Variable Selection In Additive Gene Environment Interactions with the Group Lasso

SRB

May 24, 2017

1 Background

We consider a regression model for an outcome variable $Y = (Y_1, \ldots, Y_n)$ where n is the number of subjects. Let $E = (E_1, \ldots, E_n)$ be a binary or continuous environment vector and $\boldsymbol{X} = (X_1, \ldots, X_n)^T$ be the $n \times p$ matrix of high-dimensional data where $X_i = (X_{i1}, \ldots, X_{ij}, \ldots, X_{ip})$, and $\boldsymbol{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$ a vector of errors. Consider the regression model with main effects and their interactions with E:

$$g(\boldsymbol{\mu}) = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_j + \beta_E^* E + \sum_{j=1}^{p} \alpha_j^* E X_j + \boldsymbol{\varepsilon}, \tag{1}$$

where $g(\cdot)$ is a known link function, $\boldsymbol{\mu} = \mathsf{E}\left[Y | \boldsymbol{X}, E, \boldsymbol{\beta}, \boldsymbol{\alpha}\right]$, and $\beta_0^*, \beta_j^*, \beta_E^*, \alpha_j^*$ are the true unknown model parameters for $j = 1, \ldots, p$.

Due to the large number of parameters to estimate with respect to the number of observations, one commonly-used approach is to shrink the regression coefficients by placing a constraint on the values of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. For example, the LASSO (Tibshirani, 1996) penalizes the squared loss of the data with the L_1-norm of the regression coefficients resulting in a method that performs both model selection and estimation. A natural extension of the LASSO to the interaction model (1) is given by:

$$\arg\min_{\beta_0, \boldsymbol{\beta}, \boldsymbol{\alpha}} \frac{1}{2} \left\| Y - g(\boldsymbol{\mu}) \right\|^2 + \lambda \left(\left\| \boldsymbol{\beta} \right\|_1 + \left\| \boldsymbol{\alpha} \right\|_1 \right) \tag{2}$$

where $\left\| Y - g(\boldsymbol{\mu}) \right\|^2 = \sum_i (y_i - g(\mu_i))^2$, $\left\| \boldsymbol{\beta} \right\|_1 = \sum_j |\beta_j|$, $\left\| \boldsymbol{\alpha} \right\|_1 = \sum_j |\alpha_j|$ and $\lambda \geq 0$ is a data driven tuning parameter that can set some of the coefficients to zero when sufficiently large.

However, since no constraint is placed on the structure of the model in 2, it is possible that the estimated main effects are zero while the interaction term is not. This has motivated methods that produce structured sparsity (Bach et al., 2012). For example, Bien *et al.* (Bien et al., 2013) propose a strong hierarchical lasso which forces the main effects to be included if the interaction term is non-zero. However this method and related ones are restricted to all pairwise interactions between p measured variables. Here were concern ouself with methods that impose a strong hierarchy in the context of gene environment interactions. We are interested in imposing the strong heredity principle (Chipman, 1996):

$$\hat{\alpha}_j \neq 0 \qquad \Rightarrow \qquad \hat{\beta}_j \neq 0 \qquad \text{and} \qquad \hat{\beta}_E \neq 0 \tag{3}$$

In words, the interaction term will only have a non-zero estimate if its corresponding main effects are estimated to be non-zero. One benefit brought by hierarchy is that the number of measured variables can be reduced, referred to as practical sparsity (Bien et al., 2013; She and Jiang, 2014). For example, a model involving $X_1, E, X_1 \cdot E$ is more parsimonious than a model involving $X_1, E, X_2 \cdot E$, because in the first model a researcher would only have to measure two variables compared to three in the second model. In order to address these issues, we propose to extend the model of (Choi et al., 2010) to simultaneously perform variable selection, estimation and impose the strong heredity principle in the context of high dimensional interactions with the environment (HD×E). To do so, we follow Choi and reparametrize the coefficients for the interaction terms as $\alpha_j = \gamma_j \beta_j \beta_E$. Plugging this into (1):

$$g(\boldsymbol{\mu}) = \beta_0^* + \sum_{j=1}^{p} \beta_j^* X_j + \beta_E^* E + \sum_{j=1}^{p} \gamma_j \beta_j \beta_E E X_j \tag{4}$$

This reparametrization directly enforces the strong heredity principle (3), i.e., if either main effect estimates are 0, then $\hat{\alpha}_j$ will be zero and a non-zero interaction coefficient implies non-zero $\hat{\beta}_j$ and $\hat{\beta}_E$. To perform variable selection in this new parametrization, we follow Choi et al. (2010) and penalize $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$ instead of penalizing $\boldsymbol{\alpha}$ as in (2), leading to the following penalized least squares criterion:

$$\underset{\beta_0, \boldsymbol{\beta}, \boldsymbol{\gamma}}{\arg\min} \frac{1}{2} \| Y - g(\boldsymbol{\mu}) \|^2 + \lambda_\beta \sum_{j=1}^{p} w_j |\beta_j| + \lambda_\gamma \sum_{j=1}^{p} w_{jE} |\gamma_{jE}| \tag{5}$$

where $g(\boldsymbol{\mu})$ is from (4), λ_β and λ_γ are tuning parameters and $\mathbf{w} = (w_1, \ldots, w_q, w_{1E}, \ldots, w_{qE})$ are prespecified adaptive weights. The λ_β tuning parameter controls the amount of shrinkage applied to the main effects, while λ_γ controls the interaction estimates and allows for the possibility of excluding the interaction term from the model even if the corresponding main effects are non-zero. The adaptive weights serve as a way of allowing parameters to be penalized differently. Furthermore, adaptive weighting (Zou, 2006) has been shown to construct oracle procedures (Fan and Li, 2001), i.e., asymptotically, it performs as well as if the true model were given in advance. The oracle property is achieved when the weights are a function of any root-n consistent estimator of the true parameters e.g. maximum likelihood (MLE) or ridge regression estimates. It can be shown that the procedure in (5) asymptotically possesses the oracle property (Choi et al., 2010), even when the number of parameters tends to ∞ as the sample size increases, if the weights are chosen such that

$$w_j = \left| \frac{1}{\hat{\beta}_j} \right|, \quad w_{jE} = \left| \frac{\hat{\beta}_j \hat{\beta}_E}{\hat{\alpha}_{jE}} \right| \qquad \text{for } j = 1, \ldots, q \tag{6}$$

where $\hat{\beta}_j$ and $\hat{\alpha}_j$ are the MLEs, from (1) or the ridge regression estimates when $p > n$. The rationale behind the data-dependent $\hat{\boldsymbol{w}}$ is that as the sample size grows, the weights for the truly zero predictors go to ∞ (which translates to a large penalty), whereas the weights for the truly non-zero predictors converge to a finite constant (Zou, 2006).

This can be extended to the more general additive model:

$$Y_i = \beta_0^* + \sum_{j=1}^{p} f_j^*(X_{ij}) + f_E^*(E_i) + \sum_{j=1}^{p} f_{jE}^*(X_{ij}, E_i) + \varepsilon_i \qquad i = 1, \ldots, n \tag{7}$$

As in (Radchenko and James, 2010), we can express (7) as

$$\mathbf{Y} = \sum_{j=1}^{p} \mathbf{f}_j^* + \mathbf{f}_E^* + \sum_{j=1}^{p} \mathbf{f}_{jE}^* + \boldsymbol{\varepsilon} \tag{8}$$

where $\mathbf{f}_j^* = \left(f_j^*(X_{1j}), \ldots, f_j^*(X_{nj})\right)^T$, $\mathbf{f}_{jE}^* = \left(f_{jE}^*(X_{1j}, X_{1E}), \ldots, f_j^*(X_{nj}, X_{nE})\right)^T$ and $\mathbf{f}_E^* = f_E^*(E_i)$. We consider the candidate vectors $\{\mathbf{f}_j, \mathbf{f}_E, \mathbf{f}_{jE}\}$. The general approach for fitting (8) is to minimize the following penalized regression criterion:

$$\frac{1}{2}||\mathbf{Y} - \mathbf{f}||^2 + P(\mathbf{f}) \tag{9}$$

where

$$\mathbf{f} = \sum_{j=1}^{p} \mathbf{f}_j + \mathbf{f}_E + \sum_{j=1}^{p} \mathbf{f}_{jE} \tag{10}$$

and $P(\mathbf{f})$ is a penalty function on $\mathbf{f}$

The smoothing method for variable X_j is a projection on to a set of basis functions. Consider

$$f_j(\cdot) = \sum_{\ell=1}^{p_j} \psi_{j\ell}(\cdot)\beta_{j\ell} \tag{11}$$

where the $\{\psi_{j\ell}\}_1^{p_j}$ are a family of basis functions in X_j (Hastie et al., 2015). Let

$$f_{jE}(X_j, E) = \sum_{\ell=1}^{q_j} \phi_{j\ell}(X_j, E)\alpha_{j\ell} \tag{12}$$

where the $\{\phi_{j\ell}\}_1^{q_j}$ are a family of basis functions in $X_j \cdot E$.

Following (Choi et al., 2010), we reparametrize the coefficients for the interaction terms as $\alpha_{j\ell} = \gamma_{j\ell}\beta_{j\ell}\beta_E$. Plugging this into (12):

$$f_{jE}(X_j, E) = \sum_{\ell=1}^{q_j} \phi_{j\ell}(X_j, E)\gamma_{j\ell}\beta_{j\ell}\beta_E \tag{13}$$

References

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 267–288, 1996. 1

Francis Bach, Rodolphe Jenatton, Julien Mairal, Guillaume Obozinski, et al. Structured sparsity through convex optimization. *Statistical Science*, 27(4):450–468, 2012. 1

Jacob Bien, Jonathan Taylor, Robert Tibshirani, et al. A lasso for hierarchical interactions. *The Annals of Statistics*, 41(3):1111–1141, 2013. 1, 2

Hugh Chipman. Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 24(1):17–36, 1996. 2

Yiyuan She and He Jiang. Group regularized estimation under structural hierarchy. *arXiv preprint arXiv:1411.4691*, 2014. 2

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. 2, 3

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006. 2, 3

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. 2

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. 3

Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015. 3