Journal of Statistical Software

MMMMMM YYYY, Volume VV, Issue II. doi: 10.18637/jss.v000.i00

The R Package BHAM: Fast and Scalable Bayeisan Hierarchical Additive Model for High-dimensional Data

Boyi Guo
University of Alabama at Birmingham

Nengjun Yi
University of Alabama at Birmingham

Abstract

\pkg{BHAM} is a freely avaible R pakcage that implments Bayesian hierarchical additive models for high-dimensional clinical and genomic data. The package includes functions that generlized additive model, and Cox additive model with the spike-and-slab LASSO prior. These functions implements scalable and stable algorithms to estimate parameters. \pkg{BHAM} also provides utility functions to construct additive models in high dimensional settings, select optimal models, summarize bi-level variable selection results, and visualize nonlinear effects. The package can facilitate flexible modeling of large-scale molecular data, i.e. detecting succeptable variables and inforing disease diagnostic and prognostic. In this article, we describe the models, algorithms and related features implemented in \pkg{BHAM}. The package is freely avaiable via the public GitHub repository `https://github.com/boyiguo1/BHAM`.

Keywords: additive model, spike-and-slab LASSO, scalable.

1. Introduction

High-dimensional statistics has been an indispensable area of research for its high impact in molecular and clinical data analysis. In recent year, there are continuous efforts to make high-dimensional models more flexible and interpretable, aiming to capture more complex signals. One particular family of such flexible and interpretable models is the additive models where predictors are included in a model in their functional forms. The additive models can help select predictors who have linear or nonlinear effects and provide more accurate prediction when nonlinear effects exist. Guo et al. developed Bayesian hiarchical additive models to analyze continous, categorical and survival outcomes, and demonstrated improved prediction performance compare to the state-of-the-art additive models. In this article, we introduce the R package BHAM that implements the spike-and-slab LASSO additive models and computationally efficient algorithms to fit these models.

The package BHAM provides functions for setting up and fitting various spike-and-slab LASSO additive models, including generalized additive models for various continuous and discrete otucoems and Cox survival models for censored survival outcomes. These functions are extended from previously published Bayesian Hierarchical linear models BhGLM, and develop upon commonly used R functions s in mgcv to construct additive functions. Hence, the proposed models shares similar syntax from well-developed packages and provide powerful feasures f these standard tools. The sytax can be easily followed and provide user friendliness. In addition, the algorithms implemented in BHAM is easily scalable, particularly suitable for fitting high-dimensional models. In the package, we also provide a series utility functions, for example . Hence, BHAM provides xxxx and is helpful for xxx.

1.1. Literature Review

We enlist current available packages that have similar functionality, i.e. modeling to the best of our knowledge. To note, we don't list packages that are unable of handling high-dimensional data, for example the well known R package mgcv, and high-dimensional packages that requires extra steps to construct the design matrix of functional form of predictors (Such implementation can be found with grouped sparse models, for example SGL.)

? Summarized the software development of additive models in high-dimensional data analysis before 2013.

Generalized Additive Model

- COSSO
- spikeSlabGAM
- sparseGAM

Additive Cox Proportional Hazard Model

- COSSO
- tfCox

The **BHAM** package provides a scalable solution for fitting high-dimensional generalized additive model and additive Cox model using spike-and-slab LASSO priors or other regularized priors, including continuous spike-and-slab priors, Student' T priors and double exponential priors. It fits linear, logistic, poisson and Cox regression models. The specification of the additive functions follows a popular syntax implemented in mgcv. Ancillary functions are provided, including cross-validation, model summary, and visualization.

In this article, we focus on the packages that can directly construct additive models for high-dimensional data analysis, instead of requiring additional step of constructing design matrix of functional form of the variables before fitting a sparse model.

There are other methods to model survival outcome and provides proporitonal hazards interpretation, for example **?** provides a link-based survival additive model for mixed censoring in package `GJRM`.

2. Models and algorithms

In this section, we describe the Bayeisan hiearchical additive model that `BHAM` implements. The basic idea is to impose the two-part spike-and-slab LASSO prior **?** on each additive function in the model. The choices of model includes generalized additive model and Cox proportional hazard model. The proposed two-part spike-and-slab LASSO prior consists of a spike-and-slab LASSO prior for the linear space coefficient β_j of a additive function $B_j(X_j)$ of the jth variables, and a modified group spike-and-slab LASSO prior for the nonlinear space coefficients $\beta^*_{jk}, k = 1, ..., K_j$ of the jth additive function.

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

To note, the model matrix of the additive function undergoes a reparameterization process that absorbs the smoothing penalty via eigendecomposition. Meanwhile, the reparameterization also isolate the linear and nonlinear spaces of the additive function, allowing different shrinkages on the two spaces and motivates signal selection via the linear space and function smoothing of the nonlinear space. The spike-and-slab prior use the binary indicator γ to indicate if the the corresponding variable is included in the model. Nevertheless, this selection finalized based on soft-thresholding. the spike-and-slab LASSO prior makes this selection process easier by shrinking the coefficient to exactly 0. In the two-part SSL prior, each additive function have two indicators γ_j and γ^*_j, controlling the linear and nonlinear component selection. Effect hierarchy was implemented via the conditional priors of to ensure the the linear component is more likely to be selected than the nonlinear components.

$$\gamma_j|\theta_j \sim Bin(1, \theta_j) \qquad\qquad \gamma^*_j|\gamma_j, \theta_j \sim Bin(1, \gamma_j\theta_j). \tag{2}$$

The inclusion probability parameter θ_j have a beta prior to allow adaptive shrinkage.

To fit the model in a efficient and scalable fashion, we implement the EM-coordinate descent algorithm. The EM-coordinant descent algorithm estimates maximum a posteriori of the coefficients by optimizing the log joint posterior density function. The algorithm re-writes the spike-and-slab LASSO prior as a double exponential distribution with conditional scale parameter, and leverages the relationship between double exponential prior and l_1 penalty. Hence, the log joint posterior density function can be expressed as the summation of a l_1 penalized likelihood function and log beta posterior density. Nevertheless, the nuances parameters $\boldsymbol{\gamma}$ are unknown and requires the EM algorithm to address. In each iterations of the EM procedure, we update the expectation of the log joint posterior density function with respect to the nusance parameters, calculate the penalties based on the estimation from previous iteration, and optimize the penalized likelihood and the posterior density with coordinate descent algorithm and closed-form calculation for the coefficients. The process iterates until convergence. Cross-validation is used to choose the optimal model. We defer **??**to for full description of GAM algorithm and Cox additive model algorithm.

3. Features

In this section, we demonstrate how to fit

4. Discussion

5. Reference

Affiliation:

Boyi Guo
University of Alabama at Birmingham
1665 University Blvd
Birmingham, AL 35294-0002 USA
E-mail: boyiguo1@uab.edu
URL: http://boyiguo1.github.io

Nengjun Yi
University of Alabama at Birmingham
1665 University Blvd
Birmingham, AL 35294-0002 USA
E-mail: nyi@uab.edu

Journal of Statistical Software http://www.jstatsoft.org/
published by the Foundation for Open Access Statistics http://www.foastat.org/
MMMMMM YYYY, Volume VV, Issue II *Submitted:* yyyy-mm-dd
doi:10.18637/jss.v000.i00 *Accepted:* yyyy-mm-dd