Ann Inst Stat Math (2014) 66:1–31
DOI 10.1007/s10463-013-0415-z

Bayesian nonparametric regression with varying residual density

Debdeep Pati · David B. Dunson

Received: 20 May 2010 / Revised: 7 March 2013 / Published online: 16 June 2013

Abstract We consider the problem of robust Bayesian inference on the mean regression function allowing the residual density to change flexibly with predictors. The proposed class of models is based on a Gaussian process (GP) prior for the mean regression function and mixtures of Gaussians for the collection of residual densities indexed by predictors. Initially considering the homoscedastic case, we propose priors for the residual density based on probit stick-breaking mixtures. We provide sufficient conditions to ensure strong posterior consistency in estimating the regression function, generalizing existing theory focused on parametric residual distributions. The homoscedastic priors are generalized to allow residual densities to change nonparametrically with predictors through incorporating GP in the stick-breaking components. This leads to a robust Bayesian regression procedure that automatically down-weights outliers and influential observations in a locally adaptive manner. The methods are illustrated using simulated and real data applications.

Keywords Data augmentation · Exact block Gibbs sampler · Gaussian process · Nonparametric regression · Outliers · Symmetrized probit stick-breaking process

D. Pati (✉)
Department of Statistics, Florida State University, 117 N. Woodward Ave, P.O. Box 3064330,
Tallahassee, FL 32306-4330, USA
e-mail: debdeep@stat.fsu.edu

D. B. Dunson
Department of Statistical Science, Duke University, Old Chemistry Building, P.O. Box 90251,
Durham, NC 27708, USA
e-mail: dunson@stat.duke.edu

1 Introduction

Nonparametric regression offers a more flexible way of modeling the effect of covariates on the response compared to parametric models having restrictive assumptions on the mean function and the residual distribution. Here we consider a fully Bayesian approach. The response $y \in \mathcal{Y}$ corresponding to a set of covariates $\mathbf{x} = (x_1, x_2, \ldots, x_p)' \in \mathcal{X}$ can be expressed as

$$y = \eta(\mathbf{x}) + \epsilon \tag{1}$$

where $\eta(\mathbf{x}) = \mathsf{E}(y \mid \mathbf{x})$ is the mean regression function under the assumption that the residual density has mean zero, i.e., $\mathsf{E}(\epsilon \mid \mathbf{x}) = 0$ for all $\mathbf{x} \in \mathcal{X}$. Our focus is on obtaining a robust estimate of η while allowing heavy tails to down-weight influential observations. We propose a class of models that allows the residual density to change nonparametrically with predictors $\mathbf{x}$, with homoscedasticity arising as a special case.

There is a substantial literature proposing priors for flexible estimation of the mean function, typically using basis function representations such as splines or wavelets (Denison et al. 2002). Most of this literature assumes a constant residual density, possibly up to a scale factor allowing heteroscedasticity. Yau and Kohn (2003) allow the mean and variance to change with predictors using thin plate splines. In certain applications, this structure may be overly restrictive due to the specific splines used and the normality assumption. Chan et al. (2006) also used splines for heteroscedastic regression, but with locally adaptive estimation of the residual variance and allowance for uncertainty in variable selection. Nott (2006) considered the problem of simultaneous estimation of the mean and variance function using penalized splines for possibly non-Gaussian data. Due to the lack of conjugacy, these methods rely on involved sampling techniques using Metropolis Hastings, requiring proposal distributions to be chosen that may not be efficient in all cases. The residual density is assumed to have a known parametric form and heavy-tailed distributions have not been considered. In addition, since basis function selection for multiple predictors is highly computationally demanding, additive assumptions are typically made that rule out interactions.

Gaussian process (GP) regression (Adler 1990; Ghosal and Roy 2006; Vaart and Zanten 2008, 2009; Neal 1998) is an increasingly popular choice, which avoids the need to explicitly choose the basis functions, while having many appealing computational and theoretical properties. For articles describing some of these properties, refer to Adler (1990), Cramér and Leadbetter (1967), Vaart and Zanten (2008) and Vaart and Wellner (1996). A wide variety of functions can arise as the sample paths of the GP. GP priors can be chosen that have support on the space of all smooth functions while facilitating Bayesian computation through conjugacy properties. In particular, the GP realizations at the data points are simply multivariate Gaussian. As shown by Choi and Schervish (2007), GP priors also lead to consistent estimation of the regression function under normality assumptions on the residuals. Vaart and Zanten (2009) demonstrated that a GP prior with an inverse-gamma bandwidth leads to an optimal rate of posterior convergence in a mean regression problem with Gaussian errors. Recently, Choi (2009) extended the results of Choi and Schervish (2007) to allow for non-Gaussian symmetric residual distributions (for example, the Laplace distribution) which satisfy certain regularity conditions and the induced conditional density belongs to a location-scale family. Although they require mild assumptions on the parametric scale family, the results depend heavily on parametric assumptions. In particular, their theory of posterior consistency is not applicable to an infinite mixture prior on the residual density. We extend their result allowing a rich class of residual distributions through PSB mixtures of Gaussians in Sect. 3.

There is a rich literature on Bayesian methods for density estimation using mixture models of the form

$$y_i \sim f(\theta_i), \quad \theta_i \sim P, \quad P \sim \Pi, \tag{2}$$

where $f(\cdot)$ is a parametric density and P is an unknown mixing distribution assigned a prior Π. The most common choice of Π is the Dirichlet process (DP) (Ferguson 1973; Ferguson 1974). Lo (1984) showed that DP mixtures of normals have dense support on the space of densities with respect to Lesbesgue measure, while Escobar and West (1995) developed methods for posterior computation and inference. James et al. (2005) considered a broader class of normalized random measures for Π.

To combine methods for Bayesian nonparametric regression with methods for Bayesian density estimation, one can potentially use mixture model (2) for the residual density in (1). A number of authors have considered nonparametric priors for the residual distribution in regression. For example, Kottas and Gelfand (2001) proposed mixture models for the error distributions in median regression models. To ensure identifiability of the regression coefficients, the residual distribution is constrained to have median zero. Their approach is very flexible but has the unappealing property of producing a residual density that is discontinuous at zero. In addition, the approach of mixing uniforms leads to blocky looking estimates of the residual density particularly for sparse data. Lavine and Mockus (2005) allowed both a regression function for a single predictor and the residual distribution to be unknown subject to a monotonicity constraint. A number of recent papers have focused on generalizing model (2) to the density regression setting in which the entire conditional distribution of y given $\mathbf{x}$ changes flexibly with predictors. Refer, for example, to Müller et al. (1996), Griffin and Steel (2006, 2010), Dunson et al. (2007) and Dunson and Park (2008) among others. Bush and MacEachern (1996) considered estimating the random block effects nonparametrically in an ANOVA-type mean linear-regression model with a t-residual density rather than density regression.

Although these approaches are clearly highly flexible, there are several issues that provide motivation for this article. First, to simplify inferences and prior elicitation, it is appealing to separate the mean function $\eta(\mathbf{x})$ from the residual distribution in the specification, which is accomplished by only a few density regression methods. The general framework of separately modeling the mean function and residual distribution nonparametrically was introduced by Griffin and Steel (2010). They allow the residual distribution to change flexibly with predictors using the order-based DP (Griffin and Steel 2006). On the other hand, we need to have a computationally simpler specification with straightforward prior elicitation. Chib and Greenberg (2010) develops a nonparametric model jointly for continuous and categorical responses where they model the mean of the link function and residual density separately. The mean is modeled using flexible additive splines and the residual density is modeled using a DP scale mixture of normals. However, they did not allow the residual distribution to change flexibly with the predictors. Often we have strong prior information regarding the form of the regression function. Most density regression models do not allow incorporation of prior information on the mean function separately from the residual densities. Second, in many applications, the main interest is in inference on η or in prediction, and the residual distribution can be considered as a nuisance. Third, the use of residual distribution with zero mean has rarely been attempted in the nonparametric Bayesian literature. This is one of the important contributions of the paper. Finally, we would like to provide a specification with theoretical support. In particular, it would be appealing to show strong posterior consistency in estimating η without requiring restrictive assumptions on η or the residual distribution. Current density regression models lack such theoretical support. In addition, computation for density regression can be quite involved, particularly in cases involving more than a few predictors, and one encounters the curse of dimensionality. Our goal was to obtain a computationally convenient specification that allows consistent estimation of the regression function, while being flexible in the residual distribution specification to obtain robust estimates.

We propose to place a GP prior on η and to allow the residual density to be unknown through a probit stick-breaking (PSB) process mixture. The basic PSB process specification was proposed by Chung and Dunson (2009) in developing a density regression approach that allows variable selection. On the other hand, we are concerned with robust estimation of the mean regression function, allowing the residual distribution to change flexibly with predictors. While we want to model the mean regression function nonparametrically, we also want to incorporate our prior knowledge for the regression function quite easily. Here, we propose four novel variants of PSB mixtures for the residual distribution. The first uses a scale mixture of Gaussians to obtain a prior with large support on unimodal symmetric distributions. The next is based on a symmetrized location-scale PSB mixture, which is more flexible in avoiding the unimodality constraint, while constraining the residual density to be symmetric and have mean zero. In addition, we show that this prior leads to strong posterior consistency in estimating η under weak conditions.

To allow the residual density to change flexibly with predictors, we generalize the above priors through incorporating probit transformations of GP in the weights. The last two prior specifications allow changing residual variances and tail heaviness with predictors, leading to a highly robust specification which is shown to have better performance in simulation studies and out of sample prediction. It will be shown in some small sample simulated examples that the heteroscedastic-symmetrized location-scale PSB mixture leads to even more robust inference than the heteroscedastic scale PSB mixture without compromising out of sample predictive performance.

Section 2 proposes the class of models under consideration. Section 3 shows consistency properties. Section 4 develops efficient posterior computation through an exact block Gibbs sampler. Section 5 describes measures of influence to study robustness properties of our proposed methods. Section 6 contains simulation study results, Sect. 7 applies the methods to the Boston housing data and body fat data, and Sect. 8 discusses the results. Proofs are included in the Appendix.

2 Nonparametric regression modeling

2.1 Data structure and model

Consider n observations with the ith observation recorded in response to the covariate $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})'$. Let $\mathbf{X} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)'$ be the predictor matrix for all n subjects. The regression model can be expressed as

$$y_i = \eta(\mathbf{x}_i) + \epsilon_i, \quad \epsilon_i \sim f_{\mathbf{x}_i}, \quad i = 1, \ldots, n.$$

We assume that the response $y \in \mathcal{Y}$ is continuous and $\mathbf{x} \in \mathcal{X}$, where $\mathcal{X} \subset \mathbb{R}^p$ is compact. Also, the residuals ϵ_i are sampled independently, with $f_{\mathbf{x}}$ denoting the residual density specific to predictor value $\mathbf{x}_i = \mathbf{x}$. We focus initially on the case in which the covariate space $\mathcal{X}$ is continuous, with the covariates arising from a fixed, non-random design or consisting of i.i.d realizations of a random variable. We choose a prior on the regression function $\eta(\mathbf{x})$ that has support on a large subset of $\mathcal{C}^\infty(\mathcal{X})$, the space of smooth real-valued $\mathcal{X} \to \mathbb{R}$ functions. The priors proposed for $\{f_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ will be chosen to have large support so that heavy-tailed distributions and outliers will automatically be accommodated, with influential observations down-weighted in estimating η.

2.2 Prior on the mean regression function

We assume that $\eta \in \mathcal{F} = \{g : \mathcal{X} \to \mathbb{R}$ is a continuous function$\}$, with η assigned a GP prior, $\eta \sim GP(\mu, c)$, where μ is the mean function and c is the covariance kernel. A GP is a stochastic process $\{\eta(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$ such that any finite dimensional distribution is multivariate normal, i.e., for any n and $\mathbf{x}_1, \ldots, \mathbf{x}_n$, $\eta(\mathbf{X}) := (\eta(\mathbf{x}_1), \ldots, \eta(\mathbf{x}_n))' \sim \mathrm{N}(\mu(\mathbf{X}), \Sigma^\eta)$, where $\mu(\mathbf{X}) = (\mu(\mathbf{x}_1), \ldots, \mu(\mathbf{x}_n))'$ and $\Sigma^\eta_{ij} = c(\mathbf{x}_i, \mathbf{x}_j)$. Naturally the covariance kernel $c(\cdot, \cdot)$ must satisfy, for each n and $\mathbf{x}_1, \ldots, \mathbf{x}_n$, that the matrix Σ^η is positive definite. The smoothness of the covariance kernel essentially controls the smoothness of the sample paths of $\{\eta(\mathbf{x}) : \mathbf{x} \in \mathcal{X}\}$. For an appropriate choice of c, a GP has large support in the space of all smooth functions. More precisely, the support of a GP is the closure of the reproducing kernel Hilbert space generated by the covariance kernel with a shift by the mean function (Ghosal and Roy 2006). For example, when $\mathcal{X} \subset \mathbb{R}$, the eigenfunctions of the univariate covariance kernel, $c(x, x') = \frac{1}{\tau}\mathrm{e}^{-\kappa(x-x')^2}$, span $C^\infty(\mathcal{X})$ if κ is allowed to vary freely over $\mathbb{R}^+$. Thus, we can see that the GP prior has a rich class of functions as its support and hence is appealing as a prior on the mean regression function. Refer to Rasmussen and Williams (2006) as an introductory textbook on Gaussian processes.

We follow common practice in choosing the mean function in the GP prior to correspond to a linear regression, $\mu(\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$, with $\boldsymbol{\beta}$ denoting unknown regression coefficients. As a commonly used covariance kernel, we took the Gaussian kernel $c(\mathbf{x}, \mathbf{x}') = \frac{1}{\tau}\mathrm{e}^{-\kappa||\mathbf{x}-\mathbf{x}'||^2}$, where τ and κ are unknown hyperparameters, with κ controlling the local smoothness of the sample paths of $\eta(\mathbf{x})$. Smoother sample paths imply more borrowing of information from neighboring $\mathbf{x}$ values.

2.3 Priors for residual distribution

Motivated by the problem of robust estimation of the regression function η, we consider five different types of priors for the residual distributions $\{f_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ as enumerated below. The first prior corresponds to the t distribution, which is widely used for robust modeling of residual distributions (West 1984; Lange et al. 1989; Fonseca et al. 2008), while the remaining priors are flexible nonparametric specifications.

P1. Heavy-tailed parametric error distribution: Following many previous authors, we first consider the case in which the residual distributions follow a homoscedastic Student's t distribution with unknown degrees of freedom. As the Student's t with low degrees of freedom is heavy tailed, outliers are allowed. By placing a hyperprior on the degrees of freedom, $\nu_\sigma \sim \mathrm{Ga}(a_\nu, b_\nu)$, with $\mathrm{Ga}(a, b)$ denoting the Gamma distribution with mean a/b, one can obtain a data adaptive approach to down-weighting outliers in estimating the mean regression function. However, note that this specification assumes that the same degrees of freedom and tail heaviness hold for all $\mathbf{x} \in \mathcal{X}$. Following West (1987), we express the Student's t distribution as a scale mixture of normals for ease in computation. In addition, we allow an unknown scale parameter, letting $\epsilon_i \sim \mathrm{N}(0, \sigma^2/\phi_i)$, with $\phi_i \sim \mathrm{Ga}(\nu_\sigma/2, \nu_\sigma/2)$, $\sigma^{-2} \sim \mathrm{Ga}(a, b)$.

P2. Nonparametric error distribution: Let $\mathcal{Y} = \Re$ be the response space and $\mathcal{X}$ be the covariate space which is a compact subset of $\Re^p$. Let $\mathcal{F}$ denote the space of densities on $\mathcal{X} \times \mathcal{Y}$ w.r.t. the Lebesgue measure and $\mathcal{F}_d$ denotes the space of all conditional densities subject to mean zero,

$$\mathcal{F}_d = \left\{ g : \mathcal{X} \times \mathcal{Y} \to (0, \infty), \int_{\mathcal{Y}} g(\mathbf{x}, y)\mathrm{d}y = 1, \int_{\mathcal{Y}} y g(\mathbf{x}, y)\mathrm{d}y = 0 \quad \forall \quad \mathbf{x} \in \mathcal{X} \right\}.$$

We propose to induce a prior on the space of mean zero conditional densities through a prior for collection of mixing measures $\{P_{\mathbf{x}}, \mathbf{x} \in \mathcal{X}\}$ using the following predictor-dependent mixture of kernels.

$$P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\{\mu_h(\mathbf{x}), \sigma_h\}}, \quad \mu_h \sim P_0, \sigma_h \sim P_{0,\sigma} \tag{3}$$

where $\pi_h(\mathbf{x}) \geq 0$ are random functions of $\mathbf{x}$ such that $\sum_{h=1}^{\infty} \pi_h(\mathbf{x}) = 1$ a.s. for each fixed $\mathbf{x} \in \mathcal{X}$. $\{\mu_h(\mathbf{x}), \mathbf{x} \in X\}_{h=1}^{\infty}$ are i.i.d realizations of a real-valued stochastic process, i.e., P_0 is a probability distribution over a function space $\mathcal{F}_{\mathcal{X}}$. Here $P_{0,\sigma}$ is a probability distribution on $\Re^+$. Hence, for each $\mathbf{x} \in \mathcal{X}$, $P_{\mathbf{x}}$ is a random probability measure over the measurable Polish space $(\Re \times \Re^+, \mathcal{B}(\Re \times \Re^+))$. Before proposing the prior, we first review the probit stick-breaking process specification and its relationship to the DP. Rodriguez and Dunson (2011) introduced the probit stick-breaking process in broad settings and discussed some smoothness and clustering properties. A probability measure $P \in \mathcal{P}$ on $(\mathcal{Y}, \mathcal{B}(\mathcal{Y}))$ follows a probit stick-breaking process with base measure P_0 if it has a representation of the form

$$P(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\theta_h}(\cdot), \quad \theta_h \sim P_0, \tag{4}$$

where the atoms $\{\theta_h\}_{h=1}^{\infty}$ are independent and identically distributed from P_0 and the random weights are defined as $\pi_h = \Phi(\alpha_h)\prod_{l<h}\{1-\Phi(\alpha_l)\}$, $\alpha_h \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2)$, $h = 1, \ldots, \infty$. Here $\Phi(\cdot)$ denotes the cumulative distribution function for the standard normal distribution. Note that expression (4) is identical to the stick-breaking representation (Sethuraman 1994) of the DP, but the DP is obtained by replacing the stick-breaking weight $\Phi(\alpha_h)$ with a beta$(1, \alpha)$-distributed random variable. Hence, the PSB process differs from the DP in using probit transformations of Gaussian random variables instead of betas for the stick lengths, with the two specifications being identical in the special case in which $\mu_\alpha = 0$, $\sigma_\alpha = 1$ and the DP precision parameter is $\alpha = 1$. Rodriguez and Dunson (2011) also mentioned the possibility of constructing a variety of predictor-dependent models, e.g., latent Markov random fields, spatio-temporal processes, etc. using probit transformation latent GP. Such latent GP can be updated using data augmentation Gibbs sampling as in continuation-ratio probit models for survival analysis (Albert and Chib 2001). While we follow similar computational strategies as in Rodriguez and Dunson (2011), they did not consider robust regression using predictor-dependent residual density.

Under the symmetric about zero assumption, we propose two nonparametric priors for the residual density $f_{\mathbf{x}}$ for all $\mathbf{x} \in \mathcal{X}$. The first prior is a predictor-dependent PSB scale mixture of Gaussians which enforces symmetry about zero and unimodality, and the next is a symmetrized location-scale PSB mixture of Gaussians, which we develop to satisfy the symmetric about zero assumption while allowing multimodality.

P2a. Heteroscedastic scale PSB mixtures: To allow the residual density to change flexibly with predictors, while maintaining the constraint that each of the predictor-dependent residual distributions is unimodal and symmetric about zero, we propose the following specification

$$f(\cdot) = \int \mathrm{N}(\cdot; 0, \tau^{-1}) P_{\mathbf{x}}(d\tau), \quad P_{\mathbf{x}} = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \delta_{\tau_h}, \quad \tau_h \sim \mathrm{Ga}(\alpha_\tau, \beta_\tau), \tag{5}$$

where $\pi_h(\mathbf{x}) = \Phi\{\alpha_h(\mathbf{x})\}\prod_{l<h}[1 - \Phi\{\alpha_l(\mathbf{x})\}]$ is the predictor-dependent probability weight on the hth mixture component, and the α_hs are drawn independently from zero mean GP having covariance kernel $c_\alpha(\mathbf{x}, \mathbf{x}') = \frac{1}{\tau_\alpha} e^{-\kappa_\alpha ||\mathbf{x}-\mathbf{x}'||^2}$. This implies $f_{\mathbf{x}}(\cdot) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \mathrm{N}(\cdot; 0, \tau_h^{-1})$ and is a highly flexible specification that enforces smoothly changing mixture weights across the predictor space, so that the residual densities at $\mathbf{x}$ and $\mathbf{x}'$ will tend to be similar if $\mathbf{x}$ is located close to $\mathbf{x}'$, as measured by $\kappa_\alpha ||\mathbf{x} - \mathbf{x}'||^2$.

Clearly, the specification allows the residual variance to change flexibly with predictors, as we have $\mathrm{var}(\epsilon \mid \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \tau_h^{-1}$. However, unlike the previously proposed methods for heteroscedastic non-linear regression, we do not just allow the variances to vary, but allow any aspect of the density to vary, including the heaviness of the tails. This allows locally adaptive down-weighting of outliers in estimating the mean function. Previous methods, which instead assume a single heavy-tailed residual distribution, such as a t-distribution, can lead to a lack of robustness due to global estimation of a single degree of freedom parameter. In addition, due to the form of our specification, posterior computation becomes very straightforward using a data augmentation Gibbs sampler, which involves simple steps for sampling from conjugate full conditional distributions. Even under the assumption of Gaussian residual distributions, posterior computation for heteroscedastic models tends to be complex, with Gibbs sampling typically not possible due to the lack of conditional conjugacy.

P2b. Heteroscedastic-symmetric PSB (sPSB) location-scale mixtures: The PSB scale mixture in (5) restricts the residual density to be unimodal. As this is a very restrictive assumption, it is appealing to define a prior with larger support that allows multimodal residual densities, while enforcing the symmetric about zero assumption so that the residual density is constrained to have mean zero. To accomplish this, we propose a novel symmetrized PSB process specification, which is related to the symmetrized DP proposed by Tokdar (2006). We define

$$f(\cdot) = \int \mathrm{N}(\cdot;\mu,\tau^{-1})\mathrm{d}P_{\mathbf{x}}^{s}(\mu,\tau), \quad \mathrm{d}P_{\mathbf{x}}^{s}(\mu,\tau) = \tfrac{1}{2}\mathrm{d}P_{\mathbf{x}}(-\mu,\tau) + \tfrac{1}{2}\mathrm{d}P_{\mathbf{x}}(\mu,\tau), \tag{6}$$

where the atoms (μ_h, τ_h) are drawn independently from P_0 a priori, with P_0 chosen as a product of a $\mathrm{N}(\mu_0, \sigma_0^2)$ and $\mathrm{Ga}(\alpha_\tau, \beta_\tau)$ measure. The difference between the sPSB process prior and the PSB process prior is that instead of just placing probability weight π_h on atom (μ_h, τ_h), we place probability $\pi_h/2$ on $(-\mu_h, \tau_h)$ and (μ_h, τ_h). The resulting residual density under (6) has the form $f(\cdot) = \sum_{h=1}^{\infty} \frac{\pi_h(\mathbf{x})}{2}\{\mathrm{N}(\cdot,;-\mu_h,\tau_h^{-1}) + \mathrm{N}(\cdot;\mu_h,\tau_h^{-1})\}$. Clearly, each of the realizations corresponds to a mixture of Gaussians that is constrained to be symmetric about zero. The same comments made for the heteroscedastic scale PSB mixture apply here, but (6) is more flexible in allowing multimodal residual distributions, with modality changing flexibly with predictors. Posterior computation is again straightforward, as will be shown later.

P2c. Homoscedastic scale PSB process mixture of Gaussians: A simpler homoscedastic version of (5) is to consider

$$f(\cdot) = \int \mathrm{N}(\cdot;0,\tau^{-1})P(\mathrm{d}\tau), \quad P = \sum_{h=1}^{\infty}\pi_h\delta_{\tau_h}, \quad \tau_h \sim \mathrm{Ga}(\alpha_\tau,\beta_\tau), \tag{7}$$

where the weights $\{\pi_h\}$ are specified as in

$$\pi_h = \nu_h\prod_{l<h}(1-\nu_h), \quad \nu_h = \Phi(\alpha_h), \quad \alpha_h \sim \mathrm{N}(\mu_\alpha,\sigma_\alpha^2). \tag{8}$$

This implies that $f(\cdot) = \sum_{h=1}^{\infty} \pi_h \mathrm{N}(\cdot; 0, \tau_h^{-1})$, so that the unknown density of the residuals is expressed as a countable mixture of Gaussians centered at zero but with varying variances. Observations will be automatically allocated to clusters, with outlying clusters corresponding to components having large variance (low τ_h). By choosing a hyperprior on μ_α while letting $\sigma_\alpha = 1$, we allow the data to inform more strongly about the posterior distribution on the number, sizes and allocation to clusters.

P2d. Location-scale symmetrized PSB (sPSB) mixture of Gaussians: A homoscedastic version of (6) is the following.

$$f(\cdot) = \int \mathrm{N}(\cdot; \mu, \tau^{-1}) \mathrm{d}P^s(\mu, \tau), \quad \mathrm{d}P^s(\mu, \tau) = \frac{1}{2}\mathrm{d}P(-\mu, \tau) + \frac{1}{2}\mathrm{d}P(\mu, \tau),$$
$$P = \sum_{h=1}^{\infty} \pi_h \delta_{(\mu_h, \tau_h)}, \quad (\mu_h, \tau_h) \sim P_0, \tag{9}$$

where the prior on the weights π_h are given by (8) and the prior for (μ_h, τ_h) are exactly as in 2b.

3 Consistency properties

Let $f \sim \Pi_u$ and $f \sim \Pi_s$ denote the priors for the unknown residual density defined in expressions (7) and (9), respectively. It is appealing for Π_u and Π_s to have support on a large subset of $\mathcal{S}_u$ and $\mathcal{S}_s$, respectively, where $\mathcal{S}_s$ denotes the set of densities on $\mathbb{R}$ with respect to Lebesgue measure that is symmetric about zero and $\mathcal{S}_u \subset \mathcal{S}_s$ is the subset of $\mathcal{S}_s$ corresponding to unimodal densities. We characterize the weak support of Π_u, denoted by $wk(\Pi_u) \subset \mathcal{S}_u$, in the following lemma.

Lemma 1 $wk(\Pi_u) = \mathcal{C}_m$*, where* $\mathcal{C}_m = \{f : f \in \mathcal{S}_u, h(x) = f(\sqrt{x}), x > 0$ *is a completely monotone function*$\}$.

A function $h(x)$ on $(0, \infty)$ is completely monotone in x if it is infinitely differentiable and $(-1)^m \frac{\mathrm{d}^m}{\mathrm{d}x^m} h(x) \geq 0$ for all x and for all $m \in \{1, 2, \ldots, \infty\}$. Chu (1973) proved that if f is a density on $\mathbb{R}$ which is symmetric about zero and unimodal, it can be written as a scale mixture of normals,

$$f(x) = \int \sigma^{-1} \phi(\sigma^{-1} x) g(\sigma) \mathrm{d}\sigma$$

for some density g on $\mathbb{R}$, if and only if $h(x) = f(\sqrt{x}), x > 0$, is a completely monotone function, where ϕ is the standard normal pdf. This restriction places a smoothness constraint on $f(x)$, but still allows a broad variety of densities.

Definition 1 Letting $f \sim \Pi$, f_0 is in the Kullback–Leibler(KL) support of Π if

$$\Pi\left(f : \int f_0(y) \log \frac{f_0(y)}{f(y)} \mathrm{d}y < \epsilon\right) > 0, \quad \forall \quad \epsilon > 0$$

The set of densities f in the KL support of Π is denoted by $KL(\Pi)$.

Let $\tilde{\mathcal{S}}_s$ denote the subset of $\mathcal{S}_s$ corresponding to densities satisfying the following regularity conditions.

1. f is nowhere zero and bounded by $M < \infty$
2. $\left| \int_{\Re} f(y) \log f(y) \mathrm{d}y \right| < \infty$
3. $\left| \int_{\Re} f(y) \log \frac{f(y)}{\psi_1(y)} \mathrm{d}y \right| < \infty$, where $\psi_1(y) = \inf_{t \in [y-1, y+1]} f(t)$
4. there exists $\psi > 0$ such that $\int_{\Re} |y|^{2(1+\psi)} f(y) \mathrm{d}y < \infty$

Lemma 2 $KL(\Pi_s) \supseteq \tilde{\mathcal{S}}_s$.

Remark 1 The above assumptions on f are standard regularity conditions introduced by Tokdar (2006) and Wu and Ghoshal (2008) to prove that $f \in KL(\Pi)$, where Π is a general stick-breaking prior which has all compactly supported probability distributions as its support. (1) Is usually satisfied by common densities arising in practice. (4) Imposes a minor tail restriction, e.g., t-density with $(2 + \delta)$ degrees of freedom for some $\delta > 0$ satisfies (4). (1)–(4) are satisfied by a finite mixture of t-densities or even by an infinite mixture of t-densities with $(2 + \delta)$ degrees of freedom for some $\delta > 0$ and bounded component-specific means and variances.

From Lemma 2, it follows that the sPSB location-scale mixture has KL support on a large subset of the set of densities symmetric about zero. These conditions are important in verifying that the priors are flexible enough to approximate any density subject to the noted restrictions.

We provide fairly general sufficient conditions to ensure strong and weak posterior consistency in estimating the mean regression function and the residual density, respectively. We focus on the case in which a GP prior is chosen for η and an sPSB location-scale mixture of Gaussians is chosen for the residual density as in (9). Similar results can be obtained for the homoscedastic scale PSB process mixture under stronger restrictions on the true residual density. Although showing consistent results using predictor-dependent mixtures of normals as the prior for the residual density in (5) and (6) is a challenging task, one can anticipate such results given the theory in Pati et al. (2013) and Norets and Pelenis (2010). Indeed, Norets and Pelenis (2011) showed posterior consistency of the regression coefficients in a mean linear-regression model with covariate-dependent nonparametric residuals using the kernel stick-breaking process Dunson and Park (2008). However, showing posterior consistency of the mean regression when we have a GP prior on the regression function and predictor-dependent residuals is quite challenging and is a topic of future research.

For this section, we assume $\mathbf{x}_i$ as non-random and arising from a fixed design, though the proofs are easily modified for random $\mathbf{x}_i$. When the covariate values are fixed in advance, we consider the neighborhood based on the empirical measure of the design points. Let Q_n be the empirical probability measure of the design points, $Q_n(\mathbf{x}) = \frac{1}{n} \sum_{i=1}^{n} I_{\mathbf{x}_i}(\mathbf{x})$. Based on Q_n, we define a strong L_1 neighborhood of radius $\Delta > 0$ as in Choi (2005) around the true regression function η_0. Letting $||\eta - \eta_0||_{1,n} = \int_{\mathbf{x} \in \mathcal{X}} |\eta(\mathbf{x}) - \eta_0(\mathbf{x})| \mathrm{d}Q_n(\mathbf{x})$ set,

$$S_n(\eta_0; \Delta) = \left\{ \eta : ||\eta - \eta_0||_{1,n} < \Delta \right\}. \tag{10}$$

We introduce the following notation. Let f_0 denote an arbitrary fixed density in $\tilde{\mathcal{S}}_s$, η_0 denote an arbitrary fixed regression function in $\mathcal{F}$, and

$$f_{0i} = f_0(y - \eta_0(\mathbf{x}_i)) \quad f_{\eta i} = f(y - \eta(\mathbf{x}_i)).$$

For any two densities f and g, let

$$K(f, g) = \int_{\mathbb{R}} f(y) \log\{f(y)/g(y)\} \mathrm{d}y, \quad V(f, g) = \int_{\mathbb{R}} f(y) \big[\log_+\{f(y)/g(y)\}\big]^2 \mathrm{d}y,$$

where $\log_+ x = \max(\log x, 0)$. Set $K_i(f, \eta) = K(f_{0i}, f_{\eta i})$ and $V_i(f, \eta) = V(f_{0i}, f_{\eta i})$ for $i = 1, \ldots, n$.

For technical simplicity, assume $\mathcal{X} = [0, 1]^p$, $\tau = 1$ and $\mu \equiv 0$. Denote a mean zero GP $\{W_{\mathbf{x}} : \mathbf{x} \in [0, 1]^p\}$ with covariance kernel $c(\mathbf{x}, \mathbf{x}') = \mathrm{e}^{-||\mathbf{x}-\mathbf{x}'||^2}$ by W. Rescaling the sample paths of an infinitely smooth GP is a powerful technique to improve the approximation of α-Hölder functions from the RKHS of the scaled process $\{W_{\mathbf{x}}^{\kappa} = W_{\sqrt{\kappa}\mathbf{x}} : \mathbf{x} \in [0, 1]^d\}$ with $\kappa > 0$. Intuitively, for large values of κ, the scaled process traverses the sample path of an unscaled process on the larger interval $[0, \sqrt{\kappa}]^p$, thereby incorporating more "roughness". Vaart and Zanten (2009) studied that rescaled GP, $W^{\kappa} = \{W_{\sqrt{\kappa}\mathbf{x}} : \mathbf{x} \in [0, 1]^p\}$, for a positive random variable κ was stochastically independent of W and also showed that with a Gamma prior on $\kappa^{p/2}$, one can obtain the minimax optimal rate of convergence for arbitrary smooth functions.

Assumption 1 $\eta \sim W^{\kappa}$ with the density g of $\sqrt{\kappa}$ on the positive real line satisfying

$$C_1 x^c \exp(-D_1 x \log^d x) \leq g(x) \leq C_2 x^c \exp(-D_2 x \log^d x),$$

for positive constants C_1, C_2, D_1, D_2, non-negative constants c, d, and every sufficiently large $x > 0$. Next we state the lemma on prior positivity due to Vaart and Zanten (2009).

Lemma 3 *If η satisfies Assumption 1 then $P(||\eta - \eta_0||_\infty < \epsilon) > 0 \, \forall \epsilon > 0$, if η_0 is continuous.*

To prove posterior consistency for our proposed model, we rely on a theorem of Amewou-Atisso et al. (2003), which is a modification of the celebrated Schwartz (1965) theorem to accommodate independent but not identically distributed data.

Theorem 1 *Suppose η as in Assumption 1 with $q \geq p + 2$ and $f \sim \Pi_s$, with Π_s defined in (9). In addition, assume the data are drawn from the true density $f_0(y_i - \eta_0(\mathbf{x}_i))$, with $\{\mathbf{x}_i\}$ fixed and non-random, $f_0 \in \tilde{\mathcal{S}}_s$, $\eta_0 \in \mathcal{F}$ and f_0 following the additional regularity conditions,*

1. $\int y^4 f_0(y) \mathrm{d}y < \infty$ *and* $\int f_0(y) |\log f_0(y)|^2 \mathrm{d}y < \infty$.
2. $\int_{\mathbb{R}} f_0(y) \big|\log \frac{f_0(y)}{\psi_1(y)}\big|^2 \mathrm{d}y < \infty$, *where* $\psi_1(y) = \inf_{t \in [y-1, y+1]} f_0(t)$.

Let $\mathcal{U}$ be a weak neighborhood of f_0 and $\mathcal{W}_n = \mathcal{U} \times S_n(\eta_0; \Delta)$, with $\mathcal{W}_n \subset \tilde{\mathcal{S}}_s \times \mathcal{F}$. Then the posterior probability

$$(\Pi_s \times W^{\kappa})(\mathcal{W}_n | y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n) = \frac{\int_{\mathcal{W}_n} \prod_{i=1}^n f_{\eta i}(y_i) \mathrm{d}\Pi_s(f) \mathrm{d}W^{\kappa}(\eta)}{\int_{\tilde{\mathcal{S}}_s \times \mathcal{F}} \prod_{i=1}^n f_{\eta i}(y_i) \mathrm{d}\Pi_s(f) \mathrm{d}W^{\kappa}(\eta)}$$
$$\to 1 \textit{ a.s.}$$

Theorem 1 ensures weak posterior consistency of the residual density and strong posterior consistency of the regression function η.

4 Posterior computation

We first describe the choice of hyperpriors and hyperparameters for the regression function. We choose the typical conjugate prior for the regression coefficients in the mean of the GP, $\boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \Sigma_0)$, where $\boldsymbol{\beta}_0 = 0$ and $\boldsymbol{\Sigma}_0 = c\mathbf{I}$ is a common choice corresponding to a ridge regression shrinkage prior. The prior on τ is given by $\tau \sim \mathrm{Ga}(\frac{\nu_\tau}{2}, \frac{\nu_\tau}{2})$. We let $\kappa \sim \mathrm{Ga}(\alpha_\kappa, \beta_\kappa)$ with small β_κ and large α_κ. Normalizing the predictors prior to analysis, we find that the data are quite informative about κ under these priors, so as long as the priors are not overly informative, inferences are robust. The parameter τ controls the heaviness of the tails of the prior for the regression function. In fact, choosing a $\mathrm{Ga}(\nu_\tau/2, \nu_\tau/2)$ prior induces a heavy-tailed t-process with ν_τ degrees of freedom as a prior for the regression function. We chose ν_τ to be 3. κ controls the correlation of the GP at two points in the covariate space similar to a spatial decay parameter in a spatial random effects model. Although a discrete uniform prior for κ is computationally efficient in leading to a griddy Gibbs update step, there can be sensitivity to the choice of grid. A gamma prior for κ eliminates such sensitivity at some associated computational price in terms of requiring a Metropolis-Hastings update that tends to mix slowly. We choose the parameters α_κ and β_κ so that the mean correlation is 0.1 for two points separated by a distance $\sqrt{p}$ in the covariate space.

Next we describe the hyperprior and associated hyperparameter choices for P1 and P2a–d.

1. *P1:* Since the responses are normalized and the covariates are scaled to lie in the interval [0, 1], using a single decay parameter appears to be reasonable. ν_σ controls the tail heaviness of the prior for the scaling ϕ. To accommodate outliers with the mean being fixed at 1, we assume $\phi_i \sim \mathrm{Ga}(\nu_\sigma/2, \nu_\sigma/2)$ with $\nu_\sigma \sim \mathrm{Ga}(\alpha_\nu, \beta_\nu)$. We took $\boldsymbol{\Sigma}_0 = 5\mathbf{I}$, $\alpha_\nu = 1$, $\beta_\nu = 1$. a and b are fixed at 3/2 to resemble a t distribution with 3 degrees of freedom without the scaling ϕ_i.
2. *P2a and P2b:* We assume $\kappa_\alpha \sim \mathrm{Ga}(\gamma_\kappa, \delta_\kappa)$ and $\tau_\alpha \sim \mathrm{Ga}(\frac{\nu_\alpha}{2}, \frac{\nu_\alpha}{2})$. Assuming y_i is normalized, we can expect the overall variance to be close to one, so the variance of the residuals, $Var(\epsilon \mid \mathbf{x}) = \sum_{h=1}^{\infty} \pi_h(\mathbf{x}) \tau_h^{-1}$, should be less than one. We set $\alpha_\tau = 1$ and choose a hyperprior on β_τ, $\beta_\tau \sim \mathrm{Ga}(1, k_0)$ with $k_0 > 1$ so that the prior mean of τ_h is significantly less than one. Different values of k_0 are tried out to assess robustness of the posteriors. In Sects. 5 and 6, we choose $\gamma_\kappa = 1$, $\delta_\kappa = 5$, $\nu_\alpha = 1$, $k_0 = 10$, $\mu_0 = 0, \sigma_0 = 1$.

3. *P2c and P2d:* Same choices as above except for $k_0 = 5$, $\mu_\alpha = 0, \sigma_\alpha = 1$.

For brevity, we provide details for posterior computation only for P1, P2a–b.

4.1 Gaussian process regression with t residuals (P1)

Let $\mathbf{Y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\eta} = (\eta(\mathbf{x}_1), \eta(\mathbf{x}_2), \ldots, \eta(\mathbf{x}_n))'$ and define a matrix $\mathbf{T}$ such that $\mathbf{T}_{ij} = \mathrm{e}^{-\kappa||\mathbf{x}_i - \mathbf{x}_j||^2}$. Hence, $\boldsymbol{\Sigma}^\eta = \frac{1}{\tau}\mathbf{T}$. Assume $\boldsymbol{\Omega} = \mathrm{diag}(1/\phi_i : i = 1, \ldots, n)$ and $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_n)'$. Then we have

$$\begin{aligned}
&\mathbf{Y}|\boldsymbol{\eta} \sim \mathrm{N}(\boldsymbol{\eta}, \sigma^2\boldsymbol{\Omega}), \boldsymbol{\eta}|\boldsymbol{\beta}, \tau, \kappa \sim \mathrm{N}(\mathbf{X}\boldsymbol{\beta}, \tau^{-1}\mathbf{T}), \boldsymbol{\beta} \sim \mathrm{N}(\boldsymbol{\beta}_0, \boldsymbol{\Sigma}_0)\\
&\phi_i \sim \mathrm{Ga}\Big(\frac{\nu_\sigma}{2}, \frac{\nu_\sigma}{2}\Big), \quad \nu_\sigma \sim \mathrm{Ga}(\alpha_\nu, \beta_\nu), \quad \sigma^{-2} \sim \mathrm{Ga}(a, b)\\
&\kappa \sim \mathrm{Ga}(\alpha_\kappa, \beta_\kappa), \quad \tau \sim \mathrm{Ga}\Big(\frac{\nu_\tau}{2}, \frac{\nu_\tau}{2}\Big).
\end{aligned}$$

Next we provide the full conditional distributions needed for Gibbs sampling. Due to conjugacy, $\boldsymbol{\eta}$, $\boldsymbol{\beta}$, σ^{-2}, $\boldsymbol{\phi}$ and τ have closed-form full conditional distributions, while ν_σ and κ are updated using Metropolis-Hastings steps within the Gibbs sampler. Let $V_\eta = (\tau\mathbf{T}^{-1} + \sigma^{-2}\boldsymbol{\Omega}^{-1})^{-1}$ and $V_\beta = (\tau\mathbf{X}'\mathbf{T}^{-1}\mathbf{X} + \Sigma_0^{-1})^{-1}$.

$$\begin{aligned}
\boldsymbol{\eta}|\mathbf{Y}, \boldsymbol{\beta}, \sigma^{-2}, \tau, \kappa, \nu_\sigma, \boldsymbol{\phi} &\sim \mathrm{N}\left(V_\eta(\tau\mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \sigma^{-2}\boldsymbol{\Omega}^{-1}\mathbf{Y}), V_\eta\right)\\
\boldsymbol{\beta}|\mathbf{Y}, \boldsymbol{\eta}, \sigma^{-2}, \tau, \kappa, \nu_\sigma, \boldsymbol{\phi} &\sim \mathrm{N}\left(V_\beta(\tau\mathbf{X}'\mathbf{T}^{-1}\boldsymbol{\eta} + \Sigma_0^{-1}\boldsymbol{\beta}_0), V_\beta\right)\\
\sigma^{-2}|\mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\beta}, \tau, \kappa, \nu_\sigma, \boldsymbol{\phi} &\sim \mathrm{Ga}\left(\frac{n}{2} + a, \frac{1}{2}\sum_{i=1}^n \phi_i(y_i - \eta_i)^2 + b\right)\\
\tau|\mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^{-2}, \kappa, \nu_\sigma, \boldsymbol{\phi} &\sim \mathrm{Ga}\left(\frac{n + \nu_\tau}{2}, \frac{1}{2}\left\{(\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta})'\mathbf{T}^{-1}(\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta}) + \nu_\tau\right\}\right)\\
\phi_i|\mathbf{Y}, \boldsymbol{\eta}, \boldsymbol{\beta}, \sigma^{-2}, \kappa, \nu_\sigma &\sim \mathrm{Ga}\left(\frac{\nu_\sigma + 1}{2}, \frac{1}{2}\{\sigma^{-2}(y_i - \eta_i)^2 + \nu_\sigma\}\right).
\end{aligned}$$

4.2 Heteroscedastic PSB mixture of normals (P2a)

We propose a Markov chain Monte Carlo algorithm, which is a hybrid of data augmentation, the exact block Gibbs sampler of Papaspiliopoulos (2008) and Metropolis-Hastings sampling. Papaspiliopoulos (2008) proposed the exact block Gibbs sampler as an efficient approach to posterior computation in DP mixture models, modifying the block Gibbs sampler of Ishwaran and James (2001) to avoid truncation approximations. The exact block Gibbs sampler combines characteristics of the retrospective sampler (Papaspiliopoulos and Roberts 2008) and the slice sampler (Walker 2007; Kalli et al. 2010). We included the label switching moves introduced by Papaspiliopoulos and Roberts (2008) for better mixing. Introduce $\gamma_1, \ldots, \gamma_n$ such that $\pi_h(\mathbf{x}_i) = P(\gamma_i = h), h = 1, 2, \ldots, \infty$. Then

$$\gamma_i \sim \sum_{h=1}^{\infty} \pi_h(\mathbf{x}_i)\delta_h = \sum_{h=1}^{\infty} 1(u_i < \pi_h(\mathbf{x}_i))\delta_h$$

where $u_i \sim U(0, 1)$. The MCMC steps are given below.

Step 1. Update u_is and stick-breaking random variables: Generate

$$u_i|- \sim U(0, \pi_{\gamma_i}(\mathbf{x}_i))$$

where $\pi_h(\mathbf{x}_i) = \Phi\{\alpha_h(\mathbf{x}_i)\}\prod_{l<h}[1 - \Phi\{\alpha_l(\mathbf{x}_i)\}]$. For $i = 1, \ldots, n$, introduce latent variables $Z_h(\mathbf{x}_i), h = 1, 2, \ldots$ such that $Z_h(\mathbf{x}_i) \sim \mathrm{N}(\alpha_h(\mathbf{x}_i), 1)$. Thus, $\pi_h(\mathbf{x}_i) = P(Z_h(\mathbf{x}_i) > 0, Z_l(\mathbf{x}_i) < 0 \,\text{for}\, l < h)$. Then

$$Z_h(\mathbf{x}_i)|- \sim \begin{cases} \mathrm{N}(\alpha_h(\mathbf{x}_i), 1)I_{\mathbb{R}^+}, h = \gamma_i \\ \mathrm{N}(\alpha_h(\mathbf{x}_i), 1)I_{\mathbb{R}^-}, h < \gamma_i. \end{cases}$$

Let $\mathbf{Z}_h = (Z_h(\mathbf{x}_1), \ldots, Z_h(\mathbf{x}_n))'$ and $\boldsymbol{\alpha}_h = (\alpha_h(\mathbf{x}_1), \ldots, \alpha_h(\mathbf{x}_n))'$. Letting $\left(\boldsymbol{\Sigma}_\alpha\right)_{ij} = \mathrm{e}^{-\kappa_\alpha||\mathbf{x}_i - \mathbf{x}_j||}$, $\mathbf{Z}_h \sim \mathrm{N}(\boldsymbol{\alpha}_h, \mathbf{I})$ and $\boldsymbol{\alpha}_h \sim \mathrm{N}(0, \frac{1}{\tau_\alpha}\boldsymbol{\Sigma}_\alpha)$,

$$\boldsymbol{\alpha}_h|- \sim \mathrm{N}\big((\tau_\alpha\boldsymbol{\Sigma}_\alpha^{-1} + \mathbf{I}_n)^{-1}\mathbf{Z}_h, (\tau_\alpha\boldsymbol{\Sigma}_\alpha^{-1} + \mathbf{I}_n)^{-1}\big)$$

Continue up to $h = 1, \ldots, h^* = \max\{h_1^*, \ldots, h_n^*\}$, where h_i^* is the minimum integer satisfying $\sum_{l=1}^{h_i^*} \pi_l(\mathbf{x}_i) > 1 - \min\{u_1, \ldots, u_n\}, i = 1, \ldots, n$. Now

$$\tau_\alpha|- \sim \mathrm{Ga}\left(\frac{1}{2}(nh^* + \nu_\alpha), \frac{1}{2}\left(\sum_{l=1}^{h^*} \boldsymbol{\alpha}_k'\boldsymbol{\Sigma}_\alpha^{-1}\boldsymbol{\alpha}_k + \nu_\alpha\right)\right),$$

while κ_α is updated using a Metropolis-Hastings step.

Step 2. Update allocation to atoms: Update $(\gamma_1, \ldots, \gamma_n)|-$ as multinomial random variables with probabilities

$$P(\gamma_i = h) \propto \mathrm{N}(y_i; \eta(\mathbf{x}_i), \tau_h^{-1})I(u_i < \pi_h(\mathbf{x}_i)), h = 1, \ldots, h^*.$$

Step 3. Update component-specific locations and precisions: Letting $n_l = \#\{i : \gamma_i = l\}, l = 1, 2, \ldots, h^*$,

$$\tau_l|- \sim \mathrm{Ga}\left(\frac{n_l}{2} + \alpha_\tau, \beta_\tau + \sum_{i:\gamma_i = l}(y_i - \eta(\mathbf{x}_i))^2\right), \quad l = 1, 2, \ldots, h^*$$

$$\beta_\tau|- \sim \mathrm{Ga}\left(1, \sum_{l=1}^{k^*}\tau_l + k_0\right).$$

Step 4. Update the mean regression function: Letting $\boldsymbol{\Lambda} = \mathrm{diag}(\tau_{\gamma_1}^{-1}, \ldots, \tau_{\gamma_n}^{-1})$,

$$\begin{aligned}
\boldsymbol{\eta}|- &\sim \mathrm{N}((\tau \mathbf{T}^{-1} + \boldsymbol{\Lambda}^{-1})^{-1}(\tau \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-1}\mathbf{Y}), (\tau \mathbf{T}^{-1} + \boldsymbol{\Lambda}^{-1})^{-1}) \\
\boldsymbol{\beta}|- &\sim \mathrm{N}\big((\tau \mathbf{X}'\mathbf{T}^{-1}\mathbf{X} + \tau \Sigma_0^{-1})^{-1}(\tau \mathbf{X}'\mathbf{T}^{-1}\boldsymbol{\eta} + \Sigma_0^{-1}\beta_0), (\tau \mathbf{X}'\mathbf{T}^{-1}\mathbf{X} + \Sigma_0^{-1})^{-1}\big) \\
\tau|- &\sim \mathrm{Ga}\left(\frac{n+\nu_\tau}{2}, \frac{1}{2}\big\{(\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta})'\mathbf{T}^{-1}(\boldsymbol{\eta} - \mathbf{X}\boldsymbol{\beta})' + \nu_\tau\big\}\right).
\end{aligned}$$

Step 5. Update κ in a Metropolis-Hastings step.

4.3 Heteroscedastic sPSB location-scale mixture (P2b)

We will need the following changes in the updating steps from the previous case.

Step 2. Update allocation to atoms: Update $(\gamma_1, \ldots, \gamma_n)|-$ as multinomial random variables with probabilities

$$\begin{aligned}
P(\gamma_i = h) \propto \frac{1}{2}&\left\{\mathrm{N}(y_i; \eta(\mathbf{x}_i) + \mu_h, \tau_h^{-1}) + \mathrm{N}(y_i; \eta(\mathbf{x}_i) - \mu_h, \tau_h^{-1})\right\} \\
I(u_i &< \pi_h(\mathbf{x}_i)),
\end{aligned}$$

$h = 1, \ldots, h^*$.

Step 3. Component-specific locations and precisions: Let $n_l = \#\{i : \gamma_i = l\}, l = 1, 2, \ldots, h^*$ and $m_l = \sum_{i:\gamma_i=l}(y_i - \eta_i)$. The atoms of the base measure location is updated from a mixture of normals as

$$\mu_l|- \sim p_l \mathrm{N}\left(\frac{\mu_0\sigma_0^{-2} + \tau_l n_l}{\sigma_0^{-2} + n_l\tau_l}, \frac{1}{\sigma_0^{-2} + n_l\tau_l}\right) + (1 - p_l)\mathrm{N}\left(\frac{\mu_0\sigma_0^{-2} - \tau_l n_l}{\sigma_0^{-2} + n_l\tau_l}, \frac{1}{\sigma_0^{-2} + n_l\tau_l}\right),$$

where $p_l \propto \exp\left\{\frac{1}{2}\left(\frac{\mu_0\sigma_0^{-2} + \tau_l n_l}{\sigma_0^{-2} + n_l\tau_l}\right)\right\}$.

$$\begin{aligned}
\tau_l|- \sim p_l \mathrm{Ga}&\left(\frac{n_l}{2} + \alpha_\tau, \beta_\tau + \sum_{i:\gamma_i=l}\{y_i - \eta(\mathbf{x}_i) - \mu_l\}^2\right) \\
&+ (1 - p_l)\mathrm{Ga}\left(\frac{n_l}{2} + \alpha_\tau, \beta_\tau + \sum_{i:\gamma_i=l}\{y_i - \eta(\mathbf{x}_i) + \mu_l\}2\right),
\end{aligned}$$

where $p_l \propto \left\{\frac{1}{\left(\beta_\tau + \frac{1}{2}\sum_{i:\gamma_i=l}\{y_i - \eta(\mathbf{x}_i) - \mu_l\}^2\right)}\right\}^{\frac{n_l}{2} + \alpha}$.

Step 4. Update the mean regression function: Let $\mathbf{\Lambda} = \mathrm{diag}(\tau_{\gamma_1}^{-1}, \ldots, \tau_{\gamma_n}^{-1})$, $\boldsymbol{\mu}^* = (\mu_{\gamma_1}, \mu_{\gamma_2}, \ldots, \mu_{\gamma_n})$ and $\mathbf{W} = \left(\tau \mathbf{T}^{-1} + \boldsymbol{\Lambda}^{-1}\right)^{-1}$. Hence,

$$\begin{aligned}\boldsymbol{\eta} | - p\mathrm{N}\Big(&\boldsymbol{\eta}; \mathbf{W}\{\tau \mathbf{T}^{-1}\mathbf{X}\beta + \boldsymbol{\Lambda}^{-1}(\mathbf{Y} - \boldsymbol{\mu}^*)\}, \mathbf{W}\Big)\\ &+(1-p)\mathrm{N}\Big(\boldsymbol{\eta}; \mathbf{W}\{\tau \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-1}(\mathbf{Y} + \boldsymbol{\mu}^*)\}, \mathbf{W}\Big)\end{aligned}$$

where $p \propto \exp\left[\frac{1}{2}\left\{(\tau \mathbf{T}^{-1}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-1}(\mathbf{Y} - \boldsymbol{\mu}^*))'\mathbf{W}\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\Lambda}^{-1}(\mathbf{Y} - \boldsymbol{\mu}^*) - (\mathbf{Y} - \boldsymbol{\mu}^*)'\boldsymbol{\Lambda}^{-1}(\mathbf{Y} - \boldsymbol{\mu}^*)\right\}\right]$.

5 Measures of influence

There has been limited work on sensitivity of the posterior distribution to perturbations of the data and outliers. Arellano-Vallea et al. (2000) use deletion diagnostics to assess sensitivity, but their methods are computationally expensive in requiring posterior computation with and without data deleted. Weiss (1996) proposed an alternative that perturbs the posterior instead of the likelihood, and only requires samples from the full posterior. Following Weiss (1996), let $f(y_i | \tilde{\Theta}, \mathbf{x}_i)$ denote the likelihood of the data y_i, define

$$\delta_i^*(\tilde{\Theta}) = \frac{f(y_i + \delta | \tilde{\Theta}, \mathbf{x}_i)}{f(y_i | \tilde{\Theta}, \mathbf{x}_i)},$$

for some small $\delta > 0$ and let $p_i(\tilde{\Theta}|\mathbf{Y})$ denote a new perturbed posterior,

$$p_i(\tilde{\Theta}|\mathbf{Y}) = \frac{p(\tilde{\Theta}|\mathbf{Y})\delta_i^*(\tilde{\Theta})}{\mathsf{E}(\delta_i^*(\tilde{\Theta})|\mathbf{Y})}.$$

Since the responses are normalized prior to analysis, it is reasonable to assume that the perturbation is less than 0.1. We vary δ in [0.01, 0.1] over a grid of width 0.01 and obtain the average of results. Denote by L_i the influence measure, which is a divergence measure between the unperturbed posterior $p(\tilde{\Theta}|\mathbf{Y})$ and the perturbed posterior $p_i(\tilde{\Theta}|\mathbf{Y})$,

$$L_i = \frac{1}{2}\int |p(\tilde{\Theta}|\mathbf{Y}) - p_i(\tilde{\Theta}|\mathbf{Y})| \mathrm{d}\tilde{\Theta}.$$

L_i is bounded and takes values in [0, 1]. When $p(\tilde{\Theta}|\mathbf{Y}) = p_i(\tilde{\Theta}|\mathbf{Y})$, $L_i = 0$ indicates that the perturbation δ_i^* has no influence. On the other hand, if $L_i = 1$, the supports of $p(\tilde{\Theta}|\mathbf{Y})$ and $p_i(\tilde{\Theta}|\mathbf{Y})$ are disjoint indicating maximum influence. We can define an influence measure as $L = \frac{1}{n}\sum_{i=1}^{n} L_i$. Clearly L also takes values in [0, 1] with $L = 0 \Rightarrow L_i = 0 \,\forall\, i = 1, 2, \ldots, n$. Also $L = 1 \Rightarrow L_i = 1 \,\forall\, i = 1, 2, \ldots, n$. Weiss

(1996) provided a sample version of $L_i, i = 1, \ldots, n$. Letting $\tilde{\Theta}_1, \ldots, \tilde{\Theta}_M$ be the posterior samples with B the burn-in,

$$\hat{L}_i = \frac{1}{M-B} \sum_{k=B+1}^{M} \frac{1}{2} \left| \frac{\delta_i^*(\tilde{\Theta}_k)}{\hat{E}(\delta_i^*(\tilde{\Theta}))} - 1 \right|,$$

where $\hat{\mathsf{E}}\{\delta_i^*(\tilde{\Theta})\} = \frac{1}{M-B} \sum_{k=B+1}^{M} \delta_i^*(\tilde{\Theta}_k)$. Our estimated influence measure is $\hat{L} = \frac{1}{n} \sum_{i=1}^{n} \hat{L}_i$. We will calculate the influence measure for our proposed methods and compare their sensitivity.

6 Simulation studies

To assess the performance of our proposed approaches, we consider a number of simulation examples, (i) linear model, homoscedastic error with no outliers, (ii) linear model, homoscedastic error with outliers (iii) linear model, heteroscedastic errors and outliers, (iv) non-linear model with heteroscedastic errors and outliers and (v) non-linear model with heteroscedastic errors and outliers, but with fewer true predictors. We let the heaviness of the tails and error variance change with $\mathbf{x}$ in cases (iii)–(v). We considered the following methods of assessing the performance, namely, mean squared prediction error (MSPE), coverage of 95 % prediction intervals, mean integrated squared error (MISE) in estimating the regression function at the points for which we have data, point wise coverage of 95 % credible intervals for the regression function and the influence measure ($\hat{L}$) as described in Sect. 5. We also consider a variety of sample sizes in the simulation, n=30, 60, 80 and simulate 10 covariates independently from $U(0, 1)$. Let $\mathbf{z}$ be 10-dim vector of i.i.d $U(0, 1)$ random variables independent of the covariates.

Generation of errors in heteroscedastic case and outliers: Let $f_{\mathbf{x}_i}(\epsilon_i) = p_{\mathbf{x}_i}\mathrm{N}(\epsilon_i; 0, 1) + q_{\mathbf{x}_i}\mathrm{N}(\epsilon_i; 0, 5)$ where $p_{\mathbf{x}_i} = \Phi(\mathbf{x}_i'\mathbf{z})$. The outliers are simulated from the model with error distribution $f_{\mathbf{x}_i}^{\mathrm{O}}(\cdot)$, which is a mixture of truncated normal distributions as follows. In the heteroscedastic case, $f_{\mathbf{x}_i}^{\mathrm{O}}(\epsilon_i) = p_{\mathbf{x}_i}\mathrm{TN}_{(-\infty,3)\cup(3,\infty)}(\epsilon_i; 0, 1) + q_{\mathbf{x}_i}\mathrm{TN}_{(-\infty,-3\sqrt{5})\cup(3\sqrt{5},\infty)}(\epsilon_i; 0, 5)$, where $\mathrm{TN}_{\mathcal{R}}(\cdot; \mu, \sigma^2)$ denotes a truncated normal distribution with mean μ and standard deviation σ over the region $\mathcal{R}$. We consider the following five cases:

Case (i): $y_i = 2.3 + 5.7x_{1i} + \epsilon_i, \epsilon_i \sim \mathrm{N}(0, 1)$ with no outliers.
Case (ii): $y_i = 2.3 + 5.7x_{1i} + \epsilon_i, \epsilon_i \sim 0.95\mathrm{N}(0, 1) + 0.05\mathrm{N}(0, 10)$.
Case (iii): $y_i = 1.2 + 5.7x_{1i} + 4.7x_{2i} + 0.12x_{3i} - 8.9x_{4i} + 2.4x_{5i} + 3.1x_{6i} + 0.01x_{7i} + \epsilon_i, \epsilon_i \sim f_{\mathbf{x}_i}$, with 5 % outliers generated from $f_{\mathbf{x}_i}^{\mathrm{O}}(\epsilon_i)$.
Case (iv): $y_i = 1.2 + 5.7x_{1i} + 3.4x_{1i}^2 + 4.7x_{i2} + 0.89x_{i2}^2 + 0.12x_{i3} - 8.9x_{i4}x_{i8} + 2.4x_{i5}x_{i9} + 3.1x_{i6} + x_{i6}^2 + 0.01x_{i7} + \epsilon_i, \epsilon_i \sim f_{\mathbf{x}_i}$ with 5 % outliers generated from $f_{\mathbf{x}_i}^{\mathrm{O}}(\epsilon_i)$.
Case (v): $y_i = 1.2 + 5.7\sin x_{1i} + 3.4\exp(x_{2i}) + 4.7\log|x_{i3}| + \epsilon_i, \epsilon_i \sim f_{\mathbf{x}_i}$ with 5 % outliers generated from $f_{\mathbf{x}_i}^{\mathrm{O}}(\epsilon_i)$.

For each of the cases and for each sample size n, we took the first $\frac{n}{2}$ samples as the training set and the next $\frac{n}{2}$ samples as the test set. We also compare the MSPE of the proposed methods with robust regression using M-estimation (Huber 1964), Bayesian additive regression trees (Chipman et al. 2010), and Treed GP (Gramacy and Lee 2008). The MCMC algorithms described in Sect. 5 are used to obtain samples from the posterior distribution. The results for model P1 given here are based on 20,000 samples obtained after a burn-in period of 3,000. The results for models P2a–d are based on 20,000 samples obtained after a period of 7,000. Rapid convergence was observed based on diagnostic tests of Geweke (1992) and Raftery and Lewis (1992). In addition, the mixing was very good for model P1. For models P2a–d, we use the label switching moves by Papaspiliopoulos and Roberts (2008), which lead to adequate mixing. Tables 1, 2 and 3 summarize the performance of all the methods based on 50 replicated datasets.

Tables 1, 2 and 3 clearly show that in small samples both of the heteroscedastic methods (P2a and P2b) have substantially reduced MSPE and MISE relative to the heavy-tailed parametric error model in most of the cases, interestingly even in the homoscedastic cases. This may be because discrete mixture of Gaussians better approximate a single normal than a t-distribution in small samples. Methods P2a and P2b also did a better job than method P1 in allowing uncertainty in estimating the mean regression and predicting the test sample observations. The homoscedastic versions 4 and 5 perform better than the parametric models but worse than the heteroscedastic models. In some cases, the heavy-tailed t-residual distribution results in overly conservative predictive and credible intervals. As seen from the value of the influence statistic, the heteroscedastic PSB process mixtures result in more robust inference compared to the parametric error model, the sPSB process mixture of normals being more robust than the symmetric and unimodal version. As the sample size increases, the difference in the predictive performances between the parametric and the nonparametric models is reduced and in some cases the parametric error model performs as well as the nonparametric approaches, which is as expected given the Central Limit Theorem.

Table 1 shows that in the simple linear model with normal homoscedastic errors, all the models perform similarly in terms of mean squared prediction error, though the methods P2a and P2b are somewhat better than the rest. Also, in estimating the mean regression function in case (i), methods P2a and P2b performed better than all the other methods. In case (ii) (Table 1), methods P2a and P2b are most robust in terms of estimation and prediction in the presence of outliers. However, there is no significant difference between methods P2a and P2b and methods P2c and P2d in cases (i) and (ii). In cases (iii) and (iv), when the residual distribution is heteroscedastic, methods P2a and P2b perform significantly better than the parametric model P1 and the homoscedastic models P2c and P2d in both estimation and prediction, since the heteroscedastic PSB mixture is very flexible in modeling the residual distribution. This is quite evident from the MSPE values under cases (iii) and (iv) in Table 2. Huber's M-estimation method performs similar to methods P2a–d in cases (i) and (ii) but did not do as well in estimation and prediction in cases (iii) and (iv) when the underlying mean function is actually non-linear with heteroscedastic residual distribution. Also BART failed to perform well in estimating the mean function in small samples

Table 1 Simulation results under homoscedastic residuals (cases i and ii)

	Case i					Case ii				
	MSPE	cov(y)[a]	MISE	cov(η)[b]	L	MSPE	cov(y)	MISE	cov(η)	L
n=40										
P1	0.2997	1	0.0248	1	0.0017	0.6043	1	0.0232	1	0.0027
P2a	0.2821	0.9980	0.0141	1	0.0015	0.5983	0.9740	0.0173	1	0.0019
P2b	0.2798	1	0.0144	1	0.0015	0.5987	0.9745	0.0169	1	0.0017
P2c	0.2980	1	0.0156	1	0.0016	0.5980	0.9750	0.0189	1	0.0020
P2d	0.2869	1	0.0155	1	0.0016	0.5983	0.9750	0.0190	1	0.0020
M-estimation	0.2820		0.0140			0.6013		0.0177		
BART	0.3510	0.6866	0.0714			0.7051	0.7845	0.0950		
Treed GP	0.3042	0.9134	0.0256			0.6968	0.9365	0.0803		
n=60										
P1	0.2990	1	0.0246	1	0.0019	0.5776	1	0.0242	1	0.0023
P2a	0.2769	0.9947	0.0103	1	0.0017	0.5471	0.95	0.0143	0.97	0.0016
P2b	0.2752	0.9963	0.0104	1	0.0016	0.5541	0.95	0.0141	0.98	0.0016
P2c	0.2852	0.9945	0.0176	1	0.0019	0.5664	0.95	0.0142	0.99	0.0021
P2d	0.2826	0.9960	0.0173	1	0.0018	0.5561	0.95	0.0141	0.98	0.0021
M-estimation	0.2759		0.0103			0.5623		0.0139		
BART	0.3314	0.6753	0.0539			0.6725	0.7777	0.1098		
Treed GP	0.3000	0.9193	0.0218			0.6880	0.9301	0.1198		
n=80										
P1	0.2913	1	0.0252	1	0.0021	0.5583	1	0.0172	1	0.0022
P2a	0.2592	0.9940	0.0086	1	0.0021	0.4989	0.97	0.0050	1	0.0014
P2b	0.2574	0.9956	0.0069	1	0.0020	0.4898	0.98	0.0067	1	0.0010
P2c	0.2724	0.9976	0.0187	1	0.0020	0.5104	0.98	0.0103	1	0.0017
P2d	0.2716	0.9976	0.0189	1	0.0020	0.5002	0.98	0.0097	1	0.0018
M-estimation	0.2720		0.0079			0.5431		0.0068		
BART	0.3128	0.6525	0.0437			0.6509	0.7815	0.1098		
Treed GP	0.2886	0.9301	0.0175			0.6532	0.9224	0.1031		

[a] cov(y) denotes the coverage of the 95 % predictive intervals of the test cases

[b] cov(η) denotes the coverage of the 95 % credible intervals of the mean regression function

in these cases. On the other hand, GP-based approaches perform quite well in these cases in estimating the regression function with methods P2a and P2b performing better than the rest. Treed GP performed close to method P1 in estimation and prediction as both the methods are based on GP priors on the mean function and have a parametric error distribution. In not allowing heteroscedastic error variance, BART and Treed GP underestimate uncertainty in prediction, leading to overly narrow predictive intervals.

In case (v) (Table 3), where the true model is generated using comparatively less number of true signals, BART performed slightly better in terms of prediction than

Table 2 Simulation results under heteroscedastic residuals (Cases iii and iv)

	Case iii					Case iv				
	MSPE	cov(y)	MISE	cov(η)	L	MSPE	cov(y)	MISE	cov(η)	L
$n=40$										
P1	0.4833	1	0.3612	1	0.0027	0.4416	1	0.3274	1	0.0029
P2a	0.2570	0.9990	0.1394	1	0.0025	0.2783	0.9923	0.1583	0.98	0.0023
P2b	0.2586	0.9990	0.1298	1	0.0025	0.2712	0.9867	0.1501	0.97	0.0017
P2c	0.3057	1	0.2412	1	0.0026	0.3334	0.9967	0.2213	0.99	0.0024
P2d	0.3024	1	0.2304	1	0.0026	0.3216	0.9967	0.2192	0.99	0.0022
M-estimation	0.2613		0.1376			0.2889		0.1663		
BART	0.4639	0.8444	0.3413			0.4103	0.8833	0.2675		
Treed GP	0.3320	0.7834	0.1979			0.3548	0.8268	0.2108		
$n=60$										
P1	0.2254	1	0.1154	1	0.0023	0.2367	1	0.1067	1	0.0021
P2a	0.1744	0.9973	0.0572	1	0.0020	0.2178	1	0.0562	0.97	0.0019
P2b	0.1712	0.9878	0.0567	1	0.0016	0.2099	1	0.0656	0.98	0.0017
P2c	0.1952	0.9998	0.0854	1	0.0021	0.2216	1	0.0879	0.99	0.0020
P2d	0.1934	0.9998	0.0799	1	0.0020	0.2208	1	0.0812	0.99	0.0020
M-estimation	0.1746		0.0564			0.2125		0.0678		
BART	0.3429	0.8546	0.2217			0.3385	0.9122	0.1799		
Treed GP	0.2047	0.8349	0.0779			0.2611	0.8867	0.0899		
$n=80$										
P1	0.1636	1	0.0454	1	0.0018	0.1855	1	0.0346	1	0.0019
P2a	0.1509	0.9976	0.0373	0.95	0.0015	0.1653	1	0.0321	0.9952	0.0014
P2b	0.1578	0.9931	0.0324	1	0.0013	0.1614	1	0.0312	0.9932	0.0010
P2c	0.1589	0.9960	0.0404	1	0.0017	0.1774	1	0.0329	0.9980	0.0016
P2d	0.1567	0.9969	0.0401	1	0.0017	0.1770	1	0.0320	0.9969	0.0016
M-estimation	0.1582		0.0364			0.1832		0.0325		
BART	0.2284	0.9265	0.1098			0.2491	0.9490	0.1083		
Treed GP	0.1655	0.8876	0.0427			0.2022	0.8923	0.0548		

other methods in small samples. However, as the sample size increased, BART performed poorly while the GP prior on the mean can accommodate the non-linearity resulting in substantially better predictive performances.

7 Applications

7.1 Boston housing data application

To compare our proposed approaches to alternatives, we applied the methods to a commonly used data set from the literature, the Boston housing data. The response

Table 3 Simulation results under heteroscedastic residuals (Case v)

	MSPE	cov(y)	MISE	cov(η)	L
$n=40$					
P1	0.6666	0.9800	0.5856	1	0.0033
P2a	0.5233	0.9770	0.3980	0.9812	0.0025
P2b	0.5231	0.9854	0.3745	0.9765	0.0019
P2c	0.5875	0.9850	0.4452	1	0.0029
P2d	0.5788	0.9859	0.4223	1	0.0028
M-estimation	0.5531		0.3671		
BART	0.4956	0.8980	0.4013		
Treed GP	0.7224	0.8123	0.6132		
$n=60$					
P1	0.3828	1	0.2911	0.9985	0.0031
P2a	0.3745	0.9832	0.2617	0.9840	0.0022
P2b	0.3767	0.9812	0.2601	0.9867	0.0020
P2c	0.3810	0.9900	0.2800	0.9998	0.0027
P2d	0.3800	0.9906	0.2798	0.9998	0.0026
M-estimation	0.3939		0.2824		
BART	0.3930	0.9313	0.2668		
Treed GP	0.4225	0.9023	0.3217		
$n=80$					
P1	0.3599	0.9901	0.2759	0.9998	0.0029
P2a	0.3503	0.9762	0.2582	0.9765	0.0022
P2b	0.3519	0.9712	0.2545	0.9715	0.0019
P2c	0.3560	0.9856	0.2656	0.9885	0.0025
P2d	0.3557	0.9800	0.2677	0.9881	0.0024
M-estimation	0.3905		0.2887		
BART	0.3594	0.9442	0.2867		
Treed GP	0.4489	0.9125	0.3509		

is the median value of the owner-occupied homes (measured in \$1,000) in 506 census tracts in the Boston area, and there are 13 predictors (12 continuous, 1 binary) that might help explain the variation in the median value across tracts. We predict the median value of the owner-occupied homes of which the first 253 is taken as the training set and the remaining 253 as the test set. Sample predictive performance of our three methods is compared to competitors in Table 4. The parametric model P1, and the mixture models P2a–d and the M-estimation methods perform very closely to each other in terms of prediction and did better than BART and Treed GP. Methods P1 and P2a even perform slightly better than methods P2b, P2c and P2d. As in the simulation examples, BART and Treed GP underestimate the uncertainty in prediction. On the other hand, the predictive intervals of the methods P1, P2a–d are more conservative and accommodate uncertainty in

Table 4 Boston housing data and body fat data results

Boston housing data					Body fat data			
Methods	MSPE	cov(y)	L	corr(Ytest, Ypred)[a]	MSPE	cov(y)	L	corr(Ytest, Ypred)
P1	0.0012	0.99	0.0034	0.9894	0.0055	1	0.0020	0.9972
P2a	0.0013	0.99	0.0027	0.9901	0.0031	1	0.0017	0.9984
P2b	0.0016	0.99	0.0020	0.9863	0.0029	1	0.0017	0.9989
P2c	0.0014	0.99	0.0030	0.9879	0.0034	1	0.0019	0.9969
P2d	0.0013	0.99	0.0029	0.9881	0.0032	1	0.0018	0.9975
M-estimation	0.0016			0.9858	0.0375			0.9710
BART	0.0024	0.92		0.9836	0.0355	0.95		0.9655
Treed GP	0.0053	0.91		0.9524	0.1526	0.98		0.9250

[a] corr(Ytest, Ypred) denotes the sample correlation between the test and predicted y

predicting regions with outliers quite flexibly. Also the model P2b appears to be more robust compared to models P1, P2a, P2c & P2d in terms of the influence measure.

7.2 Body fat data application

With the increasing trend in obesity and concerns about associated adverse health effects, such as heart disease and diabetes, it has become even more important to obtain accurate estimates of body fat percentage. It is well known that body mass index, which is calculated based only on weight and height, can produce a misleading measure of adiposity as it does not take into account muscle mass or variability in frame size. As a gold standard for measuring percentage of body fat, one can rely on under water weighing techniques, and age and body circumference measurements have also been widely used as additional predictors. We consider a commonly used data set from Statlib (http://lib.stat.cmu.edu/datasets/bodyfat), which contains the following 15 variables: percentage of body fat (%), body density from underwater weighing (gm/cm^3), age (year), weight (lbs.), height (inches), and ten body circumferences (neck, chest, abdomen, hip, thigh, knee, ankle, biceps, forearm, wrist, all in *cm*). Percentage of body fat is given from Siri's (1956) equation:

$$\text{Percentage of body fat} = \frac{495}{\text{Density}} - 450$$

We predict the percentage of body fat (%) taking the first 126 as the training set and the remaining 126 as the test set. We summarize the predictive performances in Table 4.

Table 4 suggests that the nonparametric regression procedures with heteroscedastic residual distribution P2a and P2b perform better than the parametric models P1 and models P2c and P2d, BART, M-estimation and Treed GP in predicting the percentage of body fat.

8 Discussion

We have developed a novel regression model that can accommodate a large range of non-linearity in the mean function and at the same time can flexibly deal with outliers and heteroscedasticity. Based on preliminary simulation results, it appears that our methods P2a and P2b can outperform contemporary nonparametric regression methods, such as Huber's M-estimation method, BART and treed GP. We also provide theoretical support for the proposed methodology when both the mean and the residuals are modeled nonparametrically.

One possible future direction is to relax the symmetry assumption on the residual distribution and introduce a model for median regression based on conditional PSB mixtures for allowing possibly asymmetric residual densities constrained to have zero median. Conditional DP mixtures are well known in the literature (Doss 1985; Burr and Doss 2005) and it is certainly interesting to extend our approach via a conditional PSB. In that way, we can hope to obtain a more robust estimate of the regression function. It is challenging to extend our theoretical results to conditional PSB and develop a fast algorithm for computation. Another possible theoretical direction is to prove posterior consistency using heteroscedastic mixtures. Currently, we only have results for the homoscedastic PSBP mixture.

Appendix: proofs of main results

Proof of Lemma 1 It follows from Chu (1973) that

$$f \in \mathcal{C}_m \Leftrightarrow f(x) = \int \sigma^{-1}\phi(\sigma^{-1}x)g(\sigma)\mathrm{d}\sigma$$

for some density g on $\mathbb{R}^+$. Recall from Ongaro and Cattaneo (2004) that a collection of random weights $\{\pi_h\}_{h=1}^{\infty}$ with $\sum_{h=1}^{\infty}\pi_h = 1$ a.s. is said to have a full support if for any $m \geq 1$, $(\pi_1, \ldots, \pi_m)$ admits a positive joint density with respect to Lebesgue measure on the simplex $\{(p_1, \ldots, p_m) : \sum_{i=1}^{m} p_i \leq 1\}$. Ongaro and Cattaneo (2004) showed that if π_hs have a full support, the weak support of

$$P = \sum_{h=1}^{\infty}\pi_h\delta_{\theta_h}, \theta_h \sim G_0$$

is the set of all probability measures whose support is contained in the support of G_0. Since

$$(\pi_1, \ldots, \pi_m) \overset{d}{=} \left(\Phi(\alpha_1), \Phi(\alpha_2)\{1 - \Phi(\alpha_1)\}, \ldots, \Phi(\alpha_m)\prod_{i=1}^{m-1}\{1 - \Phi(\alpha_i)\}\right),$$
$$\alpha_i \sim \mathrm{N}(\mu_\alpha, \sigma_\alpha^2),$$

π_hs have a full support and hence the weak support of $P = \sum_{h=1}^{\infty} \pi_h \delta_{\tau_h}$ defined in (7) is all probability measures on $\mathbb{R}^+$. It follows that the weak support of the induced prior Π_u on $\mathcal{S}_u$, denoted by $wk(\Pi_u)$, is precisely $\mathcal{C}_m$. □

Proof of Lemma 2 It follows from Tokdar (2006) that if we can show that the weak support of Π_s contains all probability measures symmetric about zero and having compact support, then $f \in \tilde{S}_s \Rightarrow f \in KL(\Pi_s)$. The argument given in Lemma 1 shows that the weak support of the PSB prior in (4) is the set of all probability measures on $\mathbb{R} \times \mathbb{R}^+$. Now we will show that an arbitrary $\tilde{P}^s$ is in a weak neighborhood of P^s if $\tilde{P}$ is in a weak neighborhood of P. We state a lemma to prove our claim. □

Lemma 4 *Let $\tilde{P}_n$ be a sequence of probability measures and $\tilde{P}$ be a fixed probability measure. Then $(\tilde{P}_n \Rightarrow \tilde{P}) \Rightarrow (\tilde{P}_n^s \Rightarrow \tilde{P}^s)$, with $\tilde{P}_n^s$ and $\tilde{P}^s$ the symmetrized versions of $\tilde{P}_n$ and $\tilde{P}$, respectively, where the symmetrizing operation is as defined in* (9).

Proof Assume $\tilde{P}_n \Rightarrow \tilde{P}$. We have to show that for any bounded function ϕ on $\mathbb{R} \times \mathbb{R}^+$,

$$\int \phi(t,\tau)\mathrm{d}\tilde{P}_n^s(t,\tau) \rightarrow \int \phi(t,\tau)\mathrm{d}\tilde{P}^s(t,\tau) \text{ as } n \rightarrow \infty.$$

Now,

$$\begin{aligned}\int \phi(t,\tau)\mathrm{d}\tilde{P}_n^s(t,\tau) &= \frac{1}{2}\int \phi(t,\tau)\mathrm{d}\tilde{P}_n(t,\tau) + \frac{1}{2}\int \phi(t,\tau)\mathrm{d}\tilde{P}_n(-t,\tau)\\ &= \int \frac{1}{2}\big\{\phi(t,\tau) + \phi(-t,\tau)\big\}\mathrm{d}\tilde{P}_n(t,\tau).\end{aligned}$$

Since $\psi(t,\tau) = \frac{1}{2}\big\{\phi(t,\tau) + \phi(-t,\tau)\big\}$ is also a bounded continuous function and $\tilde{P}_n \Rightarrow \tilde{P}$,

$$\begin{aligned}&\int \frac{1}{2}\big\{\phi(t,\tau) + \phi(-t,\tau)\big\}\mathrm{d}\tilde{P}_n(t,\tau) \rightarrow \int \frac{1}{2}\big\{\phi(t,\tau) + \phi(-t,\tau)\big\}\mathrm{d}\tilde{P}(t,\tau)\\ &= \int \phi(t,\tau)\mathrm{d}\tilde{P}^s(t,\tau)\end{aligned}$$

as $n \rightarrow \infty$. This completes the proof of Lemma 4. □

Lemma 4, in fact, shows that the weak support of Π_s contains all probability measures symmetric about zero. With an appeal to Tokdar (2006), $f \in \tilde{S}_s \Rightarrow f \in KL(\Pi_s)$. □

Proof of Theorem 1 To prove the theorem, we need the following variant of Theorem 2.1 of Amewou-Atisso et al. (2003) and Theorem 1 of Choi and Schervish (2007) which we state as Lemma 5. Existence of exponentially consistent tests is a typical tool in showing strong consistency. □

Definition 2 Let $\mathcal{W} \subset \tilde{\mathcal{S}}_s \times \mathcal{F}$. A sequence of test functions $\Phi_n\big(\{y_i, \mathbf{x}_i\}_{i=1}^n\big)$ is said to be exponentially consistent for testing

$$H_0 : (f,\eta) = (f_0,\eta_0) \text{ against } H_1 : (f,\eta) \in \mathcal{W}_n$$

if there exist constants $C_1, C_2, C > 0$ such that

1. $\mathsf{E}_{\prod_{i=1}^n f_{0i}}(\Phi_n) \le C_1 e^{-nC}$,
2. $\inf_{(f,\eta)\in\mathcal{W}_n} \mathsf{E}_{\prod_{i=1}^n f_{\eta i}}(\Phi_n) \ge 1 - C_2 e^{-nC}$.

Lemma 5 *Let $\tilde{\Pi} = (\Pi_s \times \pi)$ be the prior on $\tilde{S}_s \times \mathcal{F}$. Let U_n be a sequence of subsets of $\tilde{S}_s \times \mathcal{F}$. Suppose that there exist test functions $\{\Phi_n\}_{n=1}^{\infty}$, sets $\Theta_n \subset \tilde{S}_s \times \mathcal{F}, n \ge 1$ and constants $C_1, C_2, c_1, c_2 > 0$ such that*

1. $\sum_{n=1}^{\infty} E_{\prod_{i=1}^n f_{0i}} \Phi_n < \infty$.
2. $\sup_{(f,\eta)\in U_n^c \cap \Theta_n} E_{\prod_{i=1}^n f_{\eta i}}(1 - \Phi_n) \le C_1 e^{-c_1 n}$.
3. $\tilde{\Pi}(\Theta_n^c) \le C_2 e^{-c_2 n}$.
4. *For all $\delta > 0$ and for almost every data sequence $\{y_i, \mathbf{x}_i\}_{i=1}^{\infty}$,*

$$\tilde{\Pi}\left\{(f,\eta) : K_i(f,\eta) < \delta \ \forall i, \sum_{i=1}^{\infty} \frac{V_i(f,\eta)}{i^2} < \infty\right\} > 0.$$

Then $\tilde{\Pi}\{(f,\eta) \in U_n^c \mid (Y_1, \mathbf{x}_1), \ldots, (Y_n, \mathbf{x}_n)\} \to 0$ a.s.$[P_{f_0,\eta_0}]$.

In this case $U_n = \mathcal{W}_n = \mathcal{U} \times S_n(f_0, \Delta) \,\forall n \ge 1$. As in Vaart and Zanten (2009), we construct $\Theta_n = \mathcal{F} \times \Theta_{1n}$ where $\Theta_{1n} = \cup_{a<r_n} M_n \mathbb{H}_1^a + \epsilon \mathbb{B}_1$ where $\mathbb{H}_1$ and $\mathbb{B}_1$ are unit ball of the RKHS of W^a and unit ball of the Banach space of $C[0,1]^p$, respectively, r_n, M_n are increasing sequences to be chosen later. The nth test is constructed by combining a collection of tests, one for each of the finitely many elements of Θ_n. It follows from the proof of Theorem 3.1 in Vaart and Zanten (2009) that under Assumption 1, there exist constants $d_1, d_2, K > 0$ such that

1. $\tilde{\Pi}(\Theta_n^c) \le \exp\{-d_1 r_n^p \log^q(r_n)\} + \exp\{-M_n^2/8\}$.
2. $\log N(\epsilon, \Theta_{1n}, ||\cdot||_\infty) \le K r_n^p \left(\log \frac{M_n}{\epsilon}\right)^{p+1}$.

Choosing $M_n = O(n^{1/2})$, $r_n^p = O(n/(\log n)^{p+2})$, we observe that

1. $\tilde{\Pi}(\Theta_n^c) \le \exp\{-d_2 n\}$.
2. $\log N(\epsilon, \Theta_{1n}, ||\cdot||_\infty) = o(n)$.

for some constant $d_2 > 0$.

To verify 1 and 2 of Lemma 5, we will write $\mathcal{W}_n$ as a disjoint union of two easily tractable regions. The particular form of $\mathcal{W}_n$ that is of interest to us is $\mathcal{W}_{1n} \cup \mathcal{W}_{2n}$, where for any $\Delta > 0$,

$$\mathcal{W}_{1n} = \mathcal{U}^c \times \left\{\eta : ||\eta - \eta||_{1,n} \le \Delta\right\} \quad \mathcal{W}_{2n} = \left\{(f,\eta) : ||\eta - \eta||_{1,n} > \Delta\right\}.$$

We will establish the existence of a consistent sequence of tests for each of these regions by considering the following variants of Propositions 3.1 and 3.3 of Amewou-Atisso et al. (2003).

Proposition 1 *There exists an exponentially consistent sequence of tests for*

$$H_0 : (f, \eta) = (f_0, \eta_0) \text{ against } H_1 : (f, \eta) \in \mathcal{W}_{2n} \cap \Theta_n.$$

Proof Let $0 < t < \Delta/2$ and assume $N_t = N(t, \Theta_{1n}, || \cdot ||_\infty)$. Let $\eta^1, \ldots, \eta^{N_t} \in \Theta_{1n}$ be such that for each $\eta \in \Theta_{1n}$ there exists j such that $||\eta - \eta^j||_\infty < t$. If $||\eta - \eta_0||_{1,n} > \Delta$, $||\eta^j - \eta_0||_{1,n} > \Delta/2$. It follows from Lemma 3.2 of Amewou-Atisso et al. (2003) that there exist a set A_i^j and a constant $C > 0$ depending on f_0 such that $\alpha_i^j := P_{f_{0i}}(A_i^j) \leq \frac{1}{2} - C|\eta^j(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)|$ and $\gamma_i^j := P_{f_{\eta^j i}}(A_i) \geq \frac{1}{2}$. If $i \leq n$ and $i \notin K_n$, set $A_i = \mathbb{R}$, so that $\alpha_i^j = \gamma_i^j = 1$. Thus,

$$\liminf_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} (\gamma_i^j - \alpha_i^j) \geq C\Delta/2.$$

From Lemma 3.1 and Lemma 3.2 of Amewou-Atisso et al. (2003), it follows that there exist test functions Φ_n^j based on $\{I_{A_i^j}, i = 1, \ldots, n\}$ such that $\mathrm{E}_{\prod_{i=1}^n f_{0i}} \Phi_n^j < \mathrm{e}^{-nC_1}$ and $\mathrm{E}_{\prod_{i=1}^n f_{\eta^j i}}(1 - \Phi_n^j) < \mathrm{e}^{-nC_2}$ for constants $C_1, C_2 > 0$. Now define $\Phi_n = \max_{1\leq j\leq N_t} \Phi_n^j$. Then

$$\mathrm{E}_{\prod_{i=1}^n f_{0i}} \Phi_n \leq \sum_{j=1}^{N_t} \mathrm{E}_{\prod_{i=1}^n f_{0i}} \Phi_n^j \leq \sum_{j=1}^{N_t} \mathrm{e}^{-nC_1} \leq N_t \mathrm{e}^{-nC_1} \leq \mathrm{e}^{-nC_3}.$$

for some constant $C_3 > 0$. Clearly $\sum_{n=1}^{\infty} \mathrm{E}_{\prod_{i=1}^n f_{0i}} \Phi_n < \infty$. □

Next we consider the type II error probability. The type II error probability of Φ_n is no larger than the type II error probability of any of the $\{\Phi_n^j, j = 1, \ldots, N_t\}$ and hence exponentially small. □

Proposition 2 *There exists an exponentially consistent sequence of tests for*

$$H_0 : (f, \eta) = (f_0, \eta_0) \text{ against } H_1 : (f, \eta) \in \mathcal{W}_{1n}.$$

Proof Without loss of generality take

$$\mathcal{U} = \left\{ f : \int \Phi(y) f(y) \mathrm{d}y - \int \Phi(y) f_0(y) \mathrm{d}y < \epsilon \right\}$$

where $0 \leq \Phi \leq 1$ and Φ is Lipschitz continuous. Hence there exists $M > 0$ such that $|\Phi(y_1) - \Phi(y_2)| < M|y_1 - y_2|$. Set $\tilde{\Phi}_i(y) = \Phi\{y - \eta_0(\mathbf{x}_i)\}$. Notice that $\mathsf{E}_{f_{0i}} \tilde{\Phi}_i = \mathsf{E}_{f_0} \Phi$. Now

$$
\begin{aligned}
\mathsf{E}_{f_{\eta i}}\tilde{\Phi}_i &= \int \tilde{\Phi}_i(y) f_{\eta i}(y)\mathrm{d}y = \int \Phi(y) f[y - \{\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)\}] \\
&\geq \int \Phi[y - \{\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)\}] f[y - \{\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)\}]\mathrm{d}y \\
&\quad - \int \Big| \Phi(y) - \Phi[y - \{\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)\}] \Big| f[y - \{\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)\}]\mathrm{d}y \\
&\geq \int \Phi(y) f(y)\mathrm{d}y - M|\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)| \\
&\geq \mathsf{E}_{f_0}\Phi + \epsilon - M|\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)|.
\end{aligned}
$$

Hence $1/n \sum_{i=1}^{n} \mathsf{E}_{f_{\eta i}}\tilde{\Phi}_i \geq E_{f_0}\Phi + \epsilon - M\Delta$ for any $f \in U^c$. Now choosing $\Delta < \epsilon/M$ and applying Lemma 3.1 of Amewou-Atisso et al. (2003) we complete the proof. □

It remains to verify the second sufficient condition of Theorem 1. Under the assumptions, it follows from Lemma 2 that $f_0 \in KL(\Pi_s)$. We will present an important lemma which is similar to Lemma 5.1 of Tokdar (2006). It guarantees that $K(f_0, f_\theta)$ and $V(f_0, f_\theta)$ are continuous at $\theta = 0$. First we state and prove some properties of the prior Π_s described in (9) which will be used to prove the lemma.

Lemma 6 *If Π_s is the prior described in* (9) *and $P_0(t,\tau) = N(t;\mu_0,\sigma_0^2) \times Ga(\tau;\alpha_\tau,\beta_\tau)$, with $\alpha_\tau > 0$ and $\beta_\tau > 0$. Then,*

$$
\begin{aligned}
&\int \tau \mathrm{d}P^s(t,\tau) < \infty \text{ a.s.}, \int t^2 \mathrm{d}P^s(t,\tau) < \infty \text{ a.s.}, \\
&\int \tau t^2 \mathrm{d}P^s(t,\tau) < \infty \text{ a.s.}, -\infty < \int (\log\tau)\mathrm{d}P^s(t,\tau) < \infty \text{ a.s.}
\end{aligned}
\tag{11}
$$

Proof

$$
\begin{aligned}
&\int\int_{\tau>0,t\in\mathbb{R}} \tau \mathrm{d}P^s(t,\tau)\mathrm{d}P = \int_{\tau>0,t\in\mathbb{R}} \tau \int \mathrm{d}P^s(t,\tau)\mathrm{d}P \\
&= \frac{1}{2}\int_{\tau>0,t\in\mathbb{R}} \tau \mathrm{N}(t;\mu_0,\sigma_0^2)\mathrm{Ga}(\tau;\alpha_\tau,\beta_\tau)\mathrm{d}t\mathrm{d}\tau + \frac{1}{2}\int_{\tau>0,t\in\mathbb{R}} \\
&\tau \mathrm{N}(t;-\mu_0,\sigma_0^2)\mathrm{Ga}(\tau;\alpha_\tau,\beta_\tau)\mathrm{d}t\mathrm{d}\tau \\
&= \int_{\tau>0} \tau \mathrm{Ga}(\tau;\alpha_\tau,\beta_\tau)\mathrm{d}\tau < \infty.
\end{aligned}
$$

The proofs of $\int t^2 \mathrm{d}P^s(t,\tau) < \infty$ a.s. and $\int \tau t^2 \mathrm{d}P^s(t,\tau) < \infty$ a.s. are similar. Since $\alpha_\tau > 0$, choose an integer m large enough such that $\alpha_\tau > \frac{1}{m}$.

$$
\begin{aligned}
&\int\int_{\tau>0,t\in\mathbb{R}} (\log\tau)\mathrm{d}P^s(t,\tau)\mathrm{d}P = \int_{\tau>0} (\log\tau)\mathrm{Ga}(\tau;\alpha_\tau,\beta_\tau)\mathrm{d}\tau \\
&= C\int_{\tau>0} (\log\tau)\tau^{\alpha_\tau - 1}\mathrm{e}^{-\beta_\tau\tau}\mathrm{d}\tau = C\int_{\tau>0} (\tau^{1/m}\log\tau)\tau^{\alpha_\tau - \frac{1}{m} - 1}\mathrm{e}^{-\beta_\tau\tau}\mathrm{d}\tau > -\infty
\end{aligned}
$$

since $\tau^{1/m}\log\tau$ is bounded in $[0,1]$. Also $\int_{\tau>0}(\log\tau)\tau^{\alpha_\tau-1}\mathrm{e}^{-\beta_\tau\tau}\mathrm{d}\tau \le \int_{\tau>0}\tau\tau^{\alpha_\tau-1}\mathrm{e}^{-\beta_\tau\tau}\mathrm{d}\tau<\infty$. □

Lemma 7 *Under the conditions of the Theorem 1, if $f(\cdot)=\int N(\cdot\,;t,\tau^{-1})\mathrm{d}P^s(t,\tau)$ and $f_\theta(y)=f(y-\theta)$, then*

1. $\lim_{\theta\to 0}\int f_0(y)\log\frac{f_0(y)}{f_\theta(y)}\mathrm{d}y=\int f_0(y)\log\frac{f_0(y)}{f(y)}\mathrm{d}y$.
2. $\lim_{\theta\to 0}\int f_0(y)\big(\log_+\frac{f_0(y)}{f_\theta(y)}\big)^2\mathrm{d}y=\int f_0(y)\big(\log_+\frac{f_0(y)}{f(y)}\big)^2\mathrm{d}y$.

Proof Clearly $\tau\phi\big\{\tau(y-\theta-t)\big\}\to\tau\phi\big\{\tau(y-t)\big\}$ as $\theta\to 0$. Since $\int\tau\phi\big\{\tau(y-\theta-t)\big\}\mathrm{d}P^s(t,\tau)\le\frac{1}{\sqrt{2\pi}}\int\tau\mathrm{d}P^s(t,\tau)<\infty$, so by DCT $f_\theta(y)\to f(y)$ as $\theta\to 0$. Hence,

$$\log\frac{f_0(y)}{f_t(y)}\to\log\frac{f_0(y)}{f(y)}\text{ as } t\to 0$$
$$\left(\log_+\frac{f_0(y)}{f_t(y)}\right)^2\to\left(\log_+\frac{f_0(y)}{f(y)}\right)^2\text{ as } t\to 0.$$

To apply DCT again, we have to bound the function $|\log f_\theta(y)|$by an integrable function.

$$|\log f_\theta(y)|\le\log\sqrt{2\pi}+\left|\log\int\tau\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\right|.$$

Let $c=\int\tau\mathrm{d}P^s(t,\tau)<\infty$. Then

$$\left|\log\int\tau\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\right|\le|\log c|+\left|\log\int\frac{\tau}{c}\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\right|.$$

Now since $\int\tau\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\le c$, $\big|\log\int\frac{\tau}{c}\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\big|=-\log\int\frac{\tau}{c}\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)$. Hence, by Jensen's inequality applied to $-\log x$, we get,

$$-\log\int\frac{\tau}{c}\mathrm{e}^{-\frac{\tau}{2}(y-t-\theta)^2}\mathrm{d}P^s(t,\tau)\le\log c-\int(\log\tau)\mathrm{d}P^s(t,\tau)$$
$$+\frac{1}{2}\int\tau(y-t-\theta)^2\mathrm{d}P^s(t,\tau).$$

Now since $\theta\to 0$, w.l.o.g assume $|\theta|\le 1$. Hence

$$\int\tau(y-t-\theta)^2\mathrm{d}P^s(t,\tau)\le 4\left(y^2\int\tau\mathrm{d}P^s(t,\tau)+\int\tau t^2\mathrm{d}P^s(t,\tau)+1\right)$$
$$\Rightarrow|\log f_\theta(y)|\le\log\sqrt{2\pi}+|\log c|+\log c-\int(\log\tau)\mathrm{d}P^s(t,\tau)$$
$$+2\left(y^2\int\tau\mathrm{d}P^s(t,\tau)+\int\tau t^2\mathrm{d}P^s(t,\tau)+1\right)$$

which is clearly f_0-integrable according to the assumptions of the lemma and from the properties of Π_s proved in Lemma 6. Similarly $|\log f_\theta(y)|^2$ can be bounded by an f_0-integrable function. The conclusion of the lemma follows from a simple application of DCT. □

Lemma 2 together with the assumption (2) of the Theorem 1 guarantees $\Pi\big\{f : K(f_0, f) < \delta, V(f_0, f) < \infty\big\} > 0$ for all $\delta > 0$. Since (11) holds, we may assume

$$\Pi(\mathcal{U}) > 0,\ \text{where}\ \mathcal{U} = \left\{f : K(f_0, f) < \delta, V(f_0, f) < \infty,\ \text{(11) holds}\right\}. \quad (12)$$

Now for every $f(\cdot) = \int \mathrm{N}(\cdot; t, \tau^{-1})\mathrm{d}P^s(t, \tau) \in \mathcal{U}$, using Lemma 7, choose δ_f such that for $|\theta| < \delta_f$,

$$K(f_0, f_\theta) < 2K(f_0, f), V(f, f_\theta) < 2V(f_0, f).$$

Now if $||\eta - \eta_0|| < \delta_f$, $|\eta(\mathbf{x}_i) - \eta_0(\mathbf{x}_i)| < \delta_f$, for $i = 1, \ldots, n$. So if $f \in \mathcal{U}$ and $||\eta - \eta_0|| < \delta_f$, we have

$$K_i(f, \eta) = \int f_{0i} \log \frac{f_{0i}}{f_{\eta i}} = \int f_0 \log \frac{f_0}{f_{(\eta-\eta_0)i}} < 2K(f_0, f),$$
$$V_i(f, \eta) = \int f_{0i} \big(\log_+ \frac{f_{0i}}{f_{\eta i}}\big)^2 = \int f_0 \big(\log_+ \frac{f_0}{f_{(\eta-\eta_0)i}}\big)^2 < 2V(f_0, f).$$

From (12) and Lemma 3 we have,

$$\Pi\left\{(f, \eta) : f \in \mathcal{U}, ||\eta - \eta_0||_{1,n} < \delta_f\right\} > 0.$$

Hence

$$\Pi\left\{(f, \eta) : K_i(f, \eta) < 2\delta\,\forall\, i, \sum_{i=1}^{\infty} \frac{V_i(f, \eta)}{i^2} < \infty\right\} > 0.$$

This ensures weak consistency of the posterior of the residual density and strong consistency of the posterior of the regression function η. □

Acknowledgments This research was partially supported by grant number R01 ES017240-01 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH). The authors would like to thank Prof. Taeryon Choi for helpful suggestions on the manuscript.

References

Adler, R.J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes* (vol. 12). Hayward: Academic Press.

Albert, J., Chib, S. (2001). Sequential ordinal modeling with applications to survival data. *Biometrics*, *57*(3), 829–836.

Amewou-Atisso, M., Ghoshal, S., Ghosh, J.K., Ramamoorthi, R.V. (2003). Posterior consistency for semiparametric regression problems. *Bernoulli*, *9*(2), 291–312.

Arellano-Vallea, R.B., Galea-Rojasb, M., Zuazola, P.I. (2000). Bayesian sensitivity analysis in elliptical linear regression models. *Journal of Statistical Planning and Inference*, *86*, 175–199.

Burr, D., Doss, H. (2005). A Bayesian semiparametric model for random-effects meta-analysis. *Journal of the American Statistical Association*, *100*, 242–251.

Bush, C., MacEachern, S. (1996). A semiparametric bayesian model for randomised block designs. *Biometrika*, *83*(2), 275.

Chan, D., Kohn, R., Nott, D., Kirby, C. (2006). Locally adaptive semiparametric estimation of the mean and variance functions in regression models. *Journal of Computational and Graphical Statistics*, *15*, 915–936.

Chib, S., Greenberg, E. (2010). Additive cubic spline regression with dirichlet process mixture errors. *Journal of Econometrics*, *156*(2), 322–336.

Chipman, H.A., George, E.I., Mcculloch, R.E. (2010). Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, *4*(1), 266–298.

Choi, T. (2005). Posterior consistency in nonparametric regression problems in gaussian process priors. PhD thesis. Pittsburgh: Department of Statistics, Carnegie Mellon University.

Choi, T. (2009). Asymptotic properties of posterior distributions in nonparametric regression with non-Gaussian errors. *Annals of the Institute of Statistical Mathematics*, *61*(4), 835–859.

Choi, T., Schervish, M.J. (2007). On posterior consistency in nonparametric regression problems. *Journal of Multivariate Analysis*, *10*, 1969–1987.

Chu, K.C. (1973). Estimation and detection in linear systems with elliptical errors. *Institute of Electrical and Electronics Engineers, Transactions on Automatic Control*, *18*, 499–505.

Chung, Y., Dunson, D. (2009). Nonparametric Bayes conditional distribution modeling with variable selection. *Journal of the American Statistical Association*, *104*(488), 1646–1660.

Cramér, H., Leadbetter, M.R. (1967). *Stationary and related stochastic processes, sample function properties and their applications*. New York: John Wiley.

Denison, D., Holmes, C., Mallick, B., Smith, A.F.M. (2002). *Bayesian methods for nonlinear classification and regression*. London: Wiley.

Doss, H. (1985). Bayesian nonparametric estimation of the median; part I: computation of the estimates. *The Annals of Statistics*, *13*(4), 1432–1444.

Dunson, D., Park, J. (2008). Kernel stick-breaking processes. *Biometrika*, *95*(2), 307–323.

Dunson, D.B., Pillai, N., Park, J.H. (2007). Bayesian density regression. *Journal of the Royal Statistical Society: Series B*, *69*, 163–183.

Escobar, M.D., West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *The Annals of Statistics*, *1*(2), 209–230.

Ferguson, T.S. (1974). Prior distributions on spaces of probability measures. *The Annals of Statistics*, *2*(4), 615–629.

Fonseca, T.C.O., Ferreira, M.A.R., Migon, H.S. (2008). Objective Bayesian analysis for the Student-t regression model. *Biometrika*, *95*(2), 325–333.

Geweke, J. (1992). Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. *Bayesian Statistics*, *4*, 169–194.

Ghosal, S., Roy, A. (2006). Posterior consistency of Gaussian process prior in nonparametric binary regression. *The Annals of Statistics*, *34*(5), 2413–2429.

Gramacy, R., Lee, H. (2008). Bayesian treed Gaussian process models with an application to computer modeling. *Journal of the American Statistical Association*, *103*(483), 1119–1130.

Griffin, J., Steel, M.F.J. (2006). Order-based dependent Dirichlet processes. *Journal of the American Statistical Association, Theory and Methods*, *101*(473), 179–194.

Griffin, J.E., Steel, M.F. (2010). Bayesian nonparametric modelling with the dirichlet process regression smoother. *Statistica Sinica*, *20*(4), 1507.

Huber, P.J. (1964). Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, *35*(1), 73–101.

Ishwaran, H., James, L. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, *96*(453), 161–173.

James, L.F., Lijoi, A., Prünster, I. (2005). Bayesian nonparametric inference via classes of normalized random measures. In Technical report, International Centre for Economic Research, Applied Mathematics Working Papers Series 5/2005. Italy: University of Turin.

Kalli, M., Griffin, J., Walker, S. (2010). Slice sampling mixture models. *Statistics and Computing*, 1–13.

Kottas, A., Gelfand, A.E. (2001). Bayesian semiparametric median regression modeling. *Journal of the American Statistical Association*, *96*(456), 1458–1468.

Lange, K., Little, R.J.A., Taylor, J.M.G. (1989). Robust statistical modelling using the t distribution. *Journal of the American Statistical Association*, *84*(408), 881–896.

Lavine, M., Mockus, A. (2005). A nonparametric Bayes method for isotonic regression. *Journal of Statistical Planning and Inference*, *46*, 235–248.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates. I: density estimates. *The Annals of Statistics*, *12*(1), 351–357.

Müller, P., Erkanli, A., West, M. (1996). Bayesian curve fitting using multivariate normal mixtures. *Biometrika*, *83*(1), 67–79.

Neal, R.J. (1998). Regression and classification using Gaussian process priors. *Bayesian Statistics*, *6*, 475–501.

Norets, A., Pelenis, J. (2010). *Posterior consistency in conditional distribution estimation by covariate dependent mixtures*. USA: Princeton University (unpublished manuscript).

Norets, A., Pelenis, J. (2011). *Bayesian semiparametric regression*. USA: Princeton University (unpublished manuscript).

Nott, D. (2006). Semiparametric estimation of mean and variance functions for non-Gaussian data. *Computational Statistics*, *21*, 603–620.

Ongaro, A., Cattaneo, C. (2004). Discrete random probability measures: a general framework for nonparametric Bayesian inference. *Statistics and Probability Letters*, *67*, 33–45.

Papaspiliopoulos, O. (2008). A note on posterior sampling from Dirichlet mixture models. In Technical Report 08–20, Centre for Research in Statistical Methodology. UK: University of Warwick.

Papaspiliopoulos, O., Roberts, G. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. *Biometrika*, *95*, 169–183.

Pati, D., Dunson, D.B., Tokdar, S.T. (2013). Posterior consistency in conditional distribution estimation. *Journal of Multivariate Analysis*, *116*, 456–472.

Raftery, A.E., Lewis, S. (1992). How many iterations in the Gibbs sampler? *Bayesian Statistics*, *4*, 763–773.

Rasmussen, C., Williams, C. (2006). *Gaussian processes for machine learning (adaptive computation and machine learning)*. Cambridge: The MIT Press.

Rodriguez, A., Dunson, D. (2011). Nonparametric Bayesian models through probit stick-breaking processes. *Bayesian Analysis*, *6*(1), 145–178.

Schwartz, L. (1965). On Bayes procedures. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *4*, 10–26.

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

Tokdar, S.T. (2006). Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. *Sankhyā*, *68*, 90–110.

van der Vaart, A., van Zanten, J. (2008). Reproducing kernel Hilbert spaces of Gaussian priors. *IMS Collections*, *3*, 200–222.

van der Vaart, A., van Zanten, J. (2009). Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *The Annals of Statistics*, *37*(5B), 2655–2675.

van der Vaart, A.W., Wellner, J.A. (1996). *Weak convergence and empirical processes*. New York: Springer-Verlag.

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

Weiss, R. (1996). An approach to Bayesian sensitivity analysis. *Journal of the Royal Statistical Society Series B (Methodological)*, *58*(4), 739–750.

West, M. (1984). Outlier models and prior distributions in Bayesian linear regression. *Journal of the Royal Statistical Society Series B*, *46*(3), 431–439.

West, M. (1987). On scale mixtures of normal distributions. *Biometrika*, *74*(3), 646–648.

Wu, Y., Ghoshal, S. (2008). Kullback Leibler property of kernel mixture priors in Bayesian density estimation. *Electronic Journal of Statistics*, *2*, 298–331.

Yau, P., Kohn, R. (2003). Estimation and variable selection in nonparametric heteroscedastic regression. *Statistics and Computing*, *13*, 191–208.