PPLasso package

Wencan Zhu, Céline Lévy-Leduc, Nils Ternès

Introduction

This package provides functions for implementing the PPLasso (Prognostic Predictive Lasso) approach described in [1] to identify prognostic and predictive biomarkers in high dimensional settings. This method is designed by taking into account the correlations that may exist between the biomarkers. It consists in rewriting the initial high-dimensional linear model to remove the correlation existing between the predictors and in applying the generalized Lasso criterion. We refer the reader to the paper for further details.

We suppose that the response variable $\boldsymbol{y}$ satisfy the following linear model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1}$$

where

$$\boldsymbol{X} = \begin{bmatrix} 1 & 0 & X_{11}^1 & X_{11}^2 & \dots & X_{11}^p & 0 & 0 & \dots & 0 \\ 1 & 0 & X_{12}^1 & X_{12}^2 & \dots & X_{12}^p & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & & \vdots & & & & \\ 1 & 0 & X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & 0 & X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ 0 & 1 & 0 & 0 & \dots & 0 & X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \vdots & \vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\ 0 & 1 & 0 & 0 & \dots & 0 & X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix},$$

with $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. α_1 (resp. α_2) corresponding to the effects of treatment t_1 (resp. t_2). Moreover, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1p})'$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2p})'$) are the coefficients associated to each of the p biomarkers in treatment t_1 (resp. t_2) group. When t_1 stands for the standard treatment (placebo), prognostic (resp. predictive) biomarkers are defined as those having non-zero coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$) and non prognostic (resp. non predictive) biomarkers correspond to the indices having null coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_1 - \boldsymbol{\beta}_2$). The vector $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ are assumed to be sparse, *i.e.* a majority of its components is equal to zero. The goal of the PPLasso approach is to retrieve the indices of the nonzero components of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.

Concerning the biomarkers,

$$\boldsymbol{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \dots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix} \text{ and } \boldsymbol{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \dots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix} \tag{2}$$

are the design matrices for t_1 and t_2 groups, respectively. The rows of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are assumed to be the realizations of independent centered Gaussian random vectors having a covariance matrix equal to $\boldsymbol{\Sigma}$.

Data generation

Correlation matrix $\boldsymbol{\Sigma}$

We consider a correlation matrix having the following block structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{3}$$

where $\boldsymbol{\Sigma}_{11}$ is the correlation matrix of active variables (non null associated coefficients) with off-diagonal entries equal to a_1, $\boldsymbol{\Sigma}_{22}$ is the one of non active variables (null associated coefficients) with off-diagonal entries equal to a_3 and $\boldsymbol{\Sigma}_{12}$ is the correlation matrix between active and non active variables with entries equal to a_2. In the following example: $(a_1, a_2, a_3) = (0.3, 0.5, 0.7)$.
The first 10 variables are assumed to be active, among which the first 5 are also predictive.

In the following, $p = 50$ and $n = 50$ are used for the example but the approach can handle much larger values of n and p as it is shown in the paper describing PPLasso.

```
p <- 50 # number of variables
d <- 10 # number of actives
n <- 50 # number of samples
actives <- 1:d
nonacts <- c(1:p)[-actives]
Sigma <- matrix(0, p, p)
Sigma[actives, actives] <- 0.3
Sigma[-actives, actives] <- 0.5
Sigma[actives, -actives] <- 0.5
Sigma[-actives, -actives] <- 0.7
diag(Sigma) <- rep(1,p)
actives_pred <- 1:5
```

Generation of $\boldsymbol{X}$ and $\boldsymbol{y}$

The design matrix is then generated with the correlation matrix $\boldsymbol{\Sigma}$ previously defined by using the function `mvrnorm` and the response variable $\boldsymbol{y}$ is generated according to the linear model (1) where the non null components of $\boldsymbol{\beta}_1$ are equal to 1 and non null components of $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ are equal to 0.5, $\alpha_1 = 0$ and $\alpha_2 = 1$.

```
X_bm <- MASS::mvrnorm(n = n, mu=rep(0,p), Sigma, tol = 1e-6, empirical = FALSE)
colnames(X_bm) <- paste0("X",(1:p))
n1=n2=n/2 # 1:1 randomized
beta1 <- rep(0,p)
beta1[actives] <- 1
beta2 <- beta1
beta2[actives_pred] <- 2
beta <- c(beta1, beta2)
TRT1 <- c(rep(1,n1), rep(0,n2))
TRT2 <- c(rep(0,n1), rep(1,n2))
Y <- cbind(X_bm*TRT1,X_bm*TRT2)%*%beta+TRT2+rnorm(n,0,1)
```

Estimation of $\boldsymbol{\Sigma}$

Given $\boldsymbol{y}$ and $\boldsymbol{X}$, we can estimate the block-wise correlation matrix $\boldsymbol{\Sigma}$ containing the correlations between the columns of $\boldsymbol{X}$. We propose to use the function `cvCovEst` of the R package `cvCovEst` by keeping the default parameters.

```
cv_cov_est_out <- cvCovEst(
      dat = X_bm,
      estimators = c(
        linearShrinkLWEst, denseLinearShrinkEst,
        thresholdingEst, poetEst, sampleCovEst
```

```
    ),
    estimator_params = list(
      thresholdingEst = list(gamma = c(0.2, 0.4)),
      poetEst = list(lambda = c(0.1, 0.2), k = c(1L, 2L))
    ),
    cv_loss = cvMatrixFrobeniusLoss,
    cv_scheme = "v_fold",
    v_folds = 5
  )
Sigma_est <- cov2cor(cv_cov_est_out$estimate)
```

The optimal estimation of $\boldsymbol{\Sigma}$ can be obtained by the object `estimate` in the output.

Variable selection

With the previous $\boldsymbol{X}$ and $\boldsymbol{y}$, the function `ProgPredLasso` of the package `PPLasso` can be used to select the active variables. If the parameter `cor_matrix` (correlation matrix) is not provided, it will be automatically estimated by the function `cvCovEst` of the R package `cvCovEst` presented in the previous section. However, it can also be provided by the users. Here we use the previously estimated $\widehat{\boldsymbol{\Sigma}}$: `Sigma_est`.

```
mod <- ProgPredLasso(X1 = X_bm[1:n1, ], X2 = X_bm[(n1+1):n, ], Y = Y, cor_matrix = Sigma_est)
```

Additional arguments:

- `delta`: parameter of thresholding appearing in the method described in [1] which is set to 0.95 by default.
- `maxsteps`: integer specifying the maximum number of steps for the generalized Lasso algorithm. Its default value is 500.

Outputs:

- `lambda`: all the λ considered.
- `beta`: matrix of the estimations of $\boldsymbol{\gamma}$ for all the λ considered. Each row of `beta` corresponds to $\widehat{\gamma}$ for a given λ. More precisely, the first (resp. second) column corresponds to the estimation of treatment effect α_1 (resp. $\alpha 2$). The 3rd to $(p+2)$th columns correspond to the estimation of $\boldsymbol{\beta}_1$ and the last p columns correspond to the estimation of $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.
- `beta.min`: estimation of $\boldsymbol{\gamma}$ obtained for the λ minimizing the BIC criterion.
- `bic`: BIC criterion for all the λ considered.
- `mse`: MSE (Mean Squared Error) for all the λ considered.

Estimation of $\boldsymbol{\gamma}$

The estimation of the treatment effects α_1 and α_2 are obtained as follows:

```
#alpha1
 mod$beta.min[1]
```

```
## [1] -0.096178
```

```
#alpha2
 mod$beta.min[2]
```

```
## [1] 1.110355
```

The identified prognostic (resp. predictive) biomarkers are displayed on the left (resp. right) of Figure 1 with true prognostic or predictive biomarkers in blue and false positives in red.

To find the biomarkers identified as prognostic:

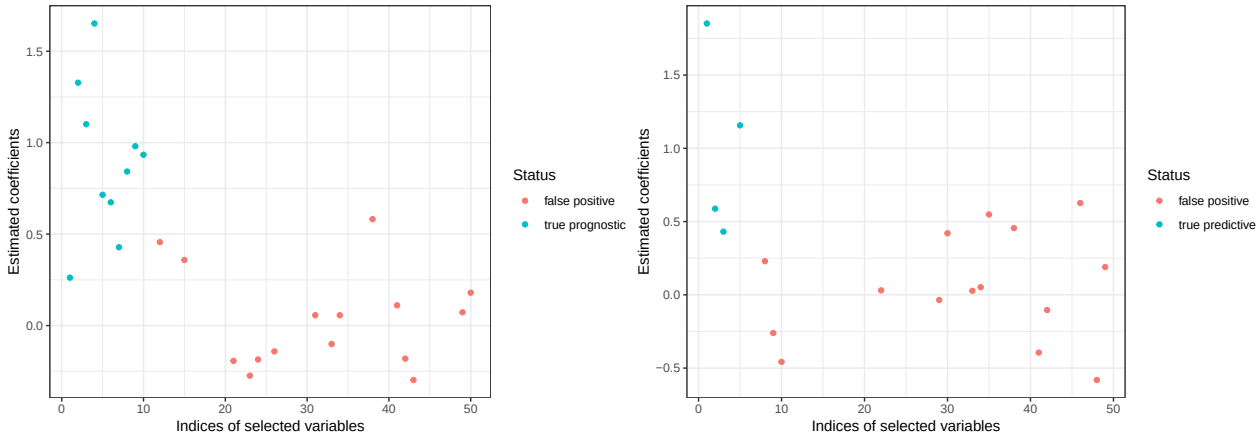

Figure 1: Left: Identified prognostic biomarkers. Right: Identified predictive biomarkers.

```
which(beta_min[1:p]!=0)
```

```
##  [1]  1  2  3  4  5  6  7  8  9 10 12 15 21 23 24 26 31 33 34 38 41 42 43 49 50
```

and the biomarkers identified as predictive:

```
which(beta_min[(p+1):(2*p)]!=0)
```

```
##  [1]  1  2  3  5  8  9 10 22 29 30 33 34 35 38 41 42 46 48 49
```

References

[1] W. Zhu, C. Lévy-Leduc, N. Ternès. Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso, 2022, Arxiv.