Spike-and-Slab LASSO Generalized Additive Models and Fast Algorithms for High-Dimensional Data Analysis

Boyi Guo and Nengjun Yi

Department of Biostatistics
University of Alabama at Birmingham

March 28th, 2022

Outline

Outline

- Background
  - Review of generalize additive model
  - Challenges in high dimensional additive model
- Objectives
- Bayesian Hierarchical Additive Model
  - Two-part Spike-and-slab LASSO Prior for Smoothing Functions
  - EM-Coordinate Descent algorithm
- Numeric Studies
- Conclusion

Background

Nonlinear Effect Modeling

> *"It is extremely unlikely that the true (effect) function f(X) (on the outcome) is actually linear in X."*
>
> *— Hastie, Tibshirani, and Friedman (2009) PP. 139*

- Traditional modeling approaches
  - Categorization of continuous variable, polynomial regression
  - Simple but may be statistically flawed
- Machine learning methods
  - Black-box algorithms: Random forests, neural network
  - Predict accurate but too complicated for interpretation

Generalized Additive Model (GAM)

Firstly formulated by Hastie and Tibshirani (1987)

$$y_i \overset{\text{iid}}{\sim} EF(\mu_i, \phi), \quad i = 1, \ldots, n$$

$$\mu_i = g^{-1}(\beta_0 + \sum_{j=1}^{p} B_j(x_j))$$

where $B_j(x_j)$ is a smoothing function, $g(\cdot)$ is a link function, ϕ is the dispersion parameter

- Objective: to estimate smoothing functions $B_j(x_j)$
- Applications in biomedical research:
  - Dose-response curve
  - Time-varying effect

High-dimensional GAM

- Grouped penalty models
  - Grouped lasso penalty (Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010), grouped SCAD penalty (Wang, Chen, and Li 2007; Xue 2009)
  - Sparse penalty induces **excess shrinkage**, causing inaccurate interpolation of nonlinear effect
- Bayesian Hierarchical Models
  - Grouped spike-and-slab priors (Scheipl, Fahrmeir, and Kneib 2012; Yang and Narisetty 2020), grouped spike-and-slab lasso prior(Bai et al. 2020; Bai 2021)
  - Mostly Markov chain Monte Carlo methods for model fitting
  - Computational inefficiency causes **scaling problems** in high-dimensional data analysis

Other challenges

- Bi-level selection
  - To detect if a smoothing function is linear and nonlinear
  - All-in-all-out selection reduces the ability of result interpretation
- Uncertainty inferences
  - Penalized models doesn't provide uncertainty measures
  - Challenging to estimate the effective degree of freedom for each smoothing functions

Objectives

- To develop statistical models that improve curve interpolation and outcome prediction
  - Local adaption of sparse penalty and smooth penalty
  - Bi-level selection for linear and nonlinear effect
- To develop a fast and scalable algorithm
- To implement a user-friendly statistical software

Bayesian Hierarchical Additive Model (BHAM)

Model

Given the data $\{\boldsymbol{X}_i, y_i\}_{i=1}^n$ where $\boldsymbol{X}_i \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ and $p >> n$, we have the generalized additive model

$$
\begin{aligned}
y_i &\overset{\text{i.i.d.}}{\sim} EF(\mu_i, \phi), \\
g(\mu_i) &= g^{-1}(\beta_0 + \sum_{j=1}^{p} B_j(x_j)), \quad i = 1, \dots, n.
\end{aligned}
$$

The smoothing function can be written in a matrix form $B_j(x_j) = \beta_j^T \boldsymbol{X}_j$, where β_j are the coefficients of the smoothing function and $\boldsymbol{X}_j$ is the basis matrix of dimension K_j.

Smoothing Function Reparameterization

- Smoothing penalty from Smoothing spline regression (Simon N. Wood 2017)

$$\lambda_j \int B_j''(x)dx = \lambda_j \boldsymbol{\beta}_j^T \boldsymbol{S}_j \boldsymbol{\beta}_j,$$

  where S_j is a known smoothing penalty matrix.

- Isolate the linear and nonlinear components via eigendecomposing S_j

$$\boldsymbol{X}\boldsymbol{\beta} = X^0\beta + \boldsymbol{X}^*\beta^*$$

- Benefits
  - Motivate bi-level selection
  - Implicit modeling of function smoothness
  - Reduce computation load with conditionally independent prior of basis coefficients

Two-part Spike-and-slab LASSO (SSL) Prior

- SSL prior for the linear coefficient and group SSL priors for nonlinear coefficients

$$\beta_j|\gamma_j, s_0, s_1 \sim DE(0, (1-\gamma_j)s_0 + \gamma_j s_1)$$

$$\beta^*_{jk}|\gamma^*_j, s_0, s_1 \overset{\text{iid}}{\sim} DE(0, (1-\gamma^*_j)s_0 + \gamma^*_j s_1), k = 1, \dots, K_j$$

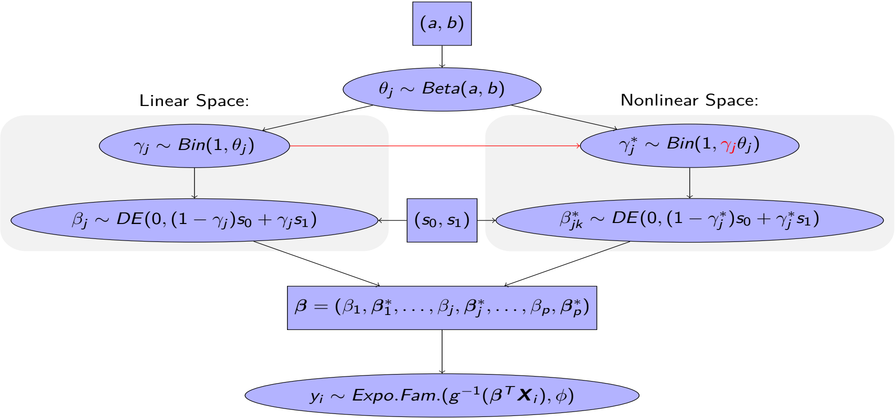
- Effect hierarchy enforced latent inclusion indicators γ_j and γ^*_j for bi-level selection

$$\gamma_j|\theta_j \sim Bin(\gamma_j|1, \theta_j), \quad \gamma^*_j|\gamma_j, \theta_j \sim Bin(1, \gamma_j\theta_j),$$

- Local adaptivity of signal sparsity and function smoothness

$$\theta_j \sim \text{Beta}(a, b)$$

Visual Representation



EM-Cooridante Descent Algrithm for Scalable Model Fitting

We are interested in estimating $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \phi\}$ using optimization based algorithm for scalability purpose

- Basic Ideas
  - Treat γs as the "missing data" in the EM procedure
  - Quantify the expectation of log posterior density function of Θ with respect to γ conditioning on $\Theta^{(t-1)}$
  - Maximize the independent parts of the objective function using Coordinate Descent algorithm and closed-form calculation
- Previous applications in high-dimensional data analysis
  - EMVS (Ročková and George 2014), Spike-and-slab lasso (Ročková and George 2018)
  - BhGLM (Yi et al. 2019)

Decomposition of Objective Function

We aim to maximize the log posterior density of Θ by averaging over all possible values of γ

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

- L_1-penalized likelihood function of $\boldsymbol{\beta}, \phi$

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

- Posterior density of θ given data points γs

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

- Q_1 and Q_2 are independent conditional on γs

Summary of EM-Coordinate Descent Algorithm

- E-step
  - Formulate $E_{\gamma|\Theta^{(t)}}[Q(\Theta, \gamma)] = E(Q_1) + E(Q_2)$
    - $E(Q_1)$ is a penalized likelihood function of β, ϕ
    - $E(Q_2)$ is a posterior density of θ given $E(\gamma)$
    - $E(Q_1)$ and $E(Q_2)$ are conditionally independent
  - Calculate $E(\gamma_j)$, $E(\gamma_j^*)$ and the penalties parameters by Bayes' theorem
- M-step:
  - Use Coordinate Descent to fit the penalized model in $E(Q_1)$ to update β, ϕ
  - Closed form calculation via $E(Q_2)$ to update θ

Tuning Parameter Selection

- s_0 and s_1 are tuning parameters
- Empirically, s_1 has extremely small effect on changing the estimates
- Focus on tuning s_0
- Consider a sequence of L ordered values $\{s_0^l\} : 0 < s_0^1 < s_0^2 < \cdots < s_0^L < s_1$
- Cross-validation to choose optimal value for s_0

Simulation Study

Simulation Study

- Follow the data generating process introduced in Bai et al. (2020).
- $n_{train} = 500$, $n_{test} = 1000$
- $p = 4, 10, 50, 200$

$$\mu = 5\sin(2\pi x_1) - 4\cos(2\pi x_2 - 0.5) + 6(x_3 - 0.5) - 5(x_4^2 - 0.3),$$

- $f_j(x_j) = 0$ for $j = 5, \ldots, p$.
- 2 types of outcome: Gaussian ($\phi = 1$), Binomial
- Splines are constructed using 10 knots
- 50 Iterations

Comparison & Metircs

- Methods of comparison
  - Proposed model BHAM
  - Linear LASSO model as the benchmark
  - mgcv (S. N. Wood 2004)
  - COSSO (Zhang and Lin 2006) and adaptive COSSO(Storlie et al. 2011)
  - Sparse Bayesian GAM (Bai 2021)
  - spikeSlabGAM (Scheipl, Fahrmeir, and Kneib 2012)
- Metrics
  - Prediction: R^2 for continuous outcomes, out-of-sample AUC for binary outcomes
  - Variable Selection Performance: positive predictive value (precision), true positive rate (recall), and Matthews correlation coefficient (MCC)

Prediction Performance

- Linear LASSO Model performs bad and mgcv performs well
- BHAM performs better than COSSO, adaptive COSSO and spikeSlabGAM
- BHAM performs better than SB-GAM in low-dimensional case but slightly worse in the high-dimensional case
- BHAM is much faster than SB-GAM in fitting models

Variable Selection Performance

- SB-GAM has the best variable selection performance
- BHAM has conservative selection
- BHAM and spikeSlabGAM have trade-offs for bi-level selection
  - spikeSlabGAM tends to select either linear or nonlinear components of the funciton
  - BHAM is more likely to select both parts

Conclusion

Conclusion

- Propose a scalable Bayesian Hierarchical Additive Model (BHAM) for high-dimensional data analysis
  - Organic balance between sparse penalty and smooth penalty
  - Bi-level selection for linear and nonlinear effects
- R package: `BHAM`
  - Ancillary functions for high-dimensional formulation
  - Model summary and variable selection
  - Website via *boyiguo1.github.io/BHAM*

References

References I

Bai, Ray. 2021. "Spike-and-Slab Group Lasso for Consistent Estimation and Variable Selection in Non-Gaussian Generalized Additive Models." *arXiv:2007.07021v5.*

Bai, Ray, Gemma E Moran, Joseph L Antonelli, Yong Chen, and Mary R Boland. 2020. "Spike-and-Slab Group Lassos for Grouped Regression and Sparse Generalized Additive Models." *Journal of the American Statistical Association*, 1–14.

Hastie, Trevor, and Robert Tibshirani. 1987. "Generalized additive models: Some applications." *Journal of the American Statistical Association* 82 (398): 371–86. https://doi.org/10.1080/01621459.1987.10478440.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

References II

Huang, Jian, Joel L Horowitz, and Fengrong Wei. 2010. "Variable Selection in Nonparametric Additive Models." *Annals of Statistics* 38 (4): 2282.

Ravikumar, Pradeep, John Lafferty, Han Liu, and Larry Wasserman. 2009. "Sparse additive models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 71 (5): 1009–30. https://doi.org/10.1111/j.1467-9868.2009.00718.x.

Ročková, Veronika, and Edward I. George. 2014. "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association* 109 (506): 828–46. https://doi.org/10.1080/01621459.2013.869223.

———. 2018. "The Spike-and-Slab LASSO." *Journal of the American Statistical Association* 113 (521): 431–44. https://doi.org/10.1080/01621459.2016.1260469.

References III

Scheipl, Fabian, Ludwig Fahrmeir, and Thomas Kneib. 2012. "Spike-and-slab priors for function selection in structured additive regression models." *Journal of the American Statistical Association* 107 (500): 1518–32. https://doi.org/10.1080/01621459.2012.737742.

Storlie, Curtis B, Howard D Bondell, Brian J Reich, and Hao Helen Zhang. 2011. "Surface Estimation, Variable Selection, and the Nonparametric Oracle Property." *Statistica Sinica* 21 (2): 679.

Wang, Lifeng, Guang Chen, and Hongzhe Li. 2007. "Group SCAD Regression Analysis for Microarray Time Course Gene Expression Data." *Bioinformatics* 23 (12): 1486–94.

References IV

Wood, S. N. 2004. "Stable and Efficient Multiple Smoothing Parameter Estimation for Generalized Additive Models." *Journal of the American Statistical Association* 99 (467): 673–86.

Wood, Simon N. 2017. *Generalized additive models: An introduction with R, second edition*. https://doi.org/10.1201/9781315370279.

Xue, Lan. 2009. "Consistent Variable Selection in Additive Models." *Statistica Sinica*, 1281–96.

Yang, Xinming, and Naveen N Narisetty. 2020. "Consistent Group Selection with Bayesian High Dimensional Modeling." *Bayesian Analysis* 15 (3): 909–35.

References V

Yi, Nengjun, Zaixiang Tang, Xinyan Zhang, and Boyi Guo. 2019. "BhGLM: Bayesian Hierarchical GLMs and Survival Models, with Applications to Genomics and Epidemiology." *Bioinformatics* 35 (8): 1419–21.

Zhang, Hao Helen, and Yi Lin. 2006. "Component Selection and Smoothing for Nonparametric Regression in Exponential Families." *Statistica Sinica*, 1021–41.