BMC Bioinformatics

RESEARCH

Open Access

Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso

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

*Correspondence: wencan.zhu@agroparistech.fr

[1] Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, 91120 Palaiseau, France

[2] Biostatistics and Programming Department, Sanofi R&D, 91380 Chilly Mazarin, France

Abstract

In clinical trials, identification of prognostic and predictive biomarkers has became essential to precision medicine. Prognostic biomarkers can be useful for the prevention of the occurrence of the disease, and predictive biomarkers can be used to identify patients with potential benefit from the treatment. Previous researches were mainly focused on clinical characteristics, and the use of genomic data in such an area is hardly studied. A new method is required to simultaneously select prognostic and predictive biomarkers in high dimensional genomic data where biomarkers are highly correlated. We propose a novel approach called PPLasso, that integrates prognostic and predictive effects into one statistical model. PPLasso also takes into account the correlations between biomarkers that can alter the biomarker selection accuracy. Our method consists in transforming the design matrix to remove the correlations between the biomarkers before applying the generalized Lasso. In a comprehensive numerical evaluation, we show that PPLasso outperforms the traditional Lasso and other extensions on both prognostic and predictive biomarker identification in various scenarios. Finally, our method is applied to publicly available transcriptomic and proteomic data.

Keywords: Variable selection, Highly correlated predictors, Genomic data

Introduction

With the development of precision medicine, there has been an increasing interest in the discovery of different types of biomarkers. A prognostic biomarker informs about a likely clinical outcome (e.g., disease recurrence, disease progression, death) in the absence of therapy or with a standard therapy that patients are likely to receive, while a predictive biomarker is associated with a response or a lack of response to a specific therapy. Fourati [3] and Clark [8] provided a comprehensive explanation and concrete examples to distinguish prognostic from predictive biomarkers, respectively. During the past decade, prognostic and predictive biomarkers showed their power in the development of precision medicine. Giannos et al. [15] identified ten prognostic gene biomarkers for non-small cell lung cancer that can be useful for improving risk prediction and therapeutic strategies. Zhao et al. [38] obtained prognostic biomarkers for 13 cancers by integrating multi-omics data and provided a reference for translational medicine researchers. He et al. [16] identified predictive biomarkers for sorafenib resistance and contributed to the guidance of individualized drug therapy. Yet, correctly identifying such biomarkers remains difficult.

Concerning the biomarker selection, the high dimensionality of genomic data is one of the main challenges as explained in Fan and Li [9]. To identify effective biomarkers in high-dimensional settings, several approaches can be considered including hypothesis-based tests described in [23], wrapper approaches proposed in [25], and penalized approaches such as Lasso designed by [30] among others. Hypothesis-based tests consider each biomarker independently and thus ignore potential correlations between them. Wrapper approaches often show high risk of overfitting and are computationally expensive for high-dimensional data as explained in [27]. More efforts have been devoted to penalized methods given their ability to automatically perform variable selection and coefficient estimation simultaneously as highlighted in [10]. However, Lasso showed some potential drawbacks when biomarkers are highly correlated. Particularly, when the Irrepresentable Condition (IC) proposed by [39] is violated, Lasso can not guarantee to correctly identify true effective biomarkers. In genomic data, biomarkers are usually highly correlated such that this condition can hardly be satisfied, see [34]. Several methods have been proposed to adress this issue. Elastic Net [41] combines the ℓ_1 and ℓ_2 penalties and is particularly effective in tackling correlation issues and can generally outperform Lasso. Adaptive Lasso [42] proposes to assign adaptive weights for penalizing different coefficients in the ℓ_1 penalty, and its oracle property was demonstrated. Wang and Leng [35] proposed the HOLP approach which consists in removing the correlation between the columns of the design matrix; Wang et al. [34] proposed to handle the correlation by assigning similar weights to correlated variables in their approach called Precision Lasso; Zhu et al. [40] proposed to remove the correlations by applying a whitening transformation to the data before using the generalized Lasso criterion designed by [31].

The challenge of finding prognostic biomarkers has been extensively explored with previously introduced methods, however, the discovery of predictive biomarkers has seen much less attention. Limited to binary endpoint, Foster et al. [13] proposed to first predict response probabilities for treatment and use this probability as the response in a classification problem to find effective biomarkers. Tian et al. [29] proposed a new method to detect interaction between the treatment and the biomarkers by modifying the covariates. This method can be implemented on continuous/binary/time-to-event endpoint. Lipkovich et al. [20] proposed a method called SIDES, which adopts a recursive partitioning algorithm for screening treatment-by-biomarker interactions. This method was further improved in [19] by adding another step of preselection on predictive biomarkers based on variable importance. The method was demonstrated with continuous endpoint. Evaluated on time-to-event data, Ternès et al.[28] proposed a framework for identifying biomarker-by-treatment interactions but not specifically in the context of correlated biomarkers. More recently, Sechidis et al. [26] applied approaches coming from information theory for ranking biomarkers on their prognostic/predictive strength.

Their method is applicable only for binary or time-to-event endpoint. Moreover, all of these methods were assessed under the situation where the sample size is relatively large and the number of biomarkers is limited, which is hardly the case for genomic data.

In the literature mentioned above, the authors focused on one of the problematic of identifying prognostic or predictive biomarkers, but rarely on both. Even if predictive biomarkers is of major importance for identifying patients more likely to benefit from a treatment, the prognostic biomarkers is also key in this context. Indeed, the clinical impact of a treatment can be judged only with the knowledge of the prognosis of a patient. It is thus of importance to reliably predict the prognosis of patients to assist treatment counseling [36]. To properly describe the two effects, the experimental treatment should be compared to a standard therapy (or a placebo), and patients receiving different treatments should be randomized. A randomized clinical trial can be ideal for such a study. In this paper, we developed a new method called PPLasso (Prognostic Predictive Lasso) to identify prognostic and at the same time predictive biomarkers in a high dimensional setting with continuous endpoints, as presented in "Methods" section . Extensive numerical experiments are given in "Numerical experiments" section to assess the performance of our approach and to compare it to other methods. PPLasso is also applied on two publicly available transcriptomic and proteomic data in "Application to transcriptomic and proteomic data" section. Finally, we give concluding remarks in "Conclusion" section .

Methods

In this section, we propose a novel approach called PPLasso (Predictive Prognostic Lasso) which consists in writing the identification of predictive and prognostic biomarkers as a variable selection problem in an ANCOVA (Analysis of Covariance) type model mentioned for instance in [12].

Statistical modeling

Let $\mathbf{y}$ be a continuous response or endpoint and t_1, t_2 two treatments. Let also $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) denote the design matrix for the n_1 (resp. n_2) patients with treatment t_1 (resp. t_2), each containing measurements on p candidate biomarkers:

$$\mathbf{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \vdots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix}, \mathbf{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \vdots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix}. \tag{1}$$

To take into account the potential correlation that may exist between the biomarkers in the different treatments, we shall assume that the rows of $\mathbf{X}_1$ (resp. $\mathbf{X}_2$) are independent centered Gaussian random vectors with a covariance matrice equal to $\mathbf{\Sigma_1}$ (resp. $\mathbf{\Sigma_2}$).

To model the link that exists between $\mathbf{y}$ and the different types of biomarkers we propose using the following model:

$$\mathbf{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \mathbf{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1p} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2p} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}, \tag{2}$$

where $(y_{i1}, \ldots, y_{in_i})$ corresponds to the response of patients with treatment t_i, i being equal to 1 or 2,

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

with α_1 (resp. α_2) corresponding to the effects of treatment t_1 (resp. t_2). Moreover, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1p})'$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2p})'$) are the coefficients associated to each of the p biomarkers in treatment t_1 (resp. t_2) group, $'$ denoting the matrix transposition and $\epsilon_{11}, \ldots, \epsilon_{2n_2}$ are standard independent Gaussian random variables independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. When t_1 stands for the standard treatment or placebo, prognostic biomarkers are defined as those having non-zero coefficients in $\boldsymbol{\beta}_1$. According to the definition of prognostic biomarkers, their effect should indeed be demonstrated in the absence of therapy or with a standard therapy that patients are likely to receive. On the other hand, predictive biomarkers are defined as those having non-zero coefficients in $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ because they aim to highlight different effects between two different treatments.

Model (2) can be written as:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{3}$$

with $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1', \boldsymbol{\beta}_2')'$. The Lasso penalty is a well-known approach to estimate coefficients with a sparsity enforcing constraint allowing variable selection by estimating some coefficients by zero. It consists in minimizing the following penalized least-squares criterion [30]:

$$\frac{1}{2}\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\|_2^2 + \lambda\|\boldsymbol{\gamma}\|_1, \tag{4}$$

where $\|\mathbf{u}\|_2^2 = \sum_{i=1}^n u_i^2$ and $\|\mathbf{u}\|_1 = \sum_{i=1}^n |u_i|$ for $\mathbf{u} = (u_1, \ldots, u_n)$. A different sparsity constraint was applied to $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ to allow different sparsity levels. Hence we propose to replace the penalty $\lambda\|\boldsymbol{\gamma}\|_1$ in (4) by

$$\lambda_1 \left\|\boldsymbol{\beta}_1\right\|_1 + \lambda_2 \left\|\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1\right\|_1. \tag{5}$$

Thus, a first estimator of $\boldsymbol{\gamma}$ could be found by minimizing the following criterion with respect to $\boldsymbol{\gamma}$:

$$\frac{1}{2}\left\|\mathbf{y} - \mathbf{X}\boldsymbol{\gamma}\right\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \boldsymbol{\gamma} \right\|_1, \tag{6}$$

where $D_1 = [\mathrm{Id}_p, \mathbf{0}_{p,p}]$ and $D_2 = [-\mathrm{Id}_p, \mathrm{Id}_p]$, with Id_p denoting the identity matrix of size p and $\mathbf{0}_{i,j}$ denoting a matrix having i rows and j columns and containing only zeros. However, since the inconsistency of Lasso biomarker selection is originated from the correlations between the biomarkers, we propose to remove the correlation by "whitening" the matrix $\mathbf{X}$. More precisely, we consider $\widetilde{\mathbf{X}} = \mathbf{X}\boldsymbol{\Sigma}^{-1/2}$, where

$$\boldsymbol{\Sigma} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \boldsymbol{\Sigma}_1 & 0 \\ 0 & 0 & 0 & \boldsymbol{\Sigma}_2 \end{bmatrix} \tag{7}$$

and define $\boldsymbol{\Sigma}^{-1/2}$ by replacing in (7) $\boldsymbol{\Sigma}_i$ by $\boldsymbol{\Sigma}_i^{-1/2}$, where $\boldsymbol{\Sigma}_i^{-1/2} = \mathbf{U}_i \mathbf{D}_i^{-1/2} \mathbf{U}_i^T$, $\mathbf{U}_i$ and $\mathbf{D}_i$ being the matrices involved in the spectral decomposition of $\boldsymbol{\Sigma}_i$ for $i = 1$ or 2. With such a transformation the columns of $\widetilde{\mathbf{X}}$ are decorrelated and Model (3) can be rewritten as follows:

$$\mathbf{y} = \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon} \tag{8}$$

where $\widetilde{\boldsymbol{\gamma}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{\gamma}$. The objective function (6) thus becomes:

$$L_{\lambda_1,\lambda_2}^{\text{PPLasso}}(\widetilde{\boldsymbol{\gamma}}) = \frac{1}{2}\left\|\mathbf{y} - \widetilde{\mathbf{X}}\widetilde{\boldsymbol{\gamma}}\right\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \boldsymbol{\Sigma}^{-1/2}\widetilde{\boldsymbol{\gamma}} \right\|_1. \tag{9}$$

Estimation of $\widetilde{\boldsymbol{\gamma}}$

Let us define a first estimator of $\widetilde{\boldsymbol{\gamma}} = (\widetilde{\alpha}_1, \widetilde{\alpha}_2, \widetilde{\boldsymbol{\beta}}_1', \widetilde{\boldsymbol{\beta}}_2')$ as follows:

$$\widehat{\widetilde{\boldsymbol{\gamma}}}_0(\lambda_1, \lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\widetilde{\boldsymbol{\beta}}}_{10}', \widehat{\widetilde{\boldsymbol{\beta}}}_{20}') = \mathrm{Argmin}_{\widetilde{\boldsymbol{\gamma}}}\, L_{\lambda_1,\lambda_2}^{\text{PPLasso}}(\widetilde{\boldsymbol{\gamma}}), \tag{10}$$

for each fixed λ_1 and λ_2. To better estimate $\widetilde{\boldsymbol{\beta}}_1$ and $\widetilde{\boldsymbol{\beta}}_2$, a thresholding was applied to $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda_1, \lambda_2) = (\widehat{\widetilde{\boldsymbol{\beta}}}_{10}(\lambda_1, \lambda_2)', \widehat{\widetilde{\boldsymbol{\beta}}}_{20}(\lambda_1, \lambda_2)')'$. For K_1 (resp. K_2) in $\{1, \ldots, p\}$, let Top_{K_1} (resp. Top_{K_2}) be the set of indices corresponding to the K_1 (resp. K_2) largest values of the components of $|\widehat{\widetilde{\boldsymbol{\beta}}}_{10}(\lambda_1, \lambda_2)|$ (resp. $|\widehat{\widetilde{\boldsymbol{\beta}}}_{20}(\lambda_1, \lambda_2)|$), then the estimator of $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1', \widetilde{\boldsymbol{\beta}}_2')$ after the correction is denoted by $\widehat{\widetilde{\boldsymbol{\beta}}}(\lambda_1, \lambda_2) = (\widehat{\widetilde{\boldsymbol{\beta}}}_1^{(\widehat{K}_1)}(\lambda_1, \lambda_2), \widehat{\widetilde{\boldsymbol{\beta}}}_2^{(\widehat{K}_2)}(\lambda_1, \lambda_2))$ where the jth component of $\widehat{\widetilde{\boldsymbol{\beta}}}_i^{(K_i)}(\lambda_1, \lambda_2)$, for $i = 1$ or 2, is defined by:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{ij}^{(K_i)}(\lambda_1, \lambda_2) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1, \lambda_2), & j \in \mathrm{Top}_{K_i} \\ K_1\text{th largest value of } |\widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1, \lambda_2)|, & j \notin \mathrm{Top}_{K_i}. \end{cases} \tag{11}$$

Note that the corrections are only performed on $\widehat{\widehat{\boldsymbol{\beta}}}_0$, the estimators $\widehat{\widehat{\alpha}}_1$ and $\widehat{\widehat{\alpha}}_2$ were not modified. The choice of K_1 and K_2 will be explained in "Choice of the parameters K_1, K_2, M_1 and M_2" section.

To illustrate the interest of using a thresholding step, we generated a dataset based on Model 3 with parameters described in "Simulation setting" section and $p = 500$. Moreover, to simplify the graphical illustrations, we focus on the case where $\lambda_1 = \lambda_2 = \lambda$. Figure 1 displays the estimation error associated to the estimators of $\widetilde{\boldsymbol{\beta}}(\lambda)$ before and after the thresholding. We can see from this figure that the estimation of $\widetilde{\boldsymbol{\beta}}(\lambda)$ is less biased after the correction. Moreover, we observed that this thresholding strongly improves the final estimation of $\boldsymbol{\gamma}$ and the variable selection performance of our method.

Estimation of $\boldsymbol{\gamma}$

With $\widehat{\widehat{\boldsymbol{\beta}}} = (\widehat{\widehat{\boldsymbol{\beta}}}_1', \widehat{\widehat{\boldsymbol{\beta}}}_2')$, the estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ can be obtained by $\widehat{\boldsymbol{\beta}}_{10} = \boldsymbol{\Sigma}_1^{-1/2}\widehat{\widehat{\boldsymbol{\beta}}}_1$ and $(\widehat{\boldsymbol{\beta}}_{20} - \widehat{\boldsymbol{\beta}}_{10}) = \boldsymbol{\Sigma}_2^{-1/2}\widehat{\widehat{\boldsymbol{\beta}}}_2 - \boldsymbol{\Sigma}_1^{-1/2}\widehat{\widehat{\boldsymbol{\beta}}}_1$. As previously, another thresholding was applied to $\widehat{\boldsymbol{\beta}}_{10}$ and $\widehat{\boldsymbol{\beta}}_{20}$: for $i = 1$ or 2,

$$\widehat{\boldsymbol{\beta}}_{ij}^{(M_i)}(\lambda_1, \lambda_2) = \begin{cases} \widehat{\boldsymbol{\beta}}_{i0j}(\lambda_1, \lambda_2), & j \in \text{Top}_{M_i} \\ 0, & j \notin \text{Top}_{M_i}, \end{cases} \tag{12}$$

for each fixed λ_1 and λ_2. The biomarkers with non-zero coefficients in $\widehat{\boldsymbol{\beta}}_1 = \widehat{\boldsymbol{\beta}}_1^{(M_1)}$ (resp. $\widehat{\boldsymbol{\beta}}_2^{(M_2)} - \widehat{\boldsymbol{\beta}}_1^{(M_1)}$) are considered as prognostic (resp. predictive) biomarkers, where the choice of M_1 and M_2 is explained in in "Choice of the parameters K_1, K_2, M_1 and M_2" section .

To illustrate the benefits of using an additional thresholding step, we used the dataset described in "Estimation of $\widetilde{\boldsymbol{\gamma}}$" section. Moreover, to simplify the graphical illustrations, we also focus on the case where $\lambda_1 = \lambda_2 = \lambda$. Additional file 1: Figure S1 displays the number of True Positive (TP) and False Positive (FP) in prognostic and predictive biomarker identification with and without the second thresholding. We

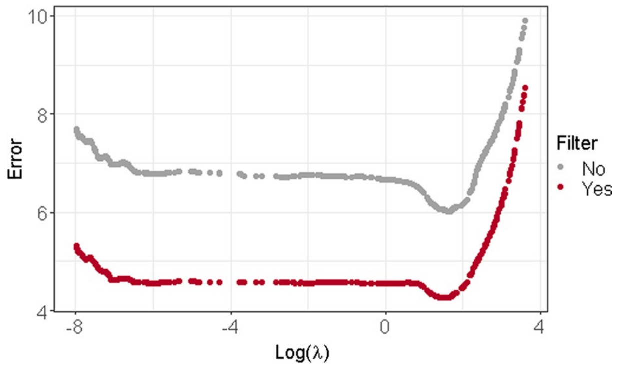

Fig. 1 Estimation error $\left\|\widehat{\widehat{\boldsymbol{\beta}}}_0(\lambda) - \widetilde{\boldsymbol{\beta}}\right\|_2$ (gray) and $\left\|\widehat{\widehat{\boldsymbol{\beta}}}(\lambda) - \widetilde{\boldsymbol{\beta}}\right\|_2$ (red) for all λ

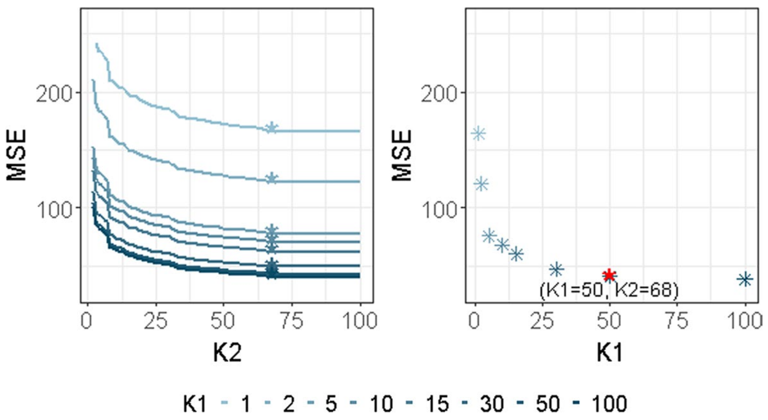

Fig. 2 Illustration of how to choose K_1 and K_2 ($\delta = 0.95$), final choice is marked with '*'

can see from this figure that the thresholding stage limits the number of false positives. Note that α_1 and α_2 are estimated by $\widehat{\widetilde{\alpha}}_1$ and $\widehat{\widetilde{\alpha}}_2$ defined in (10).

Choice of the parameters K_1, K_2, M_1 and M_2

For each (λ_1, λ_2) and each K_1, we computed:

$$\widetilde{\mathrm{MSE}}_{K_1,K_2}(\lambda_1,\lambda_2) = \|\mathbf{y} - \widetilde{\mathbf{X}}\widehat{\widetilde{\boldsymbol{\gamma}}}^{(K1,K2)}(\lambda_1,\lambda_2)\|_2^2, \tag{13}$$

where $\widehat{\widetilde{\boldsymbol{\gamma}}}^{(K1,K2)}(\lambda_1,\lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\widetilde{\boldsymbol{\beta}}}_1^{(K_1)'}, \widehat{\widetilde{\boldsymbol{\beta}}}_2^{(K_2)'})$ defined in (10) and in (11). It is displayed in the left part of Fig. 2.

For each λ_1, λ_2 and a given $\delta \in (0, 1)$, the parameter $\widehat{K_2}$ is then chosen as follows for each K_1:

$$\widehat{K_2}(\lambda_1,\lambda_2) = \mathrm{Argmin}\left\{K_2 \geq 1 \text{ s.t. } \frac{\widetilde{\mathrm{MSE}}_{(K_1,K_2+1)}(\lambda_1,\lambda_2)}{\widetilde{\mathrm{MSE}}_{(K_1,K_2)}(\lambda_1,\lambda_2)} \geq \delta\right\}.$$

The $\widehat{K_2}$ associated to each K_1 are displayed with '*' in the left part of Fig. 2. Then $\widehat{K_1}$ is chosen by using a similar criterion:

$$\widehat{K_1}(\lambda_1,\lambda_2) = \mathrm{Argmin}\left\{K_1 \geq 1 \text{ s.t. } \frac{\widetilde{\mathrm{MSE}}_{(K_1+1,\widehat{K_2})}(\lambda_1,\lambda_2)}{\widetilde{\mathrm{MSE}}_{(K_1,\widehat{K_2})}(\lambda_1,\lambda_2)} \geq \delta\right\}.$$

The values of $\widetilde{\mathrm{MSE}}_{(K_1,\widehat{K_2})}(\lambda_1,\lambda_2)$ are displayed in the right part of Fig. 2 in the particular case where $\lambda_1 = \lambda_2 = \lambda$, $\delta = 0.95$ and with the same dataset as the one used in "Estimation of $\widetilde{\boldsymbol{\gamma}}$" section. $\widehat{K_1}$ is displayed with a red star. This value of δ will be used in the following sections. However, choosing δ in the range (0.9,0.99) does not have a strong impact on the variable selection performance of our approach.

The parameters $\widehat{M}_1$ and $\widehat{M}_2$ are chosen in a similar way except that $\widetilde{\mathrm{MSE}}_{K_1,K_2}(\lambda_1,\lambda_2)$ is replaced by $\widehat{\mathrm{MSE}}_{M_1,M_2}(\lambda_1,\lambda_2)$ where:

$$\widehat{\mathrm{MSE}}_{M_1,M_2}(\lambda_1,\lambda_2) = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}^{(M_1,M_2)}(\lambda_1,\lambda_2)\|_2^2,$$

with $\widehat{\boldsymbol{\gamma}}^{(M_1,M_2)}(\lambda_1,\lambda_2) = (\widehat{\widehat{\alpha}}_1, \widehat{\widehat{\alpha}}_2, \widehat{\boldsymbol{\beta}}_1^{(M_1)'}, \widehat{\boldsymbol{\beta}}_2^{(M_2)'})$ defined in (10) and (12). In the following, $\widehat{\boldsymbol{\gamma}}(\lambda_1,\lambda_2) = \widehat{\boldsymbol{\gamma}}^{(\widehat{M}_1,\widehat{M}_2)}(\lambda_1,\lambda_2)$.

Estimation of $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$

As the empirical correlation matrix is known to be a non accurate estimator of $\boldsymbol{\Sigma}$ when p is larger than n, a new estimator has to be used. Thus, for estimating $\boldsymbol{\Sigma}$ we adopted a cross-validation based method designed by [5] and implemented in the `cvCovEst` R package [6]. This method chooses the estimator having the smallest estimation error among several compared methods (sample correlation matrix, POET [11] and Tapering [7] as examples). Since the samples in treatments t_1 and t_2 are assumed to be collected from the same population, $\boldsymbol{\Sigma}_1$ and $\boldsymbol{\Sigma}_2$ are assumed to be equal.

Choice of the parameters λ_1 and λ_2

For the sake of simplicity, we limit ourselves to the case where $\lambda_1 = \lambda_2 = \lambda$. For choosing λ we used BIC (Bayesian information criterion) which is widely used in the variable selection field and which consists in minimizing the following criterion with respect to λ:

$$\text{BIC}(\lambda) = n \log(\text{MSE}(\lambda)/n) + k(\lambda)\log(n), \tag{14}$$

where n is the total number of samples, $\text{MSE}(\lambda) = \|\mathbf{y} - \mathbf{X}\widehat{\boldsymbol{\gamma}}(\lambda)\|_2^2$ and $k(\lambda)$ is the number of non null coefficients in the OLS estimator $\widehat{\boldsymbol{\gamma}}$ obtained by re-estimating only the non null components of $\widehat{\boldsymbol{\beta}}_1$ and $\widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1$. The values of the BIC criterion as well as those of the MSE obtained from the dataset described in "Estimation of $\widetilde{\boldsymbol{\gamma}}$" section are displayed in Fig. 3.

Additional file 1: Table S1 provides the True Positive Rate (TPR) and False Positive Rate (FPR) when λ is chosen either by minimizing the MSE or the BIC criterion for this dataset. We can see from this table that both of them have TPR=1 (all true positives are identified). However, the FPR based on the BIC criterion is smaller than the one obtained by using the MSE.

Note that additional results using two different parameters λ_1 and λ_2 in the BIC criterion are provided in "Two parameters λ_1 and λ_2 v.s. λ in the BIC Criterion" section.

Numerical experiments

This section presents a comprehensive numerical study by comparing the performance of our method with other regularized approaches in terms of prognostic and predictive biomarker selection. Besides the Lasso, we also compared with Elastic Net, Adaptive Lasso and WLasso [40] since they also take into account the correlations. For these compared methods, in order to directly estimate prognostic and predictive effects, $\mathbf{X}$ and $\boldsymbol{\gamma}$ in Model (3) were replaced by

$$\mathbf{X}^* = \begin{bmatrix} \mathbf{1}_{n_1,1} & \mathbf{0}_{n_1,1} & \mathbf{X}_1 & \mathbf{0}_{n_1,p} \\ \mathbf{0}_{n_2,1} & \mathbf{1}_{n_2,1} & \mathbf{X}_2 & \mathbf{X}_2 \end{bmatrix},$$

and $\boldsymbol{\gamma}^* = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, respectively, where $\mathbf{X}_1$ and $\mathbf{X}_2$ are defined in (1), $\mathbf{0}_{i,j}$ (resp. $\mathbf{1}_{i,j}$) denotes a matrix having i rows and j columns and containing only zeros (resp. ones). Note that this is the modeling proposed by [21]. The sparsity enforcing constraint was

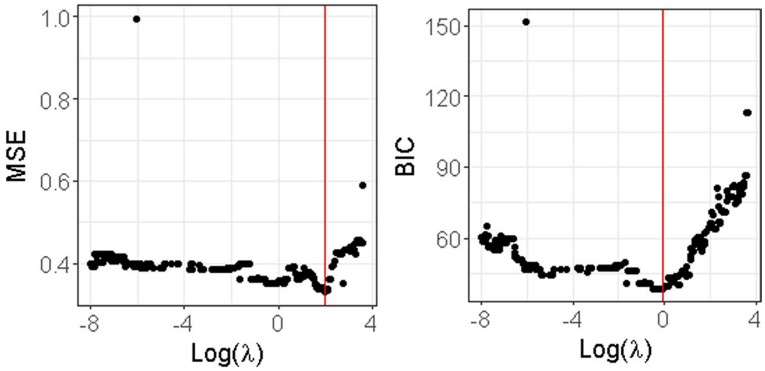

Fig. 3 MSE and BIC for all λ. The λ minimizing each criterion is displayed with a vertical line

put on the coefficients $\boldsymbol{\beta}_1^*$ and $\boldsymbol{\beta}_2^*$ which boils down to putting a sparsity enforcing constraint on $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.

Simulation setting

All simulated datasets were generated from Model (3) where the n_1 (n_2) rows of $\mathbf{X}_1$ ($\mathbf{X}_2$) are assumed to be independent Gaussian random vectors with a covariance matrix $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \boldsymbol{\Sigma}_{bm}$, and $\boldsymbol{\epsilon}$ is a standard Gaussian random vector independent of $\mathbf{X}_1$ and $\mathbf{X}_2$. We defined $\boldsymbol{\Sigma}_{bm}$ as:

$$\boldsymbol{\Sigma}_{bm} = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{12}^T & \boldsymbol{\Sigma}_{22} \end{bmatrix} \tag{15}$$

where $\boldsymbol{\Sigma}_{11}$ (resp. $\boldsymbol{\Sigma}_{22}$) are the correlation matrix of prognostic (resp. non-prognostic) biomarkers with off-diagonal entries equal to a_1 (resp. a_3). Morever, $\boldsymbol{\Sigma}_{12}$ is the correlation matrix between prognostic and non-prognostic variables with entries equal to a_2. In our simulations $(a_1, a_2, a_3) = (0.3, 0.5, 0.7)$, which is a framework proposed by [37]. We checked that the Irrepresentable Condition (IC) of [39] is violated and thus the standard Lasso cannot recover the positions of the null and non null variables. For each dataset we assumed randomized treatment allocation between standard and experimental arm with a 1:1 ratio, *i.e.* $n_1 = n_2 = 50$. We further assume a relative treatment effect of 1 ($\alpha_1 = 0$ and $\alpha_2 = 1$). The number of biomarkers p varies from 200 to 2000. The number of active biomarkers was set to 10 (*i.e.* 5 purely prognostic biomarkers with $\boldsymbol{\beta}_{1j} = \boldsymbol{\beta}_{2j} = b_1 = 1$ $(j = 1, ..., 5)$ and 5 biomarkers both prognostic and predictive with $\boldsymbol{\beta}_{1j} = b_1$ and $\boldsymbol{\beta}_{2j} = b_2 = 2$ $(j = 6, ..., 10)$).

Evaluation criteria

We considered several evaluation criteria to assess the performance of the methods in selecting the prognostic and predictive biomarkers: the TPR_{prog} as the true positive rate (i.e. rate of active biomarkers selected) and FPR_{prog} the false positive rate (i.e. rate of inactive biomarkers selected) of the selection of prognostic biomarkers, and similarly for predictive biomarkers with TPR_{pred} and FPR_{pred}. We further note TPR_{all} and FPR_{all} the criterion of overall selection among all candidate biomarkers regardless their prognostic or predictive effect. The objective of the selection is to maximize the $\mathrm{TPR}_{\mathrm{all}}$ and minimize the $\mathrm{FPR}_{\mathrm{all}}$. All metrics were calculated by averaging the results of 100 replications for each scenario.

Two parameters λ_1 and λ_2 v.s. λ in the BIC Criterion

In this section, we compare the results obtained by choosing $\lambda_1 = \lambda_2 = \lambda$ as the minimizer of the BIC criterion described in (14) with those obtained by choosing the values of λ_1 and λ_2 as those minimizing the criterion (16):

$$\mathrm{BIC}(\lambda_1, \lambda_2) = n \log(\mathrm{MSE}(\lambda_1, \lambda_2)/n) + k(\lambda_1, \lambda_2) \log(n). \tag{16}$$

Different results are presented in Fig. 4. $\mathrm{PPLasso}_{\Sigma}$ (resp. PPLasso) corresponds to the results of the method by using the true (resp. estimated) matrix $\mathbf{\Sigma}_{bm}$. For estimating $\mathbf{\Sigma}_{bm}$, we used the approach explained in "Estimation of $\mathbf{\Sigma}_1$ and $\mathbf{\Sigma}_2$" section . Different choices of parameters are also given: "optimal", "min(bic(λ))" and "min(bic(λ_1, λ_2))". The first one uses as a value of the parameters the one maximizing ($\mathrm{TPR}_{\mathrm{all}} - \mathrm{FPR}_{\mathrm{all}}$), the second one uses the approach presented in (14) and the last one uses the approach described in (16).

We observed that the results with two tuning parameters (λ_1, λ_2) were slightly better than those with a single parameter λ. However, the gap is very small and almost invisible when p increases. For this reason, we limited ourselves to a single tuning parameter λ in the following.

Biomarker selection results

In order to compare the performance of our approach to the best performance that could be reached by Elastic Net, Lasso, Adaptive Lasso and WLasso, we used for these methods the "optimal" parameters namely those maximizing ($\mathrm{TPR}_{\mathrm{all}} - \mathrm{FPR}_{\mathrm{all}}$). The first three methods were implemented with the `glmnet` R package, the best parameter α involved in Elastic Net was chosen in the set $\{0.1, 0.2, \ldots, 0.9\}$. WLasso was implemented with the `WLasso` R package. The choice of "min(bic)" is only applied to our method and corresponds to a choice of λ that could be used in practical situations. For ease of presentation, the abbreviation EN (resp. AdLasso) refers to Elastic Net (resp. Adaptive Lasso) in the following.

Figure 5 shows the selection performance of PPLasso and other compared methods in the simulation scenario presented in "Simulation setting" section. PPLasso achieved to select all prognostic biomarkers ($\mathrm{TPR}_{\mathrm{prog}}$ almost 1) even for large p, with limited false positive prognostic biomarkers selected. As compared to the optimal λ maximizing ($\mathrm{TPR}_{\mathrm{all}} - \mathrm{FPR}_{\mathrm{all}}$), the one selected with the BIC tends to select some false positives (average: 33 ($\mathrm{FPR}_{\mathrm{prog}} = 0.17$) for $p = 200$ and 10 ($\mathrm{FPR}_{\mathrm{prog}} = 0.005$) for $p = 2000$). The results obtained from the oracle and estimated $\mathbf{\Sigma}_{bm}$ are comparable. Selection performance of predictive biomarkers is slightly lowered as compared to prognostic biomarkers. Even if the false positive selection is quite similar between prognostic and predictive biomarkers, PPLasso missed some true predictive biomarkers when λ is selected with the BIC criterion (average $\mathrm{TPR}_{\mathrm{pred}} = 0.98$ and 0.80 for oracle and estimated $\mathbf{\Sigma}_{bm}$, respectively, with $p = 2000$). In this scenario where the IC is violated, PPLasso globally outperforms Lasso, Elastic Net, Adaptive Lasso and WLasso. Thanks to the whitening technique used in WLasso, it achieved higher

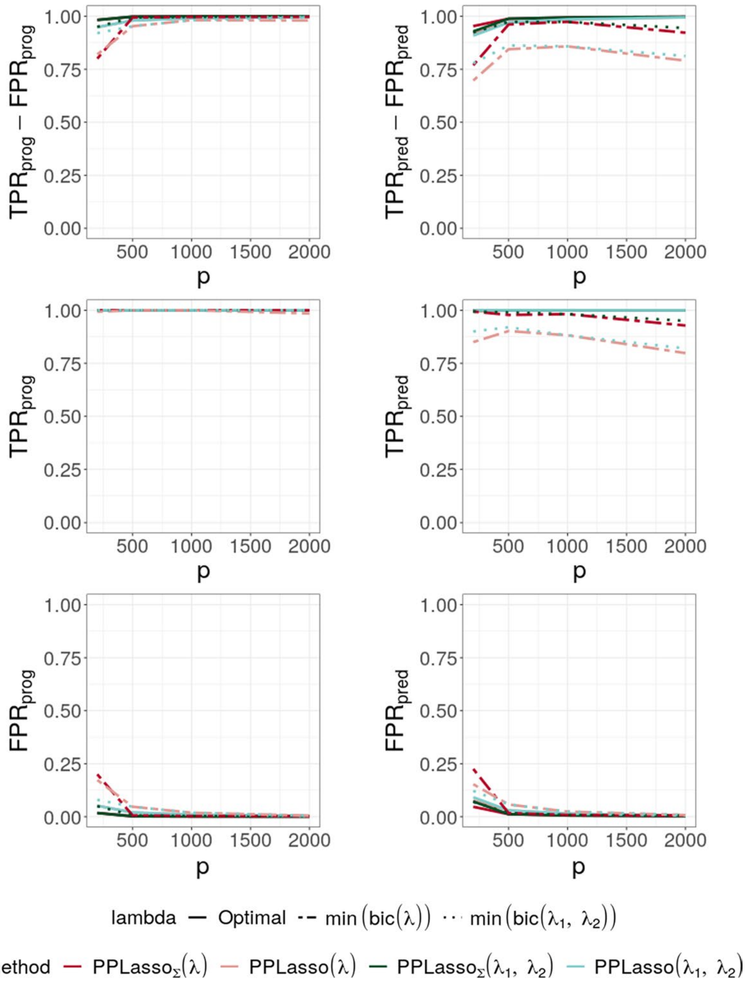

Fig. 4 Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers. Two parameters λ_1 and λ_2 v.s. λ

selection accuracy than the other three methods. Although Elastic Net showed higher TPR than Lasso and Adaptive Lasso, they all failed in selecting all truly prognostic and predictive biomarkers, and the number of missed active biomarkers increased with the dimension p. For example, for Elastic Net, $\text{TPR}_{\text{prog}} = 0.85$ and 0.53, $\text{TPR}_{\text{pred}} = 0.81$ and 0.61 for $p = 200$ and 2000, respectively.

Impact of the correlation matrix $\boldsymbol{\Sigma}$

To evaluate the impact of the correlation matrix on the selection performance of the methods, additional scenarios are presented where the IC is satisfied:

1. Compound symmetry structure where all biomarkers are equally correlated with a correlation $\rho = 0.5$;
2. Independent setting where $\boldsymbol{\Sigma}_{bm}$ is the identity matrix.

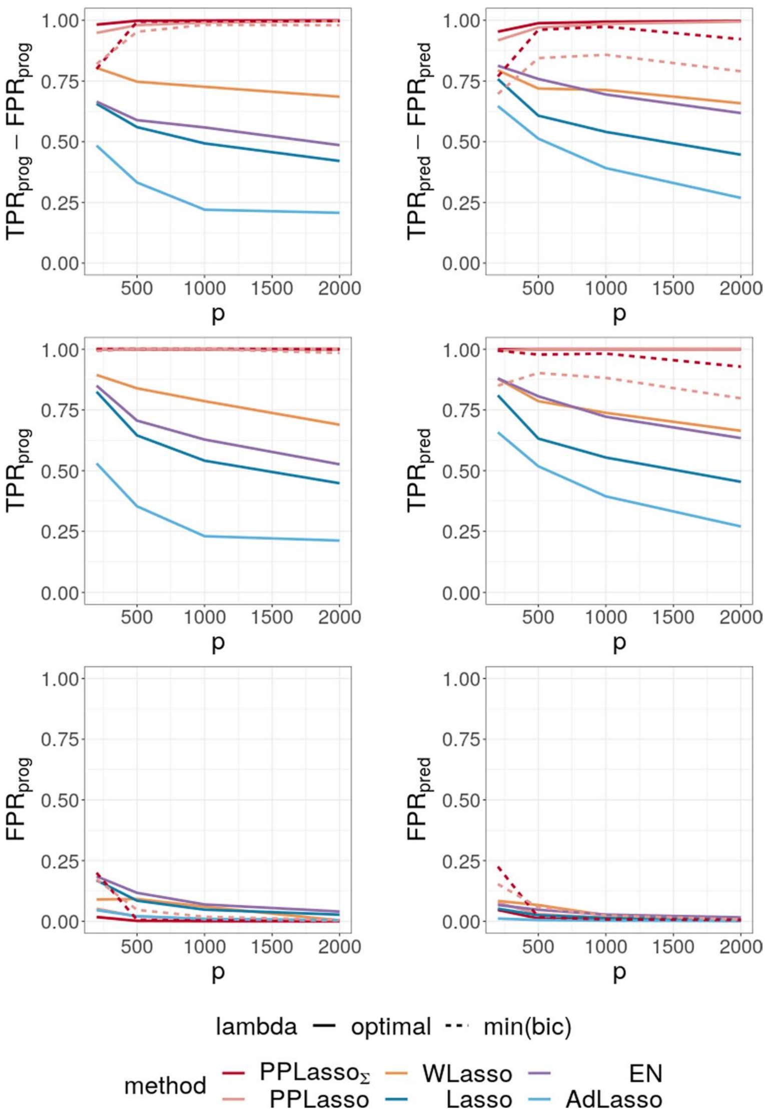

Fig. 5 Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers

For the scenario with compound symmetry structure displayed in Fig. 6, all the methods successfully identified the true prognostic biomarkers (TPR_{prog} close to 1 even for large p) with limited false positive selection. On the other hand, the compared methods (Lasso, ELastic Net, Adaptive Lasso and WLasso) missed some predictive biomarkers especially when p increases.

On the contrary, PPLasso successfully identified almost all predictive biomarkers with the optimal choice of λ. Moreover, even when λ is selected by minimizing the BIC criterion (min(bic)), $PPLasso_{est}$ outperformed Lasso and Adaptive Lasso when $p > 500$ with relatively stable TPR_{pred} and FPR_{pred} as p increases.

For the independent setting, as displayed in Fig. 7, prognostic biomarkers were globally well identified by all the compared methods with a slightly higher TPR_{prog} for Lasso and ELastic Net as compared to PPLasso but also with a slightly higher FPR_{prog}.

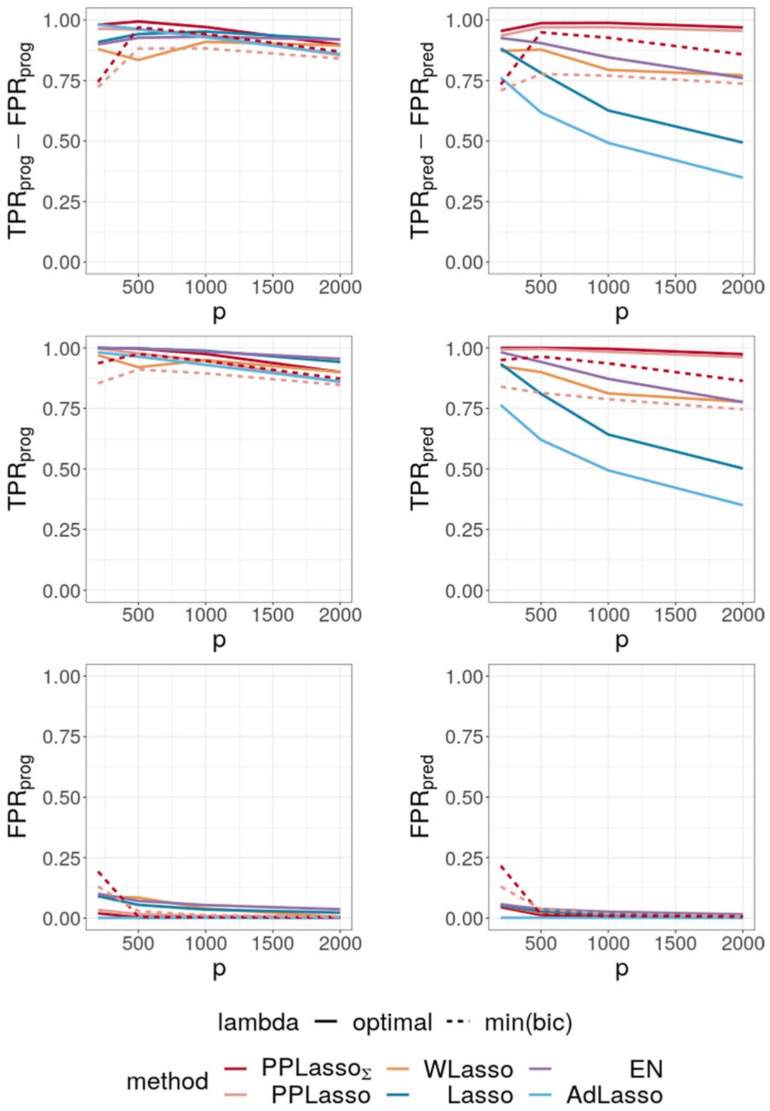

Fig. 6 Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers for the compound symmetry correlation structure

With regards to predictive biomarkers, PPLasso using $\mathbf{\Sigma}_{bm}$ (oracle) performed also similarly to the Lasso, which is reasonable since no transformation has been used in PPLasso. On the other hand, even if PPLasso with λ selected with "min(bic)" performed similarly with PPLasso with optimal λ for relatively small p, the selection performance is altered for large p and even if the performance is higher than Lasso and Adaptive Lasso, it is smaller than the one of Elastic Net.

Impact of the effect size of active biomarkers

To evaluate the impact of the effect size on biomarker selection performance, the scenario presented in "Simulation setting" section was considered with different values of b_2: 1.5, 2 and 2.5.

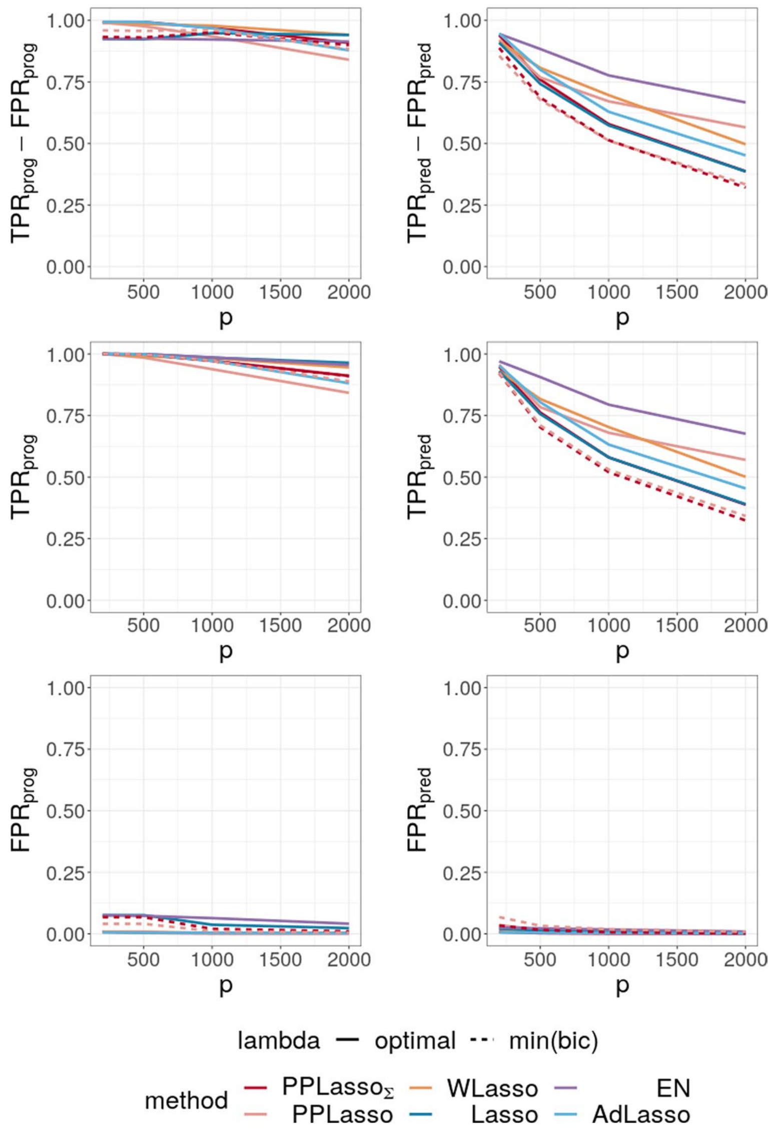

Fig. 7 Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers (independent setting)

Since the effect size of prognostic biomarkers did not change, the comparison focused on predictive biomarkers. As expected, the reduction of the effect size makes the biomarker selection harder, especially for Lasso, Elastic Net and Adaptive Lasso where the predictive biomarker selection is limited when $b_2 = 1.5$: for Lasso when $p = 2000$, $\text{TPR}_{\text{pred}} = 0.45$ (resp. 0.22) for $b_2 = 2$ (resp. 1.5), see Fig. 5 and Additional file 1: Figure S2. The selection performance of PPLasso when λ is selected with min(bic) is also reduced by decreasing b_2, especially when $\mathbf{\Sigma}_{bm}$ is also estimated. Nevertheless, the selection performance of PPLasso remains better than for most of the other compared methods for which the performance displayed are associated to the optimal value of λ. Surprisingly, WLasso performed better than PPLasso with estimated λ in this scenario. On the other hand, even with limited effect size, PPLasso with optimal λ identified all predictive biomarkers with very limited false positive selection. When b_2 was increased to 2.5, the selection performance for all methods is improved and the results for PPLasso with estimated λ was close to the ones with the optimal λ as displayed in Additional file 1: Figure S3. As compared with PPLasso, for which the selection performance remained stable as p increased, Lasso, Elastic Net, Adaptive Lasso and WLasso were more impacted by the value of p since the true positive selection decreased as p increased. As an example, for the Lasso, TPR_{pred} =0.95 (resp. 0.65) for $p = 200$ (resp. 2000).

Impact of the number of predictive biomarkers

The impact of the number of true predictive biomarkers was assessed by increasing the number of predictive biomarkers from 5 to 10 in the scenario presented in "Simulation setting" section. When the number of predictive biomarkers increased, the impact on PPLasso is almost negligible, especially for prognostic biomarker identification. However, for the other methods, we can see from Additional file 1: Figure S4 that it became even harder to identify predictive biomarkers. The impact on WLasso was less obvious, while for the other methods, TPR_{pred} decreased compared to Fig. 5, especially for large p (e.g. $\text{TPR}_{\text{pred}} = 0.12$, 0.18, and 0.02 for Lasso, Elastic Net and Adaptive Lasso respectively when $p = 2000$).

Impact of the dimension of the dataset

In this section, we studied a different sample size: $n = 50$ with $n_1 = n_2 = 25$ and a different number of biomarkers: $p = 5000$.

We can see from Additional file 1: Figure S5 that for $p = 5000$, the selection performance of PPLasso is not altered as compared with $p = 2000$ while the compared methods have more difficulties to identify both prognostic and predictive biomarkers.

When the sample size is smaller ($n = 50$), we can see from Additional file 1: Figure S6 that the ability to identify prognostic and predictive biomarkers decreased for all the methods. However, PPLasso still outperformed the others with higher TPR_{prog} and TPR_{pred} and lower FPR_{prog} and FPR_{pred}.

Application to transcriptomic and proteomic data

Application to the RV144 clinical trial transcriptomic data

We applied the previously described methods to publicly available transcriptomic data from the RV144 vaccine trial [24]. This trial showed reduced risk of HIV-1 acquisition by 31.2% with vaccination with ALVAC and AIDSVAX as compared to placebo. Transcriptomic profiles of in vitro HIV-1 Env-stimulated peripheral blood mononuclear cells (PBMCs) obtained pre-immunization and 15 days after the immunization (D15) from both 40 vaccinees and 10 placebo recipients were generated to better understand underlying biological mechanisms.

For illustration purpose, the absolute change at D15 in gene mTOR was considered as the continuous endpoint (response). mTOR plays a key role in mTORC1 signaling pathway which has been shown to be associated with risk of HIV-1 acquisition [14, 1]. The gene expression has been normalized as in the original publication of [14]. After removing non-annotated genes (LOCxxxx and HS.xxxx), the top 2000 genes with the highest empirical variances were included as candidate biomarkers for prognostic and predictive identification from PPLasso and the compared methods. The penalty parameter λ for the Lasso and Adaptive Lasso, the parameters λ and α for Elastic Net were selected through the classical cross-validation approach. For PPLasso, λ was selected based on the criterion described in "Choice of the parameters λ_1 and λ_2" section .

The estimation of $\mathbf{\Sigma}$ was obtained by comparing several candidate estimators from the `cvCovEst` R package and by selecting the estimator having the smallest estimation error. In this application, the combination of the sample covariance matrix and a dense target matrix (*denseLinearShrinkEst*) derived by [18] provides the smallest estimation error. Figure 8 (left) displays the estimated $\mathbf{\Sigma}$ and highlights the strong correlation between the genes. Additional file 1: Table S2 gives details on the compared estimators.

Prognostic and predictive genes selected by PPLasso, Lasso, Elastic Net and Adaptive Lasso are listed in Table 1. The number of genes selected are similar for all the compared methods, except for a slightly higher number of predictive genes selected by PPLasso. Lasso, Elastic Net and Adaptive Lasso selected very similar sets of prognostic and predictive genes. The intersection between PPLasso and others is moderate (2 prognostic genes (SLAMF7 and TNFRSF6B), 3 predictive genes (YTHDC1, MS4A7 and RPL21)).

To have a better overview of the prognostic and predictive genes selected by the different methods and their associated roles, pathway analysis was carried out via the REACTOME tool (https://reactome.org/), where over-representation analysis (ORA) was performed. ORA is used to determine if a set of genes shares more genes with a pathway than we would expect by chance, evaluated by a p-value. Table S3 (prognostic biomarkers) and Additional file 1: Table S4 (predictive biomarkers) showed the identified pathways with a p-value smaller than 0.01. For prognostic biomarkers, there was no pathway identified by WLasso. Most of the pathways identified by Adaptive Lasso were also identified by Elastic Net. Lasso identified a large number of pathways, but some of them may not be related to HIV. PPLasso identified three pathways (also identified by Elastic Net and Lasso). Interestingly, TNFR2 non-canonical NF-kB pathway that was already identified by Fourati et al. [14], is associated with the risk of HIV acquisition in the placebo group; the implication of regulatory T-cells on HIV-1 has also been widely discussed in the literature (e.g., [17]). For predictive biomarkers identified by different methods, Lasso, Elastic Net, and Adaptive Lasso identified comparable pathways, while PPLasso and WLasso share similar ones and different from the other methods. Among the pathways identified by PPLasso, NOD1/2 Signaling Pathway and Toll-like Receptor Cascades pathway are reported as potential targeted adjuvants for HIV-1 vaccines [22]. In addition, RIPK1-mediated regulated necrosis pathway has also been investigated as targets for HIV-1 protease activity during infection [33].

Application to the NCT01241591 clinical trial proteomic data

Baseline blood samples of 173 samples ($n = 81$ and 92) were taken from patients included in a randomized phase 3 clinical trial comparing the efficacy and safety of tofacitinib and etanercept in moderate-to-severe chronic plaque psoriasis (https://clinicaltrials.gov/ct2/show/NCT01241591) [2]. From these samples, 92 inflammation-associated

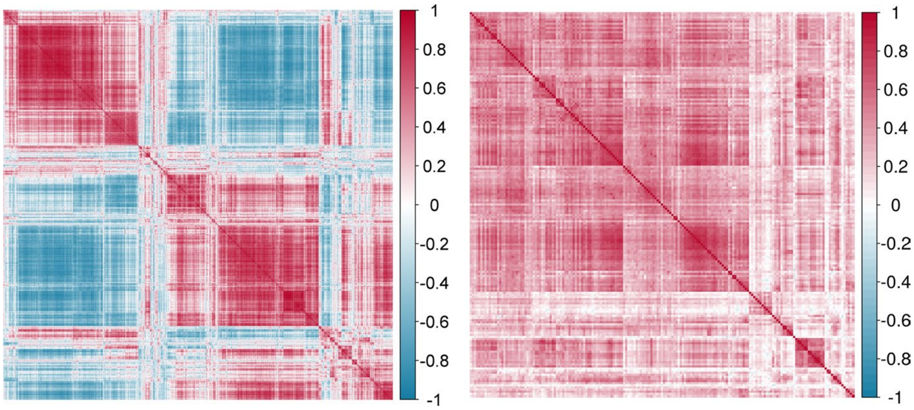

Fig. 8 Heatmaps of the correlation matrices estimated by the `cvCovEst` R package. Left: transcriptomic data from clinical trial RV144. Right: proteomic data from clinical trial NCT01241591

Table 1 Selected genes from PPLasso, Lasso, Elastic Net, Adaptive Lasso and WLasso

	Prognostic genes	Predictive genes
PPLasso	HAPLN3, **SLAMF7**, GTF3C5, FAM46A, SH3PXD2B, TM4SF1, **TNFRSF6B**, TNFRSF18, TRPM2	TLR8, **YTHDC1**, NUCKS1, **BIRC3**, SLAMF7, NFATC2IP, BOK, MGRN1, KIAA0492, SLC25A36, HMGN2, P2RY5, **RPL21**, **MS4A7**, RPL12P6
Lasso	DKFZp434K191, NUCKS1, MAFF, **SLAMF7**, HIST2H2AC, HIST1H4C, IL8, **TNFRSF6B**, TNFRSF18, SCAND1	DKFZp434K191, **YTHDC1**, VMO1, BOLA2, HIST1H4C, **RPL21**, **MS4A7**
Elastic Net	DKFZp434K191, NUCKS1,SNURF, MAFF, **SLAMF7**, IL8, ZBP1, **TNFRSF6B**, ZAK, TNFRSF18, SCAND1, NME1-NME2, DNM1L, RNF146, NPEPL1	DKFZp434K191, **YTHDC1**, PMP22, VMO1, BOLA2, HIST1H4C, **RPL21**, **MS4A7**,RAB11FIP1
Adaptive Lasso	NUCKS1,SNURF, MAFF, **SLAMF7**, IL8, ZBP1, **TNFRSF6B**, NME1-NME2, DNM1L, RNF146	**YTHDC1**, PMP22, VMO1, BOLA2, HIST1H4C, **MS4A7**, **RPL21**
WLasso	**SLAMF7**, EEF1A1P22, RPL21P87, LRRN3, MYOM2 RPS2P5, NME1-NME2, DNM1L, RNF14	**YTHDC1**, SCARA2, KSP37, **BIRC3**

Commonly selected genes are in bold

proteins and 65 cardiovascular disease-associated proteins were measured. Response to treatment was evaluated on the change in PASI score from baseline to week 12. The aim of this application is to identify potential prognostic (proteins associated with the clinical endpoint under standard therapy: etanercept) and predictive biomarkers (proteins differentially associated with the clinical endpoint between etanercept and tofacitinib, aiming to identify patients more likely to benefit from a specific treatment). Figure 8 (right) displays the estimated $\boldsymbol{\Sigma}$ and shows positive correlations between the proteins after standardization. Prognostic and predictive proteins selected by different methods are listed in Table 2. Among the identified prognostic proteins, IL-8 (identified by PPLasso and WLasso) and IL-17C (identified by Elastic Net and WLasso) both contribute to the IL-17 pathway of psoriasis pathogenesis mechanism [4]. For predictive proteins,

Table 2 Selected proteins from PPLasso, Lasso, Elastic Net, Adaptive Lasso and WLasso

	Prognostic	Predictive
PPLasso	**IL-8**, SLAMF1, **IL-24**, TIE2, NT-pro-BNP	CCL19, CD40-L, CSF-1
Lasso	–	–
Elastic Net	**IL-17C**, SCF, IL-18, IL-18R1, IL-12B, **IL-24**, CCL28, DNER, CD40-L, hK11, MMP-3, mAmP, LEP	AM, IL27-A
Adaptive Lasso	–	ECP
WLasso	**IL-8**, **IL-17C**, MCP-1, SIRT2, CCL20	ARTN, IL-20, REN, AGRP

All models are adjusted on the baseline PASI score

Proteins commonly identified by at least two methods are in bold

CSF-1, identified by PPLasso, has been shown to be in the core signatures to predict tofacitinib treatment response developed by Tomalin et al. [32]. Lasso selected no proteins. Adaptive Lasso selected only one predictive protein.

Conclusion

We propose a new method named PPLasso to simultaneously identify prognostic and predictive biomarkers. PPLasso is particularly interesting for dealing with high dimensional omics data when the biomarkers are highly correlated, which is a framework that has not been thoroughly investigated yet. From various numerical studies with or whithout strong correlation between biomarkers, we highlighted the strength of PPLasso in well identifying both prognostic and predictive biomarkers with limited false positive selection. The current method is only dedicated to the analysis of continuous responses through ANCOVA type models. However, it will be the subject of a future work to extend it to other challenging contexts, such as classification or survival analysis.

Abbreviations

Lasso	Least absolute shrinkage and selection operator
PPLasso	Prognostic Predictive Lasso
BIC	Bayesian information criterion
MSE	Mean squared error
TPR	True positive rate
FPR	False positive rate
EN	Elastic net
ANCOVA	Analysis of covariance
PASI	Psoriasis area and severity Index

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-023-05143-0.

Additional file 1. Supplementary material.

Author contributions

WZ derived the algorithms, developed the package, performed the analysis, and wrote a first version of the manuscript. CL supervised the work, was involved in theoretical investigations, gave important suggestions and participated to the writing of the paper. NT initialized the development of the algorithms, supervised the work and edited the manuscript. All authors read and approved the final manuscript.

Funding

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

Availability of data and materials

The transcriptomic dataset is available in the Gene Expression Omnibus (GEO) database, with number GSE103671 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE103671). The proteomic dataset is available in the Gene Expression Omnibus (GEO) database, with number GSE136435 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE136435). PPLasso is an R package that is freely available on the Comprehensive R Archive Network (https://cran.r-project.org/src/contrib/Archive/PPLasso/), with vignette included.

Declarations

Competing interests

The authors declare that they have no competing interests.

Received: 30 July 2022 Accepted: 9 January 2023
Published online: 23 January 2023

References

1. Akbay B, Shmakova A, Vassetzky Y, Dokudovskaya S. Modulation of mTORC1 signaling pathway by HIV-1. Cells. 2020;9:1090.
2. Bachelez H, Peter CM, Robert S, Alexey K, Fernando V, Joo-Heung L, Vladimir Y. Tofacitinib versus etanercept or placebo in moderate-to-severe chronic plaque psoriasis: a phase 3 randomised non-inferiority trial. Lancet. 2015;386(9993):552–61.
3. Ballman KV. Biomarker: predictive or prognostic? J Clin Oncol. 2015;33(33):3968–71.
4. Blauvelt A, Chiricozzi A. The immunologic role of IL-17 in psoriasis and psoriatic arthritis pathogenesis. Clin Rev Allergy Immunol. 2018;55(3):379–90.
5. Boileau P, Hejazi NS, van der Laan MJ, Dudoit S. Cross-validated loss-based covariance matrix estimator selection in high dimensions. 2021. arXiv preprint arXiv:2102.09715.
6. Boileau P, Hejazi NS, van der Laan MJ, Dudoit S. cvCovEst: cross-validated covariance matrix estimator selection and evaluation in R. J Open Source Softw. 2021;6(63):3273.
7. Cai T, Zhang C-H, Zhou H. Optimal rates of convergence for covariance matrix estimation. Ann Stat. 2010;38:2118–44.
8. Clark G. Prognostic factors versus predictive factors: examples from a clinical trial of erlotinib. Mol Oncol. 2008;1:406–12.
9. Fan J, Li R. Statistical challenges with high dimensionality: feature selection in knowledge discovery. In: Proc. Madrid Int. Congress of Mathematicians; 2006. p. 3.
10. Fan J, Lv J. A selective overview of variable selection in high dimensional feature space. Stat Sin. 2009;20(1):101–48.
11. Fan J, Liao Y, Mincheva M. Large covariance estimation by thresholding principal orthogonal complements. J R Stat Soc Ser B Stat Methodol. 2013;75:603–80.
12. Faraway JJ. Practical regression and ANOVA using R. University of Bath. 2002.
13. Foster J, Taylor J, Ruberg S. Subgroup identification from randomized clinical trial data. Stat Med. 2011;30:2867–80.
14. Fourati S, Ribeiro S, Blasco Lopes F, Talla A, Lefebvre F, Cameron M, Kaewkungwal J, Pitisuttithum P, Nitayaphan S, Rerks-Ngarm S, Kim J, Thomas R, Gilbert P, Tomaras G, Koup R, Michael N, McElrath M, Gottardo R, Sékaly R. Integrated systems approach defines the antiviral pathways conferring protection by the RV144 HIV vaccine. Nat Commun. 2019;10:1–12.
15. Giannos P, Kechagias K, Gal A. Identification of prognostic gene biomarkers in non-small cell lung cancer progression by integrated bioinformatics analysis. Biology. 2021;10(11):1200.
16. He Y, Luo Y, Huang L, Zhang D, Wang X, Ji J, Liang S. New frontiers against sorafenib resistance in renal cell carcinoma: from molecular mechanisms to predictive biomarkers. Pharmacol Res. 2021;170(9): 105732.
17. Kleinman AJ, Sivanandham R, Pandrea I, Chougnet CA, Apetrei C. Regulatory T cells as potential targets for HIV cure research. Front Immunol. 2018;9:734.
18. Ledoit O, Wolf M. The power of (non-)linear shrinking: a review and guide to covariance matrix estimation. J Financ Econom. 2020;20(323):1–32.
19. Lipkovich I, Dmitrienko A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. J Biopharm Stat. 2014;24:130–53.
20. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search (sides)—a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med. 2011;30:2601–21.
21. Lipkovich I, Dmitrienko A, D'Agostino Sr RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med. 2017;36(1):136–96.
22. Liu J, Ostrowski M. Development of targeted adjuvants for HIV-1 vaccines. AIDS Res Ther. 2017;14(1):43.
23. McDonald J. Handbook of biological statistics. 2nd ed. Baltimore: Sparky House Publishing; 2009.
24. Rerks-Ngarm S, Pitisuttithum P, Nitayaphan S, Kaewkungwal J, Chiu J, Paris R, Premsri N, Namwat C, De Souza M, Benenson M, Gurunathan S, Tartaglia J, McNeil J, Francis D, Stablein D, Birx D, Chunsuttiwat S, Khamboonruang C, Kim J. Vaccination with ALVAC and AIDSVAX to prevent HIV-1 infection in Thailand. N Engl J Med. 2009;361:2209–20.
25. Saeys Y, Inza I, Larranaga P. A review of feature selection techniques in bioinformatics. Bioinformatics. 2007;23(19):2507–17.
26. Sechidis K, Papangelou K, Metcalfe PD, Svensson D, Weatherall J, Brown G. Distinguishing prognostic and predictive biomarkers: an information theoretic approach. Bioinformatics. 2018;34(19):3365–76.

27. Smith G. Step away from stepwise. J Big Data. 2018;5(32):1–12.
28. Ternès N, Rotolo F, Heinze G, Michiels S. Identification of biomarker-by-treatment interactions in randomized clinical trials with survival outcomes and high-dimensional spaces. Biom J. 2016;59:685–701.
29. Tian L, Alizadeh A, Gentles A, Tibshirani R. A simple method for estimating interactions between a treatment and a large number of covariates. J Am Stat Assoc. 2012;109(508):1517–32.
30. Tibshirani R. Regression shrinkage and selection via the lasso. J R Stat Soc Ser B (Stat Methodol). 1996;58(1):267–88.
31. Tibshirani RJ, Taylor J. The solution path of the generalized lasso. Ann Stat. 2011;39(3):1335–71.
32. Tomalin L, Kim J, Correa da Rosa J, Lee J, Fitz L, Berstein G, Valdez H, Wolk R, Krueger J, Suárez-Fariñas M. Early quantification of systemic inflammatory proteins predicts long-term treatment response to tofacitinib and etanercept. J Investig Dermatol. 2020;140(5):1026–34.
33. Wagner RN, Reed JC, Chanda SK. HIV-1 protease cleaves the serine-threonine kinases RIPK1 and RIPK2. Retrovirology. 2015;12:74.
34. Wang H, Lengerich B, Aragam B, Xing E. Precision lasso: accounting for correlations and linear dependencies in high-dimensional genomic data. Bioinformatics. 2019;35(7):1181–7.
35. Wang X, Leng C. High dimensional ordinary least squares projection for screening variables. J R Stat. 2016;78(3):589–611.
36. Windeler J. Prognosis—what does the clinician associate with this notion? Stat Med. 2000;19:425–30.
37. Xue F, Qu A. Variable selection for highly correlated predictors. 2017. arXiv preprint arXiv:1709.04840.
38. Zhao N, Guo M, Wang K, Zhang C, Liu X. Identification of pan-cancer prognostic biomarkers through integration of multi-omics data. Front Bioeng Biotechnol. 2020;8:268.
39. Zhao P, Yu B. On model selection consistency of lasso. J Mach Learn Res. 2006;7:2541–63.
40. Zhu W, Lévy-Leduc C, Ternès N. A variable selection approach for highly correlated predictors in high-dimensional genomic data. Bioinformatics. 2021;37(16):2238–44.
41. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Ser B (Statistical Methodology). 2005;67(2):301–20.
42. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101(476):1418–29.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.