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 and separation of smoothing function linear and nonlinear spaces. A novel spike-and-slab spline prior is proposed to select components of smoothing functions. Two scalable and deterministic algorithms, EM-Coordinate Descent and EM-Iterative Weighted Least Squares, are developed for different utilities. Simulation studies and metabolomics data analyses demonstrate improved predictive or computational performance against state-of-the-art models, mgcv, COSSO and sparse Bayesian GAM. The software implementation of the proposed models is freely available via an R package BHAM.

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

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 models to the high-dimensional setting remains unclear.

Nonparametric regression models are appropriate alternatives to the "black-box" models thanks to their balance between model flexibility and interpretability. 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, Y, which is assumed to follow some exponential family distribution with mean μ and dispersion ϕ, can be modeled with the summation of smoothing functions, $\boldsymbol{B}_j(\cdot), j = 1, \ldots, p$, of a given p-dimensional vector of covariates $\boldsymbol{x}$, written as

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

where $g^{-1}(\cdot)$ is the inverse of a monotonic link function. The smoothing functions can take many forms and are estimated using a pseudo-weighted version of the backfitting algorithm[4]. Nevertheless, the classical GAMs cannot fulfill the increasing analytic demands for high-dimensional data analysis in biomedical studies.

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.[5] extended the grouped lasso[6] to additive models (AMs); Huang et al.[7] further developed adaptive grouped lasso for additive models; Wang et al.[8] and Xue[9] respectively applied grouped SCAD penalty[10] to additive models. Recently Bayesian hierarchical models are also used in the context of high-dimensional additive models. Various group spike-and-slab priors combining with computationally intensive Markov chain Monte Carlo (MCMC) algorithms[11,12] are proposed, where the application on AMs are by-products. Bai et al.[13] 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[14]. Focus on addressing the sparsity, these methods can overly penalize the basis function coefficients and produce inaccurate predictions and curve interpolation, particularly when complex signals are assumed and large number of knots are used.[15] 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.[16] proposed a spike-and-slab structure prior that address the previous challenges. 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.

To address these challenges, we propose a novel Bayesian hierarchical generalized additive model (BHAM) for high dimensional data analysis. Specifically, we incorporate smoothing penalties in the model via re-parameterization of the smoothing function to avoid overly shrinking basis function coefficients. Smoothing penalties are commonly implemented in the classical GAMs through smoothing regression splines. They are quadratic norms of the coefficients and allow locally adaptive penalties on each smoothing function. A smoothing penalty conditioning on a smoothing parameter λ_j is a function of the integration of the second derivative of the spline function, expressed mathematically as

$$\text{pen}\left[\boldsymbol{B}_j(x)\right] = \lambda_j \int \boldsymbol{B}_j''(x)dx = \lambda_j \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j, \tag{1}$$

where $\boldsymbol{S}_j$ is a known smoothing penalty matrix, and $\boldsymbol{\beta}_j$ are the basis function coefficients. Smoothing penalties were also previously used in the spike-and-slab GAM[16] and the sparsity-smoothness penalty[17]. Moreover, incorporating the smoothing penalty allows the separation of the linear space of a smoothing function from the nonlinear space. We then impose a new two-part spike-and-slab spline prior on the smoothing functions for bi-level selection such that the linear and nonlinear spaces of smoothing functions can be selected separately. The prior setup encourages a flexible solution, rendering one of three possibilities for each predictor: no effect, only linear effect, or linear and nonlinear effects. In addition, two scalable optimization-based algorithms, EM-Coordinate Descent (EM-CD) algorithm and EM-Iterative weighted least square (EM-IWLS) algorithm, are developed and implemented in an publicly available R package `BHAM` via https://github.com/boyiguo1/BHAM, making translational science more accessible.

The proposed framework, BHAM, differs from previous spike-and-slab based GAMs, i.e. the spike-and-slab GAM[16] and the SB-GAM[14] 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 coordient 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 function coefficients after the re-parameterization step, which provides some

advantage in computation. In addition, we offer an alternative algorithm, EM-IWLS, that can provide variance-covariance matrix of the coefficients. 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 two fast-fitting algorithms. In Section 3, we compare the proposed framework to state-of-the-art models, mgcv, COSSO and sparse Bayesian GAM 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 (BHAMS)

Following the GLM notation introduced in Section 1, we have a generalized additive model with link function $g(\cdot)$ and linear predictor

$$\eta = \beta_0 + \sum_{j=1}^{p} B_j(x_j) = \beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{X}_j, \tag{2}$$

with smoothing functions $B_j(x_j)$ of the variable $x_j, j = 1, \dots, p$. The outcome Y follows a exponential family distribution with density function $f(y)$, mean $\mu = g^{-1}(\eta)$ and dispersion parameter ϕ, and the data distribution is

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

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. Its corresponding smoothing penalty matrix is denoted as $\boldsymbol{S}_j$. $\boldsymbol{\beta}_j$ denotes the basis function coefficients for the jth variable such that $B_j(x_j) = \boldsymbol{\beta}_j^T \boldsymbol{X}_j$. 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, treating the parametric function as a special case of the smoothing function, e.g. $B_j(x_j) = x_j$ with the smoothing penalty matrix defined as $\boldsymbol{S}_j = [0]$.

To encourage proper smoothing of additive functions, we adopt the idea of smoothing penalties from smoothing spline models. The basic idea is to set up a smoothing penalty, described in Equation (1), in the prior density function. However, the direct integration of smoothing penalty with sparsity penalty is not obvious. Marra and Wood[18] proposed a re-parameterization procedure to accommodate the smoothing penalty implicitly. Given the smoothing penalty matrix $\boldsymbol{S}_j$ is symmetric and positive semi-definite for univariate smoothing functions, we apply eigendecomposition on 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 $\boldsymbol{U} \equiv [\boldsymbol{U}^0 : \boldsymbol{U}^*]$ 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 non-linear space. We multiply the basis function matrix $\boldsymbol{X}$ with the orthonormal matrix $\boldsymbol{U}$ for the new design matrix $\boldsymbol{X}^{\text{repa}} = \boldsymbol{X}\boldsymbol{U} \equiv [X^0 : \boldsymbol{X}^*]$. 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 $\boldsymbol{X}_j \equiv [X_j^0 : \boldsymbol{X}_j^*]$ as the basis function matrix for the jth variable after the re-parameterization. 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 (2) now is

$$E(Y|\boldsymbol{x}) = g^{-1}(\beta_0 + \sum_{j=1}^{p} B_j(x_j)) = g^{-1}(\beta_0 + \sum_{j=1}^{p} \boldsymbol{\beta}_j^T \boldsymbol{X}_j) = g^{-1}\left[\beta_0 + \sum_{j=1}^{p} (\beta_j X_j^0 + \boldsymbol{\beta}_j^{*T} \boldsymbol{X}_j^*)\right], \tag{3}$$

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

The re-parameterization step sets up the foundation of the proposed model and provides three benefits. First of all, the re-parameterization integrates the smoothing penalty into the design matrix, and encourages models to properly smooth the nonlinear function in addition to the sparse penalty for functional selection. Secondly, the eigendecomposition of the smoothing penalty allows the isolation of the linear space from the nonlinear space, improving the feasibility of bi-level functional selection. 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 greatly broadens the choices of priors and further model choices. Last but not least, the scaling step of $\boldsymbol{X}^*$ further simplifies the choice of priors, from column-specific priors to a unified prior.

2.1 | Spike-and-Slab Spline Priors

The family of spike-and-slab (SS) regression models is one of most commonly used models in high-dimensional data analysis for its utility in outcome prediction and variable selection. We defer to Bai et al.[19] for an in-depth introduction to spike-and-slab priors. To summarize, spike-and-slab priors are a family of mixture distributions that comprises a skinny spike density $f_{\text{spike}}(\cdot)$ for weak signals and a flat slab density $f_{\text{slab}}(\cdot)$ for strong signals, mathematically

$$\beta|\gamma \sim (1-\gamma)f_{\text{spike}}(\beta) + \gamma f_{\text{slab}}(\beta).$$

The most distinct feature of SS priors is that it is conditioned on a latent binary variable $\gamma \in \{0,1\}$ that indicates whether the variable x is included in the model. There are various spike-and-slab priors depending on the choice for the spike density $f_{\text{spike}}(\cdot)$ and the slab density $f_{\text{slab}}(\cdot)$, see George and McCulloch[20,21]; Chipman[22] for grouped variables; Brown et al.[23] for multivariate outcomes; Ishwaran and Rao[24]; Clyde and George[25] and reference therein. Two of most popular spike and slab priors are spike-and-slab normal prior[20] and spike-and-slab double exponential prior[26].

The major criticism of early spike-and-slab models is being computationally prohibitive.[19] Since then, many studies focus on alleviating the computational burden that sampling algorithms bear, which include EMVS based on spike-and-slab normal prior[27] and spike-and-slab Lasso (SSL)[28,26]. Particularly, the development of SSL model substantially improves the scalability of SS models, setting up the theoretical foundation for generalized models in -omics data analysis[29,30,31,32]. The SSL prior is composed of two double exponential distributions with mean 0 and different dispersion parameters, $0 < s_0 < s_1$, mathematically,

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

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,

$$\beta|\gamma \sim DE(0,S), 0 < s_0 < s_1, \tag{4}$$

with the scale parameter $S = (1-\gamma)s_0 + \gamma s_1$. The SSL also mitigates the problem of EMVS where the weak signals are not shrink to zero, and hence is preferred in high-dimensional data analysis. We notice that Bai[14] is the first to apply spike-and-slab lasso prior in the GAM framework, where the densities of the spike and slab components take the group lasso density[11] and limits to an "all-in-all-out" strategy for functional selection.

2.1.1 | Two-part Spike-and-Slab Lasso Prior

We introduce a novel prior for GAMs, particularly for high-dimensional nonlinear modeling with bi-level selection. The proposed prior extends from the spike-and-slab lasso prior described in Equation (4). Given the re-parameterized 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, we impose independent group priors on the linear space coefficients and on the nonlinear space coefficients respectively,

$$\begin{aligned} \beta_j|\gamma_j, s_0, s_1 &\sim DE(0,(1-\gamma_j)s_0 + \gamma_j s_1) \\ \beta_{jk}^*|\gamma_j^*, s_0, s_1 &\overset{\text{iid}}{\sim} DE(0,(1-\gamma_j^*)s_0 + \gamma_j^* s_1), k = 1,\dots,K_j \end{aligned} \tag{5}$$

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. A discussion of how to choose the scale parameters comes in Section 2.3. To note, this prior differs from previous group SSL priors[31,32], as the β_j and $\boldsymbol{\beta}_j^*$ have

different indicator variables γ_j, γ_j^* respectively. It is possible to add a more restrictive assumption on the priors, assuming that one indicator variable decides the inclusion of both the linear effect and nonlinear effect, i.e. $\gamma_j = \gamma_j^*$. This converges to the SB-GAM[14]. Conversely, it is also possible to relax the assumption such that each coefficient $\beta_{jk} \in \boldsymbol{\beta}_j^*$ has its own latent indicator γ_{jk}, but at the cost of complicating the bi-level functional selection. This reduces the proposed prior to the classic SSL prior.

The re-parameterization introduced in Section 2 grants the validity of the proposed prior. First of all, the smoothing function bases are linear dependent and necessitate extra attention. The eigeondecomposition remedies the problem and hence our prior can be set to be conditionally independent. Secondly, the eigenvalue scaling provides a panacea to allow unified scale parameters for all bases of all smoothing functions.

The rest of the hierarchical prior follows the traditional SSL prior: we set up hyper-priors on γ_j, γ_j^* to allow local adaption of the shrinkage using a Bernoulli distribution, written as binomial distribution of one trial. The two indicators of the jth predictor, γ_j and γ_j^*, shares the same probability parameter θ_j,

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

Equivalently, we can derive the marginal distribution of γ_j^* by integrate out γ_j (Supplement 1),

$$\gamma_j^*|\theta_j \sim Bin(1, \theta_j^2).$$

This is to leverage the fact that the probability of selecting the bases of a smoothing function should be similar, while allowing different penalty on the linear space and non-linear space of the smoothing function. The hyper prior of γ_j decides the sparsity of the model at the functional selection level, while that of γ_j^* decides the smoothness of the spline function at basis function level. Meanwhile, we specify that γ_j and γ_j^* are independently distributed for analytic simplicity. 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. For simplicity, we focus on a special case of beta distribution, uniform (0,1), i.e. $a = 1, b = 1$. When the variable have large effects in any of the bases, the parameter θ_j will be estimated large, which in turn encourages the model to include the rest of bases. Hereafter, we refer Bayesian hierarchical generalized additive models with the spike-and-slab spline prior as the BHAM, and visually presented in Figure 1.

Figure 1 here

2.1.2 | Other Priors

With the re-parameterization step of the basis function matrix $\boldsymbol{X}$, it is possible to generalized the SSL prior to other priors, for example normal priors for ridge-type regularization and mixture normal prior for spike-and-slab regularization. These priors would work better in low and medium dimensional settings where the sparse assumption is not necessary. Here we elaborate the mixture normal prior as a demonstration of applying continuous spike-and-slab prior in BHAM.

A spike-and-slab mixture normal spline prior can be expressed as

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

Similar to the spike-and-slab spline prior in Equation (5), $0 < s_0 < s_1$ are tuning parameters and can be optimized via cross-validation. One of the critics received by the spike-and-slab mixture normal prior is that the tails of a normal distribution diminishes to zero too fast, which causes problems when estimating the large effects. Distributions with heavier tails can be used as an alternative, for example mixture Student's t distribution with small degree of freedom.

2.2 | Algorithms for Fitting BHAMs

The proposed models can be fitted with MCMC algorithms. Nevertheless, the computational burden of MCMC algorithms creates scalability issues. George and McCulloch[21] examined the computation speed for various MCMC algorithms with spike-and-slab mixture normal priors, and suggested MCMC algorithms works well for medium size (p=25) of predictors with only

linear effects. However, it is not feasible for high-dimensional data analysis where the number of predictors easily exceeds 100. Specific to additive models, each predictor would expand to multiple new "predictors" via basis functions, creating greater computational demands. Scheipl et al.[15] demonstrated the computational demands of a MCMC algorithm for fitting spike-and-slab GAM grow exponentially as p increases modestly via simulation studies. Hence, we feel compelled to develop scalable algorithms for fitting Bayesian hierarchical additive models in high-dimensional settings.

As an alternative to sampling algorithms for fitting Bayesian models, optimization algorithms focus on the maximum a posteriori (MAP) estimates and speed up the model fitting process at the cost of posterior inference. The earlier work for fitting SS models using optimization algorithms includes EMVS[27]. Rockova and George[27] proposed an expectation-maximization (EM) based algorithm to fit models that use continuous mixture normal priors. In the E step, the latent binary indicators γs are treated as the missing data, and the posterior means are calculated conditioning on the current value of other parameters; in the M step, a ridge estimator was used to update the coefficients, followed by updating ϕ, θ. The same authors[28,26] further combined the EM algorithm with coordinate descent algorithm to fit SSL models. Yi and his group independently developed the EM-Iterative Weighted Least Square and EM-cyclic coordinate descent algorithms to fit models with broader class of priors, SSL included.[33] These algorithms were implemented for generalized linear models[29], Cox proportional hazards models[30] and their grouped counterparts[31,32]. Both EM based algorithms provide deterministic solutions, which becomes a popular property for reproducible research

In this section, we extend the two EM-based algorithms, EM-CD and EM-IWLS algorithms, to fit BHAMs. To note, the two proposed algorithms provides different utilities. The EM-CD algorithm is specifically for fitting BHAM with an expedited performance, recommending to use in high and ultra-high dimensional setting. A specific concern of EM-CD algorithm is that it provides no information for inference. In contrast, the EM-IWLS can estimate the variance-covariance matrix of the coefficients. Moreover, the EM-IWLS is a more general model fitting algorithm that can be used for fitting not only SSL and continuous SS priors but also Student's t-priors and double exponential priors. To note, SB-GAM[13,14] also used an EM-CD algorithm. The main difference between proposed EM-CD algorithm and that in SB-GAM is that SB-GAM uses a block CD algorithm for their group prior, while the proposed prior is pairwise independent requiring no special treatment in the CD algorithm.

2.2.1 | EM algorithms

EM algorithm is an iterative algorithm to find MAP estimates or the maximum likelihood estimates. It is commonly used when some necessary data to establish the likelihood function are missing. Instead of maximizing the the likelihood function, the algorithm maximizes the expectation of the likelihood function with respect to the "missing" data.

The recursive algorithm consists of two steps:

- E-step: to calculate the expectation of the posterior density function with respect to some "missing" data
- M-step: to maximize the expectation derived in the E-step and update parameters of interest

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. 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(1,1) 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 MAP 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 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 $\Theta^{(t)}$ that maximize $E \log f(\Theta, \boldsymbol{\gamma}|\mathbf{y}, \mathbf{X})$. The E- and M- steps are iterated until the algorithm converge.

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

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

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

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

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 $\boldsymbol{\theta}$. Depending on the choice of coefficient priors, Q_1 can be treated as penalized likelihood function and maximization of $E(Q_1)$ can be solved via CD algorithm or IWLS algorithm in each iteration. Maximization of $E(Q_2)$ can be solved via closed form equations following the beta-binomial conjugate relationship.

2.2.2 | EM-Coordinate Descent

When the prior distribution of the coefficients is set to mixture double exponential, coordinate descent algorithm can be used to estimate the parameters in the M-step. Coordinate descent is an optimization algorithm that offers extreme computational advantage, and famous for its application in optimizing the l_1 penalized likelihood function.

The density function of spike-and-slab mixture double exponential prior 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

$$E(Q_1) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) - \sum_{j=1}^{p} \left[E({S_j}^{-1})|\beta_j| + \sum_{k=1}^{K_j} E({S^*}_j^{-1})|\beta_{jk}| \right], \tag{6}$$

where $S_j = (1 - \gamma^0_j)s_0 + \gamma^0_j s_1$ and $S^*_j = (1 - \gamma^*_j)s_0 + \gamma^*_j s_1$. To calculate two unknown quantities $E({S_j}^{-1})$ and $E({S^*}_j^{-1})$, the posterior probability $p_j \equiv \Pr(\gamma^0_j = 1|\Theta^{(t-1)})$ 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 value of the vector $\boldsymbol{\beta}^*_j$ and p_j only depends on the scalar β_j. The calculation follows the equations below.

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

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

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

$$E(\gamma_j) = p_j \qquad E(\gamma^*_j) = p^*_j$$

$$E(S_j^{-1}) = \frac{1 - p_j}{s_0} + \frac{p_j}{s_1} \qquad E({S^*}_j^{-1}) = \frac{1 - p^*_j}{s_0} + \frac{p^*_j}{s_1}. \tag{7}$$

After replacing withe 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[34]. 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,

$$\theta_j = \frac{p_j + p_j^* + a - 1}{a + b}. \tag{8}$$

Totally, the proposed EM-coordinate descent 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 (6)
   b) Update $\boldsymbol{\theta}^{(t)}$ using Equation (8)

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}).

2.2.3 | EM-IWLS

Similar to the EM-CD algorithm, the EM-IWLS algorithm is an iterative EM-based algorithm where the iterative weighted least squares algorithm is used to find the estimate of $\boldsymbol{\beta}, \phi$ that maximizes $E(Q_1)$. The iterative weighted least squares algorithm was originally proposed to fit the classical generalized linear models, and generalized to fit some Bayesian hierarchical models.[35] Yi and Ma[36] formulated Student's t-distribution and double exponential distribution as hierarchical normal distributions such that generalized linear models with shrinkage priors can be easily fitted using IWLS in combination with EM algorithm. In this work, we adapt the EM-IWLS paradigm to fit BHAM with spike-and-slab spline prior .

A double exponential prior, $\beta|S \sim DE(0, S)$ can be formulated as a hierarchical normal prior with unknown variance τ^2 integrated out:

$$\begin{aligned} \beta|\tau^2 &\sim N(0, \tau^2) \\ \tau^2|S &\sim Gamma(1, 1/(2S^2)), \end{aligned}$$

For the mixture double exponential priors, we can define the scale parameter $S = (1 - \gamma)s_0 + \gamma s_1$ following Equation (4). The change in the prior formulation in turn leads to the change in the log posterior density function, as Q_1 needs to account for the hyperprior of τ^2:

$$Q_1(\boldsymbol{\beta}, \phi) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) + \sum_{j=1}^{p} \left[\log f(\beta_j|\tau_j^2) + \log f(\tau_j^2|S_j) + \sum_{k=1}^{K_j} \{\log f(\beta_{jk}^*|\tau^{*2}_{jk}) + \log f(\tau^{*2}_{jk}|S_j^*)\} \right]. \tag{9}$$

Since $\boldsymbol{\tau}^2$ are not of our primary interest, we treat them as the "missing" data in addition to the latent indicators $\boldsymbol{\gamma}$, and hence construct the expectation $E_{\boldsymbol{\gamma},\boldsymbol{\tau}^2|\Theta^{(t-1)}}(Q_1)$ in the E-step. To note, unlike the same latent indicator γ_j^* which is shared by the coefficients of the non-linear terms β_{jk}^* for $k = 1, \ldots, K_j$, τ_{jk}^2 is coefficient specific for β_{jk}^*. $E({S_j}^{-1}|\beta_j, s_0, s_1)$, $E({S^*_j}^{-1}|\boldsymbol{\beta}_j^*, s_0, s_1)$, $E(\tau_j^2|S_j, \beta_j)$ and $E(\tau^{*2}_{jk}|S_j^*, \beta_{jk}^*)$ needs to be calculated to formulate $E(Q_1)$. As neither $E({S_j}^{-1}|\beta_j, s_0, s_1)$ nor $E({S^*_j}^{-1}|\boldsymbol{\beta}_j^*, s_0, s_1)$ depends on τ^2s, they can be derived using Equation (7). On the other hand, τ^2, following gamma distributions, is a conjugate prior for the normal variance, and the conditional posterior density of τ^{-2} is an inverse Gaussian distribution. $E(\tau_j^{-2})$ and $E({\tau^*}^{-2}_{jk})$ are calculated using the closed form equation

$$E(\tau_j^{-2}|S_j, \beta_j) = {S_j}^{-1}/|\beta_j| \qquad E({\tau^*}^{-2}_{jk}|S_j^*, \beta_{jk}^*) = {S_j^*}^{-1}/|\beta_{jk}^*|,$$

where S_j and S_j^* are replaced by the expectation and βs are replaced with $\beta^{(t-1)}$. With simplification (up to constant additive terms), we have

$$E(Q_1) = \log f(\mathbf{y}|\boldsymbol{\beta}, \phi) - \sum_{j=1}^{p} \left[2E({\tau_j}^{-2}){\beta_j}^2 + \sum_{k=1}^{K_j} 2E({\tau^*_{jk}}^{-2}){\beta^*_{jk}}^2 \right]. \tag{10}$$

$2E(\tau^{-2})\beta^2$ can be seen as the kernel of a normal density with mean 0 and variance $E(\tau^2)$, and we can formulate the coefficients $\boldsymbol{\beta}$ as a multivariate normal distribution with means $\mathbf{0}$ and variance covariance matrix $\boldsymbol{\Sigma}_{\tau^2}$, where $\boldsymbol{\Sigma}_{\tau^2}$ is a diagonal matrix with $E(\tau^2)$s on the diagonal,

$$\boldsymbol{\beta} \sim \text{MVN}(0, \boldsymbol{\Sigma}_{\tau^2}).$$

Meanwhile, following the classical IWLS, we can approximate the generalized model likelihood at each iteration with a weighted normal likelihood:

$$f(\mathbf{y}|\boldsymbol{\beta}, \phi) \approx \text{MVN}(\mathbf{z}|\boldsymbol{X}\boldsymbol{\beta}, \phi\boldsymbol{\Sigma})$$

where the 'normal response' z_i and 'weight' w_i are called the pseudo-response and pseudo-weight respectively. The pseudo-response and the pseudo-weight are calculated by

$$z_i = \hat{\eta}_i - \frac{L^{'}(y_i|\hat{\eta}_i)}{L^{''}(y_i|\hat{\eta}_i)} \quad w_i = -L^{''}(y_i|\hat{\eta}_i),$$

where $\hat{\eta}_i = (\boldsymbol{X}\hat{\boldsymbol{\beta}})_i$, $L^{'}(y_i|\hat{\eta}_i, \hat{\phi})$ and $L^{''}(y_i|\hat{\eta}_i, \hat{\phi})$ are the first and second derivative of the log density, $\log f(\mathbf{y}_i|\boldsymbol{\beta}, \phi)$ with respect to η_i.

With $\boldsymbol{z} \sim \text{MVN}(\boldsymbol{X}\boldsymbol{\beta}, \phi\boldsymbol{\Sigma})$ and $\boldsymbol{\beta} \sim \text{MVN}(0, \phi\boldsymbol{\Sigma}_{\tau^2})$, we can augment the two multivariate normal distributions and update the estimates for $\boldsymbol{\beta}$ and ϕ via least squares in each iteration of the EM algorithm. We create the augmented response, augmented data, and augmented variance-covariance matrix following

$$\boldsymbol{z}_* = \begin{bmatrix} \boldsymbol{z} \\ \mathbf{0} \end{bmatrix} \qquad \boldsymbol{X}_* = \begin{bmatrix} \boldsymbol{X} \\ \boldsymbol{I} \end{bmatrix} \qquad \boldsymbol{\Sigma}_* = \begin{bmatrix} \boldsymbol{\Sigma} & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_{\tau^2}/\phi \end{bmatrix},$$

such that

$$\boldsymbol{z}_* \sim \text{MVN}(\boldsymbol{X}_*\boldsymbol{\beta}, \phi\Sigma_*).$$

Using the least squares estimators to update $\boldsymbol{\beta}$ and ϕ, we have

$$\boldsymbol{\beta}^{(t)} = (\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_*)^{-1}\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{z}_* \qquad \phi^{(t)} = \frac{1}{n}(\boldsymbol{z}_* - X_*\boldsymbol{\beta}^{(t)})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{z}_* - X_*\boldsymbol{\beta}^{(t)}).$$

To note, the variance-covariance matrix of the coefficient estimates variance-covariance matrix can be derived in the EM-IWLS algorithm and in turn can be used for statistical inferences,

$$\text{Var}(\boldsymbol{\beta}^{(t)}) = (\boldsymbol{X}_*^T\boldsymbol{\Sigma}^{-1}\boldsymbol{X}_*)^{-1}\phi^{(t)}.$$

Totally, the proposed EM-IWLS 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 $E(\tau_j^{-2})$, $E(\tau^{*-2}_{jk})$ with the estimates $\Theta^{(t-1)}$ from the previous iteration

   M-step:

   a) Based on the current value of β, calculate the pseudo-data $z_i^{(t)}$ and the pseudo-weights $w_i^{(t)}$
   b) Update $\boldsymbol{\beta}^{(t)}$ by runing the augmented weighted least squared
   c) If ϕ is present, update ϕ

Similar to EM-CD, we assess convergence by the criterion, $|d^{(t)} - d^{(t-1)}|/(0.1 + |d^{(t)}|) < \epsilon$, where ϵ is a small value (say 10^{-5}).

2.3 | Selecting Optimal Scale Values

Our proposed models, BHAM, require 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 suggested the value of slab scale s_1 have 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.[26] As a result, we focus on examining different values of spike scale s_0. Instead of the 2-D 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, 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 implemented in the widely used R package `glmnet`, which quickly fits Lasso models over a list of values of regularization parameters λ , giving a sequence of models for users to choose from.

3 | SIMULATION STUDY

In this section, we compare the performance of the proposed models to four alternative models: component selection and smoothing operator (COSSO)[37], adaptive COSSO[38], generalized additive models with automatic smoothing[39], SB-GAM[14]. 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. Generalized additive models with automatic smoothing, hereafter *mgcv*, is one of the most popular models for non-linear effect interpolation and prediction. SB-GAM is the first spike-and-slab lasso GAM. We implement COSSO and adaptive COSSO with R package `cosso 2.1-1`, generalized additive models with automatic smoothing with R package `mgcv 1.8-31`, 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. Both mgcv and proposed models allow customized smoothing functions and we choose the cubic regression spline. We controll 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 COSSO models, SB-GAM and the proposed models. 20 default candidates of tuning parameters (s_0 in BHAM, λ_0 in SB-GAM) are examined for SB-GAM and the proposed models 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[40].

Other related methods for high-dimensional GAMs also exist, notably the methods of sparse additive models by Ravikumar et al.[5] and stochastic search term selection for GAM[16]. However, we exclude these methods from current simulation study because of demonstrated inferior predictive performance compared to mgcv and scalability issues with increased number of predictors.[15]

3.1 | Monte Carlo Simulation Study

We follow the data generating process described in Bai[14]. We first generate $n = 500$ training data points with $p = 4, 10, 50, 100, 200$ predictors respectively, where the predictors X are simulated from a multivariate normal distribution $\text{MVN}_{n\times p}(0, I_P)$. 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 were simulated 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 covariates, while the rest covariates 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 performance of the chosen models, where training datasets are used to fit the models and testing datasets used to calculate assessment measures.

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 dataset, are averaged across 50 simulations. Computation time for model selection, final model fitting and prediction are recorded for all simulations.

Table 1 here
Table 2 here

The predictive performances have a consistent pattern across the two distributions of outcomes. Across all the scenarios, COSSO and adaptive COSSO have the least favorable performance among the applicable methods examined (See Table 1 and 2). To note, mgcv doesn't support high-dimensional analysis, i.e. the number of coefficients are greater than the sample size, and hence not evaluated when $p = 100, 200$. mgcv predicts well when p is small or moderate ($p = 4, 10$), and deteriorate when the number of predictors increase. Among the three fast-computing Bayesian hierarchical models, the proposed models,

BHAM-IWLS and BHAM-CD predicts better than SB-GAM when the dimension of are moderate (p=4, 10, 50). Particularly, BHAM-IWLS performs as good as mgcv if not better. However, in high dimensional case where we mimic the situation the signals are extremely sparse, SB-GAM has better performance than the proposed method. However, the BHAM-CD has extreme computational advantage over SB-GAM (see Table 3) without sacrificing much of the prediction accuracy.

4 | METABOLOMICS DATA ANALYSIS

In this section, we apply the proposed models BHAM, fitted with the EM-CD algorithm, to analyze two real-world 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 [14], while being computationally efficient (see Table **??**).

Table **??** here

4.1 | Emory Cardiovascular Biobank

We use the proposed models BHAM to analyze a metabolic dataset from a recently published research [41] studying plasma metabolomic profile on the three-year all-cause mortality among patients undergoing cardiac catheterization. The dataset is publicly available via *Dryad* [42]. 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 perform a univariate screening procedure on the features, via GAM implemented in `mgcv`, and choose the the top 200 features with smallest p-values. 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 misclassication error defined as $\frac{1}{n}\sum_{i=1}^{n} I(|y_i - \hat{y}_i| > 0.5)$ are calculated. BHAM-CD obtains superior AUC, Brier score, and misclassification error in the out-of-bag samples compared to SB-GAM (see Table 5).

Table 5 here

4.2 | Weight Loss Maintenance Cohort

We use the proposed models BHAM to analyze metabolomics data from a recently published study [43] on the association between metabolic biomarkers and weight loss, where the dataset is publicly available [44]. In this analysis, we primarily focuse on the analysis of one of the three studies included, weight loss maintenance cohort [45], due to the drastically different intervention effects. In the dataset, 765 metaboliltes in baseline plasma collected were profiled using liquid chromatography mass spectrometry. Quality control and natural log transformation are performed during metabolites data preparation. 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-CD obtains superior R^2, MSE, and MAE in the out-of-bag samples compared to SB-GAM (see Table 6).

Table 6 here

5 | DISCUSSION

In the paper, we described a novel generalized additive model using Bayesian hierarchical priors, particularly the proposed spike-and-slab spline prior for bi-level functional selection. Meanwhile, we introduced two optimization based algorithms for model fitting. The algorithms can be easily scale up to address high-dimensional data analysis in a computational efficient manner. Via simulations, we demonstrated that the proposed model provides as good, if not better, prediction performance compared to some state-of-the-art non-linear modeling devices.

The proposed model shares many commonality with an high-dimensional Bayesian GAM, SB-GAM[14], independently developed around the same time of this work. Both frameworks emphasize computational efficiency by deploying group spike-and-slab lasso type of priors and optimization-based fast and scalable algorithms. Bai provided the theoretical proof for the consistency of variable selection using group spike-and-slab lasso prior. Nevertheless, SB-GAM fails to address the bi-level functional selection and the model inference. The proposed model provides solutions to these remaining questions while maintaining the same level of prediction accuracy. Moreover, the proposed model renders additional flexibility to model specification, allowing various choices of smoothing functions and degrees of freedom.

For translational science purpose, we implemented the proposed model in a 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. In addition, in `BHAM`, we streamline the model fitting algorithms to support other popular Bayesian hierarchical prior for the smoothing functions, such as Student's T distribution, double exponential distribution, mixture normal and T distribution's. These priors could be helpful when the sparse assumption is weak or not necessary.,

There are some improvements possible for the proposed models. First of all, the proposed model achieves a bi-level selection via the two-part spike-and-slab spline prior. Nevertheless, this set-up could result in a situation that is not theoretically sound: the non-linear component is selected, but the linear component is not. We currently address it analytically by including the linear component in the model when non-linear component is selected. Another possible solution is to impose a dependent structure of γ_j^* on γ_{j^0}, i.e. $\gamma_j^*|\gamma_j^0, \theta_j$. Secondly, the computational time for fitting a BHAM model with EM-IWLS algorithms can be improved. Current implementation of the EM-IWLS algorithm jointly updates all coefficients in each iteration, which requires a lot computation resources. This jointly updating procedure can be enhanced by adapting a backfitting step[3] where each smoothing function are updated individually. Thirdly, when using the proposed model in real data analysis where a screening procedure is implemented before joint predictive modeling. we recommend to include the screening procedure in the cross-validation during model selection.[46] For the sake of model comparison, we fit the models using the same pool of predictors in Section 4.1.

Our future efforts direct to modeling survival outcomes 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[47]. However, it would be more efficient to directly fit Cox models via penalized pseudo likelihood function[48]. 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[49], 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[50].

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. Breiman L, Friedman JH. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association* 1985; 80(391): 580–598.

5. 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

6. 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.

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

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

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

10. 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.

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

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

13. 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.

14. 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.

15. 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.

16. 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

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. 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.

20. George EI, McCulloch RE. Variable Selection via Gibbs Sampling. *Journal of the American Statistical Association* 1993; 88(423): 881–889. doi: 10.1080/01621459.1993.10476353

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

22. Chipman H. Bayesian variable selection with related predictors. *Canadian Journal of Statistics* 1996; 24(1): 17–36. doi: 10.2307/3315687

23. Brown PJ, Vannucci M, Fearn T. Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society. Series B: Statistical Methodology* 1998; 60(3): 627–641. doi: 10.1111/1467-9868.00144

24. Ishwaran H, Rao JS. Spike and slab variable selection: Frequentist and Bayesian strategies. *The Annals of Statistics* 2005; 33(2): 730–773. doi: 10.1214/009053604000001147

25. Clyde M, George EI. Model uncertainty. *Statistical Science* 2004; 19(1): 81–94. doi: 10.1214/088342304000000035

26. 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

27. 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

28. 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

29. 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

30. 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

31. 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

32. 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

33. Yi N, Tang Z, Zhang X, Guo B. BhGLM: Bayesian hierarchical GLMs and survival models, with applications to genomics and epidemiology. *Bioinformatics* 2019; 35(8): 1419–1421.

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

35. Gelman A, Carlin J, Stern H, Dunson D, Vehtari A, Rubin B. Bayesian Data Analysis. 3rd editio. 2013.

36. Yi N, Ma S. Hierarchical Shrinkage Priors and Model Fitting for High-dimensional Generalized Linear Models. *Statistical Applications in Genetics and Molecular Biology* 2012; 11(6). doi: 10.1515/1544-6115.1803

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

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

39. 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.

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

41. 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.

42. 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.

43. 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.

44. 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.

45. 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.

46. Friedman JH. *The elements of statistical learning: Data mining, inference, and prediction*. springer open . 2017.

47. 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.

48. 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.

49. 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.

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

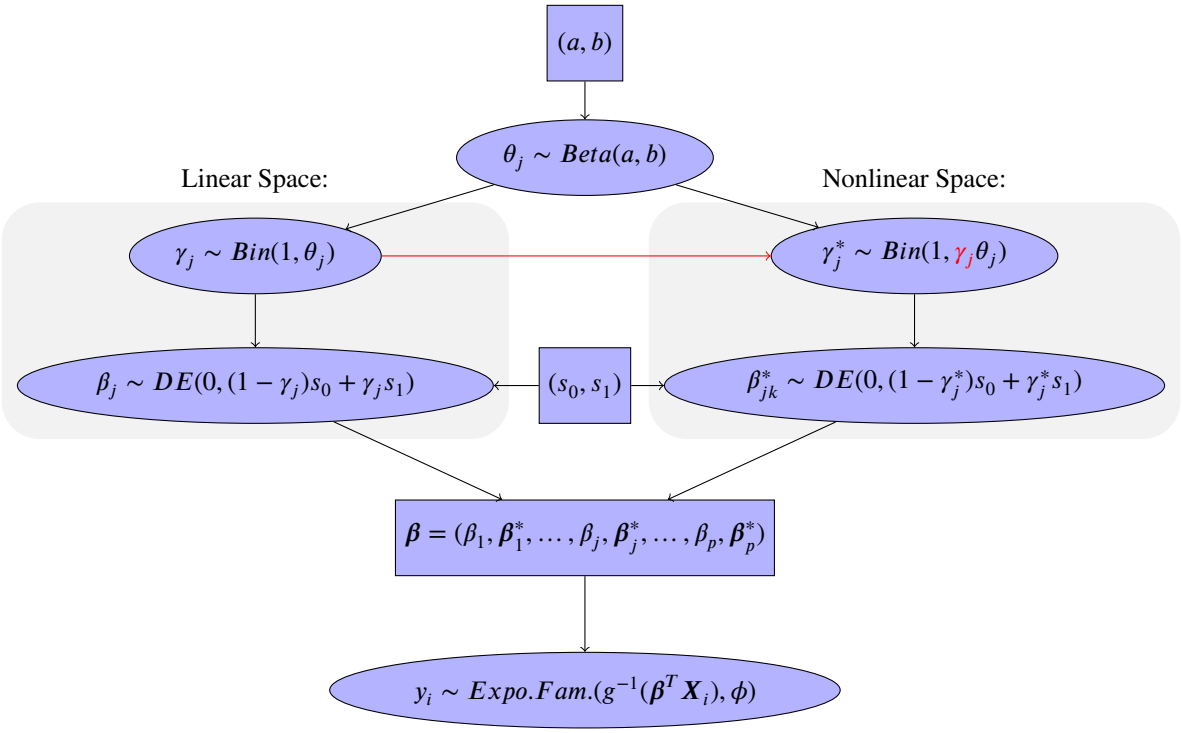

FIGURE 1 Directed acyclic graph of the proposed Bayesian hierarchical additive model with parameter expansion. Elliposes are stochastic nodes, rectangles and are deterministic nodes.

P	BHAM-IWLS	BHAM-CD	COSSO	Adaptive COSSO	mgcv	SB-GAM
4	0.90 (0.01)	0.90 (0.01)	0.75 (0.03)	0.71 (0.13)	0.90 (0.01)	0.82 (0.04)
10	0.90 (0.01)	0.88 (0.01)	0.67 (0.15)	0.76 (0.03)	0.90 (0.01)	0.82 (0.04)
50	0.88 (0.01)	0.80 (0.04)	0.43 (0.17)	0.57 (0.19)	0.86 (0.02)	0.82 (0.04)
100	0.80 (0.07)	0.73 (0.06)	0.41 (0.19)	0.51 (0.22)	-	0.81 (0.04)
200	0.72 (0.10)	0.77 (0.02)	0.33 (0.15)	0.44 (0.19)	-	0.82 (0.04)

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	BHAM-IWLS	BHAM-CD	COSSO	Adaptive COSSO	mgcv	SB-GAM
4	0.94 (0.01)	0.94 (0.02)	0.90 (0.02)	0.90 (0.01)	0.94 (0.01)	0.93 (0.01)
10	0.93 (0.01)	0.89 (0.02)	0.85 (0.04)	0.86 (0.03)	0.92 (0.04)	0.92 (0.01)
50	0.92 (0.01)	0.89 (0.01)	0.83 (0.02)	0.83 (0.02)	0.76 (0.04)	0.92 (0.01)
100	0.89 (0.02)	0.86 (0.02)	0.83 (0.02)	0.84 (0.02)	-	0.92 (0.01)
200	0.88 (0.01)	0.86 (0.02)	0.82 (0.05)	0.81 (0.08)	-	0.92 (0.01)

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.

TABLE 4 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	225.2	3.0	228.2	3,506.4	34.4	3,540.7
WLM	483.1	7.6	490.7	3,116.0	32.7	3,148.7

Distribution	P	BHAM-IWLS	BHAM-CD	COSSO	Adaptive COSSO	mgcv	SB-GAM
Binomial	4.00	5.88 (1.01)	6.96 (1.76)	2.97 (0.82)	4.60 (2.32)	0.18 (0.04)	343.24 (88.37)
Binomial	10.00	14.01 (1.86)	7.65 (1.25)	7.95 (2.70)	10.01 (3.56)	3.51 (12.03)	543.13 (133.51)
Binomial	50.00	185.26 (17.44)	53.83 (6.26)	109.72 (24.31)	128.13 (23.26)	670.01 (151.28)	1630.42 (193.61)
Binomial	100.00	1106.93 (356.73)	48.41 (5.84)	715.15 (170.55)	718.63 (193.53)	-	2783.99 (235.31)
Binomial	200.00	6923.62 (1911.59)	99.98 (7.05)	5572.51 (1295.18)	4958.69 (1970.46)	-	4780.10 (488.66)
Gaussian	4.00	4.22 (1.51)	2.51 (0.20)	0.79 (0.11)	0.67 (0.07)	0.05 (0.00)	36.02 (3.02)
Gaussian	10.00	17.59 (5.77)	33.32 (3.50)	3.49 (0.70)	3.47 (0.71)	0.32 (0.34)	72.93 (9.52)
Gaussian	50.00	329.83 (44.57)	334.53 (22.33)	35.54 (9.15)	35.44 (9.02)	73.33 (71.24)	373.42 (34.93)
Gaussian	100.00	1706.19 (225.92)	446.98 (79.31)	150.58 (46.72)	152.82 (45.07)	-	684.05 (67.80)
Gaussian	200.00	16379.27 (4825.56)	303.73 (87.27)	591.11 (147.88)	569.92 (107.45)	-	1314.12 (137.85)

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.

Methods	Deviance	AUC	Brier	Misclass
BHAM-CD	455.69	0.74	0.16	0.21
SB-GAM	1230.06	0.71	0.21	0.24

TABLE 5 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	665.63	0.07	0.94	0.75
SB-GAM	666.83	0.03	0.98	0.77

TABLE 6 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).