Generalized Group Lasso for Patient Subgroup Identification

Wenxuan Deng

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 [567]

Sparse Group Lasso [8] [9]

Structured group lasso [10]

Group lasso for logistic regression [11]

Other variable selection methods:

GUIDE: a regression tree [1213]

SIS: screening [1415]

SIR: [1617]

Stepwise selection: [18]

Methods

Model

Ordinary Linear Model

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

Where X is the baseline variables, W is the treatment variable, G is the high dimensional design matrix of genes, i.e. gene expression levels, SNP and mutations, and $W \otimes G$ is the interaction between genes and treatment. $\theta = (\beta, \tau, \alpha, \gamma)$ is the corresponding coefficients. ϵ is random error.

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

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

$$g(\theta) = \lambda_1 \sum_i \phi_i|\gamma_i| + \lambda_1 \sum_i \psi_i \sqrt{\alpha_i^2 + \gamma_i^2} + \lambda_2(\| \alpha \|_2^2 + \| \gamma \|_2^2)$$

Where $\theta = (\alpha, \tau, \alpha, \gamma)$ is the parameter, and $f(\theta|Y, X, W, G)$ is L-2 loss function. When the model is the ordinary linear model, the L-2 loss function is $\|Y - (X\beta + W\tau + G\alpha + W \otimes G\gamma)\|^2$. Penalty function $g(\theta)$ can construct a complex hierarchical selection of α and γ, that nonzero γ is a sufficient but not necessary condition for nonzero α. The contour plot for a pair of α and γ is shown in Figure 1. λ_1 and λ_2 are regularization parameters.

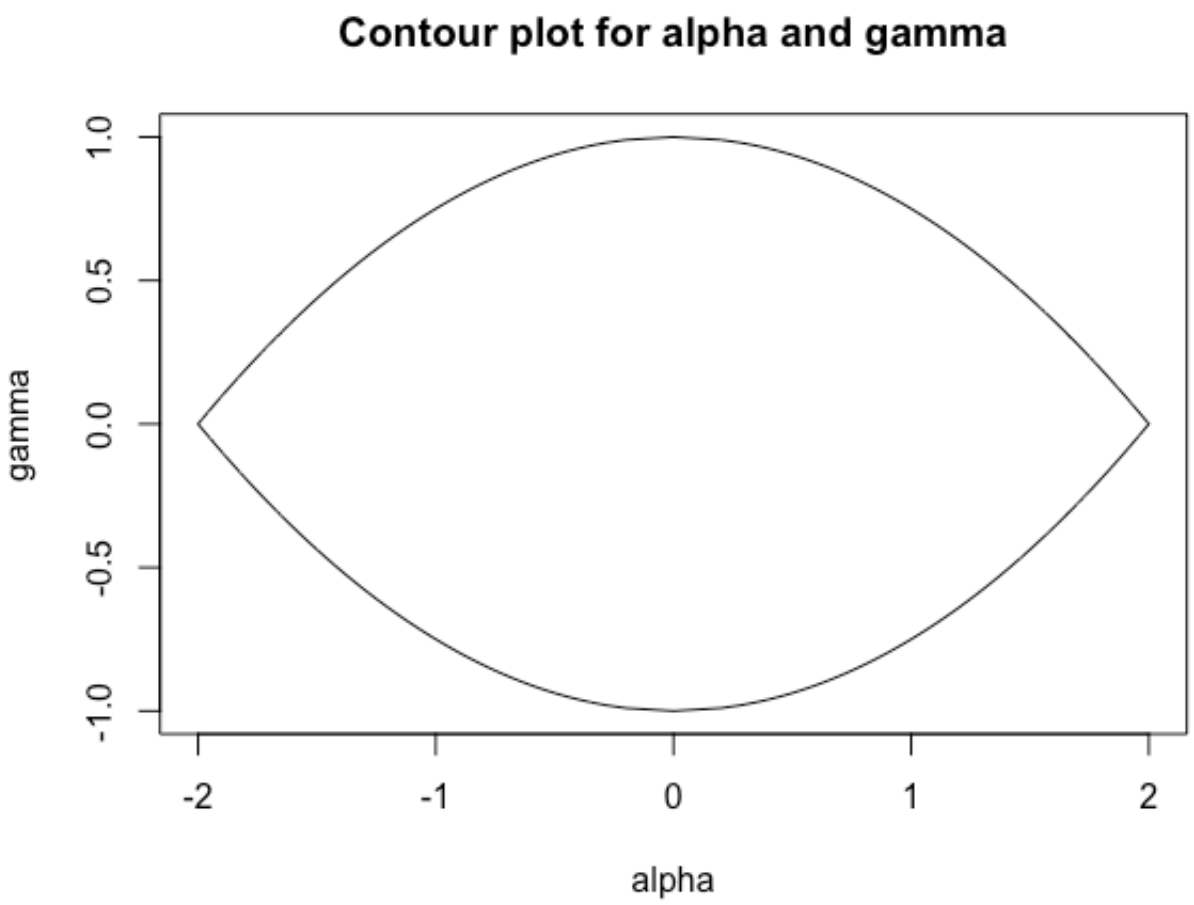

Figure 1

Criterion and Adaptive Weights

KKT [19]

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]

Reference

1. Yuan, Ming, and Yi Lin. "Model selection and estimation in regression with grouped variables." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 68.1 (2006): 49-67. ↩

2. Zou, Hui, and Trevor Hastie. "Regularization and variable selection via the elastic net." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67.2 (2005): 301-320. ↩

3. Zou, Hui, and Hao Helen Zhang. "On the adaptive elastic-net with a diverging number of parameters." Annals of statistics 37.4 (2009): 1733. ↩

4. Lim, Michael, and Trevor Hastie. "Learning interactions via hierarchical group-lasso regularization." Journal of Computational and Graphical Statistics 24.3 (2015): 627-654. ↩

5. Jacob, Laurent, Guillaume Obozinski, and Jean-Philippe Vert. "Group lasso with overlap and graph lasso." Proceedings of the 26th annual international conference on machine learning. ACM, 2009. ↩

6. Obozinski, Guillaume, Laurent Jacob, and Jean-Philippe Vert. "Group lasso with overlaps: the latent group lasso approach." arXiv preprint arXiv:1110.0413 (2011). ↩

7. Percival, Daniel. "Theoretical properties of the overlapping groups lasso." Electronic Journal of Statistics 6 (2012): 269-288. ↩

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

9. Simon, Noah, et al. "A sparse-group lasso." Journal of Computational and Graphical Statistics 22.2 (2013): 231-245. ↩

10. Zhao, Peng, Guilherme Rocha, and Bin Yu. "The composite absolute penalties family for grouped and hierarchical variable selection." The Annals of Statistics 37.6A (2009): 3468-3497. ↩

11. Meier, Lukas, 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 (2008): 53-71.z ↩

12. Loh, Wei-Yin, Xu He, and Michael Man. "A regression tree approach to identifying subgroups with differential treatment effects." Statistics in medicine 34.11 (2015): 1818-1833. ↩

13. Loh, Wei-Yin. "Regression tress with unbiased variable selection and interaction detection." Statistica Sinica (2002): 361-386. ↩

14. Fan, Jianqing, and Jinchi Lv. "Sure independence screening for ultrahigh dimensional feature space." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70.5 (2008): 849-911. ↩

15. Fan, Jianqing, Richard Samworth, and Yichao Wu. "Ultrahigh dimensional feature selection: beyond the

linear model." Journal of machine learning research 10.Sep (2009): 2013-2038. ↩

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

17. Li, Yang, and Jun S. Liu. "Robust variable and interaction selection for logistic regression and general index models." Journal of the American Statistical Association (2018): 1-16. ↩

18. Miller, Alan J. "Selection of subsets of regression variables." Journal of the Royal Statistical Society. Series A (General) (1984): 389-425. ↩

19. Tibshirani, Ryan J. "The lasso problem and uniqueness." Electronic Journal of Statistics 7 (2013): 1456-1490. ↩

20. 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. ↩

21. Liu, Jun, and Jieping Ye. "Fast overlapping group lasso." arXiv preprint arXiv:1009.0306 (2010). ↩

22. O'donoghue, Brendan, and Emmanuel Candes. "Adaptive restart for accelerated gradient schemes." Foundations of computational mathematics 15.3 (2015): 715-732. ↩

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