Bioinformatics, 37(16), 2021, 2238–2244
doi: 10.1093/bioinformatics/btab114
Advance Access Publication Date: 22 February 2021
Original Paper

Genome analysis

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

Wencan Zhu [1,2,]*, Céline Lévy-Leduc[1] and Nils Ternès[2]

[1]UMR MIA-Paris, AgroParisTech, INRAE – Université Paris-Saclay, Paris 75005, France and [2]Biostatistics and Programming Department, Sanofi R&D, Chilly Mazarin 91380, France

*To whom correspondence should be addressed.

Associate Editor: Alfonso Valencia

Received on July 24, 2020; revised on January 25, 2021; editorial decision on February 12, 2021; accepted on February 18, 2021

Abstract

Motivation: In genomic studies, identifying biomarkers associated with a variable of interest is a major concern in biomedical research. Regularized approaches are classically used to perform variable selection in high-dimensional linear models. However, these methods can fail in highly correlated settings.

Results: We propose a novel variable selection approach called WLasso, taking these correlations into account. It consists in rewriting the initial high-dimensional linear model to remove the correlation between the biomarkers (predictors) and in applying the generalized Lasso criterion. The performance of WLasso is assessed using synthetic data in several scenarios and compared with recent alternative approaches. The results show that when the biomarkers are highly correlated, WLasso outperforms the other approaches in sparse high-dimensional frameworks. The method is also illustrated on publicly available gene expression data in breast cancer.

Availabilityand implementation: Our method is implemented in the WLasso R package which is available from the Comprehensive R Archive Network (CRAN).

Contact: wencan.zhu@agroparistech.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

The identification of prognostic genomic biomarkers (i.e. biomarkers associated with a variable of interest, for example a clinical endpoint in clinical trials) has become a major concern for the biomedical research field. Indeed, prognostic biomarkers may help to anticipate the prognosis of individual patients and may also be useful to understand a disease at a molecular level and possibly guide for the development of new treatment strategies (Kalia, 2015).

To this end, statistical variable selection approaches are widely used to identify a subset of biomarkers in high-dimensional settings where the number of biomarkers p is much larger than the sample size n. Several reviews focused on this topic (Heinze *et al.*, 2018; Saeys *et al.*, 2007 for example). Commonly used techniques include hypothesis-based test: t-test (McDonald, 2009), wrapper approaches (Saeys *et al.*, 2007): forward, backward selection and penalized approaches: Lasso (Tibshirani, 1996), SCAD (Fan and Li, 2001) among others. Hypothesis tests are limited to independently consider associations for each biomarker thus neglecting potential relationships between them. Wrapper approaches often show high risk of overfitting and are computationally expensive for high-dimensional data (Smith, 2018). More efforts have been devoted to penalized methods, given the attractive feature of automatically performing variable selection and coefficient estimation simultaneously (Fan and Li, 2006). We shall thus focus on this type of approaches in the following. Let us consider the following linear regression model:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon. \tag{1}$$

where $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the variable to explain (clinical endpoint), $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p)$ is the design matrix containing the expression of biomarkers such that the correlation matrix of its columns is Σ, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_\mathrm{p})^\mathrm{T}$ is a sparse vector to estimate, namely with a majority of null coefficients, and ϵ is the error term. The Lasso penalty is a well-known approach to estimate $\boldsymbol{\beta}$ with a sparsity enforcing constraint. It consists in minimizing the following penalized least-squares criterion (Tibshirani, 1996):

$$L_\lambda(\boldsymbol{\beta}) = \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2}$$

where $\|\cdot\|_2$ is the Euclidean norm and $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$. However, the Lasso has several drawbacks in highly correlated settings (Zou and Hastie, 2005) such as the violation of the Irrepresentable Condition (IC) defined in Zhao and Yu (2006). The authors of this article proved that this condition is necessary and sufficient to recover the support of $\boldsymbol{\beta}$, namely to retrieve the null and non-null components in the vector $\boldsymbol{\beta}$ and thus to provide a sign consistent estimator. This condition is defined as follows. Let $S = \{j, \beta_j \neq 0\}$ be the set of active variables, S^c the set of non-active variables and $\mathbf{X}_A$ the submatrix of $\mathbf{X}$ containing only the indices of columns which are in the set A. Then, the design matrix $\mathbf{X}$ satisfies the Irrepresentable Condition (IC) if, for some constant $\eta \in (0, 1]$,

$$\left|\left(\mathbf{X}_{S^c}^T\mathbf{X}_S(\mathbf{X}_S^T\mathbf{X}_S)^{-1}\text{sign}(\boldsymbol{\beta}_S)\right)_j\right| \leq 1-\eta, \text{for all } j, \quad (3)$$

where sign$(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. Intuitively, this condition means that the correlation between the active and non-active explanatory variables is smaller that the correlation between the active explanatory variables. Hence, this condition is most likely to be violated when the correlations between non-active and active variables are large. In high-dimensional genomic data, this condition is difficult to guarantee as the correlation between biomarkers is usually high (Michalopoulos *et al.*, 2012). Wang *et al.* (2019) tested the Irrepresentable Condition on several publicly available genomic data and highlighted that the condition is violated in almost all the datasets investigated.

To deal with the issue of high correlations between the biomarkers, several strategies have been considered: The Elastic Net introduced by Zou and Hastie (2005) and preconditioning approaches. Elastic Net consists in using a criterion similar to the Lasso except that there is an additional penalty term $\eta\|\boldsymbol{\beta}\|_2^2$ which requires to choose properly the parameter η. The preconditioning approaches consist in transforming the given data $\mathbf{X}$ and $\mathbf{y}$ before applying the Lasso criterion. For example, Jia and Rohe (2015) and Wang and Leng (2016) proposed to left-multiply $\mathbf{X}$, $\mathbf{y}$ and thus ϵ in Model (1) by specific matrices to remove the correlations between the columns of $\mathbf{X}$. A major drawback of the latter, called HOLP (High dimensional Ordinary Least squares Projection), is that the preconditioning step may increase the variance of the error term and thus may alter the variable selection performance. Another recently published method named Precision Lasso (Wang *et al.*, 2019) proposes to handle the correlation issue by assigning similar weights to correlated variables. This approach revealed better performance than the other methods when the biomarkers were highly correlated and the sample size is relatively large. However, it failed in more favorable cases when the biomarkers are not correlated.

In this article, we propose an alternative and novel approach, called Whitening Lasso (WLasso), to take into account the correlations that may exist between the predictors (biomarkers). Our method proposes to transform Model (1) in order to remove the correlations existing between the columns of $\mathbf{X}$ and thus to 'whiten' them and make the IC valid but without changing the error term ϵ. This prevents us from noise inflation, see (4). Then, the variable (biomarker) selection is performed thanks to the generalized Lasso criterion devised by Tibshirani and Taylor (2011). The full details of our method are provided in Section 2. An extensive simulation study is presented in Section 3 to assess the selection performance of our approach and to compare it to other methods in different settings. WLasso is also applied to a publicly available dataset in breast cancer in Section 4. Finally, we discuss our findings and give concluding remarks in Section 5.

2 Materials and methods

In this section, we propose a novel variable selection approach called WLasso (Whitening Lasso) which consists in removing the correlations existing between the biomarkers (columns of $\mathbf{X}$) and in applying the generalized Lasso criterion proposed by Tibshirani and Taylor (2011) for variable selection purpose.

2.1 Model transformation

Inspired by the literature on preconditioning, we propose to rewrite Model (1) in order to remove the correlation existing between the columns of $\mathbf{X}$. More precisely, let $\Sigma^{-1/2} := \boldsymbol{U}\boldsymbol{D}^{-1/2}\boldsymbol{U}^T$ where $\boldsymbol{U}$ and $\boldsymbol{D}$ are the matrices involved in the spectral decomposition of the symmetric matrix Σ given by: $\Sigma = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. We then denote $\widetilde{\mathbf{X}} = \mathbf{X}\Sigma^{-1/2}$. Therefore, (1) can be rewritten as follows:

$$\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}} + \epsilon, \quad (4)$$

where $\widetilde{\boldsymbol{\beta}} = \Sigma^{1/2}\boldsymbol{\beta} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^{\mathrm{T}}\boldsymbol{\beta}$. With such a transformation, since the n rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\mathbf{X}$ are assumed to be independent Gaussian random vectors with a covariance matrix equal to Σ, the covariance matrix of the rows of $\widetilde{\mathbf{X}}$ is equal to identity and the columns of $\widetilde{\mathbf{X}}$ are thus uncorrelated. The advantage of such a transformation with respect to the preconditioning approach proposed by Wang and Leng (2016) is that the error term ϵ is not modified thus avoiding an increase of the noise which can overwhelm the benefits of a well-conditioned design matrix.

To illustrate the benefits of our methodology, observations $\mathbf{y}$ were generated according to Model (1) with $p = 500$, $n = 50$, $\boldsymbol{\beta}$ having 10 non-null components which are equal to 2 and with Σ defined by

$$\boldsymbol{\Sigma} = [\boldsymbol{\Sigma}_{11}\boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{12}^{\mathrm{T}}\boldsymbol{\Sigma}_{22}] \quad (5)$$

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 case where $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, which is a case where the IC is not satisfied, Figure 1 displays the percentage of components j for which the Irrepresentable Condition (3) is not satisfied from 100 replications. We can see from this figure that our approach (WLasso) dramatically improves the number of indices j for which the IC condition is satisfied. The results are even better than those obtained by the transformation proposed by HOLP (Wang and Leng, 2016).

The following illustrations of Section 2 are obtained from observations $\mathbf{y}$ generated according to the previous scenario.

2.2 Estimation of $\widetilde{\boldsymbol{\beta}}$

In order to estimate $\widetilde{\boldsymbol{\beta}}$ with a sparsity enforcing constraint on $\boldsymbol{\beta}$, we use the generalized Lasso criterion proposed by Tibshirani and Taylor (2011) which consists in minimizing the following criterion with respect to $\boldsymbol{\beta}$:

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\mathbf{D}\boldsymbol{\beta}\|_1,$$

where $\mathbf{D}$ is a specific matrix. Note that this criterion boils down to the classical Lasso criterion if $\mathbf{D}$ is the identity matrix. In Model (4), we thus propose to minimize the following criterion with respect to $\widetilde{\boldsymbol{\beta}}$:

$$L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}}) = \|\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\beta}}\|_2^2 + \lambda\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\|_1, \quad (6)$$

which guarantees a sparsity enforcing constraint on $\boldsymbol{\beta}$ thanks to the ℓ_1 penalty. We thus obtain

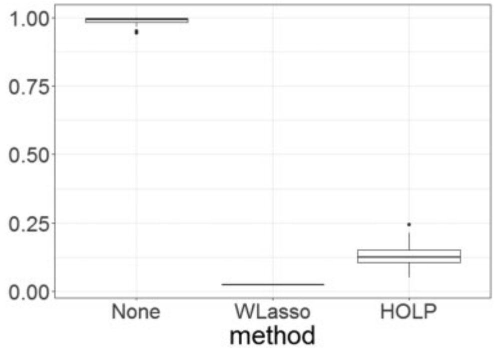

Fig. 1. Proportion of components j such that (3) is violated. These results were obtained from 100 replications

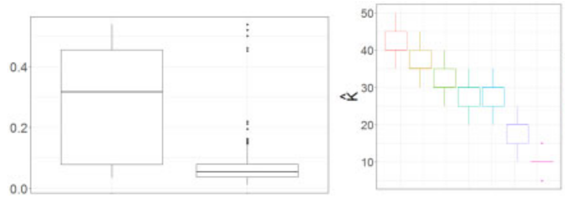

Fig. 2. Left: Boxplots of the average of $(|\hat{\tilde{\boldsymbol{\beta}}}_{0j}(\lambda) - \tilde{\boldsymbol{\beta}}_j(\lambda)|)_{1 \le j \le p}$ (left) and $(|\hat{\tilde{\boldsymbol{\beta}}}_j^{(\hat{K})}(\lambda) - \tilde{\boldsymbol{\beta}}_j(\lambda)|)_{1 \le j \le p}$ (right) for all λ obtained from 100 replications. Right: Boxplots of $\hat{K}(\lambda)$ for different values of λ obtained from 100 replications

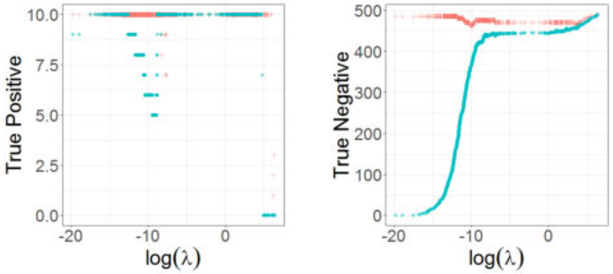

Fig. 3. Number of True Positive and True Negative for $\hat{\boldsymbol{\beta}}$ in red and $\hat{\boldsymbol{\beta}}_0$ in blue for a given vector of observations $\mathbf{y}$

$$\hat{\tilde{\boldsymbol{\beta}}}_0(\lambda) = \text{Argmin}_{\tilde{\boldsymbol{\beta}}} L_\lambda^{\text{gen}}(\tilde{\boldsymbol{\beta}}).$$

To estimate $\tilde{\boldsymbol{\beta}}$, we will not directly use $\hat{\tilde{\boldsymbol{\beta}}}_0(\lambda)$ but the following modified estimator which can be seen as a thresholding of the components of $\hat{\tilde{\boldsymbol{\beta}}}_0(\lambda)$. For K in $\{1, \dots, p\}$, let Top_K be the set of indices corresponding to the K largest values of the components of $|\hat{\tilde{\boldsymbol{\beta}}}_0|$, then the estimator of $\tilde{\boldsymbol{\beta}}$ is $\hat{\tilde{\boldsymbol{\beta}}} = (\hat{\tilde{\boldsymbol{\beta}}}_j^{(\hat{K})})_{1 \le j \le p}$ where $\hat{\tilde{\boldsymbol{\beta}}}_j^{(K)}$ is defined by:

$$\hat{\tilde{\boldsymbol{\beta}}}_j^{(K)}(\lambda) = \begin{cases} \hat{\tilde{\boldsymbol{\beta}}}_{0j}(\lambda), & j \in \text{Top}_K \\ K\text{th largest value of} |\hat{\tilde{\boldsymbol{\beta}}}_{0j}|, & j \notin \text{Top}_K. \end{cases} \tag{7}$$

The choice of $\hat{K} = \hat{K}(\lambda)$ is explained in Section 2.4. Figure 2 displays the average of $(|\hat{\tilde{\boldsymbol{\beta}}}_{0j}(\lambda) - \tilde{\boldsymbol{\beta}}_j(\lambda)|)_{1 \le j \le p}$ and $(|\hat{\tilde{\boldsymbol{\beta}}}_j^{(\hat{K})}(\lambda) - \tilde{\boldsymbol{\beta}}_j(\lambda)|)_{1 \le j \le p}$ for all the values of λ that are considered and boxplots of $\hat{K}(\lambda)$ for some λ. We can see from this figure that the thresholding improves the estimation of $\tilde{\boldsymbol{\beta}}$.

2.3 Estimation of $\boldsymbol{\beta}$

As previously, to estimate $\boldsymbol{\beta}$, we will first consider $\hat{\boldsymbol{\beta}}_0 = \Sigma^{-1/2}\hat{\tilde{\boldsymbol{\beta}}}$ and then apply a thresholding strategy. Thus, we propose to estimate $\boldsymbol{\beta}$ by $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_j^{(\hat{M})})_{1 \le j \le p}$ where $\hat{\boldsymbol{\beta}}_j^{(M)}$ is defined by:

$$\hat{\boldsymbol{\beta}}_j^{(M)}(\lambda) = \begin{cases} \hat{\boldsymbol{\beta}}_{0j}(\lambda), & j \in \text{Top}_M \\ 0, & j \notin \text{Top}_M. \end{cases} \tag{8}$$

The choice of $\hat{M}(\lambda)$ is explained in Section 2.4.

As we can see from Figure 3, more true non-null (active) components of $\boldsymbol{\beta}$ (true positive) and more true null (non-active) components of $\boldsymbol{\beta}$ (true negative) can be retrieved with $\hat{\boldsymbol{\beta}}$ than with $\hat{\boldsymbol{\beta}}_0$. We can thus conclude from Figures 2 and 3 that both thresholdings improve the variable selection.

2.4 Choice of the parameters

To choose the parameters K and M in (7) and (8) for each λ, we use a strategy based on the Mean Squared Error (MSE). We shall first explain the strategy that we used for choosing $\hat{K}$. Let

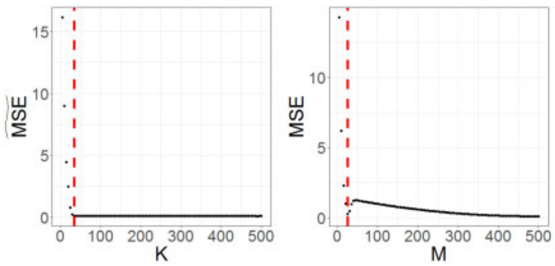

Fig. 4. $\widetilde{\text{MSE}}_K(\lambda)$ (left) and $\text{MSE}_M(\lambda)$ (right) for λ chosen thanks to the strategy explained in Section 3.2 for a given vector of observations $\mathbf{y}$ and $\gamma = 0.95$. The vertical dotted lines correspond to $\hat{K}(\lambda)$ and $\hat{M}(\lambda)$, respectively

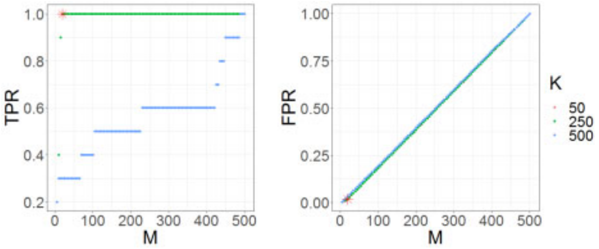

Fig. 5. TPR (left) and FPR (right) for different values of M and K. The TPR and FPR obtained for $M = \hat{M}$ and $K = \hat{K}$ are displayed with a red star ('*'). Note that the red dots are at the same position as the green dots

$$\widetilde{\text{MSE}}_K(\lambda) = \|\mathbf{y} - \tilde{\mathbf{X}}\hat{\tilde{\boldsymbol{\beta}}}^{(K)}(\lambda)\|_2^2,$$

where $\mathbf{y}$, $\tilde{\mathbf{X}}$ and $\hat{\tilde{\boldsymbol{\beta}}}^{(K)}(\lambda)$ are defined in (1, 4) and (7), respectively and

$$\hat{K}(\lambda) = \text{Argmin}\left\{K \ge 1 \text{s.t.} \frac{\widetilde{\text{MSE}}_{K+1}(\lambda)}{\widetilde{\text{MSE}}_K(\lambda)} \ge \gamma\right\}, \text{where} \gamma \in (0, 1).$$

Large values of γ will lead to large values of $\hat{K}(\lambda)$ and thus to a weak thresholding of the estimator of $\tilde{\boldsymbol{\beta}}$. In practice, as it is shown in Section 3, taking γ in (0.9,0.99) provides satisfactory and almost similar results.

For the choice of $\hat{M}(\lambda)$, we use the same procedure except that $\widetilde{\text{MSE}}_K(\lambda)$ is replaced by

$$\text{MSE}_M(\lambda) = \|\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}^{(M)}(\lambda)\|_2^2, \tag{9}$$

where $\mathbf{y}$, $\mathbf{X}$ and $\hat{\boldsymbol{\beta}}^{(M)}(\lambda)$ are defined in (1) and (8), respectively. Both criteria are displayed in Figure 4 for a value of λ which is chosen according to the strategy explained in Section 3.2.

To better understand the impact of the choice of M and K on the True Positive Rate (TPR) and the False Positive Rate (FPR) for this value of λ, Figure 5 displays the TPR and FPR for different values of K and M. We can see from this figure that our choice of M and K, displayed with a red star, guarantees a good trade-off between the TPR and FPR.

2.5 Estimation of $\boldsymbol{\Sigma}$

Since the matrix $\boldsymbol{\Sigma}$ is unknown in practice, it has to be estimated. In the particular situation where $\boldsymbol{\Sigma}$ has the block structure described in (5), we propose the following strategy. Firstly, we compute the empirical correlation matrix as follows. Let $\boldsymbol{S}$ be the sample $p \times p$ covariance matrix defined by

$$\boldsymbol{S} = \frac{1}{n-1}\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})', \quad \text{with} \overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{x}_i,$$

where $\boldsymbol{x}_i$ denotes the ith row of $\boldsymbol{X}$ defined in (1). The corresponding $p \times p$ sample correlation matrix $\boldsymbol{R} = (R_{i,j})$ is defined by:

$$\boldsymbol{R}_{i,j} = \frac{S_{i,j}}{\sigma_i \sigma_j}, \forall 1 \le i,j \le p, \qquad (10)$$

where

$$\sigma_i^2 = \frac{1}{n-1}\sum_{\ell=1}^{n}(X_{\ell,i} - \overline{X}_i)^2, \quad \text{with} \overline{X}_i = \frac{1}{n}\sum_{\ell=1}^{n} X_{\ell,i}, \forall 1 \le i \le p.$$

Secondly, the two groups (or clusters) of active and non-active biomarkers are obtained by using a hierarchical clustering with the complete agglomeration method on the matrix $\boldsymbol{R}$. Thirdly, the entries of $\hat{\boldsymbol{\Sigma}}$ are computed by averaging the values of $\boldsymbol{R}$ within the groups. More precisely, let $\rho_{i,j}$ denote the value of the entries in the block having its rows corresponding to Cluster i and its columns to Cluster j. Then, for a given clustering C:

$$\rho_{i,j} = \begin{cases} \dfrac{1}{\#C(i)\#C(j)} \displaystyle\sum_{k\in C(i), \ell \in C(j)} R_{k,\ell}, & \text{if } C(i) \ne C(j) \\ \dfrac{1}{\#C(i)(\#C(i)-1)} \displaystyle\sum_{k\in C(i), \ell \in C(i), k \ne \ell} R_{k,\ell}, & \text{if } C(i) = C(j) \end{cases}, \qquad (11)$$

where $C(i)$ denotes the cluster i, $\#C(i)$ denotes the number of elements in the cluster $C(i)$ and $R_{k,\ell}$ is the (k,ℓ) entry of the matrix $\boldsymbol{R}$ defined in (10).

We illustrate the performance of our method in Figure 6 in the case where $\boldsymbol{\Sigma}$ has the structure (5) with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. We can see from this figure that the proposed methodology for estimating the correlation coefficients within the blocks of $\hat{\boldsymbol{\Sigma}}$ is efficient.

2.6 Summary of the WLasso method

The WLasso method can be summarized as follows:

- First step: Estimation of the matrix $\boldsymbol{\Sigma}$ by $\hat{\boldsymbol{\Sigma}}$, see Section 2.5.
- Second step: Transformation of Model (1) into Model (4) to remove the correlation existing between the columns of $\mathbf{X}$, see Section 2.1 where $\boldsymbol{\Sigma}$ is replaced by $\hat{\boldsymbol{\Sigma}}$.
- Third step: Estimation of $\widetilde{\boldsymbol{\beta}}$ defined in (4), see Section 2.2.
- Fourth step: Estimation of $\boldsymbol{\beta}$ defined in (1), see Section 2.3 and identification of its null and non-null components.

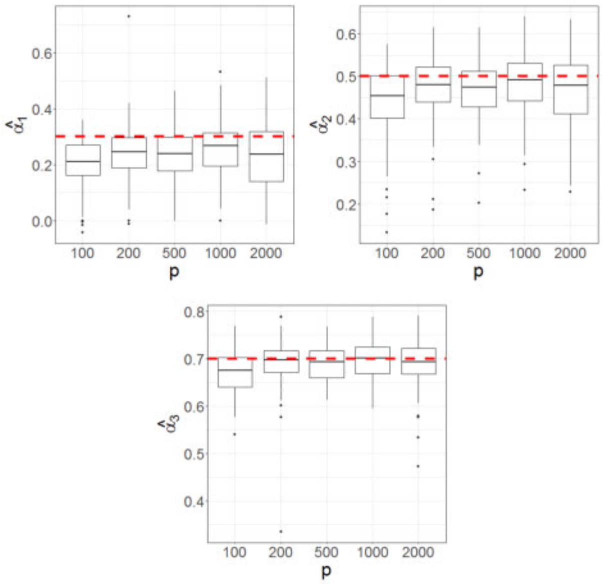

Fig. 6. Estimation of the parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. The horizontal dotted lines correspond to the true values of the parameters. These results are obtained from 100 replications for each value of p

3 Numerical experiments

We performed numerical experiments to assess the performance of the WLasso and to compare it with other recent approaches.

All simulated datasets were generated from Model (1) where the n rows of $\mathbf{X}$ are assumed to be independent Gaussian random vectors with a covariance matrix equal to $\boldsymbol{\Sigma}$ and ϵ is a standard Gaussian random vector independent of $\mathbf{X}$. Moreover, the number of predictors (biomarkers) p is equal to 100, 200, 500, 1000 or 2000 and the sample size n is equal to 50 or 100. We randomly chose 10 non-null coefficients among the p coefficients of $\boldsymbol{\beta}$ which correspond to the active biomarkers, thus considering different sparsity levels. The value b of the non-null coefficients is equal to either 0.5 or 1 to consider different signal-to-noise ratios.

Regarding the correlation matrix $\boldsymbol{\Sigma}$ which contains the correlation values between the biomarkers, namely the correlations between the columns of the design matrix $\mathbf{X}$, we mainly considered correlation structures in which the IC condition (3) was violated, this is the case of the first following correlation structure. It is indeed explained in Wang *et al.* (2019) that in almost all the publicly available genomic datasets that they have investigated the IC condition is violated.

- Block-wise correlation structure defined in (5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$ and $(0.5, 0.7, 0.9)$;
- Independent setting where $\boldsymbol{\Sigma}$ is the identity matrix.

Other correlation structures are considered in the Supplementary Material. The results that are presented are obtained from 100 replications.

3.1 Estimation of Σ

To evaluate the impact of the estimation of $\boldsymbol{\Sigma}$, simulations were performed to compare the performance of WLasso when $\boldsymbol{\Sigma}$ is known and when it is estimated. The results are displayed in Figure 7 for several values of γ (0.9, 0.95, 0.97) which is a parameter appearing in Section 2.4. In the top left part of this figure the empirical mean of the largest difference between the True Positive Rate (TPR) and False Positive Rate (FPR) over the replications is displayed for several values of p and for $n = 50$. It is obtained by selecting for each replication the value of λ achieving the largest difference between the TPR and FPR and by averaging these differences. In the top right

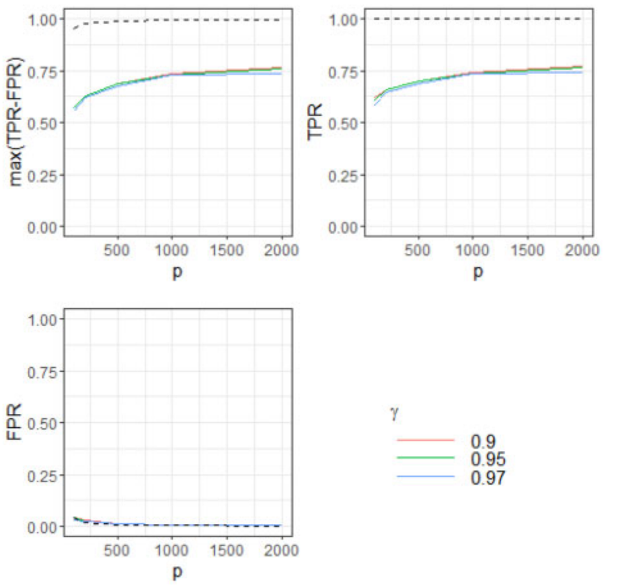

Fig. 7. Average over the replications of max(TPR-FPR) and of the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 50$. Dotted line: $\boldsymbol{\Sigma}$, solid line: $\hat{\boldsymbol{\Sigma}}$

and bottom parts of the figure, the empirical means of the corresponding TPR and FPR are displayed, respectively. We can see from this figure that for the value of λ maximizing the difference between TPR and FPR and for all the values of γ, all the active variables are properly retrieved without selecting non-active variables when $\mathbf{\Sigma}$ is known. In the case where $\mathbf{\Sigma}$ is estimated by using the approach described in Section 2.5, 75% of the active variables are recovered and less than 1% of non-active variables are wrongly estimated as active variables for p larger than 1000 and independently of the considered values of γ. Note that the results displayed in Figure 7 are obtained when $(b, n) = (0.5, 50)$ but we obtained similar conclusions for $(b, n) = (1, 50)$, $(b, n) = (0.5, 100)$ and $(1, 100)$.

3.2 Choice of λ

For tuning the parameter λ involved in our methodology, we propose choosing the value which minimizes $\mathrm{MSE}_{\hat{M}(\lambda)}(\lambda)$ defined in (9). In Figure 1 of the Supplementary Material, we compare the performance of our approach with this choice of λ called WLasso (dotted line) to the optimal one, called WLasso optimal, and obtained when λ is chosen to yield the largest difference between the TPR and the FPR (solid line). We can observe from this figure that, for the different values of γ, the TPR is 30% smaller when λ is estimated but that the FPR is quite similar. Additional comparaisons between Wlasso and Wlasso optimal can be found in Section 3.3 for $b = 0.5$ and in the Supplementary Material for $b = 1$.

3.3 Comparison with other methods

In this section, we compare our methodology with other approaches: the classical Lasso described in Tibshirani (1996) and two recently proposed methods aiming at handling the correlations between the columns of the design matrix $\mathbf{X}$: HOLP and Precision Lasso proposed by Wang and Leng (2016) and Wang *et al.* (2019), respectively. This comparison is performed by computing the TPR and FPR of these approaches for different values of the parameters involved in each of them.

The grid of λ for the classical Lasso and for our approach is provided by the glmnet and genlasso R packages, respectively. Concerning the Precision Lasso, we numerically found for each value of n and p the $\lambda_{\min}$ and $\lambda_{\max}$ leading to p non-null estimated coefficients and p null estimated coefficients, respectively. Then, we chose 100 values of λ uniformly distributed in the interval $[\lambda_{\min}, \lambda_{\max}]$ and we used the light implementation of the Precision Lasso. As for HOLP, $\boldsymbol{\beta}$ is estimated by $\hat{\boldsymbol{\beta}}_{\mathrm{HOLP}} = \mathbf{X}^T(\mathbf{X}\mathbf{X}^T)^{-1}\mathbf{y}$. Then, for each s in $\{1, \ldots, p\}$, the components of $\boldsymbol{\beta}$ which are estimated as non-null are the s largest among the $|\hat{\boldsymbol{\beta}}_{\mathrm{HOLP},j}|$, where $\hat{\boldsymbol{\beta}}_{\mathrm{HOLP},j}$ denotes the jth components of $\hat{\boldsymbol{\beta}}_{\mathrm{HOLP}}$. In this case, the parameter controlling the sparsity level of the estimator of $\boldsymbol{\beta}$ is s. It has a similar role as λ in the previous approaches.

The corresponding results are displayed in Figures 8 and 9 in the case where $n = 50$ and $b = 0.5$ and $\mathbf{\Sigma}$ has the block-wise correlation structure defined in (5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$ and $(0.5, 0.7, 0.9)$, respectively. The top left part of these figures displays displays the average over the replications of the largest difference between TPR and FPR for different values of p, which corresponds to an optimal choice of the parameters. For WLasso, we also display the results obtained when the parameter λ is chosen by using the strategy proposed in Section 3.2, $\gamma = 0.95$ and $\mathbf{\Sigma}$ is estimated using the procedure explained in Section 2.5. The corresponding TPR and FPR for each method are displayed in the top right part and bottom part of the figures, respectively. Note that we also conducted experiments in the case where $b = 1$. Since the conclusions are very similar, the corresponding figures are given in the Supplementary Material.

We can see from Figures 8 and 9 that WLasso outperforms the other methods: the TPR is one of the largest while the FPR is the smallest. HOLP has a larger TPR than WLasso. However, the associated FPR is much larger. It has moreover to be noticed that Lasso, HOLP and Precision Lasso are favored with respect to WLasso since their parameters were chosen to optimize their performance in terms of TPR and FPR whereas, in WLasso, the parameter λ was chosen by using the strategy of Section 3.2 and $\mathbf{\Sigma}$ was estimated.

Figure 10 displays the results when the sample size n is increased and equal to 100. We observe from this figure that the overall performance has been improved and that our approach outperforms the others especially in the case where p is large. Similar results are obtained in the case where $b = 1$ and $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$. We refer the reader to the Supplementary Material for further details.

Figure 11 displays the performance of the different methodologies in the case where $\mathbf{\Sigma} = \mathrm{Id}$, $n = 50$ and $b = 0.5$, that is in the case where there is no correlation between the biomarkers (columns of

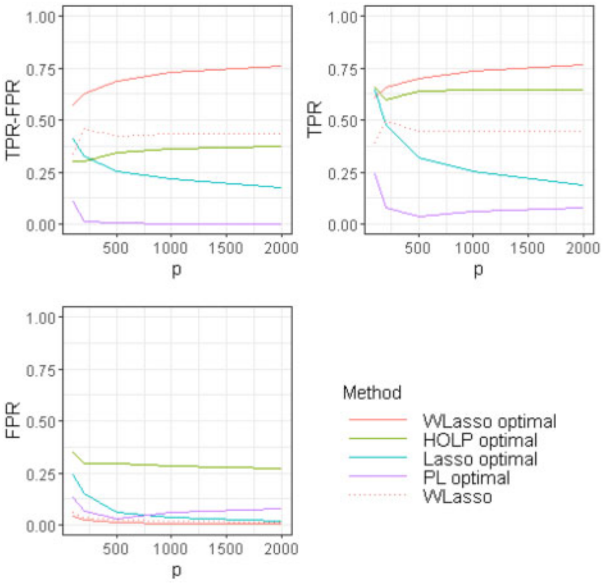

Fig. 8. Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\mathbf{\Sigma}$ 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 $n = 50$

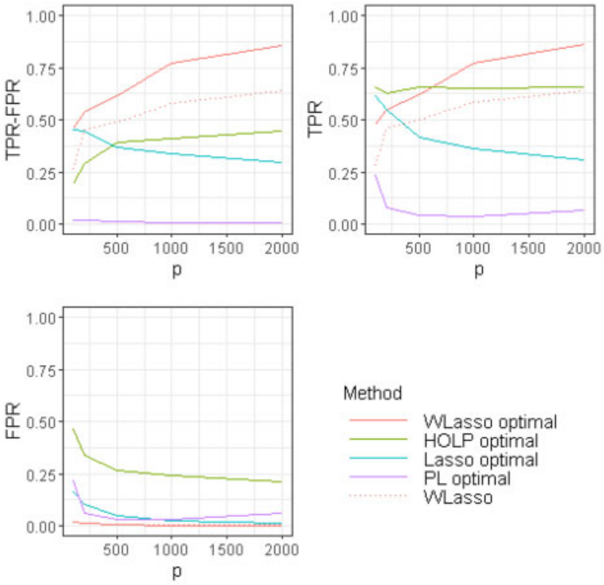

Fig. 9. Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\mathbf{\Sigma}$ 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 $n = 50$

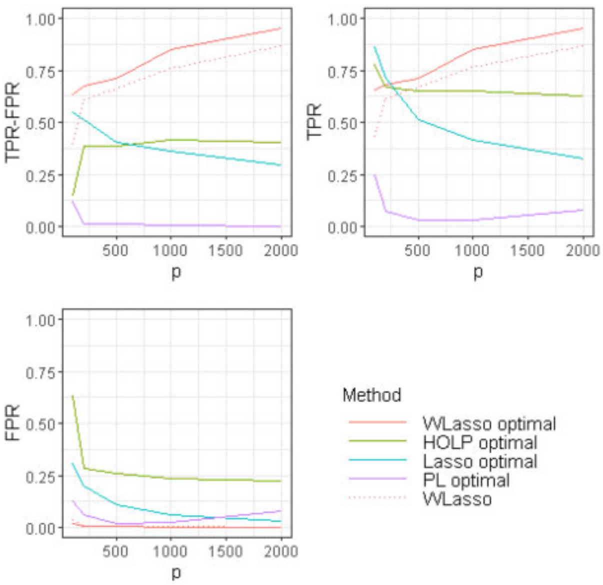

Fig. 10. Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\boldsymbol{\Sigma}$ 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 $n = 100$

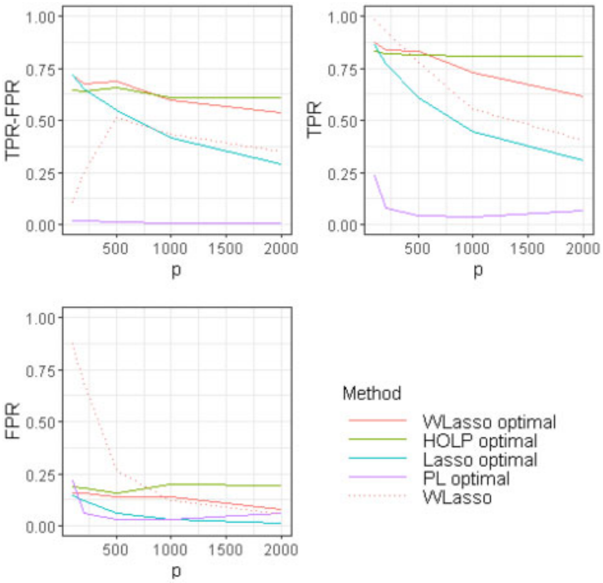

Fig. 11. Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\boldsymbol{\Sigma} = \mathrm{Id}$, $b = 0.5$ and $n = 50$

X). We can see from this figure that even in this case, our method, which is designed for handling the correlation between the biomarkers, obtains similar results as the Lasso in terms of TPR-FPR except for small values of p. In the case where $n = 100$, our approach obtains the best results in terms of TPR-FPR for p larger than 250, see the Supplementary Material.

In the correlation structures that are considered in the Supplementary Material, we show that the performance of WLasso optimal are better or at least equivalent to Lasso optimal for almost all the considered cases.

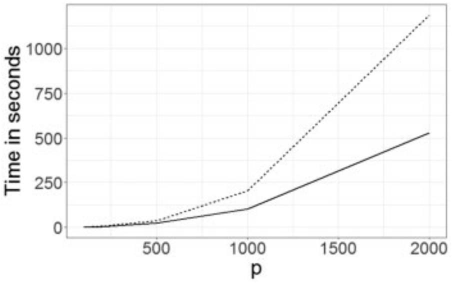

Fig. 12. Computational time of our approach WLasso when $n = 50$, ‘maxsteps’ has the default value, namely 2000 (dotted line) and maxsteps = 500 (solid line)

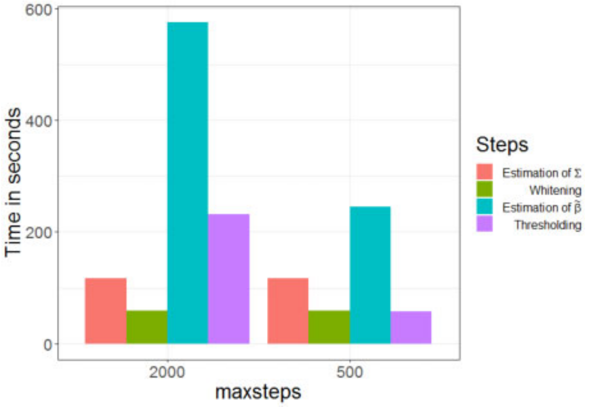

Fig. 13. Time allocation for each part of our method WLasso for $p = 2000$ and two values of the parameter ‘maxsteps’

3.4 Numerical performance

Figure 12 displays the computational times of our approach implemented in the R package WLasso for different values of p and of the parameter ‘maxsteps’ (maximum number of steps/λs considered in the algorithm) involved in the genlasso R package and $n = 50$. The timings were obtained on a workstation with 8GB of RAM and Intel Core i5 (2.4 GHz) CPU. We can see from this figure that it takes only 6 min for processing data with our approach when $p = 2000$.

Moreover, we can observe from Figure 13 that the most time consuming step of WLasso is the one where the generalized Lasso criterion is used (blue part in Fig. 13). However, the computational time of this step was divided by two when the parameter ‘maxsteps’ was changed from 2000 (default value) to 500 without changing the variable selection results. Note that all the numerical results of the previous sections were obtained with the default value of ‘maxsteps’ (2000).

4 Application to gene expression data in breast cancer

We applied the previously detailed methods to publicly available data at Gene Expression Omnibus database (www.ncbi.nlm.nih.gov/geo), with accession code GSE2990, see Sotiriou *et al.* (2006). A total of $n = 189$ tumor samples from patients with breast cancer were available and their microarray data were collected on 22 283 probes. Expression data were preprocessed and normalized as in the original publication. A filtering step based on the interquartile range (IQR) was considered to remove some probes as in Gentleman *et al.* (2005). We removed probes with IQR $<$ 1.5 and those which lack of annotation. The remaining $p = 1, 111$ probes were then standardized. The goal of the application is to identify genes potentially related to breast cancer prognosis. To this end, the ESR1 gene expression was considered as the variable $\mathbf{y}$ to explain (response variable) as it is well known to be associated with ER+ breast cancer prognosis as recently explained by Wu *et al.* (2020). The standardized 1111 probes were considered as explanatory variables.

We implemented the approaches investigated in Section 3 and illustrated their genes selection. For our approach WLasso, we used the methodology described in Section 2.5 for estimating the correlation matrix $\boldsymbol{\Sigma}$. The heatmap of the correlation between probes is provided in Supplementary Figure S15 of Supplementary Material and the coefficients α_1, α_2 and α_3 were estimated by $\hat{\alpha}_1 = 0.17, \hat{\alpha}_2 = 0.21$ and $\hat{\alpha}_3 = 0.52$, respectively. The parameter λ was chosen by cross-validation for the Lasso penalty, and the number of selected variables for Precision Lasso and HOLP was chosen in order to select approximately the same number of variables as with WLasso. Supplementary Table S1 given in the Supplementary Material provides the list of genes corresponding to the selected probes for each method. Unfortunately, HOLP could not provide any results since it requires the computation of the inverse of the matrix $\mathbf{X}\mathbf{X}^T$ which is not invertible in this case. The matrix $\mathbf{X}^T$ is indeed not full rank in this dataset. WLasso and Lasso selected almost the same number of genes, *i.e.* 63 and 66 genes respectively. Interestingly, the selection of the two methods is quite different with only 8 genes in common. Within these genes, some are already known in the literature to be associated with breast cancer prognosis such as TOP2A or NAT1 genes. In addition, some other potential prognostic genes were selected by one of the two methods, for example: BCL-2 for the Lasso, or GATA3 and CXCL12 for the WLasso. The selected genes for the Precision Lasso are also quite different from the other methods: 8 and 6 genes in common with WLasso and Lasso, respectively. Among the genes selected by the Precision Lasso, some are also known in the literature to be associated with the breast cancer prognosis such as GSTT1 or GATA3 also identified by WLasso. This application suggests that the WLasso can select meaningful variables in a context of correlated biomarker data. Nevertheless, this application can only be viewed as an illustration and cannot be used to formally compare or validate the methods in terms of variable selection.

5 Conclusion

In this article, we proposed an innovative, efficient and fully data-driven method to deal with the variable selection issue in high-dimensional frameworks where the active variables are highly correlated with the non-active ones, and is implemented in the WLasso R package. The proposed WLasso method has been assessed and compared with other methods in a simulation study with several scenarios. In the highly correlated setting, WLasso successfully identifies more true positives with limited false positives as compared with the classical Lasso. Contrary to HOLP, WLasso still works when several columns are linearly dependent and does not suffer from the inflation of noise introduced by the preconditioning. Compared with the recent Precision Lasso approach, which aims to deal with the same issue, WLasso obtained better results in terms of selection accuracy in the different settings considered. The poor performance of the Precision Lasso are consistent with the findings in Wang *et al.* (2019) in a low to moderate sample size setting even with highly correlated structure since the method selected a high number of false positives as compared to the true positives. WLasso is also very computationally efficient and demonstrated its abilities to identify some genes possibly related to breast cancer prognosis from a publicly available gene expression dataset. However, the following directions could be considered to improve its performance.

Firstly, the method that we used for estimating $\boldsymbol{\Sigma}$ could be improved by using more sophisticated approaches such as Perrot-Dockès *et al.* (2020). Secondly, our way of choosing the parameter λ for the final model selection could also be improved by considering cross-validation or stability selection. Until now, a simple approach has been considered to avoid computation time and performed quite well especially for moderate to high sample size. Thirdly, most of the computational time of our approach is spent in the application of the generalized Lasso criterion. Hence, for an application to genomic datasets having more than twenty thousands of variables, it could be worth speeding it up. This will be the subject of future work.

Funding

This work was supported by the Association Nationale Recherche Technologie (ANRT).

Conflict of Interest: none declared.

References

Fan,J. and Li,R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.*, **96**, 1348–1360.

Fan,J. and Li,R. (2006) Statistical challenges with high dimensionality: feature selection in knowledge discovery. *Proc. Madrid Int. Congress Math.*, **3**, 595–622.

Gentleman,R. *et al.* (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Springer-Verlag, Berlin, Heidelberg.

Heinze,G. *et al.* (2018) Variable selection - a review and recommendations for the practicing statistician. *Biometrical J.*, **60**, 1–19.

Jia,J. and Rohe,K. (2015) Preconditioning the lasso for sign consistency. *Electron. J. Stat.*, **9**, 1150–1172.

Kalia,M. (2015) Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism*, **64**, S16– S21.

McDonald,J. (2009). *Handbook of Biological Statistics*, 2nd edn. Sparky House Publishing, Baltimore.

Michalopoulos,I. *et al.* (2012) Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes. *BMC Res. Notes*, **5**, 265.

Perrot-Dockès,M. *et al.* (2020) Estimation of large block structured covariance matrices: Application to "multi-omic" approaches to study seed quality. arXiv:1806.10093.

Saeys,Y. *et al.* (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics*, **23**, 2507–2517.

Smith,G. (2018) Step away from stepwise. *J. Big Data*, **5**, 1–12.

Sotiriou,C. *et al.* (2006) Gene expression profiling in breast cancer: understanding the molecular basis of histologic grade to improve prognosis. *JNCI J. Natl. Cancer Inst.*, **98**, 262–272.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **58**, 267–288.

Tibshirani,R.J. and Taylor,J. (2011) The solution path of the generalized lasso. *Ann. Stat.*, **39**, 1335–1371.

Wang,H. *et al.* (2019) Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics*, **35**, 1181–1187.

Wang,X. and Leng,C. (2016) High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **78**, 589–611.

Wu,J.R. *et al.* (2020) Estrogen receptor 1 and progesterone receptor are distinct biomarkers and prognostic factors in estrogen receptor-positive breast cancer: evidence from a bioinformatic analysis. *Biomed. Pharmacother.*, **121**, 109647.

Zhao,P. and Yu,B. (2006) On model selection consistency of lasso. *J. Mach. Learn. Res.*, **7**, 2541–2563.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B (Stat. Methodol.)*, **67**, 301–320.