Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

RESEARCH ARTICLE

Spike-and-Slab Generalized Additive Models and Scalable Algorithms for High-Dimensional Data

Boyi Guo*[1] | Byron C. Jaeger[2] | AKM Fazlur Rahman[1] | D. Leann Long[1] | Nengjun Yi*[1]

[1]Department of Biostatistics, University of Alabama at Birmingham, Birmingham, USA

[2]Department of Biostatistics and Data Science, Wake Forest School of Medicine, Winston-Salem, USA

Correspondence
Boyi Guo and Nengjun Yi, Department of Biostatistics, University of Alabama at Birmingham, Birmingham, USA. Email: boyiguo1@uab.edu, Email: nyi@uab.edu

Present Address
This is sample for present address text this is sample for present address text

There are proposals that extend the classical generalized additive models (GAMs) to accommodate high-dimensional data ($p >> n$) using group sparse regularization. However, the sparse regularization may induce excess shrinkage when estimating smoothing functions, damaging predictive performance. Moreover, most of these GAMs consider an "all-in-all-out" approach for functional selection, rendering them difficult to answer if nonlinear effects are necessary. While some Bayesian models can address these shortcomings, using Markov chain Monte Carlo algorithms for model fitting creates a new challenge, scalability. Hence, we propose Bayesian hierarchical generalized additive models as a solution: we consider the smoothing penalty for proper shrinkage of curve interpolation via reparameterization. A novel two-part spike-and-slab LASSO prior for smoothing functions is developed to address the sparsity of signals while providing extra flexibility to select the linear or nonlinear components of smoothing functions. A scalable and deterministic algorithm, EM-Coordinate Descent, is implemented in an open-source R package BHAM. Simulation studies and metabolomics data analyses demonstrate improved predictive and computational performance against state-of-the-art models. Functional selection performance suggests trade-offs exist regarding the effect hierarchy assumption.

KEYWORDS:
Spike-and-Slab Priors; High-Dimensional Data; Generalized Additive Models; EM-Coordinate Decent; Scalablility; Predictive Modeling

1 | INTRODUCTION

Many modern biomedical research, e.g. sequencing data analysis, electric health record data analysis, require special treatment of high-dimensionality, commonly known as $p >> n$ problem. There is extensive literature on high-dimensional linear models via penalized models or Bayesian hierarchical models, see Mallick and Yi[1] for review. These models are built upon a restrictive and unrealistic assumption, linearity. In classical statistical modeling, many strategies and models are proposed to relax the linearity assumption with various degrees of complexity. For example, variable categorization is a simple and common practice in epidemiology, but suffers from power and interpretation issues. More complex models to address nonlinear effects include random forest and other so-called "black box" models[2]. These models are useful for statistical prediction but do not estimate parameters relevant to the data generation process that one can draw inferences from. In addition, how to generalize these "black box" models to the high-dimensional setting remains unclear.

For their easy interpretation and flexibility, nonparametric regression models serve as great alternatives to the "black-box" models in the context of prediction and variable selection. Among those, generalized additive models (GAMs), proposed in the seminal work of Hastie and Tibshirani[3], grew to be one of the most popular modeling tools. In a GAM, the response variable, which is assumed to follow some exponential family distribution, can be modeled with the summation of smoothing functions. Nevertheless, the classical GAMs cannot fulfill the increasing analytic demands for high-dimensional data analysis.

There exists some proposals to generalize the classical GAM to accommodate high-dimensional applications. The regularized models, branching out from group regularized linear models, are used to fit GAMs by accounting for the structure introduced when expanding smoothing functions. Ravikumar et al.[4] extended the grouped LASSO[5] to additive models (AMs); Huang et al.[6] further developed adaptive grouped LASSO for additive models; Wang et al.[7] and Xue[8] respectively applied grouped SCAD penalty[9] to additive models. Recently Bayesian hierarchical models are also used in the context of high-dimensional additive models, particularly within the spike-and-slab literature. Various group spike-and-slab priors[10,11] combining with computationally intensive Markov chain Monte Carlo (MCMC) algorithms are proposed, where the application on AMs are treated as a special case. Bai et al.[12] was the first to apply group spike-and-slab LASSO prior to Gaussian AMs using a fast optimization algorithm, and further generalized the framework to GAMs[13]. Focus on addressing the sparsity, these methods can excessively penalize the bases of a smoothing function and produce inaccurate predictions, particularly when complex signals are assumed and large number of knots are used.[14] In addition, these methods adapt an 'all-in-all-out' strategy, i.e. either including or excluding the variable completely, rendering no space for bi-level selection. Scheipl et al.[15] proposed a spike-and-slab structure prior that address the bi-level selection. But the model fitting relies on computational intensive MCMC algorithms and creates scalability concern. It would be of special interest to develop a fast, flexible and accurate generalized additive model framework.

We propose a novel Bayesian hierarchical generalized additive model (BHAM) for outcome prediction in the context of high-dimensional data analysis. Specifically, we incorporate smoothing penalties, derived from the smoothing spline literature[16], via reparameterization of smoothing functions to avoid excessive shrinkage on the bases. Smoothing penalties were also previously used in the spike-and-slab GAM[15] and the sparsity-smoothness penalty[17]. We then impose a new two-part spike-and-slab LASSO prior to address the sparsity of the signal. In addition, a scalable optimization-based algorithms, EM-Coordinate Descent (EM-CD) algorithm are developed. While the primary focus of this model is to improve prediction, the proposed model also provides utility in functional selection. Particularly, the two-part prior that follows the effect hierarchy principle motivates a bi-level selection, rendering one of three possibilities for each predictor: no effect, only linear effect, or linear and nonlinear effects. The proposed model is implemented in an publicly available R package `BHAM` via https://github.com/boyiguo1/BHAM.

The proposed framework, BHAM, differs from previous spike-and-slab based GAMs, i.e. the spike-and-slab GAM[15] and the SB-GAM[13] in three ways. First of all, the proposed spike-and-slab spline prior is a spike-and-slab LASSO type prior using independent mixture double exponential distribution, compared to spike-and-slab GAM that uses normal-mixture-of-inverse gamma prior. Spike-and-slab LASSO priors provide computational convenience during model fitting by using optimization algorithms instead of intensive sampling algorithms. They make fitting high-dimensional models more feasible without sacrificing performance in prediction and variable selection. Secondly, SB-GAM uses a group spike-and-slab LASSO prior with an EM-CD algorithm to fit the model. While both methods use the combination of expectation maximization algorithm and coordinate descent algorithm, there are subtle difference in the implementation due to the difference in prior specification. The proposed model sets up independent priors among basis coefficients after the reparameterization step, which provides some advantage in computation. Last but not least, the proposed model addresses the incapability of bi-level selection in SB-GAM.

In Section 2, we establish the Bayesian hierarchical generalized additive model, introduce the proposed spike-and-slab spline priors, and describe the fast-fitting EM-CD algorithm. In Section 3, we compare the proposed framework to state-of-the-art models via Monte Carlo simulation studies. Analyses of two metabolomics datasets are presented in Section 4. Conclusion and discussions are given in Section 5.

2 | BAYESIAN HIERARCHICAL ADDITIVE MODELS (BHAM)

We assume the response variable, Y, follows an exponential family distribution with density function $f(y)$, mean μ and dispersion parameter ϕ. The mean of the response variable can be modeled as the summation of smoothing functions, $B_j(\cdot), j = 1, \dots, p$, of a given p-dimensional vector of predictors $\boldsymbol{x}$, written as

$$E(Y|\boldsymbol{x}) = g^{-1}(\beta_0 + \sum_{j=1}^{p} B_j(x_j)) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{X}_j), \tag{1}$$

where $g^{-1}(\cdot)$ is the inverse of a monotonic link function. Given n data points $\{y_i, \boldsymbol{x}_i\}_{i=1}^n$, the data distribution is expressed as

$$f(\boldsymbol{Y} = \boldsymbol{y}|\boldsymbol{\beta}, \phi) = \prod_{i=1}^{n} f(Y = y_i|\boldsymbol{\beta}, \phi).$$

The basis function matrix, i.e. the design matrix derived from the smoothing function $\boldsymbol{B}_j(x_j)$, is denoted $\boldsymbol{X}_j$ for the variable x_j. The dimension of the design matrix depends on the choice of the smoothing function, and is denoted as K_j for x_j. $\boldsymbol{\beta}_j$ denotes the basis function coefficients for the jth variable such that $\boldsymbol{B}_j(x_j) = \boldsymbol{\beta}_j^T \boldsymbol{X}_j$. With slight abuse of notation, we denote vectors and matrices in bold fonts $\boldsymbol{\beta}, \boldsymbol{X}$ with conformable dimensions, where scalar and random variables are denoted in unbold fonts β, X. The matrix transposing operation is denoted with a superscript ^{T}. To note, the proposed model can include parametric forms of variables in the model, and hence considers general linear models and semiparametric regression models as special cases.

2.1 | Smoothing Function Reparameterization

To encourage proper smoothing of each additive function, we adopt the smoothing penalty from smoothing spline models[16]. A smoothing penalty is the quadratic norm of the basis coefficients and allows different shrinkage on different bases, mathematically

$$\text{pen}\left[\boldsymbol{B}_j(x)\right] = \lambda_j \int \boldsymbol{B}_j''(x)dx = \lambda_j \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j,$$

where $\boldsymbol{S}_j$ is a known smoothing penalty matrix and λ_j denotes a smoothing parameter. A linear function can be modeled as $\boldsymbol{B}_j(x_j) = x_j$ with the smoothing penalty matrix $\boldsymbol{S}_j = [0]$. Unlike previous regularized methods that either ignore the smoothing penalty completely or restrain the smoothing penalty as a component of sparse penalty which leads to a more restrictive solution, we consider an additional mechanism in pair with the proposed prior (described in Section 2.2) to address the smoothness and sparsity in signals such that the locally adaptive nature of the smoothing penalty retains.

Marra and Wood[18] proposed a reparameterization procedure to factor the smoothing penalty into the design matrix of each smoothing function. Given the smoothing penalty matrix $\boldsymbol{S}_j$ is symmetric and positive semi-definite for the univariate smoothing functions, we eigendecompose the penalty matrix $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, where the matrix $\boldsymbol{D}$ is diagonal with the eigenvalues arranged in the ascending order. To note, $\boldsymbol{D}$ can contain elements of zeros on the diagonal, where the zeros are associated with the linear space of the smoothing function. For the most popular smoothing function, cubic splines, the dimension of the linear space is one. Hereafter, we focus on discussing a uni-dimensional linear space for simplicity; however, it generalizes easily to the cases where the linear space is multidimensional. We further write the orthonormal matrix $\boldsymbol{U} \equiv \left[\boldsymbol{U}^0 : \boldsymbol{U}^*\right]$ containing the eigenvectors as columns in the corresponding order to $\boldsymbol{D}$. That is, $\boldsymbol{U}$ contains the eigenvectors U^0 with zero eigenvalues for the linear space and $\boldsymbol{U}^*$ contains the eigenvectors (as columns) for the non-zero eigenvalues, i.e. the nonlinear space. We multiply the basis function matrix $\boldsymbol{X}$ with the orthonormal matrix $\boldsymbol{U}$ for the new design matrix $\boldsymbol{X}^{\text{repa}} = \boldsymbol{X}\boldsymbol{U} \equiv \left[X^0 : \boldsymbol{X}^*\right]$. An additional scaling step is imposed on $\boldsymbol{X}^*$ by the non-zero eigenvalues of $\boldsymbol{D}$ such that the new basis function matrix $\boldsymbol{X}^*$ can receive uniform penalty on each of its dimensions. With slight abuse of the notation, we drop the superscript $^{\text{repa}}$ and denote $\boldsymbol{X}_j \equiv \left[X_j^0 : \boldsymbol{X}_j^*\right]$ as the basis function matrix for the jth variable after the reparameterization. A spline function can be expressed in the matrix form

$$\boldsymbol{B}_j(x_j) = \boldsymbol{B}_j^0(x_j) + \boldsymbol{B}_j^*(x_j) = \beta_j X_j^0 + \boldsymbol{\beta}_j^{*T} \boldsymbol{X}_j^*,$$

and the generalized additive model in Equation (1) now is

$$E(Y|\boldsymbol{x}) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \boldsymbol{B}_j(x_j)) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{X}_j) = g^{-1}\left[\beta_0 + \sum_{j=1}^{p} (\beta_j X_j^0 + \boldsymbol{\beta}_j^{*T} \boldsymbol{X}_j^*)\right], \tag{2}$$

where the coefficients $\boldsymbol{\beta}_j \equiv \left[\beta_j : \boldsymbol{\beta}_j^*\right]$ is an augmentation of the coefficient scalar β_j of linear space and the coefficient vector $\boldsymbol{\beta}_j^*$ of nonlinear space.

To summarize, the reparameterization step provides three benefits. First of all, the reparameterization integrates the smoothing penalty matrix into the design matrix, and encourages models to properly smooth the nonlinear function when sparsity penalty exists. Secondly, the eigendecomposition of the smoothing penalty matrix allows the isolation of the linear space from the nonlinear space, improving the feasibility of bi-level functional selection. Last but not least, the eigendecomposition facilitates the construction of orthonormal design matrix, which makes imposing independent priors on the coefficients possible. This reduces the computational complexity compared to using a multivariate priors, and improve the generalizability of the framework to be compatible with other choices of priors.

2.2 | Two-part Spike-and-Slab LASSO Prior for Smoothing Fucntions

The family of spike-and-slab regression models is one of most commonly used models in high-dimensional data analysis for its utility in outcome prediction and variable selection. Among all the spike-and-slab priors, the spike-and-slab LASSO (SSL) prior[19,20] is one of the most popular choices because it's highly scalable. The SSL prior is composed of two double exponential distributions with mean 0 and different dispersion parameters, $0 < s_0 < s_1$, mathematically,

$$\beta|\gamma \sim (1-\gamma)DE(0, s_0) + \gamma DE(0, s_1), 0 < s_0 < s_1.$$

The latent binary variable $\gamma \in \{0, 1\}$ indicates whether a variable x is included in the model, while the dispersion parameters s_0 and s_1 controls the shrinkage of the coefficient. Given that both double exponential distributions have a mean of 0 and the latent indicator γ can only take the value of 0 or 1, the mixture double exponential distribution can be formulated as one single double exponential density,

$$\beta|\gamma \sim DE(0, (1-\gamma)s_0 + \gamma s_1), 0 < s_0 < s_1. \tag{3}$$

Compared to other priors for high-dimensional data analysis, SSL has the following advantages. First of all, the SSL prior provides a locally adaptive shrinkage when estimating the coefficients. Secondly, the SSL prior encourages a sparse solution, making variable selection straight forward. Thirdly, the SSL prior motivates a scalable algorithm, the EM-CD algorithm, for model fitting, and hence is more feasible for high-dimensional data analysis. We defer to Bai et al.[21] for an detailed discussion.

We introduce a novel SSL-based prior for smoothing functions in GAMs. Given the reparameterized design matrix $\boldsymbol{X}_j = \left[X_j^0 \,:\, \boldsymbol{X}_j^*\right]$ for the jth variable, we impose a two-part SSL prior to the coefficients $\boldsymbol{\beta}_j = \left[\beta_j \,:\, \boldsymbol{\beta}_j^*\right]$. Specifically, the linear space coefficient has a SSL prior and the nonlinear space coefficients shares a group SSL prior,

$$\begin{aligned} \beta_j|\gamma_j, s_0, s_1 &\sim DE(0, (1-\gamma_j)s_0 + \gamma_j s_1) \\ \beta_{jk}^*|\gamma_j^*, s_0, s_1 &\overset{\text{iid}}{\sim} DE(0, (1-\gamma_j^*)s_0 + \gamma_j^* s_1), k = 1, \ldots, K_j \end{aligned} \tag{4}$$

where $\gamma_j \in \{0, 1\}$ and $\gamma_j^* \in \{0, 1\}$ are two latent indicator variables, indicating if the model includes the linear effect and the nonlinear effect of the jth variable respectively. s_0 and s_1 are scale parameters, assuming $0 < s_0 < s_1$ and given. These scale parameters s_0 and s_1 can be treated as tuning parameters and optimized via cross-validation, discussed in Section 2.4.

The proposed two-part SSL prior, particularly the group SSL prior of the nonlinear space coefficients, differs from previous group SSL priors[22,23], as the proposed prior follows the effect hierarchy principle. Effect hierarchy refers to the principle that "lower-order effects are more likely to be active than higher-order effects" defined by Chipman[24]. To implement, we consider the shared latent indicator of nonlinear coefficients γ_j^* depends on the value of the linear space latent indicator γ_j, while both latent indicators $\gamma_j and \gamma_j^*$ follow a Bernoulli distribution. While the probability of including the linear effect is θ_j, the probability of including the nonlinear effect is $\gamma_j\theta_j$.

$$\gamma_j|\theta_j \sim Bin(1, \theta_j) \qquad \gamma_j^*|\gamma_j, \theta_j \sim Bin(1, \gamma_j\theta_j).$$

This is, when the linear effect is not selected, the probability of including the nonlinear effect drops from θ_j to 0. For the computational convenience, we analytically integrate θ_j out such that $\gamma_j^*|\theta_j \sim Bin(1, \theta_j^2)$.

To allow the shrinkage to self-adapt to the sparsity and smoothing pattern of the data, we further specify the parameter θ_j follows a beta distribution with given shape parameters a and b,

$$\theta_j \sim Beta(a, b).$$

The beta distribution is a conjugate prior for the binomial distribution and hence provides some computation convenience. Having a prior distribution of θ_j enables the proposed prior to inherit the selective shrinkage property and self-adaptivity[21] from the classical SSL prior. In other words, when a smoothing function is significant, the coefficients of the smoothing function escape the overall shrinkage and produce a more accurate estimate, particular in pair with the smoothing penalty implicitly addressed via the reparameterization. Meanwhile, the hyper prior encourages information borrowing across coordinates and hence automatic adjust for different level of sparsity. Hereafter, we refer Bayesian hierarchical generalized additive models with the two-part spike-and-slab LASSO prior as the BHAM, and visually presented in Figure 1.

2.3 | Scalable EM-Coordinate Descent Algorithm

Despite the advantage to estimate posterior densities, using MCMC algorithms to fit the proposed model is computational prohibited and not feasible for high-dimensional data. Previous research shows the computation performance of MCMC algorithms

for spike-and-slab models is bottlenecked for medium size data (p=25)[25], and substantially slows as p increases modestly in the GAM context[14]. Hence, we consider the optimization algorithms that focus on the maximum a posteriori estimates at the cost of posterior inference. Specifically, we extend the EM-Coordinate Descent (EM-CD) algorithm to fit BHAMs. Similar to the EMVS algorithm[26] for spike-and-slab models, the EM-CD algorithm is based on the the expectation-maximization (EM) algorithm, integrating Coordinate Descent algorithm in each iterative step to find the posterior mode. The EM-CD algorithm has been well adapted in generalized linear models[27], Cox proportional hazards models[28], and their grouped counterparts[22,23]. The EM-CD algorithm provides deterministic solutions, which becomes a popular property for reproducible research.

For BHAMs, we define the parameters of interest as $\Theta = \{\boldsymbol{\beta}, \theta, \phi\}$ and consider the latent binary indicators $\boldsymbol{\gamma}$ as nuisance parameters of the model, in other words the "missing" data in the EM context. Our objective is to find the parameters Θ that maximize the posterior density function, or equivalently the logarithm of the density function,

$$\begin{aligned}
&\text{argmax}_{\Theta} \log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}) \\
&= \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{p} \left[\log f(\beta_j|\gamma_j) + \sum_{k=1}^{K_j} \log f(\beta_{jk}^*|\gamma_j^*) \right] \\
&+ \sum_{j=1}^{p} \left[(\gamma_j + \gamma_j^*) \log \theta_j + (2 - \gamma_j - \gamma_j^*) \log(1 - \theta_j) \right] + \sum_{j=1}^{p} \log f(\theta_j),
\end{aligned}$$

where $f(\mathbf{y}|\boldsymbol{\beta}, \phi)$ is the data distribution and $f(\theta)$ is the Beta(a, b) density. We choose non-informative prior for the intercept β_0 and the dispersion parameter ϕ; for example, $f(\beta_0|\tau_0^2) = N(0, \tau_0^2)$ with τ_0^2 set to a large value and $f(\log \phi) \propto 1$.

We use the EM algorithm to find the maximum a posteriori estimate of Θ. This is, in the E-step, we calculate the expectation of posterior density function of $\log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X})$ with respect to the latent indicators $\boldsymbol{\gamma}$ conditioning on the parameter values from previous iteration $\Theta^{(t-1)}$,

$$E_{\boldsymbol{\gamma}|\Theta^{(t-1)}} \log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}).$$

Hereafter, we use the shorthand notation $E(\cdot) \equiv E_{\boldsymbol{\gamma}|\Theta^{(t-1)}}(\cdot)$. In the M-step, we find the parameters $\Theta^{(t)}$ that maximize $E \log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X})$. The parenthesized subscription $^{(t)}$ denotes the parameter estimation at the tth iteration. The E- and M- steps are iterated until the algorithm converges.

To note here, the log-posterior density of BHAMs (up to additive constants) can be written as a two-part equation

$$\log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X}) = Q_1(\boldsymbol{\beta}, \phi) + Q_2(\boldsymbol{\gamma}, \theta),$$

where

$$Q_1 \equiv Q_1(\boldsymbol{\beta}, \phi) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{p} \left[\log f(\beta_j|\gamma_j) + \sum_{k=1}^{K_j} \log f(\beta_{jk}^*|\gamma_{jk}^*) \right]$$

and

$$Q_2 \equiv Q_2(\boldsymbol{\gamma}, \theta) = \sum_{j=1}^{p} \left[(\gamma_j + \gamma_j^*) \log \theta_j + (2 - \gamma_j - \gamma_j^*) \log(1 - \theta_j) \right] + \sum_{j=1}^{p} \log f(\theta_j).$$

Q_1 and Q_2 are respectively the log posterior density of the coefficients $\boldsymbol{\beta}$ and the log posterior density of the probability parameters $\boldsymbol{\theta}$ conditioning on $\boldsymbol{\gamma}$. Meanwhile, conditioning on $\boldsymbol{\gamma}$, Q_1 and Q_2 are independent and can be maximized separately for $\boldsymbol{\beta}, \phi$ and $\boldsymbol{\theta}$. With the proposed two-part spike-and-slab LASSO prior, Q_1 can be treated as penalized likelihood function and maximization of $E(Q_1)$ can be solved via the Coordinate Descent algorithm in each iteration. Coordinate descent is an optimization algorithm that offers extreme computational advantage, and famous for its application in optimizing the l_1 penalized likelihood function. Maximization of $E(Q_2)$ can be solved via closed form equations following the beta-binomial conjugate relationship.

The density function of the mixture double exponential prior of coefficient β can be written as

$$f(\beta|\gamma, s_0, s_1) = \frac{1}{2\left[(1-\gamma)s_0 + \gamma s_1\right]} \exp\left(-\frac{|\beta|}{(1-\gamma)s_0 + \gamma s_1}\right),$$

and $E(Q_1)$ can be expressed as a log-likelihood function with l_1 penalty

$$E(Q_1) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) - \sum_{j=1}^{p} \left[E({S_j}^{-1})|\beta_j| + \sum_{k=1}^{K_j} E({S^*}_j^{-1})|\beta_{jk}| \right], \tag{5}$$

where $S_j = (1-\gamma_j)s_0 + \gamma_j s_1$ and $S_j^* = (1-\gamma_j^*)s_0 + \gamma_j^* s_1$. To calculate two unknown quantities $E(S_j^{-1})$ and $E(S^{*-1}_j)$, the posterior probability $p_j \equiv \Pr(\gamma_j = 1|\Theta^{(t-1)})$ and $p_j^* \equiv \Pr(\gamma_j^* = 1|\Theta^{(t-1)})$ are necessary, which can be derived via Bayes' theorem. The calculation of p_j^* is slightly different from that of p_j, as p_j^* depends on the values of the vector $\boldsymbol{\beta}_j^*$ and p_j only depends on the scalar β_j. The calculation follows the equations below,

$$p_j = \frac{\Pr(\gamma_j = 1|\theta_j) f(\beta_j|\gamma_j = 1, s_1)}{\Pr(\gamma_j = 1|\theta_j) f(\beta_j|\gamma_j = 1, s_1) + \Pr(\gamma_j = 0|\theta_j) f(\beta_j|\gamma_j = 0, s_0)}$$

$$p_j^* = \frac{\Pr(\gamma_j^* = 1|\theta_j) \prod_{k=1}^{K_j} f(\beta_{jk}|\gamma_j^* = 1, s_1)}{\Pr(\gamma_j^* = 1|\theta_j) \prod_{k=1}^{K_j} f(\beta_{jk}|\gamma_j^* = 1, s_1) + \Pr(\gamma_j^* = 0|\theta_j) \prod_{k=1}^{K_j} f(\beta_{jk}|\gamma_j^* = 0, s_0)}$$

where $\Pr(\gamma_j = 1|\theta_j) = \theta_j$, $\Pr(\gamma_j = 0|\theta_j) = 1-\theta_j$, $\Pr(\gamma_j^* = 1|\theta_j) = \theta_j^2$, $\Pr(\gamma_j^* = 0|\theta_j) = 1-\theta_j^2$, $f(\beta|\gamma = 1, s_1) = \mathrm{DE}(\beta|0, s_1)$, $f(\beta|\gamma = 0, s_0) = \mathrm{DE}(\beta|0, s_0)$. It is trivial to show

$$E(\gamma_j) = p_j \qquad E(\gamma_j^*) = p_j^*$$

$$E(S_j^{-1}) = \frac{1-p_j}{s_0} + \frac{p_j}{s_1} \qquad E(S_j^{*-1}) = \frac{1-p_j^*}{s_0} + \frac{p_j^*}{s_1}.$$

After replacing the calculated quantities, $E(Q_1)$ can be seen as a l_1 penalized likelihood function with the regularization parameter $\lambda = E(S^{-1})$, and hence be optimized via coordinate descent algorithm[29]. Independently, the remaining parameters of interest $\boldsymbol{\theta}$ can be updated by maximizing $E(Q_2)$. As the beta distribution is a conjugate prior for Bernoulli distribution, $\boldsymbol{\theta}$ can be easily updated with a closed form equation,

$$\theta_j = \frac{p_j + p_j^* + a - 1}{a + b}. \tag{6}$$

Totally, the proposed EM-CD algorithm is summarized as follows:

1) Choose a starting value $\boldsymbol{\beta}^{(0)}$ and $\boldsymbol{\theta}^{(0)}$ for $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. For example, we can initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}$ and $\boldsymbol{\theta}^{(0)} = \mathbf{0}.5$

2) Iterate over the E-step and M-step until convergence

   E-step: calculate $E(\gamma_j)$, $E(\gamma_j^*)$ and $E(S_j^{-1})$, $E(S^{*-1}_j)$ with estimates of $\Theta^{(t-1)}$ from previous iteration

   M-step:

   a) Update $\boldsymbol{\beta}^{(t)}$, and the dispersion parameter $\phi^{(t)}$ if exists, using the coordinate descent algorithm with the penalized likelihood function in Equation (5)

   b) Update $\boldsymbol{\theta}^{(t)}$ using Equation (6)

We assess convergence by the criterion: $|d^{(t)} - d^{(t-1)}|/(0.1 + |d^{(t)}|) < \epsilon$, where $d^{(t)} = -2\log f(\mathbf{y}|\mathbf{X}, \boldsymbol{\beta}^{(t)}, \phi^{(t)})$ is the estimate of deviance at the tth iteration, and ϵ is a small value (say 10^{-5}).

2.4 | Selecting Optimal Scale Values

Our proposed model, BHAM, requires two preset scale parameters (s_0, s_1). Hence, we need to find the optimal values for the scale parameters such that the model reaches its best prediction performance regarding a criteria of preference. This would be achieved by constructing a two-dimensional grid, consists of different pairs of (s_0, s_1) value. However, previous research suggests the value of slab scale s_1 has less impact on the final model and is recommended to be set as a generally large value, e.g. $s_1 = 1$, that provides no or weak shrinkage.[20] As a result, we focus on examining different values of spike scale s_0. Instead of the two-dimensional grid, we consider a sequence of L decreasing values $\{s_0^l\} : 0 < s_0^1 < s_0^2 < \cdots < s_0^L < s_1$. Increasing the spike scale s_0 tends to include more non-zero coefficients in the model. A measure of preference calculated with cross-validations (CV), e.g. deviance (defined as model log-likelihood times -2, $-2\log f(\boldsymbol{y}|\hat{\boldsymbol{\beta}}, \hat{\phi})$), area under the curve (AUC), mean squared error, can be used to facilitate the selection of a final model. The procedure is similar to the LASSO implementation in the widely used R package `glmnet`, which quickly fits LASSO models over a list of values of regularization parameters λ and gives a sequence of models for users to choose from.

3 | SIMULATION STUDY

In this section, we compare the performance of the proposed model to six alternative models: linear LASSO models, component selection and smoothing operator (COSSO)[30], adaptive COSSO[31], generalized additive models with automatic smoothing (referred as *mgcv* hereafter)[32], spike-and-slab GAM[15], and SB-GAM[13]. We use linear LASSO model as the benchmark, examing the performance when linearity assumption doesn't hold. COSSO is one of the earliest smoothing spline models that consider sparsity-smoothness penalty. Adaptive COSSO improved upon COSSO by using adaptive weight for penalties such that the penalty of each functional component are different for extra flexibility. *mgcv* is one of the most popular models for nonlinear effect interpolation and prediction. Nevertheless, mgcv doesn't support analyses when the number of parameters exceeds the sample size. Spike-and-slab GAM employs a spike-and-slab prior for GAM and uses a MCMC algorithm for model fitting. SB-GAM is the first spike-and-slab LASSO GAM. We implement linear LASSO model with R package `glmnet 4.1-2`, COSSO and adaptive COSSO with R package `cosso 2.1-1`, generalized additive models with automatic smoothing with R package `mgcv 1.8-31`, spike-and-slab GAM with R package `spikeSlabGAM 1.1-15`, and SB-GAM with R package `sparseGAM 1.0`. COSSO models and SB-GAM do not provide flexibility to define smoothing functions, and hence use the default choices; mgcv, spikeSlabGAM and the proposed model allow customized smoothing functions and we choose the cubic regression spline. We control the dimensionality of each smoothing function, 10 bases, for all different choices of smoothing functions. We use 5-fold CV with the default selection criteria to select the final model for linear LASSO model, COSSO models, SB-GAM and the proposed model. 20 default candidates of tuning parameters (s_0 in BHAM, λ_0 in SB-GAM) are examined for SB-GAM and the proposed model which allow user-specification of tuning candidates. All computation was conducted on a high-performance 64-bit Linux platform with 48 cores of 2.70GHz eight-core Intel Xeon E5-2680 processors and 24G of RAM per core and R 3.6.2[33].

Other related methods for high-dimensional GAMs also exist, notably the methods of sparse additive models by Ravikumar et al.[4]. However, we exclude these methods from the current simulation study because of their demonstrated inferior predictive performance compared to *mgcv*[14].

3.1 | Monte Carlo Simulation Study

We follow the data generating process described in Bai[13]: we first generate $n = 500$ training data points with $p = 4, 10, 50, 100, 200$ predictors respectively, where the predictors $\boldsymbol{X}$ are simulated from a multivariate normal distribution $\text{MVN}_{n\times p}(0, \boldsymbol{I}_P)$. We then simulate the outcome Y from two distributions, Gaussian and binomial with the identity link and logit link $g(x) = \log(\frac{x}{1-x})$ respectively. The mean of each outcome is derived via the following function

$$\mathbb{E}(Y) = g^{-1}(5\sin(2\pi x_1) - 4\cos(2\pi x_2 - 0.5) + 6(x_3 - 0.5) - 5(x_4^2 - 0.3))$$

for Gaussian and binomial outcomes. Gaussian outcomes requires specification of dispersion, where we set the dispersion parameter to be 1. In this data generating process, we have x_1, x_2, x_3, x_4 as the active predictors, while the rest predictors are inactive, i.e. $f_j(x_j) = 0$ for $j = 4, \ldots, p$. Another set of independent sample of size $n_{test} = 1000$, are created following the same data generating process, serving as the testing data. We generate 50 independent pairs of training and testing datasets to evaluate the prediction and variable selection performance of the chosen models, where training datasets are used to fit the models and testing datasets are used to calculate metrics of interest. In addition, we consider the data generating process where all functional forms of the predictors are linear while keeping the rest of simulation parameters the same. This additional set of linear simulations is designed to investigate the flexibility of the proposed model when nonlinear assumptions are not met.

To evaluate the predictive performance of the models, the statistics, R^2 for Gaussian model and AUC for binomial model calculated based on the testing datasets, are averaged across 50 simulations. To evaluate the variable selection performance of the models, we record the set of variables each method selects and calculate the averaged positive predictive value (precision), true positive rate (recall), and Matthews correlation coefficient (MCC),

$$\begin{aligned}
\text{precision} &= \frac{TP}{PP} \\
\text{recall} &= \frac{TP}{TP + FP} \\
\text{MCC} &= \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},
\end{aligned}$$

where TP, TN, FP, FN, and PP are true positives, true negatives, false positives, false negatives, and predicted positive respectively. For the methods that don't automatically achieve variable selection, we set alpha level at 0.05 for *mgcv* that relies on hypothesis testing, and a soft-threshold at 0.5 for spikeSlabGAM given the marginal inclusion probabilities. For the two methods, BHAM and spikeSlabGAM, that are capable of bi-level selection, we record the probability that the linear and nonlinear components of each predictors are selected in the models.

3.2 | Simulation Results

3.2.1 | Prediction Performance

Among the set of simulations where the functional forms of the predictors are nonlinear, the predictive performances have a consistent pattern across the two distributions of outcomes. For simplicity, we use Gaussian simulations to exemplify the improvement of BHAM and defer to Tables 1 and 2 for detailed statistics. The proposed model, BHAM, predicts as good as, if not better than, other high dimensional additive models. Specifically, BHAM shows greate improvement over COSSO methods, resulting a median (interquartile range, IRT) 31% (131%) and 20% (129%) imporvement over COSSO and adaptive COSSO in R^2 statistics respectively. The improvement over spikeSlabGAM model is moderate, resulting in a median (IRT) 6% (10%) improvement in R^2. When comparing to SB-GAM, BHAM performs better (median (IRT) 13% (8%) improvement in R2) in lower dimensional cases ($p = 4, 10$), and equally good or slightly worse (median (IRT) 1% (9%) improvement in R2:) in high-dimensional cases ($p = 50, 100, 200$). As previously hypothesized, the linear LASSO model predicts less accurate compared to other flexible models across all scenario; mgcv performs extremely well in low-dimensional case ($p = 4, 10$), and deteriorates as the dimensionality increases until not applicable. To note, mgcv fits models but fails to converge within the default number of iterations when the sample size apporaches the number of coefficients to estimate ($p = 50$), which leads to bad performance. Even though SB-GAM has slight prediction advantage over the proposed model in high-dimensional situations, the BHAM has extreme computational advantage over SB-GAM, resulting median (IRT) 64% (39%) reduction in computation time (measured in seconds) for Gaussian simulations, without sacrificing much of the prediction accuracy (see Table 3).

We also examine the prediction performance when the functional form of predictors are linear, see Supporting Information Table S1, and S2. The proposed model, BHAM, has similar performance as the linear LASSO model regardless of the distribution. This observation demonstrates that BHAM is a flexible model, and has good prediction performance regardless the underlying functional form of predictors. spikeSlabGAM have similar prediction performance to BHAM. Surprisingly, SB-GAM doesn't perform well in high-dimensional Gaussian outcome scenarios.

3.2.2 | Variable Selection Performance

Among the set of simulations where the functional forms of the predictors are nonlinear, the proposed model, BHAM, has a consistent performance across different dimension and distribution settings (See Table 4 for Gaussian outcomes, and Support Information Table S3 for binomial outcomes): being conservative. The symptoms of conservative variable selection are high precision and low recall, where high precision means that among all the selected variables, high percentage of them are true signals; low recall means that, the model selected small subset among all the active predictions. In other words, BHAM tends to select a smaller set of variables that are truly effective to the outcome. We want to note, the variable selection performance of BHAM is plummeted and not optimized when $p = 200$. Upon further investigation, we discover it's because the generic sequence of s_0 used to tune the model doesn't contain the optimal value. Overall, among all the models examined, SB-GAM has the best performance, both high precision and high recall, and yield a high MCC. The performance of another Bayesian model, spikeSlabGAM deteriorates as the sparsity grows, particularly when ($p > 50$), or for binomial outcomes. The variable selection performance for linear simulations match with prediction performance: BHAM performs great among the Gaussian scenarios, while the performance of SB-GAM deteriorates.

Among the high-dimensional methods of comparison, there are two methods that are capable to achieve bi-level selection, the proposed BHAM and spikeSlabGAM. Among the linear simulations, both methods can accurately select the linear components and have a drastically lowered probability, close to 0, to include the nonlinear component, as anticipated. Specifically, spikeSlabGAM have smaller probability to include nonlinear component in the model than BHAM. However, this advantage of spikeSlabGAM over BHAM is less obvious among the nonlinear simulations: spikeSlabGAM performs better than BHAM when selecting components of the function forms that include only linear or nonlinear component, e.g. function forms for x_3 and x_4. However, spikeSlabGAM inclines to ignore the variable that have more complex function forms, e.g. function forms

for x_1 and x_2. In contrast, BHAM is more likely to include them in the model. This trade-off is determined by the assumption implicitly reflected via the prior hierarchy. We defer an in-depth discussion to the Section 5.

4 | METABOLOMICS DATA ANALYSIS

In this section, we apply the proposed models BHAM to analyze two published metabolomics datasets where the outcomes are binary and continuous respectively. We demonstrate the improved prediction performance compared to the other Bayesian hierarchical additive model, SB-GAM [13], while being computationally efficient (see Table **??**). BHAM requires roughly 10% of the computation time of SB-GAM to fit models.

4.1 | Emory Cardiovascular Biobank

We use the proposed models BHAM to analyze a metabolic dataset from a recently published research [34] studying plasma metabolomic profile on the three-year all-cause mortality among patients undergoing cardiac catheterization. The dataset is publicly available via *Dryad* [35]. It contains in total of 776 subjects from two cohorts. As there is a large number of non-overlapping features among the two cohorts, we use the cohort with larger sample size (N=454). There are initially 6796 features in the dataset, which is too large to be practically meaningful to analyze. Hence, we choose the the top 200 features with largest variance. We use 5-knot spline additive models for binary outcome using two different models, the proposed BHAM and the SB-GAM. 10-Fold CV are used to choose the optimal tuning parameters of each framework with respect to the default selection criterion implemented in the software. Out-of-bag samples are used for prediction performance evaluation, where deviance, AUC, Brier score defined as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and misclassification error defined as $\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5)$ are calculated. BHAM obtains superior AUC, Brier score, and misclassification error in the out-of-bag samples compared to SB-GAM (see Table 6). We plot the 33 features included in the final BHAM model in Figure 2.

4.2 | Weight Loss Maintenance Cohort

We use the proposed models BHAM to analyze metabolomics data from a recently published study [36] on the association between metabolic biomarkers and weight loss, where the dataset is publicly available [37]. In this analysis, we primarily focus on the analysis of one of the three studies included, weight loss maintenance cohort [38], due to the drastically different intervention effects. In the dataset, 765 metabolites in baseline plasma collected were profiled using liquid chromatography mass spectrometry. Quality control and natural log transformation were previously performed and documented by the study publishing team [36]. The outcome of interest are standardized percent change in insulin resistance, and hence modeled using a Gaussian model. After removing missing datapoints and addressing outliers in the data, there are p=237 features remaining in the analysis. 5-Knot spline additive models for the Gaussian outcome are constructed using two different models, the proposed BHAM and the SB-GAM. 10-Fold CV are used to choose the optimal tuning parameters of each framework with respect to the default selection criterion implemented in the software. Out-of-bag samples are used for prediction performance evaluation, where deviance, R^2, mean squared error (MSE) defined as $\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$, and mean absolute error (MAE) defined as $\frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$ are calculated. BHAM obtains superior R^2, MSE, and MAE in the out-of-bag samples compared to SB-GAM (see Table 7).

5 | DISCUSSION

In the paper, we described a novel high-dimensional generalized additive model with Bayesian hierarchical prior for the purpose of predictive modelling. In particular, we introduce a two-part spike-and-slab LASSO prior for reparameterized smoothing function and derive a scalable EM-CD algorithm for model fitting. The proposed model provides a new angle to address the excess shrinkage of smoothing functions that is commonly vulnerable to previous regularized high-dimensional GAMs, and hence improves the predictive performance. Th EM-CD algorithm, extended from previous spike-and-slab LASSO models, provides a computationally efficient alternative to the computational prohibitive MCMC algorithms, enhancing the scalability of spike-and-slab models. In addition, the two-part prior motivates the bi-level selection of predictors, selection of linear and nonlinear component. In the simulation study and real-data analyses, the proposed model demonstrates improvement in prediction and

computational advantage when compared to the state-of-the-art models. When serving the purpose of variable selection, trade-offs exist among methods of comparison. We implement the proposed model in an open-source R package `BHAM`, deposited at https://github.com/boyiguo1/BHAM. To maximize the flexibility of smoothing function specification, we deploy the same programming grammar as in the state-of-the-art package `mgcv`, in contrast to previous tools where smoothing functions are limited to the default ones. Ancillary functions are provided for model specification in high-dimensional settings, curve plotting and functional selection.

The proposed model shares many commonality with the SB-GAM[13], independently developed around the same time of the proposed work. Both frameworks emphasize computational efficiency by deploying group spike-and-slab LASSO type priors and optimization-based scalable algorithms. Bai provides the theoretical proof for the consistency of variable selection using group spike-and-slab LASSO prior. The proposed model focuses on improving prediction performance for high-dimensional GAM, with the capability of bi-level selection. Moreover, the proposed model can easily generalize to other family of priors or other family of smoothing functions if desired. Not focused in this manuscript, the generalization is described in the supporting materials.

During designing and analyzing the simulation study, we made couple interesting observations. First of all, variable selection is a delicate topic in the context predictive modelling. When prediction performance is used to tune a model, the model could possibly include noise variable in models, for example LASSO and LASSO-based models.[39] Moreover, bi-level selection is a more complex problem than variable selection. The complexity reflects on the validity of the effect hierarchy principle. While most functional forms follow that linear components exists in the nonlinear function, there are functions that don't follow it, e.g. x^2. The proposed prior and spikeSlabGAM employ different structure: the proposed prior imposes effect hierarchy while spikeSlabGAM treats the selection of linear and nonlinear components independent. The different prior setups lead to trade-offs for the purpose of bi-level selection. We recommend to use more judgement in bi-level selection, either relying on heuristic knowledge to choose appropriate prior or exploring multiple models when heuristic knowledge doesn't exist. Secondly, we find the performance of the proposed model are more sensitive to the granularity of s_0 sequence in the high-dimensional settings than in the lower dimension settings. Even though the current default sequence of s_0 can result in reasonable performance shown in the simulation studies, we recommend to fine-tune the model with granular sequence of s_0 for performance improvement.

Our future efforts direct to survival analysis and integrative analysis. While the proposed model addresses a great deal of analytic problem, analyzing the time-to-event outcome remains unsolved. An naive approach would be convert a time-to-event outcome to a Poisson outcome following Whitehead[40]. However, it would be more efficient to directly fit Cox models via penalized pseudo likelihood function[41]. Meanwhile, with growing understanding of biological structure within -omics field, it is appealing to integrate external biology information in the modeling process. The main motivation for integrative models is that biologically informed grouping of weak effects increases the power of detecting true associations between features and the outcome[42], and stabilizes the analysis results for reproducibility purpose. Such integration can be achieved by setting up a structural hyperprior on the inclusion indicator of the smoothing function null space $\boldsymbol{\gamma}^0$. The similar strategy has been used in Ferrari and Dunson[43].

References

1. Mallick H, Yi N. Bayesian Methods for High Dimensional Linear Models. *Journal of Biometrics & Biostatistics* 2013(205): 1–27. doi: 10.4172/2155-6180.S1-005

2. Breiman L. Statistical modeling: The two cultures (with comments and a rejoinder by the author). *Statistical science* 2001; 16(3): 199–231.

3. Hastie T, Tibshirani R. Generalized additive models: Some applications. *Journal of the American Statistical Association* 1987; 82(398): 371–386. doi: 10.1080/01621459.1987.10478440

4. Ravikumar P, Lafferty J, Liu H, Wasserman L. Sparse additive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2009; 71(5): 1009–1030. doi: 10.1111/j.1467-9868.2009.00718.x

5. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 2006; 68(1): 49–67.

6. Huang J, Horowitz JL, Wei F. Variable selection in nonparametric additive models. *Annals of statistics* 2010; 38(4): 2282.

7. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. *Bioinformatics* 2007; 23(12): 1486–1494.

8. Xue L. Consistent variable selection in additive models. *Statistica Sinica* 2009: 1281–1296.

9. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* 2001; 96(456): 1348–1360.

10. Xu X, Ghosh M, others . Bayesian variable selection and estimation for group lasso. *Bayesian Analysis* 2015; 10(4): 909–936.

11. Yang X, Narisetty NN, others . Consistent group selection with Bayesian high dimensional modeling. *Bayesian Analysis* 2020; 15(3): 909–935.

12. Bai R, Moran GE, Antonelli JL, Chen Y, Boland MR. Spike-and-slab group lassos for grouped regression and sparse generalized additive models. *Journal of the American Statistical Association* 2020: 1–14.

13. Bai R. Spike-and-Slab Group Lasso for Consistent Estimation and Variable Selection in Non-Gaussian Generalized Additive Models. *arXiv:2007.07021v5.* Preprint posted online June 5, 2021. https://arxiv.org/abs/2007.07021.

14. Scheipl F, Kneib T, Fahrmeir L. Penalized likelihood and Bayesian function selection in regression models. *AStA Advances in Statistical Analysis* 2013; 97(4): 349–385.

15. Scheipl F, Fahrmeir L, Kneib T. Spike-and-slab priors for function selection in structured additive regression models. *Journal of the American Statistical Association* 2012; 107(500): 1518–1532. doi: 10.1080/01621459.2012.737742

16. Wood SN. *Generalized additive models: An introduction with R, second edition* . 2017

17. Meier L, Van De Geer S, Bühlmann P. High-dimensional additive modeling. *Annals of Statistics* 2009; 37(6 B): 3779–3821. doi: 10.1214/09-AOS692

18. Marra G, Wood SN. Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis* 2011; 55(7): 2372–2387.

19. Ročková V. Bayesian estimation of sparse signals with a continuous spike-and-slab prior. *The Annals of Statistics* 2018; 46(1): 401–437. doi: 10.1214/17-AOS1554

20. Ročková V, George EI. The Spike-and-Slab LASSO. *Journal of the American Statistical Association* 2018; 113(521): 431–444. doi: 10.1080/01621459.2016.1260469

21. Bai R, Rockova V, George EI. Spike-and-Slab Meets LASSO: A Review of the Spike-and-Slab LASSO. *arXiv:2010.06451*. Preprint posted online July 1, 2021. https://arxiv.org/abs/2010.06451.

22. Tang Z, Shen Y, Li Y, et al. Group spike-And-slab lasso generalized linear models for disease prediction and associated genes detection by incorporating pathway information. *Bioinformatics* 2018; 34(6): 901–910. doi: 10.1093/bioinformatics/btx684

23. Tang Z, Lei S, Zhang X, et al. Gsslasso Cox: A Bayesian hierarchical model for predicting survival and detecting associated genes by incorporating pathway information. *BMC Bioinformatics* 2019; 20(1): 1–15. doi: 10.1186/s12859-019-2656-1

24. Chipman H. Prior distributions for Bayesian analysis of screening experiments. In: Springer. 2006 (pp. 236–267).

25. George EI, McCulloch RE. Approaches for Bayesian variable selection.. *Statistica Sinica* 1997; 7(2): 339–373.

26. Ročková V, George EI. EMVS: The EM approach to Bayesian variable selection. *Journal of the American Statistical Association* 2014; 109(506): 828–846. doi: 10.1080/01621459.2013.869223

27. Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso generalized linear models for prediction and associated genes detection. *Genetics* 2017; 205(1): 77–88. doi: 10.1534/genetics.116.192195

28. Tang Z, Shen Y, Zhang X, Yi N. The spike-and-slab lasso Cox model for survival prediction and associated genes detection. *Bioinformatics* 2017; 33(18): 2799–2807. doi: 10.1093/bioinformatics/btx300

29. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software* 2010; 33(1): 1.

30. Zhang HH, Lin Y. Component selection and smoothing for nonparametric regression in exponential families. *Statistica Sinica* 2006: 1021–1041.

31. Storlie CB, Bondell HD, Reich BJ, Zhang HH. Surface estimation, variable selection, and the nonparametric oracle property. *Statistica Sinica* 2011; 21(2): 679.

32. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)* 2011; 73(1): 3-36.

33. R Core Team . R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing* 2021. https://www.R-project.org/.

34. Mehta A, Liu C, Nayak A, et al. Untargeted high-resolution plasma metabolomic profiling predicts outcomes in patients with coronary artery disease. *PloS one* 2020; 15(8): e0237579.

35. Mehta A, Liu C, Uppal K, Quyyumi A. Data from: Metabolomics - Emory Cardiovascular Biobank. *Dryad, Dataset* Retrived online August 17, 2021. https://doi.org/10.5061/dryad.866t1g1mt.

36. Bihlmeyer NA, Kwee LC, Clish CB, et al. Metabolomic profiling identifies complex lipid species and amino acid analogues associated with response to weight loss interventions. *Plos one* 2021; 16(5): e0240764.

37. Bihlmeyer NA, Kwee LC, Clish CB, et al. Metabolomic profiling identifies complex lipid species and amino acid analogues associated with response to weight loss interventions. *Zenodo* Retrived online August 18, 2021. https://doi.org/10.5281/zenodo.4767969.

38. Svetkey LP, Stevens VJ, Brantley PJ, et al. Comparison of strategies for sustaining weight loss: the weight loss maintenance randomized controlled trial. *Jama* 2008; 299(10): 1139–1148.

39. Wu J, Witten D. Flexible and Interpretable Models for Survival Data. *Journal of Computational and Graphical Statistics* 2019; 28(4): 954–966. doi: 10.1080/10618600.2019.1592758

40. Whitehead J. Fitting Cox's regression model to survival data using GLIM. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 1980; 29(3): 268–275.

41. Simon N, Friedman J, Hastie T, Tibshirani R. Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of statistical software* 2011; 39(5): 1.

42. Peterson CB, Stingo FC, Vannucci M. Joint Bayesian variable and graph selection for regression models with network-structured predictors. *Statistics in medicine* 2016; 35(7): 1017–1031.

43. Ferrari F, Dunson DB. Identifying main effects and interactions among exposures using Gaussian processes. *The Annals of Applied Statistics* 2020; 14(4): 1743–1758.

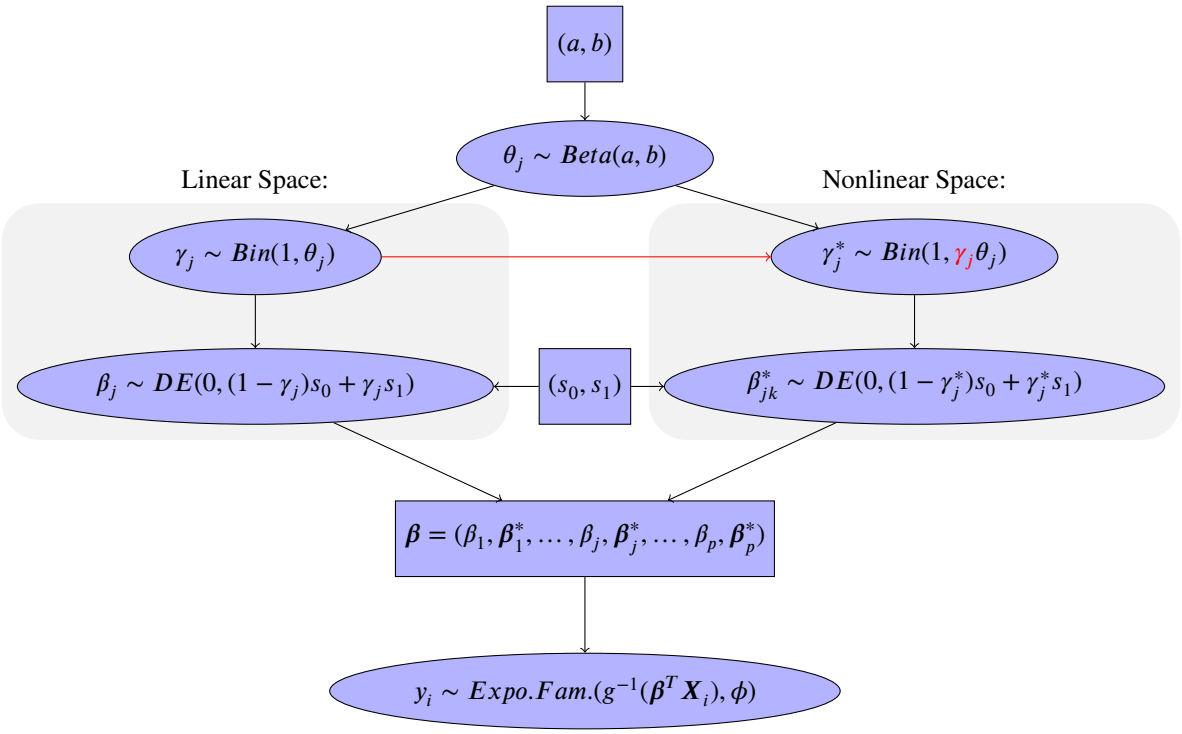


FIGURE 1 Directed acyclic graph of the proposed Bayesian hierarchical additive model with parameter expansion. Elliposes are stochastic nodes, rectangles and are deterministic nodes.

P	mgcv	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	0.90 (0.01)	0.33 (0.01)	0.71 (0.13)	0.72 (0.11)	0.90 (0.01)	0.79 (0.04)	0.80 (0.00)
10	0.90 (0.01)	0.33 (0.01)	0.66 (0.21)	0.77 (0.02)	0.89 (0.01)	0.79 (0.04)	0.79 (0.00)
50	0.86 (0.02)	0.32 (0.01)	0.46 (0.19)	0.57 (0.18)	0.80 (0.02)	0.78 (0.05)	0.78 (0.01)
100	-	0.32 (0.01)	0.41 (0.23)	0.48 (0.25)	0.79 (0.01)	0.79 (0.05)	0.77 (0.01)
200	-	0.32 (0.01)	0.39 (0.19)	0.40 (0.17)	0.79 (0.01)	0.78 (0.04)	0.75 (0.01)

TABLE 1 The average and standard deviation of the out-of-sample R^2 measure for Gaussian outcomes over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM) fitted with Iterative Weighted Least Square (BHAM-IWLS) and Coordinate Descent (BHAM-CD) algorithms, component selection and smoothing operator (COSSO), adaptive COSSO, mgcv and sparse Bayesian generalized additive model (SB-GAM). mgcv doesn't provide estimation whe number of parameters exceeds sample size i.e. p = 100, 200.

P	mgcv	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	0.94 (0.01)	0.83 (0.01)	0.90 (0.01)	0.90 (0.01)	0.92 (0.01)	0.92 (0.01)	0.90 (0.00)
10	0.92 (0.03)	0.83 (0.00)	0.86 (0.04)	0.86 (0.03)	0.92 (0.01)	0.92 (0.01)	0.90 (0.00)
50	0.76 (0.03)	0.83 (0.01)	0.83 (0.02)	0.84 (0.02)	0.90 (0.01)	0.92 (0.01)	0.89 (0.01)
100	-	0.83 (0.01)	0.83 (0.02)	0.81 (0.09)	0.90 (0.01)	0.92 (0.01)	0.88 (0.01)
200	-	0.83 (0.01)	0.81 (0.06)	0.82 (0.05)	0.88 (0.02)	0.92 (0.01)	0.87 (0.02)

TABLE 2 The average and standard deviation of the out-of-sample area under the curve measures for binomial outcomes over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM) fitted with Iterative Weighted Least Square (BHAM-IWLS) and Coordinate Descent (BHAM-CD) algorithms, component selection and smoothing operator (COSSO), adaptive COSSO, mgcv and sparse Bayesian generalized additive model (SB-GAM). mgcv doesn't provide estimation whe number of parameters exceeds sample size i.e. $p = 100, 200$.

Distribution	P	mgcv	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
Binomial	4	0.18 (0.04)	3.16 (1.39)	5.51 (4.07)	2.73 (0.22)	347.17 (89.43)	8.41 (0.91)
Binomial	10	3.46 (11.06)	8.30 (1.70)	10.66 (5.30)	4.08 (0.29)	539.05 (135.55)	20.36 (2.16)
Binomial	50	660.31 (141.53)	103.41 (20.00)	118.82 (18.45)	14.22 (0.58)	1590.09 (142.19)	236.73 (14.83)
Binomial	100	-	662.61 (125.00)	672.65 (185.09)	31.61 (2.56)	2720.53 (250.43)	967.97 (186.85)
Binomial	200	-	5325.66 (995.60)	4963.93 (1482.11)	82.17 (3.29)	4788.76 (420.64)	3371.88 (194.02)
Gaussian	4	0.05 (0.01)	0.75 (0.09)	0.75 (0.11)	8.78 (1.57)	38.82 (2.74)	1.84 (0.18)
Gaussian	10	0.32 (0.39)	3.42 (0.24)	3.41 (0.23)	20.77 (3.95)	76.12 (5.55)	5.93 (0.57)
Gaussian	50	72.03 (57.99)	33.98 (2.88)	34.35 (2.86)	285.73 (12.53)	374.76 (23.79)	65.18 (8.12)
Gaussian	100	-	117.79 (3.33)	119.63 (3.66)	372.01 (56.92)	640.44 (21.91)	194.14 (8.09)
Gaussian	200	-	518.86 (40.78)	524.76 (39.15)	471.46 (72.23)	1300.70 (72.74)	738.52 (62.76)

TABLE 3 The average and standard deviation of computation time in seconds, including cross-validation and final model fitting, over 50 iterations. The models of comparison include the proposed Bayesian hierarchical additive model (BHAM) fitted with Iterative Weighted Least Square (BHAM-IWLS) and Coordinate Descent (BHAM-CD) algorithms, component selection and smoothing operator (COSSO), adaptive COSSO, mgcv and sparse Bayesian generalized additive model (SB-GAM). mgcv doesn't provide estimation whe number of parameters exceeds sample size i.e. p = 100, 200.

P	Metric	mgcv	LASSO	COSSO	Adaptive COSSO	BHAM	SB-GAM	spikeSlabGAM
4	Precision	1.00	1.00	1.00	1.00	1.00	1.00	1.00
10	Precision	0.89	0.76	0.84	0.93	0.62	0.86	1.00
50	Precision	0.36	0.48	0.70	0.69	0.88	0.75	1.00
100	Precision		0.43	0.61	0.59	0.99	0.79	0.99
200	Precision		0.36	0.61	0.47	0.28	0.75	0.99
4	Recall	1.00	0.53	0.49	0.53	0.88	0.99	0.51
10	Recall	1.00	0.40	0.52	0.52	0.83	1.00	0.50
50	Recall	1.00	0.35	0.40	0.48	0.37	1.00	0.50
100	Recall		0.33	0.36	0.40	0.30	0.99	0.50
200	Recall		0.32	0.33	0.35	0.52	1.00	0.50
4	MCC							
10	MCC	0.90	0.32	0.49	0.57	0.46	0.86	0.61
50	MCC	0.54	0.31	0.47	0.53	0.50	0.83	0.69
100	MCC		0.32	0.41	0.45	0.53	0.87	0.70
200	MCC		0.28	0.41	0.38	0.36	0.85	0.70

TABLE 4

TABLE 5 Model fitting time in seconds for two metabolomics data analyses, from Emory Cardiovascular Biobank (ECB) and Weight Loss Maintenance Cohort (WLM). It tabulates the computation time for cross-validation step (CV) and optimal model fitting step (Final), and total computation time (Total) for the proposed model BHAM with EM-CD algorithm (BHAM-CD) and the model of comparison SB-GAM.\label{tab:time_real_data}

Data	BHAM-CD			SB-GAM		
	CV	Final	Total	CV	Final	Total
ECB	100.8	3.5	104.4	2,659.0	20.9	2,679.9
WLM	365.4	6.0	371.4	3,116.0	32.7	3,148.7

Methods	Deviance	AUC	Brier	Misclass
BHAM-CD.1	510.99	0.61	0.19	0.24
BHAM-CD.2	510.99	0.61	0.19	0.24
SB-GAM	636.56	0.56	0.22	0.30

TABLE 6 Prediction performance of BHAM fitted with Cooridnate Descent algorithm (BHAM-CD) and SB-GAM models for Emory Cardiovascular Biobank by 10-fold cross-validation, including deviance, area under the curve (AUC), Brier score, and misclassification error (Misclass) where class labels are defined using threshold = 0.5.

Methods	Deviance	R^2	MSE	MAE
BHAM-CD	668.01	0.07	0.93	0.76
SB-GAM	666.83	0.03	0.98	0.77

TABLE 7 Prediction performance of of BHAM fitted with Cooridnate Descent algorithm (BHAM-CD) and SB-GAM models for Weight Loss Maintenance Cohort by 10-fold cross-validation, including deviance, R^2, mean squared error (MSE), and mean absolute error (MAE).

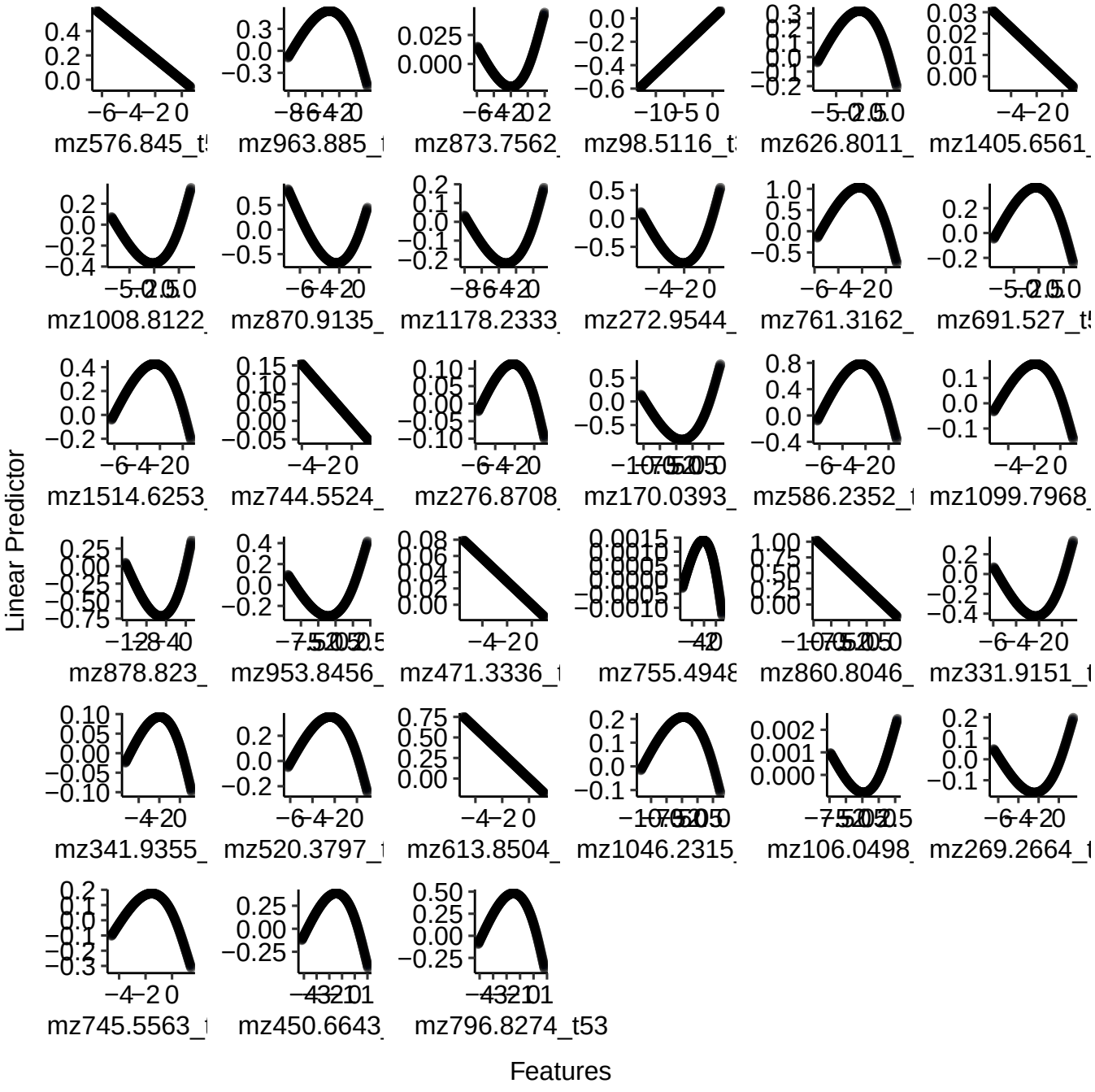
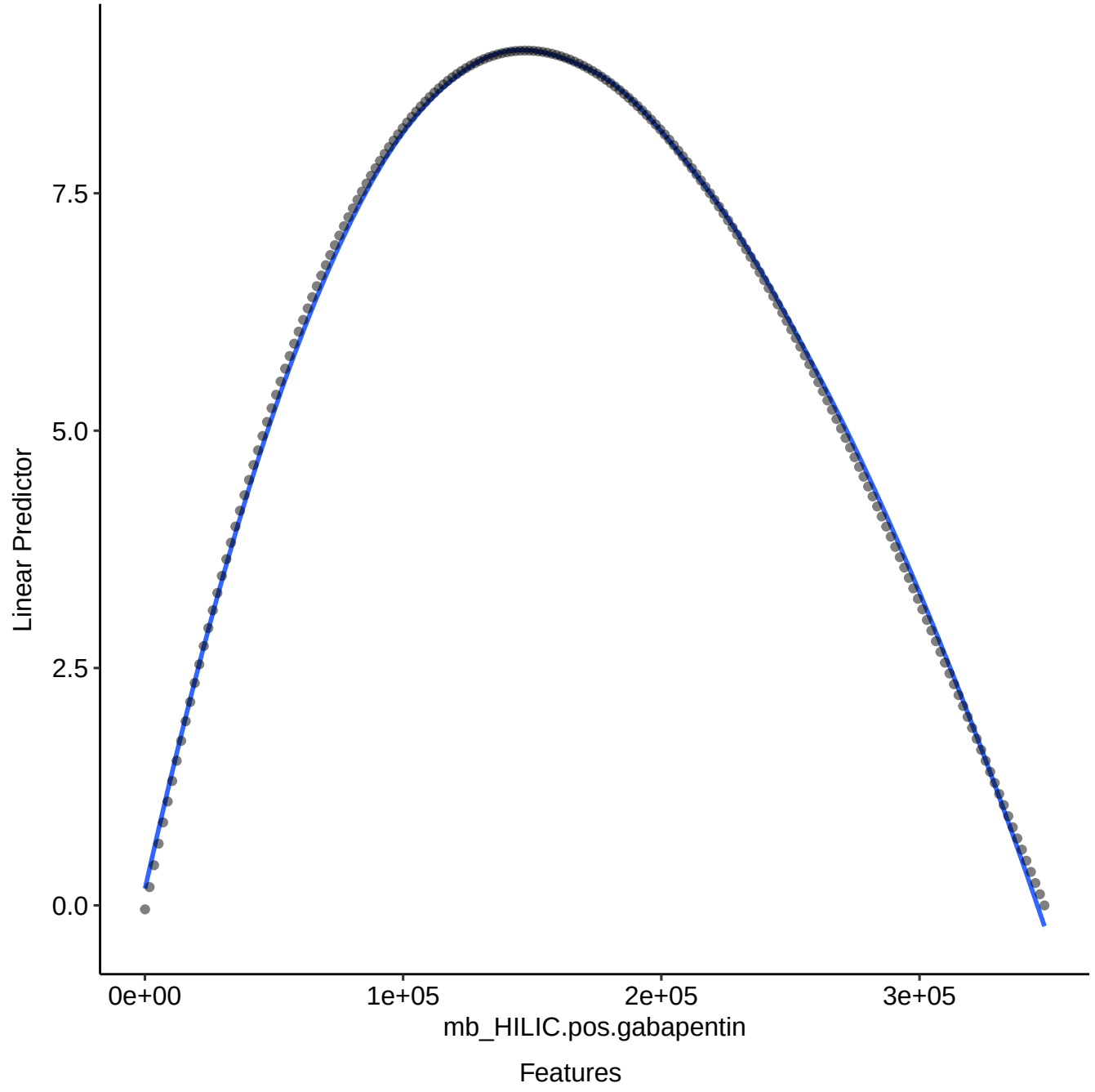


FIGURE 2 TODO: change this caption


7.5
5.0
2.5
0.0
Linear Predictor
0e+00
1e+05
2e+05
3e+05
mb_HILIC.pos.gabapentin
Features