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

Boyi Guo and Nengjun Yi

Department of Biostatistics
University of Alabama at Birmingham

August 8th, 2021

Outline

Outline

- Background
  - Challenges in higher dimensional additive models
- Objectives
- Bayesian Hierarchical Additive Model
  - Spike-and-slab Spline Prior
  - EM algorithms
- Numeric Studies
- Conclusion

Background

Curve Interpolation

- Traditional modeling approach
  - Categorization of continuous variable
  - Polynomial regression
  - Simple but may be statistically flawed
- Machine learning methods
  - Random forests, neural network
  - Black-box algorithms
  - Accurate but too complicated for interpretation

Generalized Additive Model

Firstly formalized by Hastie and Tibshirani (1987)

$$y_i \stackrel{\text{i.i.d.}}{\sim} EF(\mu_i, \phi), \quad i = 1, \dots, n$$

$$\mu_i = g^{-1}(a + \sum_{j=1}^{p} f_j(x_{ij}))$$

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

- Objective: to estimate smoothing functions $f_j(\cdot)$
- Applications:
  - 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 non-linear 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 identify linear- and non-linear effects
  - All-in-all-out selection reduces the ability of result interpretation
- Uncertainty measures
  - Penalized models doesn't provide uncertainty measures
  - Bayesian models with MCMC algorithms are not scalable enough

Objectives

- To develop statistical models that improve curve interpolation in high-dimensional data analysis
  - Local adaption of sparse penalty and smooth penalty
  - Bi-level selection for linear- and non-linear effect
- To develop fast and scalable algorithms
  - Uncertainty measures
- To develop user-friendly statistical softwares

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

$$y_i \overset{\text{i.i.d.}}{\sim} EF(\mu_i, \phi),$$

$$g(\mu_i) = \sum_{j=1}^{p} f_j(X_{ij}), \quad i = 1, \dots, n.$$

We express smoothing functions in the matrix form using reparameterization

$$g(\mu_i) = \sum_{j=1}^{p} f_j(X_{ij}) = \sum_{j=1}^{p} \left[{\beta_j^0}^T X_{ij}^0 + {\beta_j^{pen}}^T X_{ij}^{pen} \right].$$

Reparameterization

- Introduced in Wood (2011)
- Smoothing penalty

$$\lambda_j \int f_j''(x)dx = \lambda_j \beta_j^T \boldsymbol{S}_j \beta_j$$

- Re-parameterization based on eigen-decomposition of S_j
  - $\boldsymbol{S} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$
  - $\boldsymbol{U} \equiv [\boldsymbol{U}^{\text{pen}} : \boldsymbol{U}^0]$ and $\boldsymbol{D} \equiv [\boldsymbol{D}^{\text{pen}} : \boldsymbol{0}]$
  - $\boldsymbol{X}\beta = \boldsymbol{X}\boldsymbol{U}\boldsymbol{U}^T\beta = X^0\beta^0 + \boldsymbol{X}^{\text{pen}}\beta^{\text{pen}}$
- Benefits
  - Isolate linear parts from the polynomial parts of smoothing functions
  - Independent prior for the penalized part

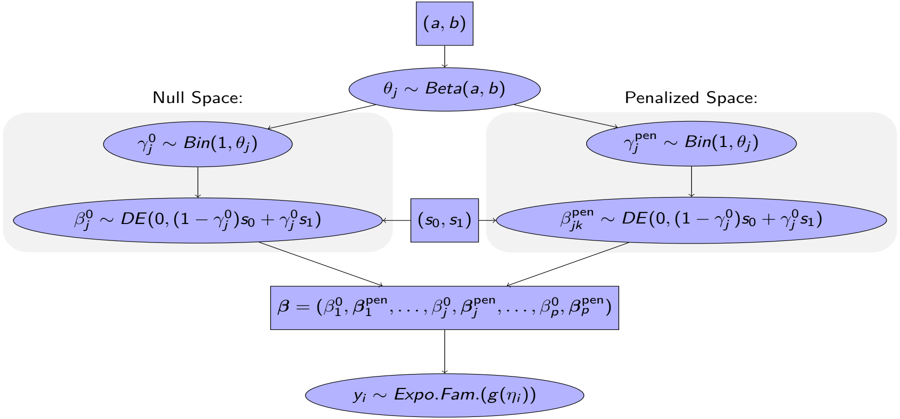
Spike-and-slab Spline Prior

We propose a two-part spike-and-slab lasso prior, mixture double exponential prior

$$
\begin{aligned}
\beta_j^0|\gamma_j^0, s_0, s_1 &\sim DE(0, (1-\gamma_j^0)s_0 + \gamma_j^0 s_1),\\
\beta_{jk}^{\text{pen}}|\gamma_j^{\text{pen}}, s_0, s_1 &\sim DE(0, (1-\gamma_j^{\text{pen}})s_0 + \gamma_j^{\text{pen}} s_1),\\
\gamma_j^0|\theta_j &\sim Bin(\gamma_j^0|1, \theta_j),\\
\gamma_j^{\text{pen}}|\theta_j &\sim Bin(\gamma_j^{\text{pen}}|1, \theta_j),\\
\theta_j &\sim \text{Beta}(a, b)
\end{aligned}
$$

$\boldsymbol{\beta}_j$ for curve interpolation, $\gamma_j^0, \gamma_j^{\text{pen}}$ for bi-level selection, θ_j for local adaption

Visual Representation



Fast Computing Algorithms

Fast Computing Algorithms

We are interested in estimate $\Theta = \{\boldsymbol{\beta}, \boldsymbol{\theta}, \phi\}$

- Two optimization based algorithms are proposed
  - EM - Coordinate descent algorithm
    - Sparse Solution and faster computation
  - EM - Iterative weighted least square
    - Uncertainty inference
- Successful history in high-dimensional data analysis
  - EMVS (Ročková and George 2014), Spike-and-slab lasso (Ročková and George 2018)
  - BhGLM (Yi et al. 2019)

EM algorithm

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

$$Q(\Theta, \gamma) \equiv \log p(\Theta, \gamma | \mathbf{y}, \mathbf{X})$$
$$= \log p(\mathbf{y}|\boldsymbol{\beta}, \phi) + \log p(\phi) + \sum_{j=1}^{p} \left[\log p(\beta_j^0 | \gamma_j^0) + \sum_{k=1}^{K_j} \log p(\beta_{jk}^{pen} | \gamma_{jk}^{pen}) \right]$$
$$+ \sum_{j=1}^{p} \left[(\gamma_j^0 + \gamma_j^{pen}) \log \theta_j + (2 - \gamma_j^0 - \gamma_j^{pen}) \log(1 - \theta_j) \right] + \sum_{j=1}^{p} \log p(\theta_j)$$

EM algorithms

- 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 $E(\gamma_j^0)$ and $E(\gamma_j^{pen})$, and penalties by Bayes' theorem
- M-step:
  - Use algorithms to fit penalized model in $E(Q_1)$to update β, ϕ
    - Coordinate descent
    - Iterative weighted least square
  - 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
- Instead of the 2-D grid, We 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 with EM-CD and EM-IWLS
  - mgcv (Wood 2004)
  - COSSO (Zhang and Lin 2006) and adaptive COSSO(Storlie et al. 2011)
  - Sparse Bayesian GAM (Bai 2021)
- Metrics
  - out-of-sample R^2 for continuous outcomes
  - out-of-sample AUC for binary outcomes

Out-of-sample AUC

- The proposed method works better in low, medium, high settings than other state-of-art methods
- SB-GAM works better in ultra-high setting

p	EM-IWLS	EM-CD	COSSO	ACOSSO	mgcv	SB-GAM
4	**0.94 (0.01)**	0.89 (0.04)	0.90 (0.02)	0.90 (0.02)	0.94 (0.01)	0.93 (0.01)
10	**0.93 (0.01)**	0.87 (0.03)	0.87 (0.03)	0.85 (0.03)	0.92 (0.04)	0.92 (0.01)
50	**0.92 (0.01)**	0.87 (0.02)	0.83 (0.02)	0.83 (0.02)	0.76 (0.04)	0.92 (0.01)
200	0.88 (0.01)	0.86 (0.02)	0.81 (0.06)	0.81 (0.08)	-	**0.92 (0.01)**

Conclusion

Conclusion

- Proposed fast and scalable high dimensional GAM
  - Organic balance between sparse penalty and smooth penalty
  - Bi-level selection for linear- and non-linear effects
  - Uncertainty measures provided
- R package: `BHAM`
  - Ancillary functions for high-dimensional formulation
  - Model summary and variable selection
  - Covariate adjustment without penalty
  - 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.

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

References II

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.

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.

References III

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.

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.

References IV

———. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society (B)* 73 (1): 3–36.

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.

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.

References V

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