A Scalable and Flexible Cox Proportional Hazards Model for High-Dimensional Survival Prediction and Functional Selection

Boyi Guo and Nengjun Yi

Department of Biostatistics
University of Alabama at Birmingham

August 8th, 2022

Background

Background

"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

Question

How to model nonlinear effects for survival outcome in **high-dimensional** setting?

Additive Cox Proportional Hazards Model

Following all necessary assumptions, a Cox proportional hazards model with event time t_i and predictors $x_{ij}, j = 1, \ldots, p$, is expressed as

$$h(t_i) = h_0(t_i) \exp(\sum_{j=1}^{p} B_j(x_{ij})), \quad i = 1, \ldots, n.$$

- Spline functions

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

$b_k(x)$ are the *basis functions*, possibly truncated power basis and b-spline basis. (Simon N. Wood 2017)

Function Smoothing

- 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 Partial Likelihood Function

$$\hat{h}_0(t_i|\beta) = d_i / \sum_{i' \in R(t_i)} exp(X_{i'}\beta).$$

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

High-dimensional Additive Cox Model

Primary Challenges:

- Jointly model signal sparsity versus function smoothness
- Adaptive shrinkage
- Bi-level selection that simultaneously answers
    - if a variable is predictive to the outcome, $B_j(X_j) = 0$
    - if a variable has a nonlinear relationship with the outcome, $B_j(X_j) = \beta_j X_j$

Bayesian Hierarchical Additive Model

Bayesian Hierarchical Additive Model

- Two-part spike-and-slab LASSO prior for spline functions
  - Variable selection via inclusion indicator
  - Bi-level selection accounting for effect hierarchy
  - Adaptive shrinkage via Bayesian regularization
- EM-Coordinate Descent algorithm
  - Expedited computation
  - Seamless variable selection via sparse solution

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

Follow xxx, a spline function $B(X) = \boldsymbol{X}^T\beta$ can be decomposed to linear and nonlinear components with respect to the smoothing penalty matrix S

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

- Proposed spike-and-slab LASSO prior

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

  - γ_j controls the inclusion of linear component
  - γ^*_j controls the inclusion of nonlinear component

Effect Hierarchy and Adaptive Shrinkage

Effect hierarchy assumes lower-order effects are more likely to be active than higher-order effects

- Structured prior on latent indicators γ_j and γ_j^*

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

  - Simplification via analytic integration

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

- Adaptive shrinkage

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

EM-Cooridante Descent Algrithm

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 two parts of the objective function independently

Decomposition of Objective Function

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

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

- L_1-penalized likelihood function of β, ϕ

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

- 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 conditioning on γs

Summary of EM-Coordinate Descent Algorithm

- E-step
  - Formulate $E_{\boldsymbol{\gamma}|\Theta^{(t)}}[Q(\Theta,\boldsymbol{\gamma})] = E(Q_1) + E(Q_2)$
    - $E(Q_1)$ is a l_1 penalized partial 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

Numeric Studies

Simulation Study

- $n_{train} = 500$, $n_{test} = 1000$
- $p = 4, 10, 50, 100, 200$
- Survival and censoring time follow Weibull distribution

$$\log \eta = (x_1 + 1)^2/5 + \exp(x_2 + 1)/25 + 3\sin(x_3)/2 + (1.4x_4 + 0.5)/2$$

- Censoring rate is controlled at $\{0.15, 0.3, 0.45\}$
- Splines are constructed using 10 knots
- 50 Iterations

Comparison & Metircs

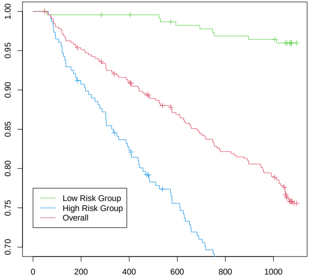
- Methods of comparison
  - Proposed model BHAM
  - Linear LASSO model as the benchmark
  - mgcv (S. N. Wood 2004)
  - COSSO (Zhang and Lin 2006) and adaptive COSSO(Storlie et al. 2011)
- Metrics
  - Out-of-sample deviance & Concordance

Prediction Performance

- Linear LASSO Model performs bad in general
- Low dimensional settings:
  - mgcv performs the best
  - BHAM performs as good as mgcv
- High dimensional setting:
  - BHAM performs better than COSSO models as p increases and more censoring events

Emory Cardiovascular Biobank

- All-cause mortality among patents undergoing cardiac catheterization
- Sample size N=454 and number of features p=200
- 5-knot cubic spline



Conclusion

Conclusion

- A scalable and flexible Cox Model for high-dimensional survival data analysis
  - Two-part spike-and-slab LASSO prior for spline functions
    - Jointly model signal sparsity and function smoothness with adaptive regularization
    - Bi-level selection that accounts the effect hierarchy principle
  - EM-Coordinate Descent algorithm
    - Computation advantage and sparse solution
- R package: `BHAM`
  - Ancillary functions for high-dimensional formulation
  - Model summary and variable selection
  - Website via *boyiguo1.github.io/BHAM*

References

References I

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

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.

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.

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