WLogit package

Wencan Zhu

Introduction

This package provides functions for implementing the variable selection approach in high-dimensional linear models called WLogit described in Zhu et al. (2022). This method is designed for taking into account the correlations that may exist between the predictors (columns of the design matrix). It consists in rewriting the initial high-dimensional logistic regression 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.

Given a design matrix $\mathbf{X}$ of size $n \times p$, $X_j^{(i)}$ corresponds to the measurement of the jth biomarker on sample i, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is the vector of effect size for each biomarker, with most components equal to zero. We assume that the binary response $y_1, y_2, ..., y_n$ are independent random variables having a Bernoulli distribution with parameter $\pi_{\boldsymbol{\beta}}(X^{(i)})$ ($y_i \sim Bernoulli(\pi_{\boldsymbol{\beta}}(X^{(i)}))$), where for all i in $\{1, \ldots, n\}$,

$$\pi_{\boldsymbol{\beta}}(X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}. \tag{1}$$

The rows of $\boldsymbol{X}$ are assumed to be the realizations of independent centered Gaussian random vectors having a covariance matrix equal to $\boldsymbol{\Sigma}$. The vector $\boldsymbol{\beta}$ is assumed to be sparse, *i.e.* a majority of its components is equal to zero. The goal of the WLoigt approach is to retrieve the indices of the nonzero components of $\boldsymbol{\beta}$, also called active variables.

Data generation

Correlation matrix $\boldsymbol{\Sigma}$

We consider a correlation matrix having the following block structure:

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{2}$$

where $\boldsymbol{\Sigma}_{11}$ is the correlation matrix of active variables with off-diagonal entries equal to α_1, $\boldsymbol{\Sigma}_{22}$ is the one of non active variables with off-diagonal entries equal to α_3 and $\boldsymbol{\Sigma}_{12}$ is the correlation matrix between active and non active variables with entries equal to α_2. In the following example: $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$.

The first 10 variables are active variables among the $p = 500$ variables and $n = 100$.

```
p <- 500 # number of variables
d <- 10 # number of actives
n <- 100 # number of samples
actives <- c(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)
```

Generation of X and y

The design matrix is then generated with the correlation matrix $\mathbf{\Sigma}$ previously defined by using the function `mvrnorm` and the response variable $\boldsymbol{y}$ is generated according to model (1) where the non null components of $\boldsymbol{\beta}$ are equal to 1.

```
X <- MASS::mvrnorm(n = n, mu=rep(0,p), Sigma, tol = 1e-6, empirical = FALSE)
beta <- rep(0,p)
beta[actives] <- 1
pr <- CalculPx(X,beta=beta)
y <- rbinom(n,1,pr)
```

Variable selection

With the previous $\boldsymbol{X}$ and $\boldsymbol{y}$, the function `WhiteningLogit` of the package can be used to select the active variables.

```
mod <- WhiteningLogit(X = X, y = y)
```

Additional arguments:

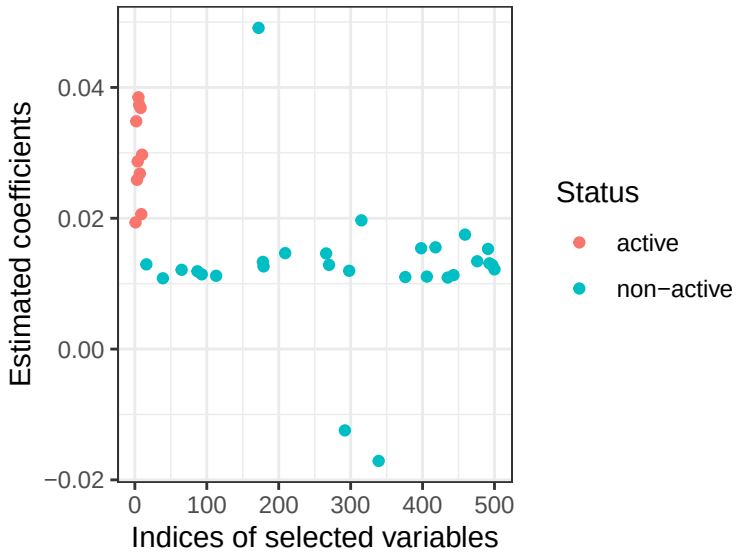
- `nlambda`: number of lambda to be considered, the default value is 50.
- `gamma`: parameter described in the paper.
- `maxit`: integer specifying the maximum number of steps for the iteration in the Iterative Re-weighted Least Square algorithm. Its default value is 100.

Outputs:

- `beta`: matrix of the estimations of $\boldsymbol{\beta}$ for all the λ considered.
- `beta.min`: estimation of $\boldsymbol{\beta}$ which maximize the log-likelihood.
- `log.likelihood`: Log-likelihood for all the λ considered.
- `lambda`: All λ considered.

Estimation of β by $\widehat{\beta}(\lambda)$ which maximizes the log-likelihood

```
beta_min <- mod$beta.min
df_beta <- data.frame(beta_est=beta_min, Status = ifelse(beta==0, "non-active", "active"))
df_plot <- df_beta[which(beta_min!=0), ]
df_plot$index <- which(beta_min!=0)
ggplot2::ggplot(data=df_plot, mapping=aes(y=beta_est, x=index, color=Status))+geom_point()+
  theme_bw()+ylab("Estimated coefficients")+xlab("Indices of selected variables")
```



True Positive Rate: 1 (all active variables identified)

False Positive Rate: $28/490 = 0.0571429$

References

[1] W. Zhu, C. Lévy-Leduc, N. Ternès. Variable selection in high-dimensional logistic regression models using a whitening approach, 2022, Arxiv: 2202.01970.