Spike-and-Slab Additive Models And Fast Algorithms For High-Dimensional Data Analysis

Boyi Guo

Department of Biostatistics
University of Alabama at Birmingham

July 12th, 2022

Outline

Outline

- Background
  - Spline Model Development
  - Bayesian Regularization
  - Bayesian Variable Selection
- Dissertation
  - Two-part Spike-and-slab LASSO Prior for Spline Functions
  - EM-Coordinate Descent Algorithms
  - Empirical Performance of Prediction & Selection
- Future Research
  - Structured Additive Regression with Spike-and-Slab LASSO prior
  - Spatially Variable Genes Screening
  - Other Questions of Interest

Background

Spline Model Development

Spline Model Development

"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

Spline Functions

A *spline* function is a piece-wise polynomial function

$$B(x) = \sum_{k=1}^{K} \beta_k b_k(x) \equiv \boldsymbol{X}^T \beta$$

$b_k(x)$ are the *basis functions*, possibly truncated power basis and b-spline basis.

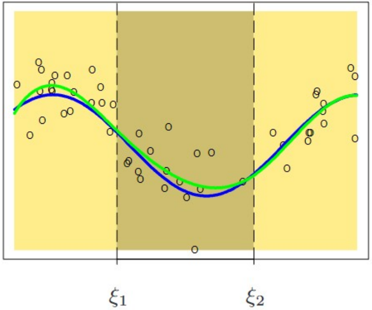

Figure 1: A cubic spline function with 2 knots (courtesy of Hastie, Tibshirani, and Friedman (2009))

Generalized Additive Models with Splines

Generalized additive model (Hastie and Tibshirani 1987) is expressed

$$y_i \overset{\text{iid}}{\sim} EF(\mu_i, \phi), \quad i = 1, \dots, n$$
$$g(\mu_i) = \beta_0 + B(x_i) = \beta_0 + \boldsymbol{X}_i^T \beta, \quad \mathbb{E}\left[B(X)\right] = 0$$

where $B(x_i)$ is the spline function, $g(\cdot)$ is a link function, ϕ is the dispersion parameter

- Model fitting follows the generalized linear models, e.g. ordinary least square for Gaussian outcome
$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \left[y_i - \beta_0 - \boldsymbol{X}_i^T \beta\right]^2$$

Problem: Function Smoothness

The estimation of $B(X)$ can be wiggly when the underlying function is smooth, particularly as the number of bases ,K, increases.

[TODO: add two plots, overfitting and not overfitting]

Bayesian Regularization

Smoothing Spline Model

- Smoothing penalty $\lambda \int B^{''}(X)^2 dx = \lambda \beta^T \boldsymbol{S} \beta$
  - The smoothing penalty matrix $\boldsymbol{S}$ is known given $\boldsymbol{X}$
  - $\boldsymbol{S}$ is symmetric and positive semi-definite
- Penalized Least Square for Gaussian Outcome

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \sum_{i=1}^{n} \left[y_i - \beta_0 - \boldsymbol{X}_i^T \beta \right]^2 + \lambda \beta^T \boldsymbol{S} \beta$$

- The smoothing parameter λ is a tuning parameter, selected via cross-validation

Problem: Multiple Predictor Model

When a model contains multiple spline functions for variables $X_1, \dots, X_p$, the penalized least square estimator is

$$\hat{\beta} = \arg\min \sum_{i=1}^{n} \sum_{i=1}^{n} \left[y_i - \beta_0 - \sum \boldsymbol{X}_{ij}^T \beta_j \right]^2 + \lambda_j \beta_j^T \boldsymbol{S}_j \beta_j$$

How to decide λ_i?

- Global smoothing, i.e. $\lambda_1 = \cdots = \lambda_p$ assumes all functions shares the same shape
- Adaptive smoothing, i.e. examining λ_i combination, are computationally intensive

Bayesian Regularization

- Bayesian Regularization is the Bayesian analogy of penalized models by using regularizing priors
  - Bayesian ridge via normal prior

$$\beta \sim N(0, \tau^2) \rightarrow \lambda = \sigma^2/\tau^2$$

- Adaptive shrinkage with hierarchical priors

$$\tau_j^2 \overset{\text{iid}}{\sim} IG(a, b)$$

  - Adaptive Smoothing
    - Random walk prior on b-spline bases with IG hyperprior
    - Normal prior on truncated power bases with a log-normal spline model for variance

Bayesian Variable Selection

Problem: Functional Selection

In the context of variable selection and high-dimensional statistics, we always assume some variables are not effective or predictive to the outcome.

How to statistically detect

- if a variable is predictive to the outcome, $B_j(X_j) = 0$
- if a variable has a nonlinear relationship with the outcome, $B_j(X_j) = \beta_j X_j$

Bi-level selection is the procedure that simultaneously addresses the two questions above

Spike-and-Slab Priors

Spike-and-slab priors are a family of mixture distributions that deploys a characterizing structure

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

- Latent indicator γ follows a Bernoulli distribution with probability θ
- Spike density $f_{spike}(x)$ concentrates around 0 for small effects
- Slab density $f_{slab}(x)$ is a flat density for large effects
- Natural procedure to select variables via posterior distribution of γ
- Markov chain Monte Carlo is not compelling for high-dimensional data analysis

Spike-and-Slab LASSO Priors

- Double exponential distributions as the spike and slab distributions

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

  - Seamless variable selection as coefficients shrinkage to 0
  - Computation advantages via Expectation-Maximization (EM) algorithms
- Group spike-and-slab LASSO
  - Structure underlying predictors, e.g. gene pathways, bases of a spline function
  - Structured prior on γ

$$\gamma_k|\theta_j \; Binomial(1, \theta_j), k \in j$$

Problem: High-dimensional Spline Model

How to jointly model signal sparsity and function smoothness, while capable of bi-level selection?

- Excess shrinkage due to ignoring smooth penalty completely
  - Group lasso penalty (Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010), group SCAD penalty (Wang, Chen, and Li 2007; Xue 2009)
  - Global penalty VS adaptive penalty
- All-in-all-out selection
  - Can not detect if a function is linear, e.g. spike-and-slab grouped LASSO prior (Bai et al. 2020; Bai 2021)
  - Failed to select function as whole, e.g. group spike-and-slab LASSO prior
- Computational prohibitive algorithms
  - MCMC algorithms doesn't scale well for high-dimensional models (Scheipl, Fahrmeir, and Kneib 2012)

Dissertation

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

Scope

Scope of this dissertation * BHAM * Survival Model * R package `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.

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.

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 III

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.