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
Notice: Accepted version subject to English editing.	

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, the statistical challenges are to identify the subset of significant covariates and to estimate their joint effect. One example is an epidemiological study that analyzes the effects of exposure variables on a health response. In order to make inference on the covariate effects, we propose a Bayesian additive nonparametric regression model with a multivariate continuous shrinkage prior to address model uncertainty and identify important covariates. Our general approach is to decompose the response function as the sum of nonlinear main effects and two-way interaction terms, and apply the computationally-advantageous Bayesian variable selection method to identify important effects. The proposed Bayesian method is a multivariate Dirichlet-Laplace prior that aggressively shrinks many of the terms towards zero, thus mitigating the noise of including unimportant exposures and isolating the effects of important covariates. The theoretical studies demonstrate asymptotic prediction and variable selection consistency properties, and the numerical simulations evaluate model performance of prediction and variable selection under practical scenarios. The method is applied to a neurobehavioral data set from 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 neurobehavioral responses, while restricting the number of covariates included in the model through variable selection.

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

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 mixture of multiple chemicals and a suite of health endpoints to increase statistical power and paint a more realistic picture of health risk. As a motivating application, we analyze the 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 neurobehavioral health endpoints for each of the 701 farmers from Iowa and North Carolina. In previous work, Starks et al. (2012a, b) use a linear regression model to examine the associations between the indicator of ever having a pesticide exposure event and each health endpoint of NB tests separately. They conduct conventional hypothesis testing, and conclude that two of the nine neurobehavioral end points are observed to have negative association with pesticide exposure.

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

While fully nonparametric models are robust to model assumptions, they suffer from a lack of interpretability and are hard to fit in high dimensions. In order to reduce complexity, an additive model can be assumed to decompose the joint effect function into the summation of individual effects and interaction effects. For example, the smoothing spline ANOVA (SS-ANOVA, Gu 2002) models nonlinear main effects and higher-order interactions among predictors. Reich et al. (2009) implement the 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, Fahrmeir and Kneib (2012) propose structured additive regression model for nonlinear functions with spike-and-slab prior. Their method aims to select relevant covariates and determine the effects under different scenarios such as Gaussian and non-Gaussian models. One of the disadvantages of these methods is their computational burdens for large problems. Under similar model setting, Curtis et al. (2014) use a multivariate Laplace prior on the basis coefficients in their additive nonparametric model, which relies on large sample approximation for parameter sampling. In this paper, we apply the same additive regression model with basis expansion technique as in Curtis et al. (2014) but use a different Bayesian variable selection method.

Variable selection techniques under Bayesian framework have been studied extensively, especially for the high-dimensional linear regression model. 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. As a complex nonparametric model that requires heavy computation, we are interested in a Bayesian method that can substantially alleviate the computational burden. As a computationally efficient alternative to SSVS priors, shrinkage priors are continuous distributions imposed on model parameters and mimic the behavior of SSVS priors with dominant peak near zero and heavy tails. Various options of shrinkage priors such as Horseshoe prior (Carvalho et al. 2010), normal-gamma prior (Griffin and Phillip, 2010), double Pareto prior (Armagan et al. 2013a) and Dirichlet-Laplace prior (Bhattacharya et al. 2015) are proposed and shown to fall in the family of Gaussian global-local scale mixtures. Theoretically, shrinkage priors obtain almost the same contraction rate as the point-mass prior for recovering the model parameters and the true subset of covariates in the model, for both low-dimensional (Armagan et al. 2013b) and high-dimensional (Song and Liang, 2017) model.

In this paper, an additive nonparametric regression model is assumed for both main effect and interaction effect between covariates. We consider multivariate continuous shrinkage prior on the block of B-spline basis coefficients for variable selection. The contributions to the literature are in two major ways. First, we address the model uncertainty in nonparametric additive regression setting by incorporating block variable selection on B-spline basis expansion coefficients. Therefore, the concentration of posterior distributions of the block of basis coefficients near zero is reported

to identify the significant main effects and interactions. Second, we expand the notion of the computationally-efficient Dirichlet-Laplace (DL) prior introduced in Bhattacharya et al. (2015) to multidimensional vectors to achieve simultaneous shrinkage on the basis coefficient vector for each main or interaction effect function. In theoretical research, we expand the current prediction consistency and variable selection consistency results for shrinkage prior under linear regression model to the proposed additive nonparametric model, where induced bias from B-spline basis approximation and shrinkage on multidimensional vectors pose the major challenges. Further, we use neurobehavioral data from AHS to explore the health effects of multiple pesticides measurements and how the effects on each health endpoint differ from the effects on overall neurobehavioral system.

2 Model Description and Prior Specification

2.1 Main-effect-only model

We first describe the main-effect-only model that assumes the regression mean function as the summation of 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 all the covariates are standardized to lie in 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 as the sum of 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 individual covariate is sufficiently smooth, it can be approximated using B-spline basis expansion with a predetermined number of basis functions, m. For the covariate matrix X with elements all scaled to 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 $B_r(\cdot)$, $r = 1, \ldots, m$, are 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), effect quantity and model uncertainty of each covariate X_j are addressed through the basis coefficients $\boldsymbol{\beta}_j$. The m-vector coefficients $\boldsymbol{\beta}_j$s are assigned multivariate normal priors with mean zero 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_js determine the shrinkage on the m-vector basis coefficients so that the problem of selecting important main effects reduces to the shrinkage of the λ_js. Using the Dirichlet-Laplace (DL) prior in Bhattacharya et al. (2015), λ_j follows an exponential distribution with mean $\tau\phi_j$, where τ is the global factor that determines the tail of the marginal distribution of λ_j's and $\phi_j > 0$, with $\sum_{j=1}^{p} \phi_j = 1$, is the proportion of variance allocated to covariate X_j. Further, 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 Dirichlet distribution controls the level of shrinkage so that a smaller α gives more concentration around zero and thus a sparser model. The uncertainty on each nonparametric main effect function is addressed by the implementation of 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 regression. As a shrinkage prior on the B-spline basis coefficients, the proposed multivariate DL method leads to a slightly different prior from 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, we get the marginal prior distribution of β_j given (ϕ_j, τ) is 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, for this case, the original DL prior places more mass near zero than the proposed multivariate extension with only one basis term. Despite the differences in quantities, our proposed multivariate DL prior for nonparametric regression is an extension of DL prior in linear regression and they

share similar shrinkage properties.

2.2 Main and interaction effects model

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

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

where the second-order term $f_{kl}(X_k, X_l)$ represents the interaction effects on health endpoint between covariates X_k and X_l. We consider only two-way interactions in this model since higher order interactions are less interpretable and including them will make 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 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, and 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 main effects and interactions.

2.3 Identifiability constraints

We propose identifiability constraints on the functions f_j and f_{kl} in the additive model (6) so that all the model parameters can be uniquely determined from the distribution of observations (X, Y). For example, adding a constant to f_j and subtracting the same constant from $f_{j'}$ for $j \neq j'$ gives the same mean regression function but different individual main effect functions. The lack of identifiability does not translate into a completely different qualitative relations between the predictors and the response variable, but the constraints are imposed for easier interpretation and presentation, especially in theoretical analysis. We choose to restrict the main effect and interaction functions to integrate to zero so that the shrinkage prior would encourage shrinkage towards zero as is customary in the variable selection literature. For the main effect functions, we assume that $\int_0^1 f_j(x_j)dx_j = 0$, $j = 1, \ldots, p$. For 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, so 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

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

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 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):

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

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

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) 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 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

Due to the properties of continuous shrinkage priors, the factors $f_j(\cdot)$ will never equal to zero exactly. Thus, a post-processing procedure is needed in order to determine zero and nonzero effects. We implement a thresholding technique based on the idea of choosing a subset of predictors so that the corresponding deterioration in prediction accuracy in terms of "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 for are calculated as

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

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 only includes the covariates with 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 so that the $(1-\alpha_0) \times 100\%$ credible interval of $V_{(k)}$ includes the posterior mean of 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 the combination of Gibbs sampling and direct sampling. The sampler cycles through (i) $\boldsymbol{\beta}|\lambda, \phi, \tau, Y, X$, (ii) $\lambda|\boldsymbol{\beta}, \phi, \tau, Y, X$ and (iii) $\phi, \tau|\lambda, \boldsymbol{\beta}, Y, X$. Within step (iii), it follows

direct sampling of (iiia) $\tau|\phi,\lambda$ and (iiib)$\phi|\lambda$.

(i) Given λ_j, Y and X_j, the conditional posterior distribution for $\boldsymbol{\beta}_j$ 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 basis coefficients in (12), we sample $\boldsymbol{\beta}_j \big| \sum_{r=1}^{m} \beta_{jr} D_r = 0$ from 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 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 $\lambda_1, \ldots, \lambda_p$, first sample T_j, $j = 1, \ldots, p$, 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 individual main effects with the multivariate DL prior. Since predictors are considered deterministic in our setting, extension to include interactions is similar except that the full model will have more terms.

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 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 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 the matrix of the value of B-spline basis functions, $\boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, \ldots, \boldsymbol{\beta}_{p_n}^T)^T$ is the $p_n m_n$-vector with each m_n-dimensional components corresponding to each covariate X_j and ε is an n-dimensional standard normal vector. We study a Bayesian approach with continuous shrinkage prior for the m_n-dimensional vectors $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_{p_n}$. After integrating out 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 parameters, $\xi^* \subset \{1, \ldots, p_n\}$ be the indices of 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 will be defined later. These assumptions are similar as those in Song and Liang (2016), where they presented the asymptotic properties of shrinkage priors in the linear regression model. The difference lies in the fact that we are dealing with the design matrix of basis expansions, and the additive nonparametric functions are estimated by B-spline basis expansion so that an estimation bias is introduced. Let $a \prec b$ mean $\lim a/b = 0$ and $a \asymp b$ mean "$\lim a/b$" is bounded by constants.

- $A_1(1)$: 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
- $A_2(1)$: $\{\epsilon_n\}$ sequence is assumed to satisfy that $s m_n \log p_n \prec n\epsilon_n^2$ 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 exists 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 qm_n columns of $B(X)$

The assumption of high dimensionality in $A_1(1)$ is mainly used for concise representation of certain bounds. With lower dimensionality of the parameter space, the same asymptotic properties can be obtained but the following assumptions and proofs will be treated slightly differently. To save space and avoid monotonicity of arguments, it is customary to forgo separate arguments for the lower dimensional situation. In $A_1(2)$, κ defines the minimum smoothness for all additive functions, so that the bias induced by m_n-dimensional basis expansion is $\delta \asymp m_n^{-\kappa}$. The assumption $A_1(3)$ is an extension of linear regression model in Song and Liang (2016), by replacing covariate matrix X to 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, so $A_1(3)$ is specified. The assumption in $A_2(1)$ restricts ϵ_n to be $\max(\sqrt{sm_n \log p_n/n}, m_n^{-\kappa})$.

Since the nonparametric regression function assumes no model parameter, 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 prediction performance of $B(X)\boldsymbol{\beta}$ is asymptoticly 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 that* $\|\delta\| \lesssim \sqrt{n}m_n^{-\kappa}$, *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 prior density* $\pi_\alpha(\cdot)$ *as 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 posterior concentration rate $\sqrt{n}\epsilon_n$ for 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 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 posterior variable selection consistency given that $a_n \prec \log p_n/np_n$ and the prior density is moderately flat at nonzero basis coefficients $\boldsymbol{\beta}_j^*/\sigma^*$. The proof of the theorem is given in the supplementary material.

5 Simulation Study

We first consider the additive nonparametric regression model that only includes individual main effects as in Section 2.1, and then 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^*, $j = 1, \ldots, p$, from Gaussian distribution with $\mathrm{E}(X_j^*) = 0$, $\mathrm{Var}(X_j^*) = 1$ and $\mathrm{Cov}(X_j^*, X_k^*) = 0$ for mutually independent case or $\mathrm{Cov}(X_j^*, X_k^*) = 0.5^{|j-k|}$ for 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

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

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 a testing data:

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

where $f(\cdot)$ is the true mean function, $\hat{f}$ is the estimated mean function, $X'_i = (x'_{i1},\ldots,x'_{ip})^T$ is a new data point randomly sampled from the proposed covariate distribution, $i = 1,\ldots,500$. Note that X'_i's are not used for model fitting but only for evaluating the prediction performance. Therefore, we can compare the prediction performance of each method on future observations.

We also record the variable selection performance for correctly identifying the four significant covariates. We evaluate variable selection results in terms of the percentage of unimportant variables selected (False Positive) and the percentage of important variables excluded (False Negative), averaged over all the simulated data sets under each scenario. We further include the proportion

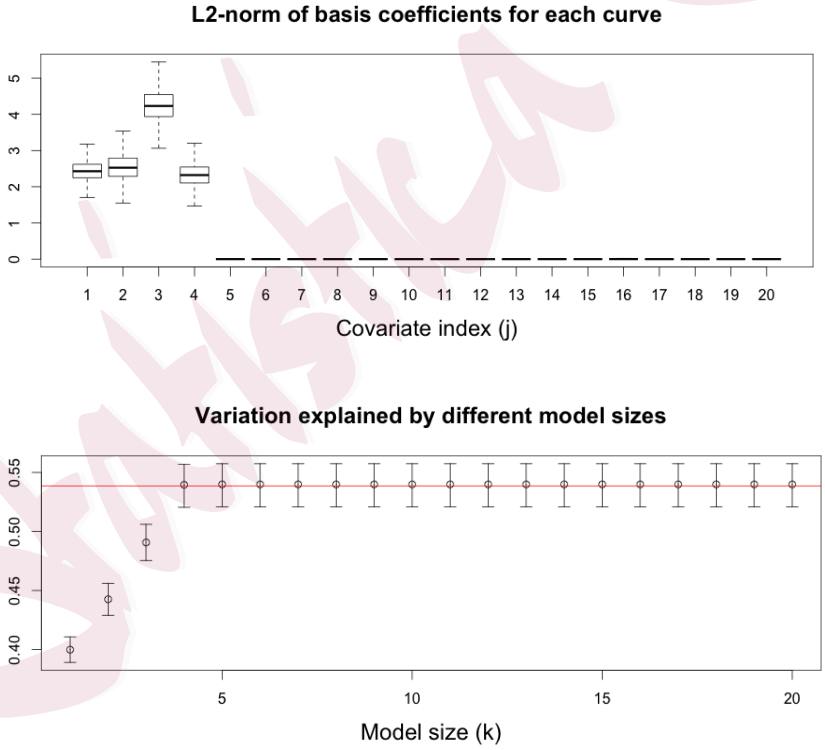

Figure 1: (a) Box-plot of the posterior distribution of the $\mathcal{L}_2$-norm $\|\boldsymbol{\beta}_j/\sigma\|$ for a single simulated data set; (b) Variation-explained plot at different model sizes for a single data set (the horizontal red line is the full model "variation-explained" measurement).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

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 the box-plot of posterior samples of $\|\boldsymbol{\beta}_j\|$, $j = 1, \ldots, p$. 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 variable selection is determined by the smallest k so that the 80% interval of $V_{(k)}$ includes the posterior mean of full model $V_{(50)}$ (red line in Figure 1(b)). We only include model sizes less than 20 as 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 competitors with similar model assumptions, we fit the additive nonparametric model with a multivariate DL prior (DL method) to each data set and compare with four competing methods assuming additive non-parametric functions: approximate Bayesian method ("ABayes") by Curtis et al. (2014); the Bayesian smoothing splines ANOVA ("BSS-ANOVA") by Reich et al. (2009); component selection and smoothing operator ("COSSO") by Lin and Zhang (2006) and multivariate adaptive regression splines model ("MARS") by Friedman (1991). We also implement the "Oracle" selection in DL method which includes the four covariates with the largest posterior medians of $\|\boldsymbol{\beta}_j/\sigma\|$. Given the correct information on the number of significant covariates, this DL oracle method demonstrates the ability of the general DL method to rank important covariates and formally select significant ones. For MARS, we use the function `polymars`$(\cdot)$ in the R package `polspline`. For 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. As for the prediction performance, our proposed DL method is better than the competing methods, even though all the methods make similar assumptions about the regression model. The DL method achieves the smallest MSE under all the scenarios, and MARS has good prediction accuracy under the independent case. For the other methods such as BSS-ANOVA and COSSO, their performances are similar with BSS-ANOVA being slightly better than COSSO. The ABayes method does not provide prediction in their available code so there are no results for this method.

$n = 200$		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 mean squared errors ("MSE"), False Positive ("FP"), False Negative ("FN") and True model ("True") with their standard errors ("SE") under independent and autoregressive covariate covariance in parentheses. All values except the MSE's are given in percentages (%).

As expected, the DL Oracle method performs the best under all scenarios for variable selection. The DL method with data-driven threshold, DL(VarExp), is slightly worse than DL(Oracle), but still outperforms other methods. In conclusion, the thresholding policy defined in (15) adds uncertainty in determining the number of important covariates so that when the number is fixed like in the Oracle method, the proposed method correctly ranks the covariates through shrinkage and

achieves better variable selection accuracy. For the others, ABayes is too conservative in choosing subset of covariates with false positive equal to zero under every scenario and thus large false negative proportion. 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 the 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 with small sample size and correlated predictors.

5.3 Model with main and interaction effects

We also conduct simulation for additive models with both main effects and interactions. Since 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 standard normal distribution and functions f_1, f_2 and f_3 are the same as 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 $f_j(X_j)$ with 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, similar to the main-effect-only model. The DL method is compared with ABayes, BSS-ANOVA, COSSO and MARS methods on both prediction performance and variable selection. The accuracy of model prediction is evaluated by computing the mean squared error (MSE) for a newly generated covariate matrix, while variable selection performance is determined by False Negative, False Positive and the percentages of correctly identifying the main effects, interactions and the complete model. In ABayes and COSSO where interactions are not considered in the available

				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 simulation study with main and interaction effects. Methods are compared in terms of mean squared errors ("MSE"), False Positive ("FP"), False Negative ("FN") and correct selection of main effects, interactions, as well as complete model with their standard errors ("SE") in parentheses. All values except the MSE's are given in percentages (%).

code, 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, it is actually in favor of these methods by providing extra information on the correct format of interaction effects. The simulation results are summarized in Table 2.

From Table 2, the simulation results show that our proposed method improves the prediction on the newly-generated data set compared to other nonparametric methods. In particular, by correctly addressing the joint effect through decomposition of 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 shrinkage prior on 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 for selecting the correct model that includes both nonzero main and interaction effects. The DL method has the highest true model proportion among all 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 for ABayes and COSSO where the true format of the interaction term is specified, they cannot outperform the DL method under either scenario. The second best method is the BSS-ANOVA, but this requires 5 times more computing time than the DL method.

6 Analysis of the AHS Neurobehavioral Dataset

6.1 Description of the neurobehavioral data

We demonstrate our method using data from a neurobehavioral (NB) sub-study of the Agricultural Health Study (AHS; `http://aghealth.nih.gov/`). The goal of the NB study (data version number: AHS44436) is to examine the association between pesticide exposures and neurobehavioral function of the central nervous system (CNS). From 2006 to 2008, $n = 701$ male farmers from Iowa or North Carolina took neurobehavioral 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 focus on the analysis of these 8 continuous CNS response variables. In Starks et al. (2012a, b), they conclude that participants with one or more pesticide exposures are more likely to have adverse CNS outcomes but they have not investigated the association of each individual pesticide effects on the overall neurobehavioral system. In this application, the exposure variables are the lifetime-specific pesticide use information for $p = 20$ pesticides from AHS questionnaires and interviews. Each exposure covariate is quantified as the days of applications of 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 $X_{ij} \in (0, 1)$ so 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 unit interval, while still allowing for a wide range of regression relationships between covariates X and response Y via the additive nonparametric function. All response measurements are standardized to have mean zero and variance 1, 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 indicated better performance.

6.2 Multivariate extension

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

For CNS response variables, we model the confounding variable age (Z_1) as having a nonparametric effect on health response but the other confounding variables ($Z_2, \ldots, Z_q$) as linear effects. The additive nonparametric model for response variable Y_b, $b = 1, \ldots, N$, on confounder Z and pesticide exposures 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 the 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 through basis expansions as in (6) and (7).

To build the connection between CNS health responses, we specify the Bayesian hierarchical model so that the model coefficients for each response variable share the 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 coefficients for the nonparametric effect of covariate X_j on the overall neurobehavioral system (overall effect curve). The same multivariate extension is applied on the interaction effect functions so that $\boldsymbol{\mu}_{kl}$ determines the joint effect of X_l and X_k on overall neurobehavioral 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 by Bayesian hierarchical model with DL prior on both the

response-specific curves and overall effect curves. However, we do not use shrinkage priors for the mean confounder coefficients ν_{1r} and ν_l to conservatively account for their effects in our study. The following prior distribution structure is assumed for the model parameters in pesticide main effects, while similar structure can be followed for the parameters in interaction effect functions:

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

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}, and thus each response-specific curve $f_j^b(X_j)$ is shrunk towards the average curve for the overall effects of exposures: $\bar{f}_j(X_j) \approx \sum_{r=1}^m \mu_{jr}B_r(X_j)$. Small λ_j shrinks all N curves towards $\bar{f}_j(X_j)$; large λ_j allows for variations among the response-specific curves $f_j^b(X_j)$ for $j = 1, \ldots, N$. As for the overall effects curve that reflects the average main effect across health responses, small ω_j shrinks the average main effect function for exposure X_j towards zero so that jth pesticide does not influence the overall neurobehavioral system significantly; large ω_j, however, allows for a significant association between X_j and neurobehavioral system through 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 are shrunk toward zero.

6.3 Neurobehavioral data analysis

We use 5-fold 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 linear regression model using least squares (MSE=1.023), generalized additive model by 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 $\mathcal{L}_2$-norm of mean curve coefficients, $\|\boldsymbol{\mu}_j/\sigma\|$. Using

the variation explained measurements in Figure 2(b), the size of the final model is chosen by the smallest model for which the 80% credible interval includes the median variation explained value of 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 coefficients of individual main effects as 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 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

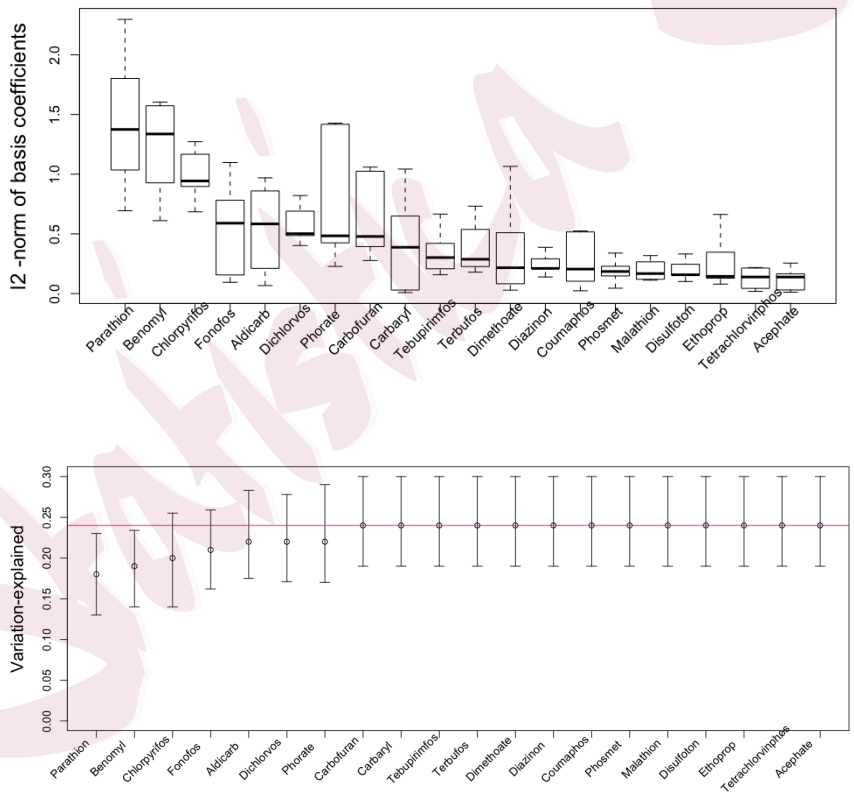

Figure 2: (a) Box-plot for the $\mathcal{L}_2$-norm of mean curve basis coefficients $\|\boldsymbol{\mu}_j/\sigma\|$.; (b) Variation-explained plot at different model sizes (the horizontal red line is the full model "variation-explained" measurement).

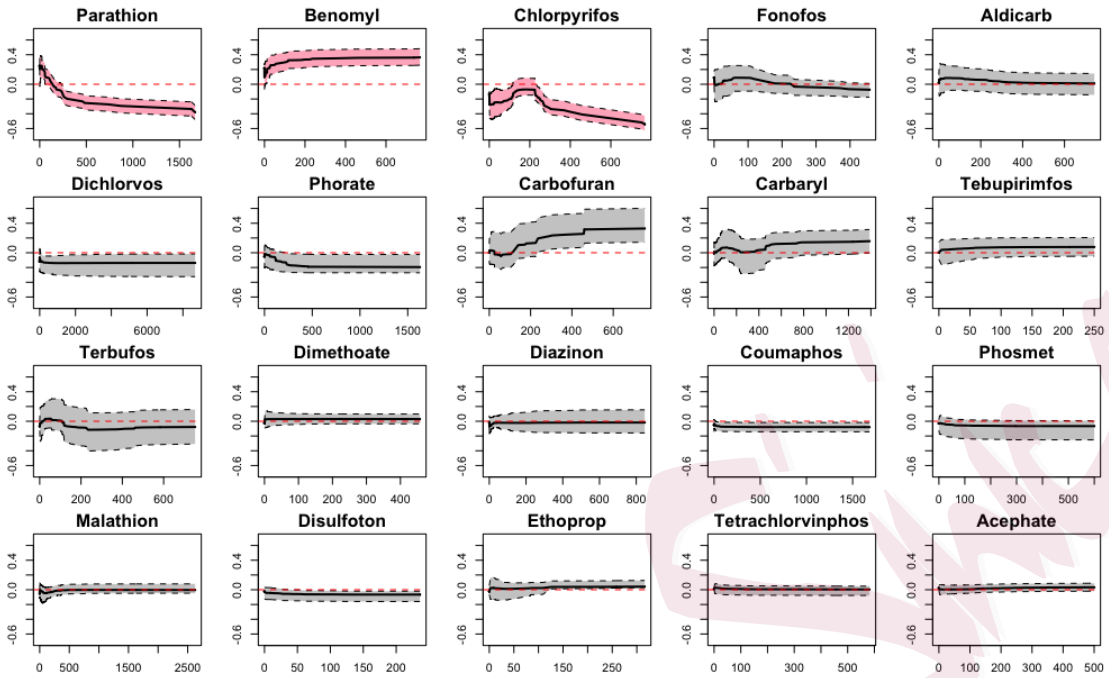

Figure 3: Average (over Central Nervous System tests) exposure-response curves $\bar{f}_j(X)$ for each pesticide in the Agricultural Health Study. The x-axis is the cumulative number of pesticide applications. The solid lines are the posterior means and the dashed lines are point-wise 95% credible intervals.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

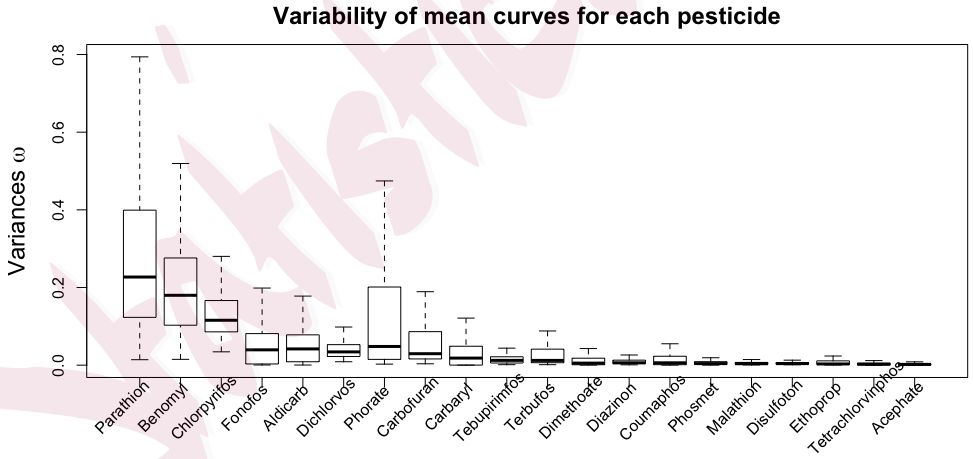

Figure 4: Posterior distribution of normal variance ω_j for average coefficients across CNS responses.

and Chlorpyrifos show decreasing pattern in the mean curves, which implies 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 curves across pesticide exposure j, λ_j, and the variance of the average curve for pesticide j, ω_j, illustrate the overall importance of each pesticide on health response. The posterior samples of all λ_js are concentrated near zero, which shows 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 variance ω_js indicate the average effects of each pesticide on the overall CNS responses and thus determine the covariates to be included in the model through variable selection technique as described in Section 2.1.

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

7 Summary

In this paper, we propose a nonparametric regression model with additivity assumption on 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 B-spline basis expansion with multivariate extension of shrinkage prior on individual functions. Further we show posterior consistency of model prediction and variable selection for the additive nonparametric regression model. In its application on the neurobehavioral data from AHS, the proposed method achieves good prediction accuracy and identifies subset of pesticide exposures that contribute the most to the neurobehavioral function.

As a limitation, the proposed method deals with continuous response measurements only. Since there are binary or count NB responses in the data sets, it will be useful to consider the extension of the additive nonparametric regression model into categorical response variables. Brezger and

Lang (2006) propose generalized structured additive regression for nonlinear effects of continuous covariates. Their MCMC simulation methods can be combined with multivariate shrinkage priors on B-spline basis coefficients and therefore implemented as a non-Gaussian extension of our proposed model. Extensions to address multiple time or spatial measurements are also desirable.

Supplementary Materials

The proof of Theorems are in the supplementary materials.

Acknowledgement

We thank Dr. Fred Gerr of the University of Iowa for providing the neurobehavioral data. The data for this work was supported by 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.

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.

[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 .