Statistica Sinica Preprint No: SS-2017-0315	
Title	Sparse Bayesian Additive Nonparametric Regression with Application to Health Effects of Pesticides Mixtures
Manuscript ID	SS-2017-0315
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202017.0315
Complete List of Authors	Ran Wei Brian J. Reich Jane A. Hoppin and Subhashis Ghosal
Corresponding Author	Ran Wei
E-mail	rwei@ncsu.edu

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

Sparse Bayesian Additive Nonparametric Regression with Application to Health Effects of Pesticides Mixtures

Ran Wei, Brian J. Reich, Jane A. Hoppin, Subhashis Ghosal

North Carolina State University

Abstract

In many practical problems that simultaneously investigate the joint effect of covariates, we first need to identify the subset of significant covariates, and then estimate their joint effect. An example is an epidemiological study that analyzes the effects of exposure variables on a health response. In order to make inferences on the covariate effects, we propose a Bayesian additive nonparametric regression model with a multivariate continuous shrinkage prior to address the model uncertainty and to identify important covariates. Our general approach is to decompose the response function into the sum of the nonlinear main effects and the two-way interaction terms. Then we apply the computationally advantageous Bayesian variable selection method to identify the important effects. The proposed Bayesian method is a multivariate Dirichlet–Laplace prior that aggressively shrinks many terms toward zero, thus mitigating the noise of including unimportant exposures and isolating the effects of the important covariates. Our theoretical studies demonstrate asymptotic prediction and variable selection consistency properties. In addition, we use numerical simulations to evaluate the model performance in terms of prediction and variable selection under practical scenarios. The method is applied to a neurobehavioral data set from the Agricultural Health Study that investigates the association between pesticide usage and neurobehavioral outcomes in farmers. The proposed method shows improved accuracy in predicting the joint effects on the neurobehavioral responses, while restricting the number of covariates included in the model through variable selection.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

Keywords: Additive nonparametric regression; Bayesian variable selection; Continuous shrinkage prior; Environmental epidemiology; Posterior consistency.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

1 Introduction

Traditional epidemiological studies in toxicology analyze the correlation between chemical exposures and a single health endpoint. As data become more complex, advanced statistical methods are needed to estimate the relationships between mixtures of multiple chemicals and a suite of health endpoints to increase the statistical power of models and to create a more realistic picture of health risks. As a motivating application, we analyze neurobehavioral (NB) data collected as part of the Agricultural Health Study (AHS; `http://aghealth.nih.gov/`). The data include measurements of 20 organophosphate pesticides and 12 NB health endpoints for each of 701 farmers from Iowa and North Carolina. In previous works, Starks et al. (2012a, b) use a linear regression model to separately examine the associations between an indicator for having experienced a pesticide exposure event and each health endpoint of the NB tests . They conduct conventional hypothesis testing, concluding that two of the nine NB endpoints have a negative association with pesticide exposure.

Because pesticide exposure may have complex nonlinear associations with health endpoints, nonparametric models are preferable, considering their robustness to model assumptions. Previous works in the literature consider various nonparametric regression models to delineate the relationships between the covariates and the response variables and to overcome the limitations of a linear regression. Friedman (1991) develops a multivariate regression splines (MARS) method that defines a nonparametric regression model using splines. Lin and Zhang (2006) propose the component selection and smoothing operator (COSSO) technique, which uses a penalized regression to select variables. Under a Bayesian framework, Bobb et al. (2014) implement a Bayesian kernel machine regression model that assumes nonparametric associations between mixtures and health responses.

Whereas fully nonparametric models are robust to model assumptions, they suffer from a lack of interpretability and are difficult to fit in high dimensions. In order to reduce the complexity, an additive model can be assumed to decompose the joint effect function into the summation of the individual effects and the interaction effects. For example, the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

smoothing spline ANOVA (SS-ANOVA, Gu 2002) method models nonlinear main effects and higher-order interactions between predictors. Reich et al. (2009) implement an SS-ANOVA model with Gaussian process priors, and search for the best model using stochastic search variable selection (SSVS, George and McCulloch, 1993) in MCMC sampling. As a more general nonparametric regression model, Scheipl et al. (2012) propose a structured additive regression model for nonlinear functions, with a spike-and-slab prior. Their method aims to select relevant covariates and determine their effects under different scenarios, such as Gaussian and nonGaussian models. One of the disadvantages of these methods is their computational burden for large problems. Under a similar model setting, Curtis et al. (2014) use a multivariate Laplace prior on the basis coefficients in their additive nonparametric model, which relies on a large-sample approximation for parameter sampling. In this study, we apply the same technique of an additive regression model with a basis expansion as that of Curtis et al. (2014), but use a different Bayesian variable selection method.

Variable selection techniques under Bayesian frameworks have been studied extensively, especially for high-dimensional linear regression models. One commonly used method for variable selection is SSVS, which defines a two-component mixture prior on linear coefficients. The first component is concentrated at zero, which takes care of unimportant predictors, and the second is a diffuse normal distribution, which models active signals. Here, we are interested in a Bayesian method that can substantially alleviate the computational burden of the complex nonparametric model. As a computationally efficient alternative to SSVS priors, shrinkage priors are continuous distributions imposed on the model parameters. Furthermore, they mimic the behavior of SSVS priors with a dominant peak near zero and heavy tails. Various shrinkage priors have been proposed, including the Horseshoe prior (Carvalho et al. 2010), normal-gamma prior (Griffin and Phillip, 2010), double Pareto prior (Armagan et al. 2013a), and Dirichlet–Laplace (DL) prior (Bhattacharya et al. 2015), and have been shown to fall within the family of Gaussian global–local scale mixtures. Theoretically, shrinkage priors obtain almost the same contraction rate as the point-mass prior when

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

recovering the model parameters and the true subset of covariates in both low-dimensional (Armagan et al. 2013b) and high-dimensional (Song and Liang, 2017) models.

In this study, we assume an additive nonparametric regression model for both the main effect and the interaction effects between the covariates. We consider a multivariate continuous shrinkage prior on the block of B-spline basis coefficients for the variable selection. We make two major contributions to the literature. First, we address the model uncertainty in a nonparametric additive regression setting by incorporating block variable selection on B-spline basis expansion coefficients. Therefore, the concentration of the posterior distributions of the block of basis coefficients near zero is used to identify the significant main effects and the interactions. Second, we expand the notion of the computationally efficient DL prior introduced in Bhattacharya et al. (2015) to multidimensional vectors in order to achieve a simultaneous shrinkage on the basis coefficient vector of each main or interaction effect function. In our theoretical research, we expand the current prediction consistency and variable selection consistency results for a shrinkage prior under a linear regression model to the proposed additive nonparametric model. Here, the induced bias from the B-spline basis approximation and the shrinkage on multidimensional vectors pose the major challenges. Furthermore, we use NB data from the AHS to explore the health effects of multiple pesticide measurements, and how the effects on each health endpoint differ from those on the on overall NB system.

2 Model Description and Prior Specification

2.1 Main-effect-only model

We first describe the main-effect-only model, which assumes the regression mean function is a summation of the main effect functions for each individual covariate. Let the data be (Y, X), where Y is the response variable denoting the health endpoint and $X_{n\times p} = (X_1, \ldots, X_p)^T$ is a p-vector of chemical exposure measurements. Without loss of generality, we assume that

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

all covariates are standardized to lie within the unit interval, $(0, 1)$. For the nonparametric regression model of Y on the covariates X, we assume that

$$Y = \mu + f(X_1, \ldots, X_p) + \varepsilon, \tag{1}$$

where μ is the intercept and $\varepsilon \sim \mathrm{N}(0, \sigma^2)$ is the error term. Assuming that the covariates in the data affect the response variable in an additive manner through unknown functions, the joint effect is decomposed into the sum of the individual main effect functions: $f(X_1, \ldots, X_p) = \sum_{j=1}^{p} f_j(X_j)$, where $f_j(X_j)$ is the univariate nonparametric function of X_j.

If each main effect function of an individual covariate is sufficiently smooth, it can be approximated using a B-spline basis expansion with a predetermined number of basis functions, m. For the covariate matrix X, with all elements scaled to the unit interval, each main effect function is approximated by $f_j(X_j) \approx \sum_{r=1}^{m} B_r(X_j)\beta_{jr}$, where $B_r(\cdot)$, for $r = 1, \ldots, m$, denote B-spline basis terms. This approximation transforms the nonlinear effect of X_j into a linear combination of its basis terms, with the m-vector basis coefficients $\boldsymbol{\beta}_j = (\beta_{j1}, \ldots, \beta_{jm})^T$. The regression model in (1) is written as

$$Y = \mu + \sum_{j=1}^{p} \sum_{r=1}^{m} B_r(X_j)\beta_{jr} + \varepsilon. \tag{2}$$

For the regression model defined in (2), the effect quantity and model uncertainty of each covariate X_j are addressed through the basis coefficients $\boldsymbol{\beta}_j$. The m-vector coefficients $\boldsymbol{\beta}_j$ are assigned multivariate normal priors with a zero mean and different variance factors across $j = 1, \ldots, p$: $\boldsymbol{\beta}_j \overset{ind}{\sim} \mathbb{N}_m(0, \sigma^2\lambda_j\mathbb{I}_m)$. The local variance factors λ_j determine the shrinkage on the m-vector basis coefficients, such that the problem of selecting important main effects reduces to the shrinkage of λ_j. Using the DL prior of Bhattacharya et al. (2015), λ_j follows an exponential distribution with mean $\tau\phi_j$. Here τ is the global factor that determines the tail of the marginal distribution of λ_j and $\phi_j > 0$, with $\sum_{j=1}^{p} \phi_j = 1$, is the proportion of the variance allocated to covariate X_j. Furthermore, a Dirichlet distribution on $\boldsymbol{\phi} = (\phi_1, \ldots, \phi_p)^T$ and a gamma distribution on τ are imposed. The hyper-parameter α in the Dirichlet distribution controls the level of shrinkage, such that a smaller α yields

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

a greater concentration around zero and, thus, a sparser model. The uncertainty on each nonparametric main effect function is addressed by implementing a multivariate DL prior on the B-spline basis coefficients. The full Bayesian model on the model parameters is

$$\boldsymbol{\beta}_j|\lambda_j,\sigma^2 \overset{ind}{\sim} \mathbb{N}_m(0,\sigma^2\lambda_j\mathbb{I}_m),\ j=1,\ldots,p, \tag{3}$$

$$\lambda_j|\phi_j,\tau \overset{ind}{\sim} \text{Exp}(\phi_j\tau),\ j=1,\ldots,p, \tag{4}$$

$$\phi \perp \tau,\ \boldsymbol{\phi}=(\phi_1,\ldots,\phi_p)^T \sim \text{Dirichlet}(\alpha,\ldots,\alpha),\ \tau \sim \text{Gamma}(p\alpha,2). \tag{5}$$

The proposed Bayesian hierarchical method for the additive nonparametric model is a multivariate extension of the DL prior in linear regressions. As a shrinkage prior on the B-spline basis coefficients, the proposed multivariate DL method leads to a slightly different prior to the original univariate DL prior of Bhattacharya et al. (2015). When $m = 1$ and $\sigma^2 = 1$, our multivariate DL prior is $\beta_j|\lambda_j \sim \text{N}(0,\lambda_j)$ and $\lambda_j|\phi_j,\tau \sim \text{Exp}(\phi_j\tau)$. After integrating out λ_j, the marginal prior distribution of β_j, given (ϕ_j,τ), is a double exponential distribution, denoted as $\text{DE}(\sqrt{\phi_j\tau/2})$. In the linear regression model, however, the DL prior on the coefficient is $\beta_j|\phi_j,\tau \sim \text{DE}(\phi_j\tau)$. Therefore, in this case, the original DL prior places more mass near zero than does the proposed multivariate extension with only one basis term. Despite the differences in quantities, our proposed multivariate DL prior for a nonparametric regression is an extension of the DL prior in a linear regression, both of which share similar shrinkage properties.

2.2 Main and interaction effects model

If the effects of the covariates on the response are not only additive in the main effects, but also have two-way interaction terms, the underlying joint function includes both the main effects and the interactions,

$$Y = \mu + \sum_{j=1}^{p} f_j(X_j) + \sum_{k=1}^{p-1}\sum_{l=k+1}^{p} f_{kl}(X_k,X_l) + \varepsilon, \tag{6}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

where the second-order term $f_{kl}(X_k, X_l)$ represents the interaction effects on the health endpoint between covariates X_k and X_l. Note that we consider only two-way interactions in this model, because higher-order interactions are less interpretable, and including them increases the computational complexity beyond manageable limits.

Using the B-spline basis expansion in Section 2.1 to represent each individual main effect function, we incorporate the outer product of the B-spline basis terms for the interaction effect functions:

$$f_{kl}(X_k, X_l) \approx \sum_{s=1}^{m^*} \sum_{t=1}^{m^*} B_s^*(X_k) B_t^*(X_l) \beta_{klst}. \tag{7}$$

In order to ensure this approximation is valid, we assume the two-way interaction functions have the same smoothness along both coordinate axes. We use m^* terms for the interaction effects, as opposed to m terms for the main effects; thus the basis functions $B_s^*(X)$ may also differ from the main effect basis functions $B_s(X)$. We propose a similar multivariate DL prior on the basis coefficients for the interaction effect function. Then, normal priors are placed on the coefficients $\beta_{klst} \overset{ind}{\sim} \mathrm{N}(0, \sigma^2 \lambda_{kl})$, for $s, t = 1, \ldots, m^*$. The DL prior is imposed on the local variance factors λ_{kl}:

$$\boldsymbol{\beta}_{kl} | \lambda_{kl}, \sigma^2 \overset{ind}{\sim} \mathbb{N}_{m^* \times m^*}(0, \sigma^2 \lambda_{kl} \mathbb{I}_{m^* \times m^*}), l > k, k = 1, \ldots, p-1, \tag{8}$$

$$\lambda_{kl} | \phi_{kl}^*, \tau^* \overset{ind}{\sim} \mathrm{Exp}(\phi_{kl}^* \tau^*), l > k, k = 1, \ldots, p-1, \tag{9}$$

$$\boldsymbol{\phi^*} = (\phi_{12}, \ldots, \phi_{1p}, \ldots, \phi_{p-1,p})^T \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \tag{10}$$

$$\tau^* \sim \mathrm{Gamma}\left(\frac{p(p-1)}{2}\alpha, 2\right), \tag{11}$$

where we assume the same sparsity level α for both the main effects and the interactions.

2.3 Identifiability constraints

We propose identifiability constraints on the functions f_j and f_{kl} in the additive model (6) such that all model parameters can be determined uniquely from the distribution of observations (X, Y). For example, adding a constant to f_j and subtracting the same constant

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

from $f_{j'}$ for $j \neq j'$ gives the same mean regression function, but different individual main effect functions. This lack of identifiability does not translate into different qualitative relations between the predictors and the response variable, but the constraints are imposed for easier interpretation and presentation, especially in the theoretical analysis. We choose to restrict the main effect and interaction functions to integrate to zero such that the shrinkage prior encourages shrinkage toward zero, as is customary in the variable selection literature. For the main effect functions, we assume that $\int_0^1 f_j(x_j)dx_j = 0$, for $j = 1, \ldots, p$. For the interactions, we assume that the bivariate functions integrate to zero in both directions, $\int_0^1 f_{kl}(x_k, x_l)dx_k = 0$ for all x_l, and $\int_0^1 f_{kl}(x_k, x_l)dx_l = 0$ for all x_k, such that the main effect functions are not in the linear span of the interaction functions.

The restrictions on the main effect functions can be written as follows:

$$\int_0^1 f_j(x)dx = \int_0^1 \left[\sum_{r=1}^m \beta_{jr} B_r(x)\right] dx = \sum_{r=1}^m \beta_{jr} \left[\int_0^1 B_r(x)dx\right] \stackrel{\text{def}}{=} \sum_{r=1}^m \beta_{jr} D_r = 0, \tag{12}$$

where $D_r \stackrel{\text{def}}{=} \int_0^1 B_r(x)dx$. Therefore, the integral restriction is equivalent to a linear restriction on the basis coefficient $\boldsymbol{\beta}_j$, $\sum_{r=1}^m \beta_{jr} D_r = 0$. For the B-spline basis,

$$D_r = \begin{cases} r/[d(m-d+1)], & r = 1, \ldots, (d-1), \\ 1/(m-d+1), & r = d, \ldots, (m-d+1), \\ (m-r+1)/[d(m-d+1)], & r = (m-d+2), \ldots, m, \end{cases}$$

where m is the number of B-spline basis functions and d is the degree of the B-spline basis.

The restrictions on each bivariate function of interactions are also treated as linear constraints on the coefficients for the B-spline expansion in (7):

$$\int_0^1 f_{kl}(x_k, x_l)dx_k \stackrel{\text{def}}{=} \sum_{t=1}^{m^*} B_t^*(x_l) \left[\sum_{s=1}^{m^*} D_s^* \beta_{klst}\right] = 0, \tag{13}$$

$$\int_0^1 f_{kl}(x_k, x_l)dx_l \stackrel{\text{def}}{=} \sum_{s=1}^{m^*} B_s^*(x_k) \left[\sum_{t=1}^{m^*} D_t^* \beta_{klst}\right] = 0, \tag{14}$$

where we assume $D_s^* \stackrel{\text{def}}{=} \int_0^1 B_s^*(x)dx$, for $s = 1, \ldots, m^*$. The restrictions in (13) and (14)

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

hold for all x_k and x_l if an only if

$$\sum_{s=1}^{m^*} D_s^* \beta_{klst} = 0, \text{ for all } t = 1, \ldots, m^*;$$
$$\sum_{t=1}^{m^*} D_t^* \beta_{klst} = 0, \text{ for all } s = 1, \ldots, m^*.$$

These restrictions on the interaction functions $f_{kl}(x_k, x_l)$ are composed of $2m^*$ linear combinations of $\boldsymbol{\beta}_{kl}$, for $k < l$ and $k = 1, \ldots, p-1$.

2.4 Thresholding

Owing to the properties of continuous shrinkage priors, the factors $f_j(\cdot)$ will never equal zero. Thus, a post-processing procedure is needed to determine the zero and nonzero effects. We implement a thresholding technique where we choose a subset of predictors such that the corresponding deterioration in prediction accuracy in terms of the "variation-explained" can be tolerated (Hahn and Carvalho, 2015). Given the posterior samples of coefficients $\boldsymbol{\beta}_j$, the posterior samples of the "variation-explained" values are calculated as follows:

$$V_{(k)} = \frac{\| \sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j \|^2}{\| \sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j \|^2 + n\sigma^2 + \| \sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j - \sum_{j=1}^{p} B(X_j)\boldsymbol{\beta}_j^{(k)} \|^2}, \tag{15}$$

where $\boldsymbol{\beta}_j^{(k)} = 0$ if $\|\boldsymbol{\beta}_j/\sigma\|$ is among the k smallest terms, and $\boldsymbol{\beta}_j^{(k)} = \boldsymbol{\beta}_j$ otherwise. Specifically, $V_{(k)}$ represents the percentage of information explained by the reduced model that includes only those covariates with the k biggest $\|\boldsymbol{\beta}_j/\sigma\|$, for $k = 1, \ldots, p$. Given the posterior samples of $V_{(k)}$, the level of sparsity k is determined by choosing the smallest k such that the $(1-\alpha_0) \times 100\%$ credible interval of $V_{(k)}$ includes the posterior mean of the full model $V_{(p)}$.

3 Posterior Computation

We now describe the computational algorithm for the main-effect-only model. The regression model with both main and interaction effects in the mean function is very similar. We sample the parameters using a combination of Gibbs sampling and direct sampling. The sampler

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

cycles through (i) $\boldsymbol{\beta}|\lambda, \phi, \tau, Y, X$, (ii) $\lambda|\boldsymbol{\beta}, \phi, \tau, Y, X$ and (iii) $\phi, \tau|\lambda, \boldsymbol{\beta}, Y, X$. Step (iii) follows direct sampling of (iiia) $\tau|\phi, \lambda$ and (iiib)$\phi|\lambda$.

(i) Given λ_j, Y, and X_j, the conditional posterior distribution of $\boldsymbol{\beta}_j$ is the m-dimensional multivariate normal, with mean $\boldsymbol{\mu}_{\beta_j}$ and variance matrix $\boldsymbol{\Sigma}_{\beta_j}$, where

$$\boldsymbol{\mu}_{\beta_j} = \left(B(X_j)^T B(X_j) + \frac{\mathbb{I}_m}{\lambda_j} \right)^{-1} B(X_j)^T \left(Y - \sum_{l=1, l\neq j}^{p} B(X_l)\beta_l \right), \tag{16}$$

$$\boldsymbol{\Sigma}_{\beta_j} = \left(B(X_j)^T B(X_j) + \frac{\mathbb{I}_m}{\lambda_j} \right)^{-1} \cdot \sigma^2. \tag{17}$$

For the restrictions on the basis coefficients in (12), we sample $\boldsymbol{\beta}_j \big| \sum_{r=1}^{m} \beta_{jr} D_r = 0$ from the conditional multivariate normal $\mathbb{N}_m(\boldsymbol{\mu}^*_{\beta_j}, \boldsymbol{\Sigma}^*_{\beta_j})$, where

$$\boldsymbol{\mu}^*_{\beta_j} = \boldsymbol{\mu}_{\beta_j} - \boldsymbol{\Sigma}_{\beta_j} \boldsymbol{D} (\boldsymbol{D}^T \boldsymbol{\Sigma}_{\beta_j} \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{\mu}_{\beta_j}, \tag{18}$$

$$\boldsymbol{\Sigma}^*_{\beta_j} = \boldsymbol{\Sigma}_{\beta_j} - \boldsymbol{\Sigma}_{\beta_j} \boldsymbol{D} (\boldsymbol{D}^T \boldsymbol{\Sigma}_{\beta_j} \boldsymbol{D})^{-1} \boldsymbol{D}^T \boldsymbol{\Sigma}_{\beta_j}, \tag{19}$$

and $\boldsymbol{D} = (D_1, \ldots, D_m)^T$.

(ii) Given ϕ_j, τ, and $\boldsymbol{\beta}_j$, the variance component λ_j is sampled from the generalized inverse Gaussian distribution $\text{GiG}\left(1 - \frac{m}{2}, \frac{2}{\phi_j \tau}, \frac{\boldsymbol{\beta}_j^T \boldsymbol{\beta}_j}{\sigma^2}\right)$.

(iiia) Given $\phi_1, \ldots, \phi_p$ and $\lambda_1, \ldots, \lambda_p$, the global parameter τ in the DL prior is sampled from the generalized inverse Gaussian distribution $\text{GiG}\left(p(\alpha - 1),\ 1,\ 2\sum_{j=1}^{p} \frac{\lambda_j}{\phi_j}\right)$.

(iiib) Now, given $\lambda_1, \ldots, \lambda_p$, we first sample T_j, for $j = 1, \ldots, p$, independently from the generalized inverse Gaussian distribution $\text{GiG}(\alpha - 1,\ 1,\ 2\lambda_j)$, and then let $\phi_j = T_j / \sum_{l=1}^{p} T_l$.

4 Asymptotic Properties

Next, we study the asymptotic properties of the additive nonparametric regression model for the individual main effects with the multivariate DL prior. Because the predictors are considered deterministic in our setting, the extension to include interactions is similar, except that the full model has a greater number of terms.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

Notation. For a fixed $n \times p_n$ covariate matrix $X = (X, \ldots, X_{p_n})$, we consider the additive nonparametric model $Y = \sum_{j=1}^{p_n} f_j(X_j) + \sigma\varepsilon$, where each additive function is approximated by an m_n-dimensional B-spline basis expansion: $f_j(X_j) = B(X_j)\boldsymbol{\beta}_j + \sigma\delta$, and δ denotes the bias induced from the basis expansion approximation. Therefore, the true additive regression model is

$$Y = \sum_{j=1}^{p_n} B(X_j)\boldsymbol{\beta}_j + \sigma\delta + \sigma\varepsilon \stackrel{\text{def}}{=} B(X)\boldsymbol{\beta} + \sigma\delta + \sigma\varepsilon,$$

where $B(X) = [B(X_1), \ldots, B(X_{p_n})]$ is a matrix of the values of the B-spline basis functions, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{p_n}^T)^T$ is a $p_n m_n$-vector, with each m_n-dimensional component corresponding to a covariate X_j, and ε is an n-dimensional standard normal vector. We study a Bayesian approach with a continuous shrinkage prior for the m_n-dimensional vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p_n}$. After integrating out the parameters ϕ_j, the hierarchical Bayesian model in Section 2.1 is represented by, for $j = 1, \ldots, p_n$,

$$\boldsymbol{\beta}_j | \lambda_j, \sigma^2 \stackrel{ind}{\sim} \mathbb{N}_m(0, \lambda_j\sigma^2\mathbb{I}_m), \tag{20}$$

$$\lambda_j | \psi_j \stackrel{ind}{\sim} \text{Exp}(\psi_j),\ \psi_j \sim \text{Gamma}(\alpha, 2). \tag{21}$$

We let $f_j^*(\cdot)$ be the true function for covariate X_j, $\boldsymbol{\beta}^*$ and σ^* be the true values of the parameters, and $\xi^* \subset \{1, \ldots, p_n\}$ be the indices of the covariates with nonzero effects such that $f_j^*(X_j) \neq 0$ for $j \in \xi^*$. The true sparsity level is $s = |\xi^*|$.

Assumptions. We first state some regularity conditions on the eigen structure of the B-spline basis expansion matrix $B(X)$ with respect to a sequence $\{\epsilon_n\}$, which is defined later. These assumptions are similar to those in Song and Liang (2016), who present the asymptotic properties of the shrinkage priors in linear regression models. The difference between the assumptions lies in the fact that we are dealing with a design matrix of the basis expansions. In addition, the additive nonparametric functions are estimated using a B-spline basis expansion, such that an estimation bias is introduced. Let $a \prec b$ mean that $\lim a/b = 0$, and let $a \asymp b$ mean that "$\lim a/b$" is bounded by constants.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

- $A_1(1)$: The number of parameters in the linear expansion satisfies $m_n p_n \geq n$.
- $A_1(2)$: All main effect functions of the additive model are κ-times continuously differentiable.
- $A_1(3)$: The rank of $B(X)$ is n and $B(X)^T B(X)$ has n positive eigenvalues, denoted as $nd_1/m_n, \ldots, nd_n/m_n$, where $d_1, \ldots, d_n$ are bounded away from zero.
- $A_2(1)$: The sequence $\{\epsilon_n\}$ is assumed to satisfy $sm_n \log p_n \prec n\epsilon_n^2$ and $\epsilon_n \succ m_n^{-\kappa}$.
- $A_2(2)$: $\min_{j\in\xi^*}\{\frac{\|\boldsymbol{\beta}_j^*/\sigma^*\|}{\sqrt{m_n}}\} \succ \epsilon_n$ and $\max_j\{\frac{\|\boldsymbol{\beta}_j^*/\sigma^*\|}{\sqrt{m_n}}\} \leq \gamma_3 E$, for fixed $\gamma_3 \in (0,1)$, and E is nondecreasing with n.
- A_3: There exist an integer $\bar{p}$, which depends on n and p_n, and two constants d and d_0', such that $\bar{p}m_n \log p_n \succ n\epsilon_n^2$ and $nd_0/m_n \geq d_{\max}\left(\tilde{B}(X)^T\tilde{B}(X)\right) \geq d_{\min}\left(\tilde{B}(X)^T\tilde{B}(X)\right) \geq nd_0'/m_n$, for any $q \leq \bar{p}$ and any sub-matrix $\tilde{B}(X)$ consisting of qm_n columns of $B(X)$.

The assumption of high dimensionality in $A_1(1)$ is mainly used for the concise representation of certain bounds. Using a lower dimensionality of the parameter space, the same asymptotic properties can be obtained, but the following assumptions and proofs need to be treated slightly differently. To save space and avoid monotonicity of the arguments, it is customary to forgo separate arguments for the lower-dimensional case. In $A_1(2)$, κ defines the minimum smoothness for all additive functions, such that the bias induced by the m_n-dimensional basis expansion is $\delta \asymp m_n^{-\kappa}$. Assumption $A_1(3)$ is an extension of the linear regression model in Song and Liang (2016), where we replace the covariate matrix X with the B-spline basis design matrix $B(X)$. From Lemma $A.9$ in Yoo and Ghosal (2016), we combine the B-spline property with the linear regression model assumption to specify $A_1(3)$. Assumption in $A_2(1)$ restricts ϵ_n to $\max(\sqrt{sm_n \log p_n/n}, m_n^{-\kappa})$.

Because the nonparametric regression function assumes there are no model parameters, we mainly want to demonstrate the prediction consistency of the joint effects. Therefore, the following theorem proves that when the B-spline basis expansion is implemented, the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

prediction performance of $B(X)\boldsymbol{\beta}$ is asymptotically concentrated around the additive mean function under the truth.

Theorem 4.1 *For the regression model* $Y = \sum_{j=1}^{m_n} f_j(X_j) + \sigma\varepsilon = B(X)\boldsymbol{\beta} + \sigma\delta + \sigma\varepsilon$, *the basis expansion bias satisfies* $\|\delta\| \lesssim \sqrt{n} m_n^{-\kappa}$, *where* κ *is the degree of smoothness. Let* A_1, A_2, *and* A_3 *hold for the design matrix* $B(X)$, *and let the basis coefficients* $\boldsymbol{\beta}_j$ *for covariate* X_j *follow the prior density* $\pi_\alpha(\cdot)$ *defined in (20) and (21), with* $\alpha \asymp p_n^{-(1+\nu)}$ *for* $\nu > 0$. *Then,*

$$\mathrm{P}^*\left(\pi\left(\left\|B(X)\boldsymbol{\beta} - \sum_{j=1}^{p_n} f_j^*(X_j)\right\| \geq c_0\sqrt{n}\epsilon_n \middle| X, Y\right) \geq e^{-c_1 n\epsilon_n^2}\right) \leq e^{-c_2 n\epsilon_n^2}, \tag{22}$$

for some constants c_0, c_1, *and* c_2.

Theorem 4.1 shows the posterior concentration rate $\sqrt{n}\epsilon_n$ for the predictions at the observation points. Therefore, given the matrix X of covariates, the predictor $B(X)\boldsymbol{\beta}$ obtained from the regression model concentrates around the true mean function $\sum_{j=1}^{p_n} f_j^*(X_j)$, with a concentration rate close to $\sqrt{n}\max\{\sqrt{sm_n \log p_n/n},\ m_n^{-\kappa}\}$. The proof of the theorem is provided in the Supplementary Material.

Theorem 4.2 *Define the sub-model* $\xi(a_n) = \{j : \|\boldsymbol{\beta}_j/\sigma\| > a_n\}$ *corresponding to a threshold* a_n, *where* $na_np_n \prec \log p_n$. *Assume that the conditions of Theorem 4.1 hold and* $\min_{j\in\xi^*} \|\boldsymbol{\beta}_j^*\|/\sqrt{m_n} \succ \epsilon_n$. *Then,*

$$\mathrm{P}^*\left(\pi(\xi(a_n) = \xi^*|X, Y) > 1 - p^{-\mu''}\right) > 1 - p_n^{-\mu'}, \tag{23}$$

for some positive constants μ' *and* μ''.

Theorem 4.2 shows the posterior variable selection consistency, given that $a_n \prec \log p_n/np_n$ and the prior density is moderately flat at the nonzero basis coefficients $\boldsymbol{\beta}_j^*/\sigma^*$. The proof of the theorem is given in the Supplementary Material.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

5 Simulation Study

We first consider the additive nonparametric regression model that includes individual main effects only as in Section 2.1,. Then we expand the regression model to include two-way interaction effects, as in Section 2.2.

5.1 Simulation description

For the simulated data, the number of covariates is $p = 50$ and the sample size is fixed at either $n = 200$ or $n = 500$. For the matrix X of the covariate values, we first sample X_j^* for $j = 1, \ldots, p$, from a Gaussian distribution, with $\mathrm{E}(X_j^*) = 0$, $\mathrm{Var}(X_j^*) = 1$, and $\mathrm{Cov}(X_j^*, X_k^*) = 0$ for the mutually independent case or $\mathrm{Cov}(X_j^*, X_k^*) = 0.5^{|j-k|}$ for the autoregressive case. Next, the simulated random vectors are rescaled onto the unit interval by $X_j = \frac{X_j^* - \min(X_j^*)}{\max(X_j^*) - \min(X_j^*)}$. Given X, the response Y is generated from normal distribution with mean $f(X) \stackrel{\text{def}}{=} f_1(X_1) + f_2(X_2) + f_3(X_3) + f_4(X_4)$ and variance $\sigma^2 = 1.5$, where

$$f_1(x) = \exp(1.1x^3) - 2, \quad f_2(x) = 2x - 1,$$
$$f_3(x) = \sin(4\pi x), \qquad f_4(x) = \log\{(e^2 - 1)x + 1\} - 1.$$

The remaining $p - 4$ predictors have no effect on the response.

Under each scenario of different sample sizes ($n = 200$ or $n = 500$) and dependence structures (independent or autoregressive), we simulate 100 data sets. For each method implemented, we compare the prediction accuracy and variable selection performance. Specifically, the prediction accuracy is evaluated by the mean squared error (MSE) on the test data, as follows:

$$\mathrm{MSE} = \frac{1}{500} \sum_{i=1}^{500} \left[f(x'_{i1}, \ldots, x'_{ip}) - \hat{f}(x'_{i1}, \ldots, x'_{ip}) \right]^2, \tag{24}$$

where $f(\cdot)$ is the true mean function, $\hat{f}$ is the estimated mean function, and $X'_i = (x'_{i1}, \ldots, x'_{ip})^T$ is a new data point randomly sampled from the proposed covariate distribution, for $i = 1, \ldots, 500$. Note that X'_i is not used for model fitting, but is used to evaluate the prediction

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

performance. Therefore, we can compare the prediction performance of each method on future observations.

We also record the variable selection performance in terms of correctly identifying the four significant covariates. We evaluate these results by examining the percentage of unimportant variables selected (False Positive) and the percentage of important variables excluded (False Negative), averaged over all simulated data sets under each scenario. We further include the proportion of the simulated data sets in which the true model is selected (Truth). In order to identify the sub-model of nonzero covariates, we follow the proposed thresholding method of "variation-explained." Figure 1(a) shows a box-plot of posterior samples of $\|\boldsymbol{\beta}_j\|$, for $j = 1, \ldots, p$. Figure 1(b) shows the posterior median of $V_{(k)}$, with 80% intervals at different model sizes, $k = 1, \ldots, 20$. The shrinkage level for the variable selection is determined by the

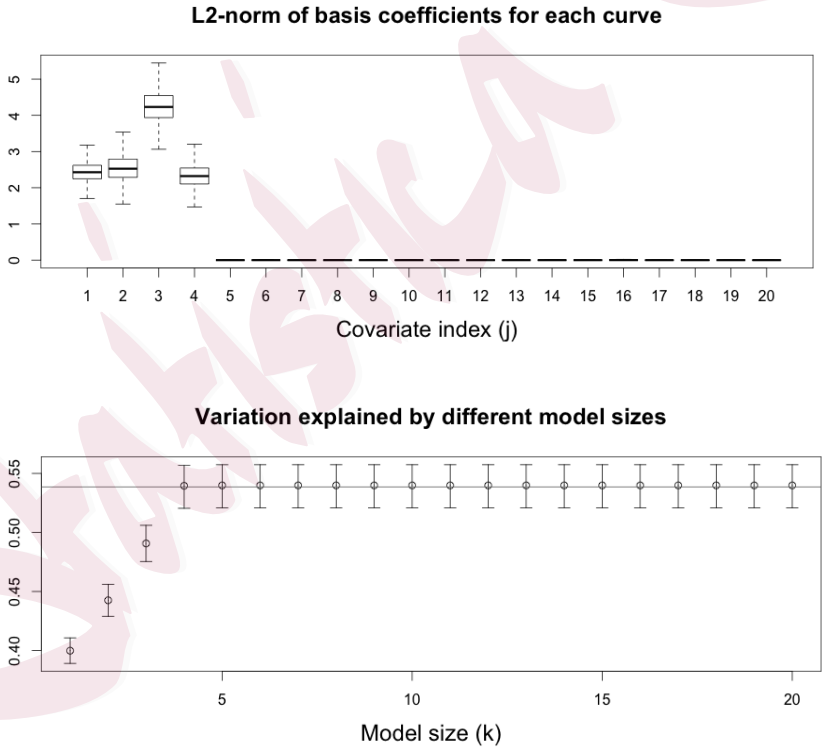

Figure 1: (a) Box-plot of the posterior distribution of the $\mathcal{L}_2$-norm $\|\boldsymbol{\beta}_j/\sigma\|$ for a single simulated data set; (b) Variation-explained plot at different model sizes for a single data set (the horizontal line is the full model "variation-explained" measurement). Note that only the first 20 model sizes are shown in the figure.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

$n = 200$		MSE (SE)	FP (SE)	FN (SE)	True (SE)
Independent	DL(Oracle)	1.85(0.13)	0.00(0.00)	0.00(0.00)	100(0)
	DL(VarExp)	1.85(0.13)	0.88(0.22)	0.00(0.00)	86(3)
	ABayes	-	0.00(0.00)	15.25(1.91)	56(5)
	BSS-ANOVA	2.86(0.20)	4.88(0.58)	0.00(0.00)	47(5)
	COSSO	2.90(0.18)	2.94(0.54)	7.50(1.26)	46(5)
	MARS	2.31(0.17)	1.50(0.28)	1.00(0.49)	73(4)
AR(1)	DL(Oracle)	2.36(0.11)	4.81(0.44)	19.25(1.77)	39(5)
	DL(VarExp)	2.36(0.11)	2.50(0.38)	23.50(1.94)	25(4)
	ABayes	-	0.00(0.00)	40.25(1.77)	5(2)
	BSS-ANOVA	3.51(0.15)	3.69(0.47)	20.21(1.65)	17(4)
	COSSO	3.83(0.19)	4.87(0.87)	26.50(1.94)	10(3)
	MARS	3.90(0.14)	0.62(0.21)	38.00(1.90)	8(3)
$n = 500$		MSE (SE)	FP (SE)	FN (SE)	True (SE)
Independent	DL(Oracle)	1.52(0.09)	0.00(0.00)	0.00(0.00)	100(0)
	DL(VarExp)	1.52(0.09)	0.25(0.12)	0.00(0.00)	96(2)
	ABayes	-	0.00(0.00)	0.00(0.00)	100(0)
	BSS-ANOVA	2.17(0.10)	2.06(0.39)	0.00(0.00)	74(3)
	COSSO	2.12(0.11)	2.81(0.51)	0.25(0.25)	70(5)
	MARS	1.95(0.08)	0.44(0.18)	0.00(0.00)	94(2)
AR(1)	DL(Oracle)	2.06(0.14)	1.94(0.30)	7.75(1.21)	70(5)
	DL(VarExp)	2.06(0.14)	0.12(0.09)	13.25(1.40)	59(5)
	ABayes	-	0.00(0.00)	10.25(1.38)	60(5)
	BSS-ANOVA	2.89(0.13)	1.94(0.22)	2.00(0.68)	51(5)
	COSSO	3.01(0.10)	3.62(0.49)	7.75(1.21)	36(5)
	MARS	2.93(0.11)	0.75(0.22)	8.00(1.37)	55(5)

Table 1: Summary of the main-effect-only simulation study. Methods are compared in terms of their mean squared error ("MSE"), False Positive ("FP"), False Negative ("FN"), and True model ("True"), with their standard errors ("SE") under independent and autoregressive covariance shown in parentheses. All values except the MSE are given in percentages (%).

smallest k such that the 80% interval of $V_{(k)}$ includes the posterior mean of the full model $V_{(50)}$ (horizontal line in Figure 1(b)). Note that we only include model sizes less than 20, because model sizes larger than that do not affect the value of "variation-explained."

In order to demonstrate the advantages of our proposed method over other methods

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

with similar model assumptions, we first fit the additive nonparametric model with a multivariate DL prior (DL method) for each data set. Then we compare our method with the following four competing methods, assuming additive nonparametric functions: the approximate Bayesian method ("ABayes") of Curtis et al. (2014); the Bayesian smoothing splines ANOVA ("BSS-ANOVA") of Reich et al. (2009); the component selection and smoothing operator ("COSSO") of Lin and Zhang (2006), and the multivariate adaptive regression splines model ("MARS") of Friedman (1991). We also implement the "Oracle" selection in the DL method, which identifies the four covariates with the largest posterior medians of $\|\boldsymbol{\beta}_j/\sigma\|$. Given correct information on the number of significant covariates, this method demonstrates the ability of the general DL method to rank the important covariates and formally select those that are significant. For MARS, we use the function `polymars`$(\cdot)$ in the R package `polspline`. For the DL and BSS-ANOVA methods, where MCMC sampling is implemented, $15,000$ samples are drawn in total, with $5,000$ burn-in steps.

5.2 Simulation results: Main-effect-only model

Table 1 summarizes the simulation results for the additive regression model with individual main-effect functions. In terms of prediction performance, our proposed DL method outperforms the competing methods, even though the methods all make similar assumptions about the regression model. The DL method achieves the smallest MSE under all scenarios, and MARS has good prediction accuracy under the independent case. The other methods perform similarly, with the BSS-ANOVA being slightly better than the COSSO. The ABayes method does not provide predictions so there are no results for this method.

As expected, the DL Oracle method performs best under all scenarios in terms of variable selection. The DL method with a data-driven threshold, DL(VarExp), performs slightly worse than DL(Oracle), but still outperforms the other methods. In conclusion, the thresholding policy defined in (15) adds uncertainty in determining the number of important covariates. Thus when the number is fixed, as in the Oracle method, the proposed method

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

correctly ranks the covariates through shrinkage and achieves better variable selection accuracy. For the other methods, ABayes is too conservative when choosing a subset of covariates, with a false positive equal to zero under every scenario and, thus, a large proportion of false negatives. In the dependent case, ABayes has perfect selection for the larger sample size, but the true model proportions drop considerably when the sample size decreases. BSS-ANOVA and COSSO have similar variable selection performance. Their performance for independent data is worse than that of their competitors, and the computation time for BSS-ANOVA is more than three times that of the DL method. MARS, on the other hand, performs poorly for correlated data, as in the AR(1) case. Overall, the DL method outperforms the competitors, especially in the more difficult setting of a small sample size and correlated predictors.

5.3 Model with main and interaction effects

We also conduct simulations for the additive models with both main effects and interactions. Because adding interactions increases the dimensionality significantly, we reduce the number of covariates to $p = 10$ and only investigate the independent case. Therefore, there are 10 main effects and 45 interaction effects. The response variable Y is simulated as:

$$Y = f_1(X_1) + f_2(X_3) + f_3(X_1X_3) + \epsilon,$$

where ϵ is a random number generated from a standard normal distribution, and the functions f_1, f_2, and f_3 are defined in Section 5.1. As in the simulation for the main-effect-only model, we let the sample size be $n = 200$ or $n = 500$.

To fit the nonparametric regression model in (6), we approximate the main-effect function $f_j(X_j)$ using B-spline basis functions of order $m = 10$ and the two-way interaction functions $f_{kl}(X_k, X_l)$, with the outer product of basis terms of order $m^* = 5$. We implement the DL method as described in Section 2.2 and select the important main and interaction effects using the "variation-explained" criterion, similarly to the main-effect-only model. The DL

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

				Correct selection (SE)		
$n = 200$	MSE (SE)	FP (SE)	FN (SE)	Main	Interaction	Model
DL(Oracle)	1.21(0.08)	7.17(0.39)	28.67(1.57)	19(4)	38(5)	19(4)
DL(VarExp)	1.21(0.08)	0.08(0.08)	41.33(1.84)	46(5)	38(5)	15(4)
ABayes	-	0.17(0.12)	49.00(1.86)	3(2)	97(2)	2(1)
BSS-ANOVA	2.33(0.12)	0.17(0.12)	37.67(1.81)	53(5)	30(5)	9(3)
COSSO	2.71(0.13)	8.58(0.67)	41.67(2.70)	66(5)	11(3)	10(3)
MARS	2.39(0.16)	0.25(0.14)	38.33(1.29)	81(4)	1(1)	1(1)
				Correct selection (SE)		
$n = 500$	MSE (SE)	FP (SE)	FN (SE)	Main	Interaction	Model
DL(Oracle)	1.19(0.09)	0.58(0.21)	2.33(0.85)	93(3)	93(3)	93(3)
DL(VarExp)	1.19(0.09)	0.00(0.00)	2.33(0.85)	93(3)	100(0)	93(3)
ABayes	-	0.42(0.18)	16.67(1.87)	53(5)	95(2)	50(5)
BSS-ANOVA	2.08(0.11)	0.08(0.08)	2.33(0.85)	92(3)	90(1)	80(4)
COSSO	2.51(0.12)	3.08(0.51)	7.00(1.73)	94(2)	68(5)	68(5)
MARS	2.30(0.15)	0.17(0.12)	20.67(1.63)	98(1)	38(5)	37(5)

Table 2: Summary of the simulation study with main and interaction effects. These methods are compared in terms of mean squared errors ("MSE"), False Positive ("FP"), False Negative ("FN"), and the correct selection of the main effects, interactions, and complete model, with their standard errors ("SE") in parentheses. All values except the MSE are given in percentages (%).

method is compared with the ABayes, BSS-ANOVA, COSSO, and MARS methods in terms of both prediction performance and variable selection. The accuracy of the model prediction is evaluated by computing the MSE for a newly generated covariate matrix. The variable selection performance is determined by False Negative, False Positive, and the percentages of correctly identifying the main effects, interactions, and complete model. In ABayes and COSSO, where interactions are not considered, the interaction terms are represented as the product of two covariates; that is, we define 45 new covariates $X_l \cdot X_k$ and then use the main-effects model with 55 additive predictors. Therefore, providing extra information on the correct format of the interaction effects actually favors these methods. The simulation results are summarized in Table 2.

From Table 2, the simulation results show that our proposed method improves the pre-

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

diction on the newly generated data set compared with the other nonparametric methods. In particular, by correctly addressing the joint effect by decomposing main-effect functions and interaction-effect functions, our method improves the prediction accuracy by more than 50% in terms of the MSE. Furthermore, the inclusion of a shrinkage prior on the basis expansion coefficients addresses the model uncertainty and improves the performance in identifying the correct sub-model. The $n = 200$ scenario is challenging for all methods, especially when selecting the correct model that includes both nonzero main and interaction effects. The DL method has the highest true model proportion among the methods. When $n = 500$, the DL method successfully selects the true interaction effects for all data sets. The competing methods perform worse than the DL method. Even the ABayes and COSSO, where the true format of the interaction term is specified, cannot outperform the DL method under the two scenarios. The second-best method is the BSS-ANOVA, but this requires five times the computing time of the DL method.

6 Analysis of the AHS NB Data Set

6.1 Description of the NB data

We demonstrate our method using data from an NB sub-study of the Agricultural Health Study (AHS; `http://aghealth.nih.gov/`). The goal of the study (data version number: AHS44436) is to examine the association between pesticide exposure and the NB function of the central nervous system (CNS). From 2006 to 2008, $n = 701$ male farmers from Iowa or North Carolina took NB tests. There are 12 response variables, including $N = 8$ CNS tests that assess memory, motor speed, sustained attention, verbal learning and visual scanning, and processing. For this implementation, we analyze these eight continuous CNS response variables. Starks et al. (2012a, b) conclude that participants with one or more pesticide exposures are more likely to have adverse CNS outcomes, but they do not investigate the individual pesticide effects on the overall NB system. In this application, the exposure

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

variables are the lifetime-specific pesticide use information for $p = 20$ pesticides from the AHS questionnaires and interviews. Each exposure covariate is quantified as the number of days of applying a certain pesticide over the participant's lifetime. We also include $q = 6$ confounding variables Z for age (years), testing site (1 if North Carolina, 0 if Iowa), farm size (acres), smoking status (packs per year), drinking status (drinks per year), and highest level of education (years).

The B-spline basis expansion requires $X_{ij} \in (0, 1)$. Therefore, we apply a rank transformation. For example, $X_{ij} = x$ means that subject i applied pesticide j more days than $100x\%$ of the study participants. This transformation makes the covariates uniformly distributed over the unit interval, while still allowing for a wide range of regression relationships between the covariates X and the response Y via the additive nonparametric function. All response measurements are standardized to have mean zero and variance one, and some response variables (continuous performance test, digit symbol latency, sequence A and sequence B latencies) are multiplied by -1, as appropriate, so that higher values indicate better performance.

6.2 Multivariate extension

Because the data include multiple response variables, we extend the model to account for multiple health responses rather than analyzing the $N = 8$ NB responses individually. This multivariate analysis is preferred because we are more interested in the pesticide effects on overall NB performance than in the individual tests. Furthermore, borrowing strength across response measurements should improve the statistical power of identifying important exposures and estimating their exposure-response curves. The main structure of the multivariate extension is still consistent with the model proposed in Section 2.

For the CNS response variables, we model the confounding variable age (Z_1) as having a nonparametric effect on the health response, but model the other confounding variables $(Z_2, \ldots, Z_q)$ as linear effects. The additive nonparametric model for the response variable

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

Y_b, for $b = 1, \ldots, N$, on confounder Z and pesticide exposure X is

$$Y_b = \mu + g_{1,b}(Z_1) + \sum_{l=2}^{q} Z_l \gamma_{l,b} + f_b(X_1, \ldots, X_p) + \varepsilon_b, \tag{25}$$

where μ is the intercept term and ε_b is normally distributed with mean zero and variance σ^2. For the nonparametric function of age, we use the B-spline basis expansion $g_{1,b}(Z_1) \approx \sum_{r=1}^{m} \gamma_{1r,b} B_r(Z_1)$. The joint-effect function on the bth health response, $f_b(X_1, \ldots, X_p)$, is decomposed into main-effect and interaction-effect functions and approximated using basis expansions, as in (6) and (7).

To build the connection between CNS health responses, we specify the Bayesian hierarchical model so that the coefficients for each response variable share a common prior distribution, with a global mean across $b = 1, \ldots, N$. Therefore, for each covariate index j, $\boldsymbol{\beta}_j^b \overset{ind}{\sim} \mathbb{N}_m(\boldsymbol{\mu}_j, \lambda_j \sigma^2 \mathbb{I}_m)$, where the normal mean $\boldsymbol{\mu}_j$ is treated as the basis coefficient for the nonparametric effect of covariate X_j on the overall NB system (overall effect curve). The same multivariate extension is applied to the interaction-effect functions, such that $\boldsymbol{\mu}_{kl}$ determines the joint effect of X_l and X_k on the overall NB functions. In the confounding effects, a similar method is implemented. We assume $\boldsymbol{\gamma}_1^b \sim \mathbb{N}_m(\boldsymbol{\nu}_1, \sigma^2 \mathbb{I}_m)$ and $\gamma_l^b \overset{ind}{\sim} \mathrm{N}(\nu_l, \sigma^2)$ for $l = 2, \ldots, q$. The model uncertainty is then addressed using a Bayesian hierarchical model with a DL prior on both the response-specific curves and the overall effect curves. However, we do not use shrinkage priors for the mean confounder coefficients ν_{1r} and ν_l in order to conservatively account for their effects in our study. The following prior distribution structure is assumed for the model parameters in the pesticide main effects. A similar structure can be used for the parameters in the interaction-effect functions:

$$\begin{aligned}
&\boldsymbol{\beta}_j^b | \boldsymbol{\mu}_j, \lambda_j, \sigma^2 \overset{ind}{\sim} \mathbb{N}_m(\boldsymbol{\mu}_j, \lambda_j \sigma^2 \mathbb{I}_m), \quad \boldsymbol{\mu}_j | \omega_j \overset{ind}{\sim} \mathbb{N}_m(0, \omega_j \mathbb{I}_m), \\
&\lambda_j | \phi_j', \tau' \overset{ind}{\sim} \mathrm{Exp}(\phi_j' \tau'), \quad \boldsymbol{\phi'} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau' \sim \mathrm{Gamma}(p\alpha, 2), \\
&\omega_j | \phi_j, \tau \overset{ind}{\sim} \mathrm{Exp}(\phi_j \tau), \quad \boldsymbol{\phi} \sim \mathrm{Dirichlet}(\alpha, \ldots, \alpha), \quad \tau \sim \mathrm{Gamma}(p\alpha, 2), \\
&\sigma^2 \sim \mathrm{InvGamma}(0.01, 0.01), \quad \nu_{1r}, \nu_l \overset{iid}{\sim} \mathrm{N}(0, 10^2).
\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

This hierarchical model centers all N exposure-response curves around the overall effect curve by shrinking the response-specific coefficient β_{jr}^b to μ_{jr}. Thus each response-specific curve $f_j^b(X_j)$ shrinks toward the average curve for the overall effects of the exposures: $\bar{f}_j(X_j) \approx \sum_{r=1}^m \mu_{jr} B_r(X_j)$. A small λ_j shrinks all N curves toward $\bar{f}_j(X_j)$; a large λ_j allows for variations among the response-specific curves $f_j^b(X_j)$, for $j = 1, \ldots, N$. For the overall effects curve, which reflects the average main effect across health responses, a small ω_j shrinks the average main-effect function for exposure X_j toward zero. Thus, the jth pesticide does not influence the overall NB system significantly. However, a large ω_j allows for a significant association between X_j and the NB system through the nonparametric function $\bar{f}_j(X_j)$. If one pesticide is not associated with any of the response variables, such that the overall effects are negligible, both λ_j and ω_j are small and all curves shrink toward zero.

6.3 NB data analysis

We use fivefold cross validation to select the number of basis functions (we consider $m_1 \in \{5, 10, 15, 20\}$) and the Dirichlet parameter (we consider $\alpha \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$). The best prediction performance is achieved (MSE = 0.908) with $m_1 = 10$ and $\alpha = 0.5$. We also find that this nonparametric additive model outperforms the linear regression model using least squares (MSE = 1.023), generalized additive model using the restricted maximum likelihood method (MSE = 0.993), MARS (MSE = 0.973), and COSSO (MSE = 0.965) when the responses are analyzed separately.

Figure 2(a) shows the posterior samples for the $\mathcal{L}_2$-norm of the mean curve coefficients, $\|\boldsymbol{\mu}_j/\sigma\|$. Using the variation-explained measurements in Figure 2(b), the size of the final model is chosen as the smallest model for which the 80% credible interval includes the median variation-explained value of the full size model. Three pesticides are selected: Parathion, Benomyl, and Chlorpyrifos. In the additive nonparametric regression model that includes both main effects and interactions, the thresholding method selects three main effects of pesticides, and excludes the interaction effects. Therefore, we only show the results for the

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

coefficients of the individual main effects, because the interaction effects are negligible.

Figure 3 plots the average exposure-response functions $\bar{f}_j(X_j)$ for each pesticide main effect. Each individual curve is a function of the cumulative number of pesticide applications (i.e., the original measurement before the rank transformation). The mean curves plateau for large exposures because these exposures are rare in the sampled subjects. Among the selected pesticides, Parathion and Chlorpyrifos show a decreasing pattern in the mean curves, implying that these pesticides have negative overall effects on central nervous systems. The mean curve for Benomyl shows a positive effect, but that might be because of the collinearity with other pesticides.

The variance of the curves across pesticide exposure j, λ_j, and the variance of the av-

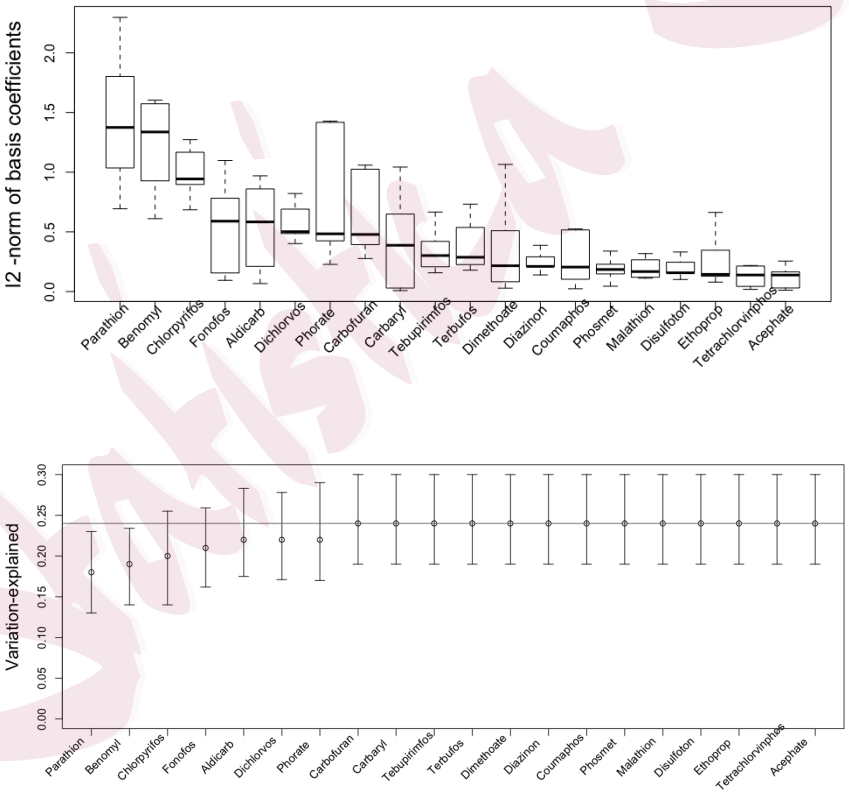

Figure 2: (a) Box-plot for the $\mathcal{L}_2$-norm of the mean curve basis coefficients $\|\boldsymbol{\mu}_j/\sigma\|$; (b) Variation-explained plot at different model sizes (the horizontal line is the full model "variation-explained" measurement).

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

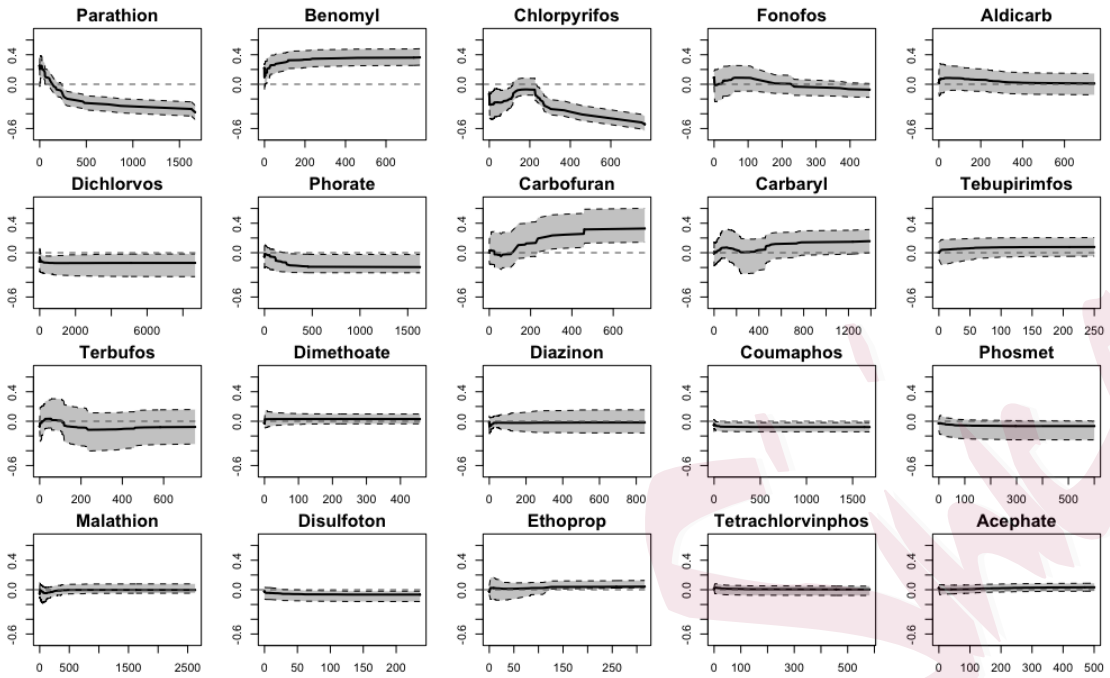

Figure 3: Average (over Central Nervous System tests) exposure-response curves $\bar{f}_j(X)$ for each pesticide in the Agricultural Health Study. The x-axis is the cumulative number of pesticide applications. The solid lines are the posterior means and the dashed lines are point-wise 95% credible intervals. The first three plots in are the pesticide covariates selected by the DL model.

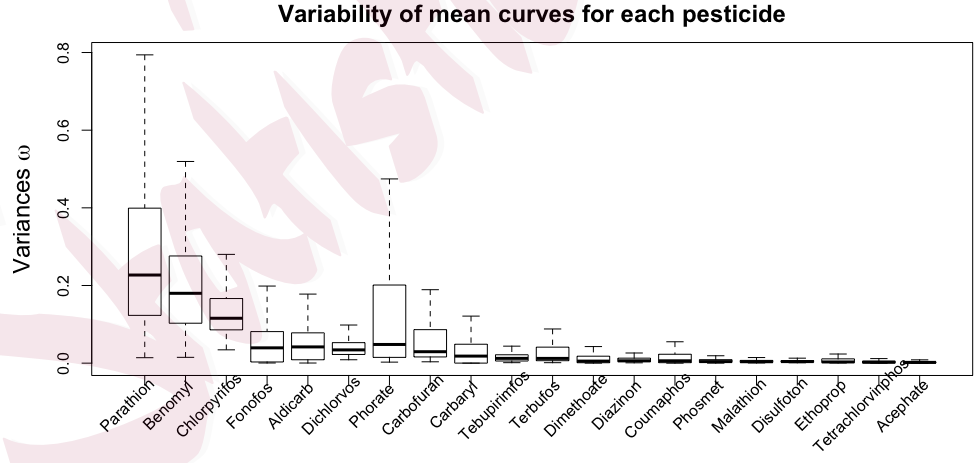

Figure 4: Posterior distribution of the normal variance ω_j for the average coefficients across CNS responses.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

erage curve for pesticide j, ω_j, illustrate the overall importance of each pesticide on health response. The posterior samples of all λ_j are concentrated near zero, indicating that there is no significant difference between the response-specific main-effect functions for each CNS measurement. Therefore, the average exposure-response curves $\bar{f}_j(X)$ are sufficient to delineate the associations between pesticides and the overall NB test results. For the posterior samples of ω_j, we present the box-plot in Figure 4. The values of the variance ω_j indicate the average effects of each pesticide on the overall CNS responses and, thus, determine the covariates to be included in the model using the variable selection technique described in Section 2.1.

In conclusion, we detect significant associations between three pesticide chemicals and CNS overall functions using the nonparametric model with a multivariate DL prior. In contrast, the other parametric or nonparametric methods cannot find associations from the data. Compared with the simple linear regression analysis results of Starks et al. (2012a), our proposed method chooses a sparse model and demonstrates nonlinear effects on the overall performance of CNSs. By integrating the CNS response variables, we may have greater utility for those outcome measurements, because an individual NB test may fail to capture the overall impact.

7 Summary

We propose a nonparametric regression model with an additivity assumption on the main-effect and interaction-effect functions, motivated by a study of multiple pesticide exposures. The additive nonparametric functions in the decomposed model are approximated by a B-spline basis expansion with a multivariate extension of the shrinkage prior on individual functions. Furthermore, we show the posterior consistency of the model prediction and variable selection for the additive nonparametric regression model. We apply the model to NB data from the AHS, showing that the proposed method achieves good prediction

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

accuracy and identifies the subset of pesticide exposures that contribute most to the NB function.

A limitation of this study is that the proposed method deals with continuous response measurements only. Because there are binary or count NB responses in the data sets, it would be useful to extend the additive nonparametric regression model to include categorical response variables. Brezger and Lang (2006) propose a generalized structured additive regression for nonlinear effects of continuous covariates. Their MCMC simulation methods can be combined with the multivariate shrinkage priors on the B-spline basis coefficients and, therefore, implemented as a nonGaussian extension of our proposed model. Extensions to address multiple time or spatial measurements are also desirable.

Supplementary Material

The proofs of the theorems are provided in the Supplementary Material.

Acknowledgements

We thank Dr. Fred Gerr of the University of Iowa for providing the neurobehavioral data. The data for this work were supported by the National Institute of Environmental Health Sciences (NIEHS), grant R01-ES013067-03, and the Intramural Research Program of the National Institutes of Health (National Institute of Environmental Health Sciences Z01ES04903 and National Cancer Institute Z01CP010119). Reich was partially supported by NIH grants R21ES025374, R21ES022795-01A1, and R01ES014843-02. Ghosal's research was partially supported by NSF grant DMS-1510238.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

References

[1] Armagan, A., Dunson, D. and Lee, J. (2013). Generalized double Pareto shrinkage. *Statistica Sinica.* **23**, 119–143.

[2] Armagan, A., Dunson, D., Lee, J., Bajwa, W. and Strawn, N. (2013). Posterior consistency in linear models under shrinkage priors. *Biometrika.* **100**, 1011–1018.

[3] Bhattacharya, A., Pati, D., Pillai, N. and Dunson, D. (2015). Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association.* **110**, 1479–1490.

[4] Bobb, J., Valeri, L., Henn, B. C., Christiani, D. C., Wright, R. O., Mazumdar, M., Godleski, J. J. and Coull B. A. (2015). Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. *Biostatistics.* **16**, 493–508.

[5] Brezger, A. and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis.* **50-4**, 967–991.

[6] Carvalho, C. and Polson, N. (2010). The horseshoe estimator for sparse signals. *Biometrika.* 97, 465–480.

[7] Castillo, I., Schmidt-Hieber, J. and van der Vaart, A. (2015). Bayesian linear regression with sparse priors. *The Annals of Statistics.* **43**, 1986–2018.

[8] Curtis, S., Banerjee, S. and Ghosal S. (2014). Fast Bayesian model assessment for nonparametric additive regression. *Computational Statistics and Data Analysis.* **71**, 347–358.

[9] Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics.* **19**, 1–141.

[10] George, E. and McCulloch, R.E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association.* **88**, 881–889.

Statistica Sinica: Preprint
doi:10.5705/ss.202017.0315

[11] Griffin, J. E. and Philip, J. B. (2010). Inference with normal-gamma prior distributions in regression problems. *Bayesian Analysis. 5*, 171-188.

[12] Gu, C. (2002). Smoothing spline ANOVA models. *Springer*.

[13] Hahn, R. and Carvalho, C. (2015). Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective. *Journal of the American Statistical Association.* **110**, 435–448.

[14] Lin, Y. and Zhang, H. (2006). Component selection and smoothing in multivariate nonparametric regression. *The Annals of Statistics.* **34**, 2272–2297.

[15] Linkletter, C., Bingham, D. Hengartner N., Higdon, D. and Ye, K. (2006). Variable selection for Gaussian process models in computer experiments. *Technometrics.* **48**, 478–490.

[16] Reich, B., Storlie, C. and Bondell, H. (2009). Variable selection in Bayesian smoothing spline ANOVA models: Application to deterministic computer codes. *Technometrics.* **51**, 110–120.

[17] Savitsky, T., Vannucci, M. and Sha, N. (2011). Variable selection for nonparametric Gaussian process priors: models and computational strategies. *Statistical Science.* **26**, 130–149.

[18] Scheipl, F., Fahrmeir, L. and Kneib, T. (2012). Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association.* **107**, 1518–1532.

[19] Song, Q. and Liang, F. (2016). Nearly optimal Bayesian shrinkage for high dimensional regression. *Preprint.*

[20] Starks, S., Gerr, F., Kamel, F., Lynch, C., Alavanja, M., Sandler, D. and Hoppin, J. (2012). High pesticide exposure events and central nervous system function among

pesticide applicators in the Agricultural Health Study. *Int Arch Occup Environ Health.* **85**, 505–515.

[21] Starks, S., Hoppin, J., Kamel, F., Lynch, C., Jones, M., Alavanja, M., Sandler, D. and Gerr, F. (2012). Peripheral nervous system function and Organophosphate pesticide use among licensed pesticide applicators in the Agricultural Health Study. *Environmental Health Perspectives.* **120**, 515–520.

[22] Wood, S., Shively, T. and Jiang, W. (2002). Model selection in spline nonparametric regression. *Journal of Royal Statistical Society: Series B.* **64**, 119–139.

[23] Yoo, W. and Ghosal, S. (2016). Supremum norm posterior contraction and credible sets for nonparametric multivariate regression. *The Annals of Statistics.* **44**, 1069–1102 .