A variable selection approach for highly correlated predictors in high-dimensional genomic data

Wencan Zhu, Céline Lévy-Leduc, Nils Ternès

Bioinformatics, 22 February 2021

Data

- biomarkers associated with a variable of interest
- helpful with the prognosis of clinical endpoint for individual patient
- understanding a disease at a molecular level

Motivation

- the number of biomarkers p is generally much larger than the sample size n
- the relationships between biomarkers should be taken into account

Lasso approach

We consider linear regression model:

$$\mathrm{y} = \mathrm{X}\beta + \epsilon$$

where X is the design matrix containing the expression of biomarkers such that the correlation matrix of its columns is Σ. To estimate the sparse vector β we minimize the penalized least-squares criterion

$$L_\lambda(\beta) = ||y - X\beta||_2^2 + \lambda||\beta||_1$$

where

$$||\beta||_1 = \sum_{k=1}^{p} |\beta_k|.$$

Irrepresentable condition (IC)

Let $S = \{j, \beta_j \neq 0\}$ be the set of active variables, S^c the set of non-active variables and X_S the submatrix of X containing those columns with indices from S. Then, for some constant $\nu \in (0, 1]$,

$$|(X_{S^c}^T X_S (X_S^T X_S)^{-1} \text{sign}(\beta_S))_j| \leq 1 - \nu.$$

Whitening Lasso (WLasso)

Model Transformation

Using the eigendecomposition of matrix $\Sigma = UDU^T$ we define $\tilde{X} = X\Sigma^{-1/2}$ and $\tilde{\beta} = \Sigma^{1/2}\beta$ where $\Sigma^{\pm 1/2} = UD^{\pm 1/2}U^T$. Thus, $\tilde{X}\tilde{\beta} = X\beta$ and we rewrite the considered model as follows

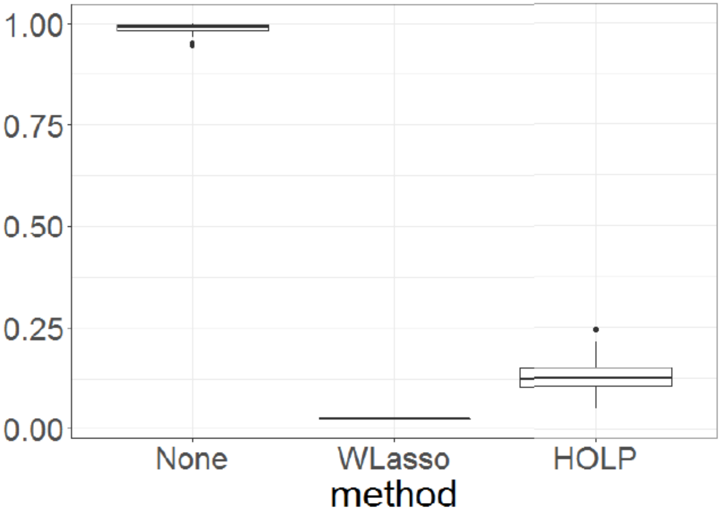
$$y = \tilde{X}\tilde{\beta} + \epsilon.$$

Block structure of Σ

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}$$

- Σ_{11} - correlation matrix of active variables
- Σ_{12} - correlation matrix between active and non active variables
- Σ_{22} - correlation matrix of non active variables

Irrepresentable condition (IC)



Estimation of $\tilde{\beta}$

The following function of $\tilde{\beta}$ is minimized

$$L_{\lambda}^{gen}(\tilde{\beta}) = ||y - \tilde{X}\tilde{\beta}||_2^2 + \lambda||\Sigma^{-1/2}\tilde{\beta}||_1$$

based on Tibshirani and Taylor (2011) who provided the solution for specific linear transformations of β.

Thresholding strategy

$$\hat{\tilde{\beta}}_j^K(\lambda) = \begin{cases} \hat{\tilde{\beta}}_{0j}(\lambda), & j \in Top_K \\ \text{Kth largest value of } |\hat{\tilde{\beta}}_{0j}(\lambda)| & j \notin Top_K \end{cases}$$

$$\hat{\beta}_j^M(\lambda) = \begin{cases} \hat{\beta}_{0j}(\lambda), & j \in Top_M \\ 0 & j \notin Top_M \end{cases}$$

- chose K and M by minimizing MSE

Summary

1. Estimation of the matrix Σ by $\hat{\Sigma}$
2. Transformation of Model to remove correlation
3. Estimation of $\tilde{\beta}$
4. Estimation of β

Summary

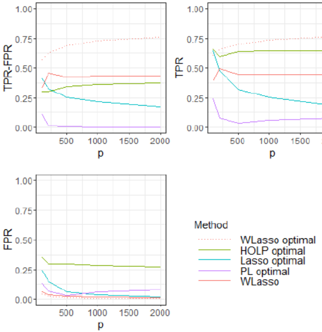


FIGURE 7. Top left: max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL) and (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (solid line). Results obtained for the optimal choice of λ for WLasso (dotted line). Corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 50$.

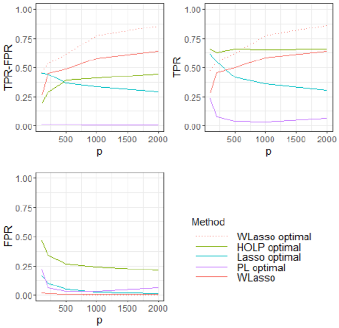


FIGURE 8. Top left: max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL) and (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (solid line). Results obtained for the optimal choice of λ for WLasso (dotted line). Corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$, $b = 0.5$ and $n = 50$.

Summary

- handling correlation in case of the specific structure of Σ
- fully data-driven method
- overall performance is good