Generalized Group Elastic Net for Predictive Biomarker Identification

Wenxuan Deng[1]

[1]Department of Biostatistics, Yale University,

Predictive biomarker identification is a significant problem when targeting patient subpopulation who gets an enhanced benefit under treatment. This paper proposed a new method, Predictive Effects Net (PEN), based on group lasso and special hierarchical structure for figuring out predictive effects. The new approach takes predictive biomarkers as interaction effects between treatment and biomarker. To show PEN has an supreme performance, this paper shows simulations in different scenerios and comparisons with several other variable selection methods.

Introduction

Prognostic biomarkers and predictive biomarkers.

Why decision trees is not workable? Because the sample size in real clinical datasets is too small, typically no more than 100 patients.

Group lasso (*1*)

Elastic net (*2*) adaptive weights for elastic net (*3*)

Hierarchical Group lasso for interactions (*4*)

Overlapping group lasso (*5*) (*6*) (*7*)

Sparse Group Lasso (*8*) (*9*)

Structured group lasso (*10*) Group lasso for logistic regression (*11*)

Other variable selection methods:

GUIDE: a regression tree (*12*) (*13*)

SIS: screening (*14*) (*15*)

SIR: (*16*) (*17*)

Stepwise selection: (*18*)

Methods

Model

$$Y = X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2 + \epsilon$$

Where X_0 is the baseline variables, X_T is the treatment variable, X_1 is the high dimensional design matrix of genes, i.e. gene expression levels, SNP and mutations, and $X_T \otimes X_1$ is the interaction between genes and treatment. $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$ is the corresponding coefficients. ϵ is random error.

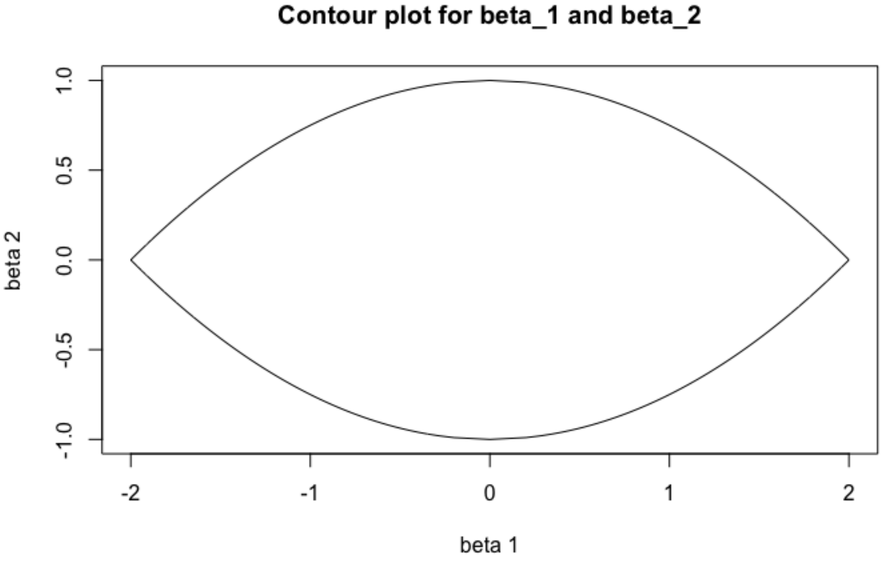
Let

$$X = [X_0, X_T, X_1^{(1)}, \ldots, X_m^{(1)}, X_T X_1^{(1)}, \ldots, X_T X_1^{(m)}]$$

and

$$\beta = [\beta_0, \beta_\tau, \beta_1^{(1)}, \ldots, \beta_1^{(m)}, \beta_2^{(1)}, \ldots, \beta_2^{(m)}]$$

For each gene l, its prognostic and predictive design matrix is denoted as $X^{(l)} = [X_1^{(l)}, X_T X_1^{(l)}]$ and its corresponding coefficients are $\beta^{(l)} = [\beta_1^{(l)}, \beta_2^{(l)}]$



Loss Function

We used group lasso and elastic net for variables selection when $n \ll p$, and assumed the hierarchical relationship between prognostic biomarkers and predictive biomarkers, that is the predictive biomarkers should be a prognostic biomarkers. The loss function is

$$\min_{\theta} f(\beta|Y, X_0, X_T, X_1) + g(\beta)$$

$$g(\beta) = \lambda_1 \sum_i \phi_i |\beta_2^{(i)}| + \lambda_1 \sum_i \psi_i \sqrt{(\beta_1^{(1)})^2 + (\beta_2^{(1)})^2} + \lambda_2(\| \beta_1 \|_2^2 + \| \beta_2 \|_2^2)$$

Where $\beta = (\beta_0, \beta_\tau, \beta_1, \beta_2)$ is the parameter, and $f(\beta|Y, X_0, X_T, X_1)$ is L-2 loss function. When the model is the ordinary linear model, the L-2 loss function is $\| Y - (X_0\beta_0 + X_T\beta_\tau + X_1\beta_1 + X_T \otimes X_1\beta_2) \|^2$. Penalty function $g(\beta)$ can construct a complex hierarchical selection of β_1 and β_2, that nonzero β_2 is a sufficient but not necessary condition for nonzero β_1. The contour plot for a pair of β_1 and β_2 is shown in Figure 1. λ_1 and λ_2 are regularization parameters.

Criterion and Adaptive Weights

KKT conditions

KKT (*19*)

For group $\hat{\beta^{(l)}}$, the KKT condition is

$$X^{(l)^T}(Y - X\hat{\beta}) = \lambda_1\phi_l \begin{bmatrix}0\\ v\end{bmatrix} + \lambda_1\psi_l u + \frac{1}{2}\lambda_2\hat{\beta}^{(l)} \tag{1}$$

where

$$v = \begin{cases}\text{sign}(\hat{\beta}_2^{(l)}) & \text{if } \hat{\beta_2}^{(l)} \neq 0\\ \in \{v : |v|_1 \leq 1\} & \text{if } \hat{\beta_2}^{(l)} = 0\end{cases} \tag{2}$$

$$u = \begin{cases}\hat{\beta}^{(l)}/ \parallel \hat{\beta}^{(l)} \parallel_2 & \text{if } \hat{\beta}^{(l)} \neq 0\\ \in \{u :\parallel u \parallel_2 \leq 1\} & \text{if } \hat{\beta}^{(l)} = 0\end{cases} \tag{3}$$

- $\hat{\beta}^{(l)} = 0$ if

  $$S(X_1^{(l)^T} r_{(-l)}, 0)^2 + S(X_2^{(l)^T} r_{(-l)}, \lambda_1\phi_l)^2 \leq \lambda_1^2\psi_l^2 \tag{4}$$

  where

  $$S(z, a) = \text{sign}(z)(|z| - a)_+ \tag{5}$$

  and

  $$r_{(-l)} = Y - X^{(-l)^T}\hat{\beta}^{(-l)} \tag{6}$$

We can find

$$u = \begin{bmatrix} \frac{X_1^{(l)} r_{(-l)}}{\lambda_1 \psi_l} \\ \frac{S({X_2^{(l)}}^T r_{(-l)}, \lambda_1 \phi_l)}{\lambda_1 \psi_l} \end{bmatrix} \tag{7}$$

$$v = \begin{bmatrix} 0 \\ \frac{{X^{(l)}}^T r_{(-l)} - S({X_2^{(-l)}}^T r_{(-l)}, \lambda_1 \phi_l)}{\lambda_1 \psi_l} \end{bmatrix} \tag{8}$$

Thus, we have $\parallel u \parallel_2 \leq 1$ and $|v|_\infty \leq 1$, so that subgradient equation 1 was satisfied with $\hat{\beta}^{(l)} = 0$

- if $\hat{\beta}^{(l)} \neq 0$ but $\hat{\beta}_2^{(l)} = 0$, we need to satisfy

  $${X^{(l)}}^T (Y - {X^{(-I(l))}}^T \hat{\beta}^{(-I(l))}) = \lambda_1 \phi_l \begin{bmatrix} 0 \\ v \end{bmatrix} + \lambda_1 \psi_l \frac{\hat{\beta}^{(l)}}{\parallel \hat{\beta}^{(l)} \parallel_2} + \frac{1}{2} \lambda_2 \hat{\beta}^{(l)} \tag{9}$$

  we have $\hat{\beta}_2^{(l)} = 0$ if

  $${X_2^{(l)}}^T r_{(-I(l))} \leq \lambda_1 \phi_l \tag{10}$$

  and $r_{(-I(l))} = Y - {X^{(-I(l))}}^T \hat{\beta}^{(-I(l))}$ and $I(l)$ is the interaction effect nested in biomarker lth group.

- $\hat{\beta}^{(l)} \neq 0$ as well as $\hat{\beta}_2^{(l)} \neq 0$, we have

  $${X^{(l)}}^T (Y - {X^{(-I(l))}}^T \hat{\beta}^{(-I(l))}) = \lambda_1 \phi_l \begin{bmatrix} 0 \\ \text{sign}(\hat{\beta}_2^{(l)}) \end{bmatrix} + \lambda_1 \psi_l \frac{\hat{\beta}^{(l)}}{\parallel \hat{\beta}^{(l)} \parallel_2} + \frac{1}{2} \lambda_2 \hat{\beta}^{(l)} \tag{11}$$

  Then we get

  $$\begin{aligned} \hat{\beta}_2^{(l)} &= \frac{{X_2^{(l)}}^T (Y - X^T \hat{\beta}^{(-I(l))}) - \lambda_1 \phi_l \text{sign}(\hat{\beta}_2^{(l)})}{{X_2^{(l)}}^T X_2^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2} \\ &= \frac{S({X_2^{(l)}}^T r_{(-I(l))}, \lambda_1 \phi_l)}{{X_2^{(l)}}^T X_2^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2} \lambda_2} \end{aligned} \tag{12}$$

$$\hat{\beta}_1^{(l)} = \frac{{X_2^{(l)}}^T r_{(-G(l))}}{{X_1^{(l)}}^T X_1^{(l)} + \frac{\lambda_1 \psi_l}{\|\hat{\beta}^{(l)}\|_2} + \frac{1}{2}\lambda_2} \tag{13}$$

$$\hat{\beta}_0 = \frac{X_0^T r_{(-0)}}{X_0^T X_0} \tag{14}$$

$$\hat{\beta}_\tau = \frac{X_T^T r_{(-T)}}{X_T^T X_T} \tag{15}$$

Adaptive Weights

To give each biomarker equal probability to be prognostic and predictive, we define adaptive weights via a null model that the residual $\epsilon = r_{(-I(l))}$ is a normal random error where $\epsilon \sim N(0, \sigma^2)$ when $\beta_2^{(l)} = 0$. Let

$$\begin{aligned} \| {X_2^{(l)}}^T r_{(-I(l))} \|_2 &= \lambda_1 \phi_l \\ E[({X_2^{(l)}}^T r_{(-I(l))})^2] &= \lambda_1^2 \phi_l^2 \end{aligned} \tag{16}$$

Thus, we can get $\lambda_1^2 \phi_l^2 = \text{Var}({X_2^{(l)}}^T r_{(-I(l))})$ and

$$\phi_l \propto \| X_2^{(l)} \|_2 \tag{17}$$

Since λ_1 is regularization parameter, we define $\phi_l = \| X_2^{(l)} \|_2$ without loss generality.

On the other hand, based on inequality 4 and results from formula 16, we let

$$\mathbb{E}[S({X_1^{(l)}}^T r_{(-l)}, 0)^2 + S({X_2^{(l)}}^T r_{(-l)}, \lambda_1 \phi_l)^2] = \lambda_1^2 \psi_l^2 \tag{18}$$

and assume $r_{(-l)} = \epsilon \sim N(0, \sigma^2)$ if $\beta^{(l)} = 0$, thus $\epsilon_1 = {X_2^{(l)}}^T r_{(-l)} \sim N(0, \lambda_1^2\phi_l^2)$ and $\epsilon_0 = \epsilon_1/\lambda_1\phi_l \sim N(0,1)$.

$$\begin{aligned}\mathbb{E}[S({X_2^{(l)}}^T r_{(-l)}, \lambda_1\phi_l)^2] &= \mathbb{E}[\| \epsilon_1 \|_2^2 \ \mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}] - 2\lambda_1\phi_l\mathbb{E}[|\epsilon_1|\mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}] \\ &\quad + \lambda_1^2\phi_l^2\mathbb{E}[\mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}]\end{aligned} \tag{19}$$

$$\begin{aligned}\mathbb{E}[\| \epsilon_1 \|_2^2 \ \mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}] &= \mathbb{E}[\| \epsilon_1 \|_2^2 \ (1 - \mathbb{1}_{|\epsilon_1|\leq\lambda_1\phi_l})] \\ &= \lambda_1\phi_l^2(1 - \mathbb{E}[\| \epsilon_0 \|_2^2]\mathbb{1}_{|\epsilon_0|\leq 1}) \\ &\approx \lambda_1\phi_l^2(1 - (0.68 - 2\frac{1}{\sqrt{2\pi}}\exp(-0.5))) \\ &= (0.32 + \sqrt{\frac{2}{\pi}}\exp(-0.5))\lambda_1^2\phi_l^2\end{aligned} \tag{20}$$

$$\begin{aligned}\mathbb{E}[|\epsilon_1|\mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}] &= \lambda_1\phi_l\mathbb{E}[|\epsilon_0|\mathbb{1}_{|\epsilon_0|>1}] \\ &= \lambda_1\phi_l(\mathbb{E}|\epsilon_0| - \mathbb{E}|\epsilon_0|\mathbb{1}_{|\epsilon_0|\leq 1}) \\ &= \lambda_1\phi_l(\sqrt{\frac{2}{\pi}} - \sqrt{\frac{2}{\pi}}(1 - \exp(-0.5))) \\ &= \sqrt{\frac{2}{\pi}}\exp(-0.5)\lambda_1\phi_l\end{aligned} \tag{21}$$

$$\mathbb{E}[\mathbb{1}_{|\epsilon_1|>\lambda_1\phi_l}] = \mathbb{P}(|\epsilon_0| > 1) \approx 0.32 \tag{22}$$

Take equations 20 - 22 to equation 19, we can get

$$\mathbb{E}[S({X_2^{(l)}}^2 r_{(-l)}, \lambda_1\phi_l)^2] \approx (0.64 - \sqrt{\frac{2}{\pi}}\exp(-0.5))\lambda_1^2\phi_l^2 \tag{23}$$

Insert results of equation (23) into (18), we define

$$\psi_l = \sqrt{\parallel X_1^{(l)} \parallel_2 + \{0.64 - \sqrt{\frac{2}{\pi}}\exp(-0.5)\} \parallel X_2^{(l)} \parallel_2} \tag{24}$$

such that

$$\lambda_1^2 \psi_l^2 = \lambda_1^2 [\parallel X_1^{(l)} \parallel_2 + \{0.64 - \sqrt{\frac{2}{\pi}}\exp(-0.5)\} \parallel X_2^{(l)} \parallel_2] \tag{25}$$

Algorithms

Fast iterative shrinkage-thresholding algorithm with backtracking (*20*)

Proximal operator for group lasso (*21*)

Adaptive restart for rippling behavior (*22*)

Adaptive stepwise of cyclic Barzilai-Borwein spectral approach (*23*)

```
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

Experiments

Signal to noise ratio: $SNR = \frac{Var(X\beta)}{Var(\epsilon)}$

Future Steps

check KKT condition

References and Notes

1. Ming Yuan and Yi Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.

2. Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.

3. Hui Zou and Hao Helen Zhang. On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics*, 37(4):1733, 2009.

4. Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, 24(3):627–654, 2015.

5. Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert. Group lasso with overlap and graph lasso. In *Proceedings of the 26th annual international conference on machine learning*, pages 433–440. ACM, 2009.

6. Daniel Percival et al. Theoretical properties of the overlapping groups lasso. *Electronic Journal of Statistics*, 6:269–288, 2012.

7. Guillaume Obozinski, Laurent Jacob, and Jean-Philippe Vert. Group lasso with overlaps: the latent group lasso approach. *arXiv preprint arXiv:1110.0413*, 2011.

8. Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A note on the group lasso and a sparse group lasso. *arXiv preprint arXiv:1001.0736*, 2010.

9. Noah Simon, Jerome Friedman, Trevor Hastie, and Robert Tibshirani. A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245, 2013.

10. Peng Zhao, Guilherme Rocha, Bin Yu, et al. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, 37(6A):3468–3497, 2009.

11. Lukas Meier, Sara Van De Geer, and Peter Bühlmann. The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(1):53–71, 2008.

12. Wei-Yin Loh, Xu He, and Michael Man. A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833, 2015.

13. Wei-Yin Loh. Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386, 2002.

14. Jianqing Fan and Jinchi Lv. Sure independence screening for ultrahigh dimensional feature space. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70(5):849–911, 2008.

15. Jianqing Fan, Richard Samworth, and Yichao Wu. Ultrahigh dimensional feature selection: beyond the linear model. *Journal of machine learning research*, 10(Sep):2013–2038, 2009.

16. Bo Jiang and Jun S Liu. Sliced inverse regression with variable selection and interaction detection. *arXiv preprint arXiv:1304.4056*, 652, 2013.

17. Yang Li and Jun S Liu. Robust variable and interaction selection for logistic regression and general index models. *Journal of the American Statistical Association*, pages 1–16, 2018.

18. Alan J Miller. Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, pages 389–425, 1984.

19. Ryan J Tibshirani et al. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

20. Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

21. Jun Liu and Jieping Ye. Fast overlapping group lasso. *arXiv preprint arXiv:1009.0306*, 2010.

22. Brendan O'donoghue and Emmanuel Candes. Adaptive restart for accelerated gradient schemes. *Foundations of computational mathematics*, 15(3):715–732, 2015.

23. Stephen J Wright, Robert D Nowak, and Mário AT Figueiredo. Sparse reconstruction by separable approximation. *IEEE Transactions on Signal Processing*, 57(7):2479–2493, 2009.