Generalized Group Lasso for Patient Subgroup Selection

Wenxuan Deng

Takeda Pharmaceuticals U.S.A., Inc.

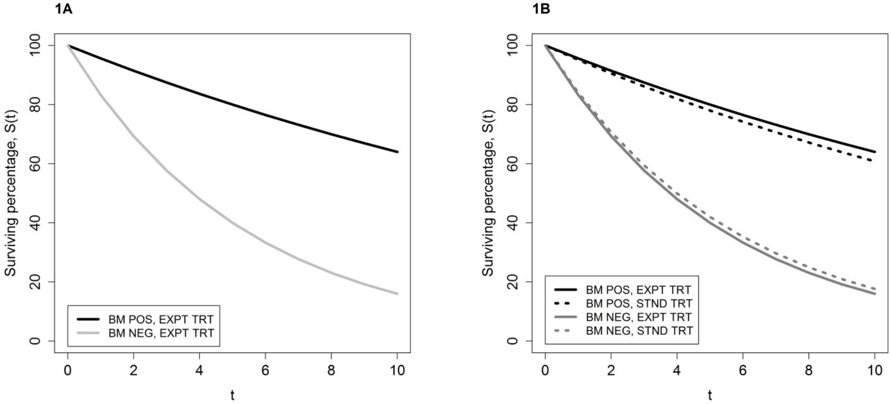
Wenxuan.Deng@takeda.com

August 16, 2018

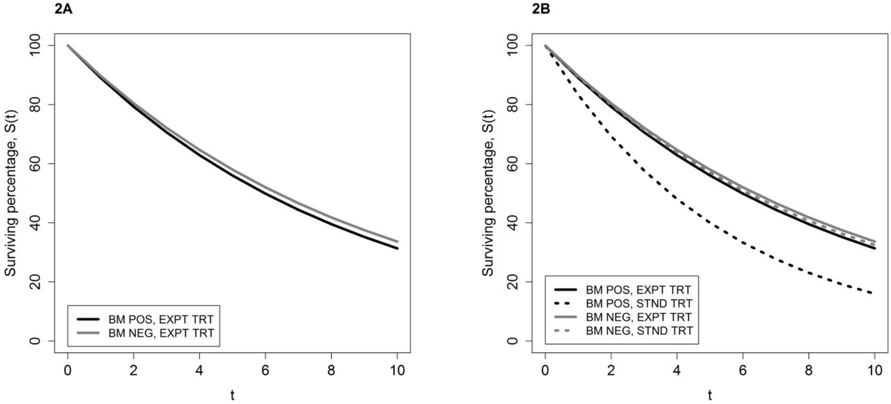
Overview

1. Introduction
   - Prognostic and Predictive Biomarkers
   - Why not regression trees?
2. Methods
3. Algorithm
   - Algorithm Framework
4. Simulation

Prognostic Biomarkers



Predictive Biomarkers



Tree-based Methods

Regression trees GUIDE[Loh, 2018]:

- piecewise-linear Model
- examine residual patterns for each treatment level

Cannot repeat even the most naive simulation in GUIDE paper.

Reason: limited sample size. Even two splits will results in small sample size in each branch. The results would be highly unstable. Tree-based method is not appropriate to clinical trial dataset and identify prognostic and predictive biomarkers.

Ordinary Linear Model

$$Y = X\beta + W\tau + G\alpha + W \otimes G\gamma + \epsilon$$

- X: Baseline variables
- W: Treatment variables
- G: Main effects of genes, i.e. expression levels, SNP or mutation
- $W \otimes G$: Interaction effects of genes and treatment
- ϵ: Random errors

Group lasso

We choose group lasso for its ability to

- handle high dimensional data
- allow hierarchical structure

However, the current group lasso based methods

- penalize on all parameters
- have no efficient adaptive penalty weights
- do not specifically target on patients treatment subgroup identification

Loss Function

We assume the hierarchical relationship between prognostic biomarkers and predictive biomarkers, that is a predictive biomarker should be a prognostic biomarker.

The loss function is

$$\min_{\theta} f(\theta|Y,X,W,G) + g(\theta)$$

where

$$g(\theta) = \lambda \sum_i \phi_i |\gamma_i| + \lambda \sum_i \theta_i \sqrt{\alpha_i^2 + \gamma_i^2} + \rho(||\alpha||^2 + ||\gamma||^2)$$

where $f(\theta|Y,X,W,G) = \| Y - (X\beta + W\tau + G\alpha + W \otimes G\gamma) \|^2$ is L-2 loss function, i.e. sum of squared errors for ordinary linear model.

$\theta = (\beta, \tau, \alpha, \gamma)$ is parameter vector.

Loss function for ordinary linear model

$$\begin{aligned}\min_{\theta} \parallel Y-(X\beta+W\tau+G\alpha+W\otimes G\gamma)\parallel^2 \\ +\lambda\sum_i \phi_i|\gamma_i|+\lambda\sum_i \theta_i\sqrt{\alpha_i^2+\gamma_i^2} \\ +\rho(||\alpha||^2+||\gamma||^2)\end{aligned} \tag{1}$$

Denote $X^{(i)}=[G_i, W\otimes G_i]$ is the lth group of the main and interaction effects of gene l. Then, based on KKT conditions, we let

$$\phi_i = \parallel X^{(i)} \parallel_2$$

$$\theta_i = \sqrt{\parallel G_i \parallel_2^2 + 1.4(1-\sqrt{\frac{2}{\pi}}) \parallel W\otimes G_i \parallel_2^2}$$

$$\min_{\theta} \parallel Y - (X\beta + W\tau + G\alpha + W \otimes G\gamma) \parallel^2 + \lambda \sum_i \phi_i |\gamma_i| + \lambda \sum_i \theta_i \sqrt{\alpha_i^2 + \gamma_i^2} \tag{2}$$

Theorem

Let p_0 be the dimension of baseline and treatment covariates, and p_1 be the dimension of main effect covariates. The total dimension is $p = p_0 + 2p_1$. Then only $\min\{p_1, n - p_0\}$ genes have nonzero main effects in equation (1).

Remark: When $p > 2n$, the number of selected genes is bounded by sample size.

Optimization Stratgies

- Fast iterative shrinkage-thresholding algorithm with backtracking[Beck and Teboulle, 2009]
- Adaptive restart for rippling behavior [O'Donoghuet and Candes, 2009]
- Adaptive stepsize of cyclic Barzilai-Borwein spectral approach[Wright, 2009]
- Warm start in cross validation

Proximal Operator

Definition

Let

$$Q_{\tau_i,g}(t,u) = g(t) + \frac{1}{2\tau} \parallel t - u \parallel^2$$

then the proximal operator is defined as

$$\tilde{t} = arg \min Q_{\tau_i,g}(t,u)$$

For convenience, we denote $P_{\tau_i,g}(u) = \tilde{t}$

Remark: Proximal operator is a point that compromises between minimizing g and being near to u.

Algorithm

```
initialization $\theta 0 = 0$ or warm start from previous run, $\tau_0 = 0.1$, stepsize $\eta = 0.5$;
while $i \leq k$ do
    $u_i = \theta_{i-1} - \tau_i \bigtriangledown f(\theta_{i-1})$ Find the smallest nonnegative integers $s_i$ such that with $\tau_i = \eta^{s_{i-1}} \tau_{i-1}$,
    $(f + g)(P_{\tau_i,g}(u_i)) \leq Q_{\tau_i,g}(P_{\tau_i,g}(u_i), u_i)$;
    Then, we compute $t_i = P_{\tau_i,g}(u_i)$ And accelarate the computation by setting if $f(\theta_i + g(\theta_i)) > f(\theta_{i-1}) + g(\theta_{i-1})$ then
        $\rho_i = 1$
    else
        $\rho_i = \frac{1+\sqrt{1+4\rho_{i-1}^2}}{2}$
    end
    $\theta_i = t_i + (\frac{\rho_{i-1}-1}{\rho_i})(t_i - t_{i-1})$ and find $\tau_{i+1}$ that $\tau_{i+1} I$ can mimic the Hessian $\bigtriangledown^2 f(\theta_i)$
end
```

Algorithm 1: Patient Subgroup Identification Group Lasso Algorithm

References

Loh, Wei-Yin, Michael Man, and Shuaicheng Wang.
"Subgroups from regression trees with adjustment for prognostic effects and postselection inference."
Statistics in medicine (2018).

Beck, Amir, and Marc Teboulle.
"A fast iterative shrinkage-thresholding algorithm for linear inverse problems."
SIAM journal on imaging sciences 2.1 (2009): 183-202. (2009).

Wright, Stephen J., Robert D. Nowak, and Mário AT Figueiredo.
"Sparse reconstruction by separable approximation."
IEEE Transactions on Signal Processing 57.7 : 2479-2493.(2009)

O'Donoghue, B., and E. Candes.
"Adaptive restart for accelerated gradient schemes."
Foundations of computational mathematics 15.3: 715-732. (2015)

The End