution of the black male case shows broadly skewed large variance. For these data, a regular GLM might not be enough to capture the large variance. Thus, instead of considering the whole dataset, in this paper, we used part of the dataset which is for *black males with violence and alcohol history under criminal justice status*. There are $n = 840$ cases meeting this new criterion.

These data are used to detect significant factors with controlling hidden structures and latent information. GP regression and tree GP regression implement Bayesian regression models of varying complexity and allow for the explicit estimation of predictive uncertainty when considering neighbor points. Thus GP regression and tree GP regression are not meaningful for the purpose of the data analysis in this section.

For this smaller dataset, we consider a linear mixed effect model (LMEM) and semiparametric model. For LMEM, we use $\texttt{TEC}_{10}$ as a random effect term. We observe from LMEM that most demographical variables are significant at $\alpha = 0.05$, but 8 prevalence and treatment effect estimated variables are not significant. Modality 1 is strongest abuse/dependence treatment; thus we might expect that it decreases the number of drug-involved offenders. However, if you fit LMEM to the dataset of *black males with violence and alcohol history under criminal justice status*, Long-Term Residential Treatment has no statistically reliable effect with `ESIZE` at the 0.95 level. Also, other treat-

	RSS	Number of knots
LMEM	153.67	–
Natural Cubic Spline	267.53	–
Small valued	270.17	10
Large valued	270.78	228

Table 4: Residual sum of squares and number of knots from crime reducing benefits data. Small and large valued parameters of the Dirichlet distribution in MCMC are compared. Also, linear mixed effect model (LMEM) with random effect TEC_{10} and natural cubic spline fit on TEC_{10} are compared.

ment methods are not statistically reliable either. It might be due to large variances (various variabilities of estimation or scaling) for these 8 estimated variables. Thus, we consider a semiparametric model with natural cubic spline (NCS) on $\texttt{TEC}_{10}$. For the semiparametric fit, we consider the regular NCS method based on generalized cross validation and our proposed Gibbs sampling with large and small valued parameters in the Dirichlet distribution.

From Table 4, we observe that the RSS of LMEM is smaller than the RSSs of the regular NCS semiparametric model and our proposed models. Coefficients of LMEM are estimated based on the restricted maximum likelihood method to reduce variance in this dataset, thus we might expect the RSS of LMEM to be smallest compared to other models. For these data, the number of knots has little effect on the fit, because RSS with the small valued parameter of the Dirichlet distribution is close to the RSS with the large valued parameter. However, from Figure 5, we observe that our proposed semiparametric model with a Dirichlet prior distribution has uniformly smaller 95% highest posterior density (HPD) intervals than the standard NCS semiparametric model and LMEM. This variance reduction property in linear Dirichlet random effects models has been proved theoretically by Kyung *et al.* (2009). They showed that if the data vector does not consist of a contrast within observations between each position of knots, the mean of the posterior distribution of the variance from the Dirichlet random effect model is smaller than that of the regular normal random effects model. However, in most cases we might not be able to find such contrast. The left panel in Figure 5 shows 95% intervals for the demographical variables, `AGE`, `OFFENSE`, `THIST`, `GEO` and `AHIST`. HPD intervals of the proposed semiparametric model (small and large) are noticeably smaller than intervals from regular models (LMEM and NCS). The right panel

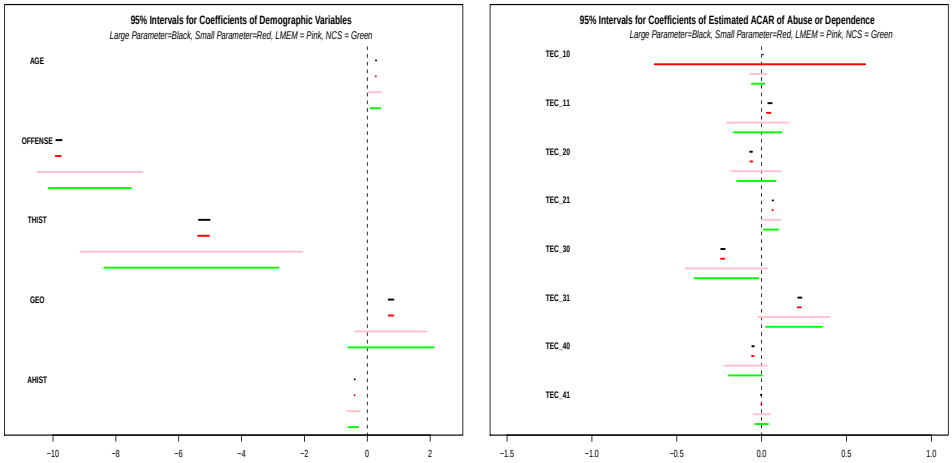

Figure 5: 95% highest posterior density (HPD) Intervals. The HPD intervals for LMEM and NCS models are given in Pink and Green, respectively, and intervals for the proposed semiparametric model with small valued and large valued parameter models are given in Red and Black, respectively.

in Figure 5 shows 95% intervals of eight estimated abuse/dependence variables based on the four different treatment modalities. The $\texttt{TEC}_{10}$ variable is used in the smoothing method, so the credible intervals for the coefficients include zero for all models. However, all other coefficients in our proposed semiparametric model are reliably different from zero. Therefore, our proposed method gets rid of unexpected and hidden variance and correlation under the data structure efficiently and provides a smoother curve fit with small variance compared to other regular methods.

From Figure 5, we observe that for black males with violence and alcohol history also under criminal justice status, if the current offense case is violence (`OFFENSE`=VIOL) and if drug abuse/dependent has been treated (`THIST`=YES), the number of arrestees (`ESIZE`) tends to decrease compared to drug offense (`OFFENSE`=DRUG) and no treatment history (`THIST`=NO). Also, as the number of previous arrests (`AHIST`) increases, the estimated number of following arrests tends to decrease. Considering four different drug abuse/dependent treatments, treatment for drug abuse in each of the domains (modalities) ($\texttt{TEC}_{20}$, $\texttt{TEC}_{30}$, $\texttt{TEC}_{40}$) substantially reduces the number of arrestees. Specifically, outpatient methadone treatment (Modality 3) is shown to be the most efficient method to reduce the number of arrests. However, for those at greatest risk of drug dependence ($\texttt{TEC}_{11}$, $\texttt{TEC}_{21}$, $\texttt{TEC}_{31}$, $\texttt{TEC}_{41}$), reverse treatment in each of the modalities tends

to create an increase in the number of arrests. This might be the reason that, compared to abuse, the drug dependence treatment is not intensive, and the standard categorization method of abuse or dependence is not restricted. Thus, by trying more intensive drug abuse/dependence care, drug-involved recidivism should be reduced. When we use a standard method to fit a nonlinear model, we are not able to detect this hidden information, but by using the proposed method, we reduce the variance in the data structures and find better fitting models with greater predictive qualities. Note that using the Dirichlet distribution to generate the positions of the knots gives us a powerful implementation of a Stein-rule like estimator.

6 Discussion

For synthetic data, in the aggregation process hidden structures and latent information are created in unexpected ways. What this suggests is that the usual set of parametric models will be inadequate for explaining the key underlying relationships in such data. Thus we developed a new Bayesian semiparametric regression model that balances user-defined restrictions against pure data information, and which is sufficiently flexible that it has the ability to capture unusual or hidden features of the data that would ordinarily be missed by conventional approaches. For the analysis of synthetic data of crime-reducing benefits in treating drug-involved offenders who are black males with violence and alcohol history under criminal justice status, when we used standard methods to fit a nonlinear model, we were not able to detect this hidden information. However by using our proposed method, we remove more variance in the data structure and we find a more smooth and informative model for better prediction.

To control the number of knots, we use the Dirichlet distribution with large valued parameter for a large number of knots and with small valued parameter for a small number of knots. For the function with a small number of change points, the number of knots has a big effect on the fit with a preference on a small number of knots; therefore a Dirichlet distribution with a small valued parameter for the number and positions of knots in MCMC performs better compared to that with a large valued parameter based on the mean squared error. From simulation, we observe that the estimated curve with a small valued parameter is close to the true function, but the estimated function with large valued parameter is close to the data points. Also, from data analysis, we could observe that the large valued curve is more sensitive to the data points, but less smooth. This may lead to the poor smooth for the large valued parameter seen in

Figure 2. Comparing to a full nonstationary model with coupled stationary GP with tree partitioning, a new Bayesian semiparametric regression is not much less effective for fitting arbitrary functions or surfaces and it is simpler and faster. One of the advantages of the proposed semiparametric model is that it removes more hidden variance in the data structure and also provides better estimates of model parameters for the smoother curve fit by giving us a powerful implementation of a Stein-rule like estimator.

If we believe that there are not a large number of knots, in other words the distribution of data is smooth and not degenerated, we consider a Dirichlet distribution with small valued parameter for the number and positions of knots using MCMC. However, if there exist many change points (partition points) in the data, the independent local regression in each partition will be the best curve fit method. A semiparametric regression spline with a Dirichlet distribution with large valued parameter for the number and positions of knots is not as effective as a semiparametric model with small valued parameter for the smoother curves. However, if the distribution of data is skewed or a mixture of curves with degenerated points, our proposed semiparametric regression with a Dirichlet distribution with large valued parameter tends to fit the smooth curve on data faster without considering all possible partitions. Also, if our intention for curve fitting is better prediction over the sample space, our proposed semiparametric models with a Dirichlet distribution prior will provide a more informative model with simpler and faster implementation, even with a Dirichlet distribution with large valued parameter. For prediction, the choice of valued parameters (large or small) in the Dirichlet prior should be considered differently depending on the data.

The convergence properties of the proposed Bayesian semiparametric regression are discussed based on the Halpern's Bayesian spline model setting with conjugate priors (Halpern 1973). For the number and positions of knots, a discrete distribution is assigned as a prior and our chain is updated to a new number and position of knots with higher posterior probability. This implies conditionally consistent multivariate proper normal priors on coefficients $\boldsymbol{\gamma}$ of the basis functions; then the posterior is proper and the set of multivariate normal posteriors on coefficients will be conditionally consistent. Also, with this setting, the posterior mean of coefficients $\boldsymbol{\gamma}$ in our model is the optimal predictor under the chosen number (K) and positions ($\boldsymbol{\kappa}_K$) of knots.

We have provided a new semiparametric regression approach but we can extend our method to more of a nonparametric regression strategy. We express a semiparametric model as an addition of parametric and nonparametric parts. If we consider the parametric part as a smoother, semiparametric regression is an additive model of non-

parametric regressions. With a linear mixed model representation, a semiparametric regression has the same form as a nonparametric regression, but with more parameters in the fixed effects. Thus, we can easily use the proposed sampler with fully nonparametric regression.

Generalization of the proposed semiparametric regression can be considered with regression in exponential families with multivariate nonparametric structure. However, these are open issues within the proposed semiparametric regression and we leave these further explorations as a part of our future research.

From synthetic data analysis, we observe that a linear mixed effects model and a regular semiparametric model with natural cubic spline are not enough to explain effects of treatments for the number of arrestees among black males with violence and alcohol history under criminal justice status. However, our proposed semiparametric models with a Dirichlet distribution prior remove the unexpected and hidden variabilities and correlations under the data structure efficiently and provide a smoother curve fit with small variance compared to other regular methods. Thus, with uniformly smaller 95% highest posterior density intervals for the demographical variables than the standard NCS semiparametric model and LMEM, our methods provide that by trying more intensive drug abuse/dependence care, drug-involved recidivism could be reduced. By using our described method, we could remove more variance in the data structure and we could find a smoother model for better prediction.

Appendix

1. Generating the Model Parameters

The joint posterior distribution can be written as

$$\begin{aligned}
&\pi\left(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\gamma}, \sigma^2, K, \boldsymbol{\kappa}_K | \mathbf{y}\right) \\
&\quad \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n+K+1+p}{2}+a_1+1} \exp\left(-\frac{b_1}{\sigma^2}\right) |\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}|^{-1/2} \\
&\quad \times \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}}\right)' {\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*}^{-1}\left(\boldsymbol{\gamma}-\tilde{\boldsymbol{\gamma}}\right)\right\} \exp\left\{-\frac{1}{2\sigma^2}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1}\left(\boldsymbol{\beta}-\tilde{\boldsymbol{\beta}}\right)\right\} \\
&\quad \times \exp\left\{-\frac{1}{2d\sigma^2}\left(\mathbf{b}-\tilde{\mathbf{b}}\right)' \boldsymbol{\Sigma}_{\mathbf{b}}^{-1}\left(\mathbf{b}-\tilde{\mathbf{b}}\right)\right\} \qquad (17) \\
&\quad \times \exp\left[-\frac{1}{2\sigma^2}\mathbf{y}'\left\{\boldsymbol{\Sigma}_*^{-1} - \boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\left(\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\right\}\mathbf{y}\right],
\end{aligned}$$

where

$$\begin{aligned}
{\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*}^{-1} &= \mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1} \qquad \tilde{\boldsymbol{\gamma}} = \left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right)^{-1}\mathbf{S}'\left(\mathbf{y} - \boldsymbol{X}\boldsymbol{\beta}\right) \\
\boldsymbol{\Sigma}_*^{-1} &= \mathbf{I} - \mathbf{S}\left(\mathbf{S}'\mathbf{S} + \boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\right)^{-1}\mathbf{S}' \\
\boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} &= \frac{1}{d}\mathbf{I} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}, \qquad \tilde{\boldsymbol{\beta}} = \left(\frac{1}{d}\mathbf{I} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1}\left(\frac{1}{d}\mathbf{b} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\mathbf{y}\right) \\
\boldsymbol{\Sigma}_{\mathbf{b}}^{-1} &= \mathbf{I} - \frac{1}{d}\left(\frac{1}{d}\mathbf{I} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1} \\
\tilde{\mathbf{b}} &= \left\{\mathbf{I} - \frac{1}{d}\left(\frac{1}{d}\mathbf{I} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1}\right\}^{-1}\left(\frac{1}{d}\mathbf{I} + \boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{\Sigma}_*^{-1}\mathbf{y}.
\end{aligned}$$

Then for fixed K and $\boldsymbol{\kappa}_K$, a Gibbs sampler of $(\boldsymbol{\beta}, \mathbf{b}, \boldsymbol{\gamma}, \sigma^2)$ is

$$\begin{aligned}
\boldsymbol{\gamma}|\boldsymbol{\beta}, \mathbf{b}, \sigma^2, K, \boldsymbol{\kappa}_K, \mathbf{y} &\sim N_{K+1}\left(\tilde{\boldsymbol{\gamma}}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^*\right) \\
\boldsymbol{\beta}|\boldsymbol{\gamma}, \mathbf{b}, \sigma^2, K, \boldsymbol{\kappa}_K, \mathbf{y} &\sim N_p\left(\tilde{\boldsymbol{\beta}}, \sigma^2\boldsymbol{\Sigma}_{\boldsymbol{\beta}}\right) \\
\mathbf{b}|\boldsymbol{\gamma}, \boldsymbol{\beta}, \sigma^2, K, \boldsymbol{\kappa}_K, \mathbf{y} &\sim N_p\left(\tilde{\mathbf{b}}, d\sigma^2\boldsymbol{\Sigma}_{\mathbf{b}}\right) \\
\sigma^2|\boldsymbol{\gamma}, \boldsymbol{\beta}, \mathbf{b}, K, \boldsymbol{\kappa}_K, \mathbf{y} &\sim \mathrm{IG}\left(\frac{n+K+p}{2} + a_1,\right. \\
&\qquad \left.\frac{1}{2}|\mathbf{y} - \boldsymbol{X}\beta - \mathbf{S}\boldsymbol{\gamma}|^2 + \frac{1}{2}\boldsymbol{\gamma}'\boldsymbol{\Sigma}_{\boldsymbol{\gamma}}^{-1}\boldsymbol{\gamma} + \frac{1}{2d}|\boldsymbol{\beta} - \mathbf{b}|^2 + b_1\right).
\end{aligned}$$

References

Aronszajn, N., 1950. "Theory of Reproducing Kernels." *Transactions of the American Mathematical Society* **68**, 337-404. 797

Bhati, A. S. and Roman J., 2009. Empirical Investigation of "Going to Scale" in Drug Interventions in the United States, 1990, 2003 [Computer file]. ICPSR26101-v1. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2009-08-26. doi:10.3886/ICPSR26101 793, 794, 795

Biller, C., 2000. "Adaptive Bayesian Regression Splines in Semiparametric Generalized Linear Models." *Journal of Computational and Graphical Statistics* **9**, 122-140. 798

Blei, D. M. and Jordan, M. I., 2006. "Variational Inference for Dirichlet Process Mixtures." *Bayesian Analysis* **1**, 121-144. 803

Breiman, L., 1991. "The $\prod$ Method for Estimating Multivariate Functions from Noisy Data." *Technometrics* **33**, 125-160. 811, 812

Brinkman, N. D., 1981. "Ethanol Fuel - A Single Cylinder Engine Study of Efficiency and Exhaust Emissions." *SAE Transactions* **90**, 1410-1424. 811

Carroll, R., 1982. "Adapting for Heteroscedasticity In Linear Models." *The Annals of Statistics* **10**, 1224-1233. 797

Claeskens, G., Krivobokova, T., and Opsomer, J. D., 2009. "Asymptotic Properties of Penalized Spline Estimators." *Biometrika* **96**, 529-544. 798

Cleveland, W. S. and Devlin, S. J., 1988. "Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting." *Journal of the American Statistical Association* **83**, 596-610. 811

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M., 1998. "Automatic Bayesian Curve Fitting." *Journal of the Royal Statistical Society. Series B* **60**, 333-350. 798, 807

DiMatteo, I., Genovese, C. R., and Kass, R. E., 2001. "Bayesian Curve-Fitting with Free-Knot Splines." *Biometrika* **88**, 1055-1071. 798, 801, 807

Eilers, P. H. C. and Marx, B. D., 1996. "Flexible Smoothing with B-splines and Penalties." *Statistical Science* **11**, 89-102. 796, 797

Escobar, M. D. and West, M., 1995. "Bayesian Density Estimation and Inference Using Mixtures." *Journal of the American Statistical Association* **90**, 577-588. 803

Fahrmeir L. and Lang, S., 2001. "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors." *Journal of the Royal Statistical Society. Series C (Applied Statistics)* **50**, 201-220. 794

French, J. L., Kammann, E. E. and Wand, M.P., 2001. Comment on "Semiparametric Nonlinear Mixed-Effects Models and Their Applications" by Ke and Wang. *Journal of the American Statistical Association* **96**, 1285-1288. 795, 797, 798

Friedman, J. H. and Silverman, B. W., 1989. "Flexible Parsimonious Smoothing and Additive Modeling." *Technometrics* **31**, 3-21. 798

Girón, F. J., Moreno, E., and Casella, G., 2007. "Objective Bayesian Analysis of Multiple Changepoints for Linear Models." *Bayesian Statistics 8* (J. M. Bernardo, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West, eds.) Oxford University Press 1-27. 799

Gramacy, R. B., 2007. "tgp: An R Package for Bayesian Nonstationary, Semiparametric Nonlinear Regression and Design by Treed Gaussian Process Models." *Journal of Statistical Software* **19**, Issue 9. 807, 809

Gramacy, R. B. and Lee, H. K. H., 2008. "Bayesian Tree Gaussian Process Models With an Application to Computer Modeling." *Journal of the American Statistical Association* **103**, 1119-1130. 807

Gray, R. J., 1994. "Spline-Based Tests in Survival Analysis. " *Biometrics* **50**, 640-652. 797

Green, P.J., 1995. "Reversible Jump Markov Chain Monte Carlo Computation and Bayesian Model Determination." *Biometrika* **82**, 711-732. 798

Green, P.J. and Silverman, B.W., 1994. *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Chapman & Hall, London. 796

Gu, C., 2002. *Smoothing Spline ANOVA Models*. Springer. 811, 813

Halpern, E. F., 1973. "Bayesian Spline Regression When the Number of Knots is Unknown." *Journal of the Royal Statistical Society, Series B* **2**, 347-360. 805, 806, 819

Härdle, W., Müller, M., Sperlich, S. and Werwatz, A., 2004. *Nonparametric and Semiparametric Models*, Springer. 794

Harville, D., 1976. "Extension Of The Gauss-Markov Theorem To Include The Estimation Of Random Effects." *The Annals of Statistics* **4**, 384-395. 797

Hastie, T. J., 1996. "Pseudosplines." *Journal of the Royal Statistical Society, Series B* **58**, 379-396. 797

Hastie, T. J. and Tibshirani, R. J., 1990. *Generalized Additive Models*, Chapman & Hall/CRC 798

Hobert, J. P. and Marchev, D., 2008. "A Theoretical Comparison of the Data Augmentation, Marginal Augmentation and PX-DA Algorithms." *The Annals of Statistics* **36**, 532-554. 805

Holmes, C. C. and Mallick, B. K., 2001. "Bayesian Regression with Multivariate Linear Splines." *Journal of the Royal Stiatistical Society, Series B* **63**, 3-17. 798

Holmes, C. C. and Mallick, B. K., 2003. "Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines." *Journal of the American Statistical Association* **98**, 352-368. 798

Huang, S. Y. and Lu, H. H.-S. (2001). "Extended Gauss-Markov theorem for nonparametric mixed-effects models. " *Journal of Multivariate Analysis* **76**, 249-266 797

Kauermann, G., Krivobokova, T., and Fahrmeir, L., 2009. "Some Asymptotic Results on Generalized Penalized Spline Smoothing." *Journal of the Royal Statistical Society, Series B* **71**, 487-503. 797

Ke, C. and Wang, Y., 2001. "Semiparametric Nonlinear Mixed-Effects Models and Their Applications." *Journal of the American Statistical Association* **96**, 1272-1281. 794

Kelly, C. and Rice, J., 1990. "Monotone Smoothing with Application to Dose-Response Curves and the Assessment of Synergism." *Biometrics* **46**, 1071-1085. 797

Kyung, M, Gill, J,, and Casella G, 2009. "Characterizing the Variance Improvement in Linear Dirichlet Random Effects Models." *Statistics and Probability Letters* **79**, 2343-2350. 816

Kyung, M, Gill, J,, and Casella G., 2010. "Estimation in Dirichlet Random Effects Models." *Annals of Statistics* **38**, 979-1009. 800, 803, 804, 805

Leitenstorfer, F. and Tutz, G., 2007. "Knot Selection by Boosting Techniques." *Computational Statistics and Data Analysis* **51**, 4605-4621. 798

Lindley, D. V., 1968. "The Choice of Variables in Multiple Regression." *Journal of the Royal Statistical Society, Series B* **1**, 31-66. 806

Maity, A., Carroll, R. J., Mammen, E., and Chatterjee, N., 2009. "Testing in Semiparametric Models with Interaction, with Applications to Gene-Environment Interaction." *Journal of the Royal Statistical Society, Series B* **71**, 75-96. 794

Moreno, E., Casella, G., and Garcia-Ferrer, A., 2005. "An Objective Bayesian Analysis of the Change Point Problem." *Stochastic Environmental Research and Risk Assessment* **19**, 191-204. 799

O'Sullivan, F., 1986. "A Statistical Perspective on Ill-Posed Inverse Problems" *Statistical Science* **1**, 502-518. 797

Parker, R. L. and Rice, J. A., 1985. Discussion of "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" by Silverman. *Journal of the Royal Statistical Society, Series B* **47**, 40-42. 797

Pfeffermann, D., 1984. "On Extensions of the Gauss-Markov Theorem to the Case of Stochastic Regression Coefficients." *Journal of the Royal Statistical Society, Series B* **46**, 139-148. 797

Ripley, B.D., 1996. *Pattern Recognition and Neural Networks*. University Press, Cambridge. 796

Robinson, P. M., 1988. "Root-N-Consistent Semiparametric Regression." *Econometrica* **56**, 931-954. 797

Rubin, D. B., 1993. "Discussion: Statistical disclosure limitation." *Journal of Official Statistics* **2**, 461-468. 794

Ruppert, D., 2002. "Selecting the Number of Knots for Penalized Splines." *Journal of Computational and Graphical Statistics* **11**, 735-757. 798

Ruppert, D. and Carroll, R. J., 2000. "Spatially-Adaptive Penalties for Spline Fitting." *Australian and New Zealand Journal of Statistics* **42**, 205-224. 797

Ruppert, D., Wand, M. P. and Carroll, R. J., 2003. *Semiparametric Regression*, Wiley, New York 794, 797, 798

Silverman, B. W., 1985. "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting." *Journal of the Royal Statistical Society, Series B* **47**, 1-52. 795

Stone, C. J., 1985. "Additive Regression and Other Nonparametric Models." *The Annals of Statistics* **13**, 689-705. 795

Stone, C. J., Hansen, M. H., Kooperberg, C., and Truong, Y. K., 1997. "Polynomial Splines and Their Tensor Products in Extended Linear Modeling." *The Annals of Statistics* **25**, 1371-1470. 798

United States Department of Health and Human Services., 2006. Substance Abuse and Mental Health Services Administration. Office of Applied Studies. National Survey on Drug Use and Health, 2003 [Computer file]. ICPSR04138-v2. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2006-10-17. doi:10.3886/ICPSR04138 814

United States Department of Health and Human Services., 2010. National Institutes of Health. National Institute on Drug Abuse. Drug Abuse Treatment Outcome Study (DATOS), 1991-1994: [United States] [Computer file]. ICPSR02258-v5. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2010-02-16. doi:10.3886/ICPSR02258 814

U.S. Dept. of Justice, National Institute of Justice., 2004. ARRESTEE DRUG ABUSE MONITORING (ADAM) PROGRAM IN THE UNITED STATES, 2003 [Computer file]. ICPSR version. Washington, DC: U.S. Dept. of Justice, National Institute of Justice [producer], 2004. Ann Arbor, MI: Inter-university Consortium for Political and Social Research [distributor], 2004. doi:10.3886/ICPSR04020 814

Wahba, G., 1977. "Practical Approximate Solutions To Linear Operator Equations When The Data Are Noisy." *SIAM Journal on Numerical Analysis* **14**, 651-667. 795

Wand, M.P., 2003. "Smoothing and Mixed Models." *Computational Statistics* **18**, 223-249. 798

Woods, S., 2006. *Generalized Additive Models: An Introduction with R*, Chapman & Hall/CRC 798

Yin, G., Li, H., and Zeng, D., 2008. "Partially Linear Additive Hazards Regression With Varying Coefficients." *Journal of the American Statistical Association* **103**, 1200-1213. 794

Zeng, D. and Lin, D. Y., 2007. "Maximum Likelihood Estimation in Semiparametric Regression Models with Censored Data." *Journal of the Royal Statistical Society, Series B* **69**, 507-564. 794

Zhang D. and Davidian, M., 2001. "Linear Mixed Models with Flexivle Distributions of Random Effects for Longitudinal Data." *Biometrics* **57**, 795-802. 794

Acknowledgments

The author is grateful to a referee, the associate editor and the editor for their careful reading of the paper and their constructive comments. The author also thanks Professor George Casella, Department of Statistics, University of Florida and Professor Jeff Gill, Department of Political Science, Washington University in St. Louis for comments and suggestions that led to a much improved version of the paper. Supported by National Science Foundation Grants DMS-0631632 and SES-0631588.