Variable Selection In Additive Gene Environment Interactions with the Group Lasso

Sahir Bhatnagar and Yi Yang

January 31, 2017

1 Introduction

We consider a regression model for an outcome variable $\mathbf{Y} = (Y_1, \ldots, Y_n)$ where n is the number of subjects. Let $E = (E_1, \ldots, E_n)$ be a binary or continuous environment vector and $\mathbf{X} = (X_1, \ldots, X_n)^T$ be the $n \times p$ matrix of high-dimensional data where $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{ip}) \in [0, 1]^p$. Consider the regression model with main effects and their interactions with E:

$$Y_i = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_{ij} + \beta_E^* E_i + \sum_{j=1}^{p} \alpha_j^* E_i X_j + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

were $\beta_0^*, \beta_j^*, \beta_E^*, \alpha_j^*$ are the true unknown model parameters for $j = 1, \ldots, p$. This can be extended to the more general additive model:

$$Y_i = \beta_0^* + \sum_{j=1}^{p} f_j^*(X_{ij}) + f_E^*(E_i) + \sum_{j=1}^{p} f_{jE}^*(X_{ij}, E_i) + \varepsilon_i \qquad i = 1, \ldots, n \tag{2}$$

As in (Radchenko and James, 2010), we can express (2) as

$$\mathbf{Y} = \sum_{j=1}^{p} \mathbf{f}_j^* + \mathbf{f}_E^*(E_i) + \sum_{j=1}^{p} \mathbf{f}_{jE}^* + \varepsilon_i \tag{3}$$

where $\mathbf{f}_j^* = \left(f_j^*(X_{1j}), \ldots, f_j^*(X_{nj})\right)^T$, $\mathbf{f}_{jE}^* = \left(f_{jE}^*(X_{1j}, X_{1E}), \ldots, f_j^*(X_{nj}, X_{nE})\right)^T$

Bibliography

Marcel Dettling, Edward Gabrielson, and Giovanni Parmigiani. Searching for differentially expressed gene combinations. *Genome biology*, 6(10):R88, 2005.

Mitsunori Kayano, Ichigaku Takigawa, Motoki Shiga, Koji Tsuda, and Hiroshi Mamitsuka. Ros-det: robust detector of switching mechanisms in gene expression. *Nucleic acids research*, 39(11):e74–e74, 2011.

Robert Tibshirani, Guenther Walther, and Trevor Hastie. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

Peter Langfelder, Bin Zhang, and Steve Horvath. Defining clusters from a hierarchical cluster tree: the dynamic tree cut package for r. *Bioinformatics*, 24(5):719–720, 2008.

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

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.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, pages 894–942, 2010.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Peter Bühlmann, Philipp Rütimann, Sara van de Geer, and Cun-Hui Zhang. Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143(11):1835–1858, 2013.

Jerome Friedman, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*, volume 1. Springer series in statistics Springer, Berlin, 2001.

William M Rand. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336):846–850, 1971.

Alexandros Kalousis, Julien Prados, and Melanie Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1): 95–116, 2007.

Karl Pearson. Note on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London*, pages 240–242, 1895.

Charles Spearman. The proof and measurement of association between two things. *The American journal of psychology*, 15(1):72–101, 1904.

Paul Jaccard. The distribution of the flora in the alpine zone. *New phytologist*, 11(2):37–50, 1912.

Qing Mai and Hui Zou. The kolmogorov filter for variable screening in high-dimensional binary classification. *Biometrika*, page ass062, 2012.

Qing Mai and Hui Zou. The fused kolmogorov filter: A nonparametric model-free screening method. *The Annals of Statistics*, 43(4):1471–1497, 2015.

Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996.

Yiyuan She and He Jiang. Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*, 2014.

Jacob Bien, Jonathan Taylor, Robert Tibshirani, et al. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013.

Nam Hee Choi, William Li, and Ji Zhu. Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, 105(489):354–364, 2010.

Peng Zhao, Guilherme Rocha, and Bin Yu. The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, pages 3468–3497, 2009.

Peter Radchenko and Gareth M James. Variable selection using adaptive nonlinear interaction structures in high dimensions. *Journal of the American Statistical Association*, 105 (492):1541–1553, 2010. 1

Michael Lim and Trevor Hastie. Learning interactions via hierarchical group-lasso regularization. *Journal of Computational and Graphical Statistics*, (just-accepted):00–00, 2014.

Asad Haris, Daniela Witten, and Noah Simon. Convex modeling of interactions with strong heredity. *arXiv preprint arXiv:1410.3517*, 2014.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1, 2010.