Generalized Group Lasso for Patient Subgroup Selection

Wenxuan Deng

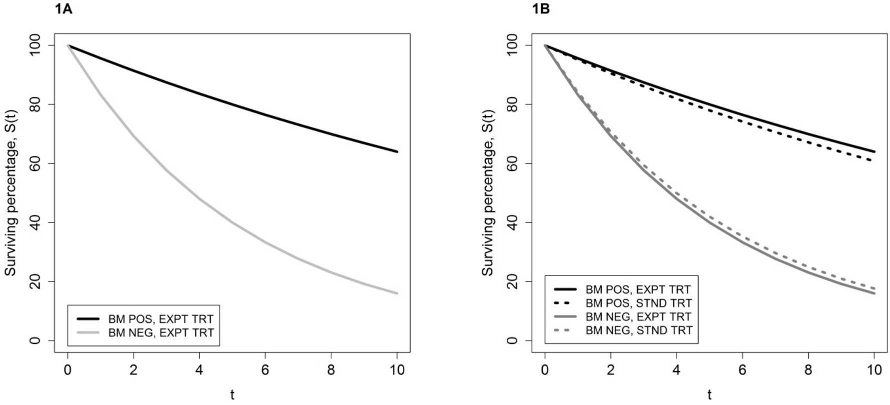
Takeda Pharmaceuticals U.S.A., Inc.

Wenxuan.Deng@takeda.com

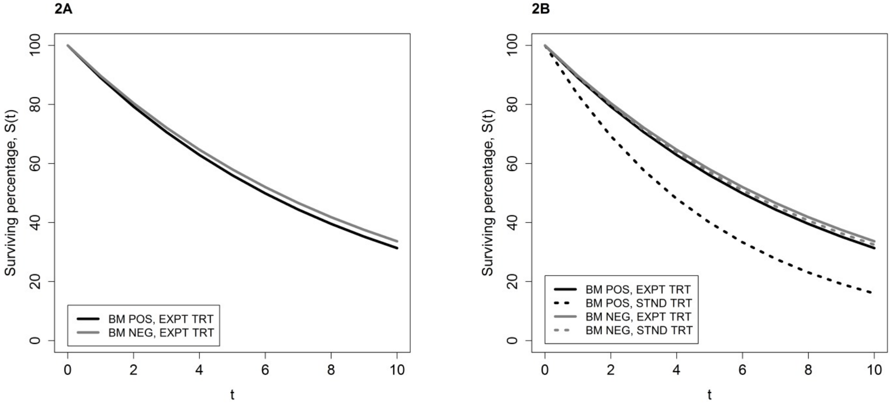
July 16, 2018

Overview

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

$$g(\theta) = \lambda \sum_i \phi_i^I |\gamma_i| + \lambda \sum_i \phi_i^M \sqrt{\alpha_i^2 + \gamma_i^2}$$

where $f(\theta | Y, X, W, G)$ 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

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

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

$$\phi_i^I = \| X^{(i)} \|_2$$

Optimization Stratgies

- Fast iterative shrinkage-thresholding algorithm with backtracking[Beck and Teboulle, 2009]
- Adaptive restart for rippling behavior [ODonoghuet 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 τ_{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

Theorem

Because $Q_{\tau_i,g}(t,u)$ is separable, so we have

$$arg\min Q_{\tau_i,g}(t,u) = [arg\min \frac{1}{2} \parallel t^{(i)} - u^{(i)} \parallel^2 + \lambda\phi_i^I|\gamma_i| + \lambda\phi^M\sqrt{\alpha_i^2+\gamma_i^2}]_{1\leq i\leq}$$

Theorem

$E = mc^2$

Multiple Columns

Heading

1. Statement
2. Explanation
3. Example

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Integer lectus nisl, ultricies in feugiat rutrum, porttitor sit amet augue. Aliquam ut tortor mauris. Sed volutpat ante purus, quis accumsan dolor.

Table

Treatments	**Response 1**	**Response 2**
Treatment 1	0.0003262	0.562
Treatment 2	0.0015681	0.910
Treatment 3	0.0009271	0.296

Table: Table caption

Theorem

Theorem (Mass–energy equivalence)

$E = mc^2$

Verbatim

Example (Theorem Slide Code)

```
\begin{frame}
\frametitle{Theorem}
\begin{theorem}[Mass--energy equivalence]
$E = mc^2$
\end{theorem}
\end{frame}
```

References

Loh, WeiYin, 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 Mrio AT Figueiredo.
"Sparse reconstruction by separable approximation."
IEEE Transactions on Signal Processing 57.7 : 2479-2493.(2009)

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

The End