Development of machine learning approaches in precision medicine for the identification of prognostic and predictive biomarkers based on high-dimensional omics data

Wencan Zhu

université PARIS-SACLAY

Development of machine learning approaches in precision medicine for the identification of prognostic and predictive biomarkers based on high-dimensional omics data

Développement de méthodes d'apprentissage statistique pour l'identification de biomarqueurs pronostiques et prédictifs à l'aide de données "-omiques" de grande dimension dans le domaine de la médecine de précision

Thèse de doctorat de l'université Paris-Saclay

École doctorale n° 574, mathématiques Hadamard (EDMH)
Spécialité de doctorat : Mathématiques appliquées
Graduate School : Mathématiques
Référent : Faculté des sciences d'Orsay

Thèse préparée dans l'**UMR MIA Paris-Saclay** (Université Paris-Saclay, AgroParisTech, INRAE), sous la direction de **Céline LÉVY-LEDUC**, professeure de statistique à AgroParisTech, et la co-supervision de **Nils TERNÈS**, statisticien biomarqueur à Sanofi R&D

Thèse soutenue à Paris-Saclay, le 26 septembre 2022, par

Wencan ZHU

Composition du jury

Liliane BEL Professeure, AgroParisTech/INRAE MIA Paris-Saclay	Présidente
Anne-Laure BOULESTEIX Professeure, Faculté de médecine, Université Ludwig-Maximilians (LMU) Munich	Rapporteur & Examinatrice
Pierre NEUVIAL DR CNRS, Institut de Mathématiques de Toulouse (IMT)	Rapporteur & Examinateur
Riccardo DE BIN Maître de conférences, Département de mathématiques, Université d'Oslo	Examinateur
Céline LÉVY-LEDUC Professeure, AgroParisTech/INRAE MIA Paris-Saclay	Directrice de thèse

THESE DE DOCTORAT
NNT : 2022UPASM021

université PARIS-SACLAY

ÉCOLE DOCTORALE
de mathématiques
Hadamard (EDMH)

Titre : Développement de méthodes d'apprentissage statistique pour l'identification de biomarqueurs pronostiques et prédictifs à l'aide de données "-omiques" de grande dimension dans le domaine de la médecine de précision

Mots clés : données de grande dimension ; sélection de variables ; biomarqueurs pronostiques/prédictifs ; méthodes régularisées ; apprentissage automatique ; médecine de précision.

Résumé : Avec la révolution génomique et l'arrivée de la médecine de précision, l'identification de biomarqueurs qui sont explicatifs (biomarqueurs actifs) d'une réponse clinique devient de plus en plus importante dans la recherche clinique. Ces biomarqueurs sont utiles pour mieux comprendre la progression d'une maladie (biomarqueurs pronostiques) et pour mieux identifier les patients les plus susceptibles de bénéficier d'un traitement donné (biomarqueurs prédictifs). Les données relatives aux biomarqueurs (génomique, transcriptomique et protéomique, par exemple) sont en général de grande dimension, le nombre de biomarqueurs mesurés (variables) étant beaucoup plus important que la taille de l'échantillon. Cependant, seule une fraction des biomarqueurs est réellement active, d'où la nécessité de sélectionner les variables. Parmi les divers algorithmes d'apprentissage statistique, les approches régularisées telles que le Lasso sont très utilisées pour faire de la sélection de variables dans des contextes de grande dimension en raison de leurs performances statistiques et numériques. Cependant, la consistance de leur sélection n'est pas garantie lorsque les biomarqueurs sont fortement corrélés. Au cours de ma thèse, plusieurs nouvelles approches ont été développées pour effectuer la sélection de variables dans ce contexte difficile. Plus précisément, quatre méthodes sont mises en place sous différents modèles statistiques (modèle de régression linéaire, modèle de type ANCOVA et modèle de régression logistique). L'idée principale est de supprimer les corrélations en blanchissant la matrice de design. Pour l'une d'entre elles, des résultats de la consistance en signe ont été obtenus sous des hypothèses peu restrictives. Les approches proposées ont été évaluées par des études de simulation et appliquées à des données publiques. Les résultats montrent que les performances statistiques de nos méthodes sont meilleures que celles de l'état de l'art. Nos méthodes sont implémentées dans les packages R suivants : `WLasso`, `PPLasso`, et `WLogit`.

université PARIS-SACLAY | ÉCOLE DOCTORALE de mathématiques Hadamard (EDMH)

Title: Development of machine learning approaches in precision medicine for the identification of prognostic and predictive biomarkers based on high-dimensional omics data

Keywords: high-dimensional data; variable selection; prognostic/predictive biomarkers; regularized approaches; machine learning; precision medicine.

Abstract: With the genomic revolution and the new era of precision medicine, the identification of biomarkers that are informative (i.e. active) for a response (endpoint) is becoming increasingly important in clinical research. These biomarkers are beneficial to better understand the progression of a disease (prognostic biomarkers) and to better identify patients more likely to benefit from a given treatment (predictive biomarkers). Biomarker data (e.g. genomics, transcriptomics, and proteomics) usually have a high-dimensional nature, with the number of measured biomarkers (variables) much larger than the sample size. However, only a fraction of biomarkers is truly active, therefore raising the need for variable selection. Among various statistical learning approaches, regularized methods such as Lasso have become very popular for high-dimensional variable selection due to their statistical and numerical performance. However, their selection consistency is not guaranteed when the biomarkers are highly correlated. Throughout my PhD, several novel regularized approaches were developed to perform variable selection in this challenging context. More precisely, four methods were proposed in different statistical models (linear regression model, ANCOVA-type model, and logistic regression model). The main idea is to remove the correlations by whitening the design matrix. For one of the methods, results of the sign consistency were established under mild conditions. The proposed approaches were evaluated through simulation studies and applications on publicly available datasets. The results suggest that our approaches are more performant than compared methods for selecting prognostic and predictive biomarkers in high-dimensional (correlated) settings. Three of our methods are implemented in the R packages: `WLasso`, `PPLasso`, and `WLogit`, available from the CRAN (Comprehensive R Archive Network).

Remerciements

J'aimerais tout d'abord remercier ma directrice et mon encadrant de thèse.

Je remercie Céline Lévy-Leduc pour avoir dirigé mes recherches. Je la remercie pour toute la patience et la disponibilité dont elle a fait preuve à mon égard. Ses conseils et remarques constructives m'ont permis d'améliorer grandement la qualité de mes travaux et de ce mémoire. Je lui suis également reconnaissant pour le temps conséquent qu'elle m'a accordé, ses qualités pédagogiques et scientifiques, sa franchise et sa sympathie. J'ai beaucoup appris à ses côtés et je lui adresse ma gratitude pour tout cela.

Je remercie Nils Ternès pour m'avoir encadré. Je le remercie pour la confiance qu'il m'a témoignée tout au long de ces années, qui a été moteur pour moi. Je le remercie également pour tous ses conseils et remarques constructives qui ont été prépondérants pour la bonne réussite de cette thèse. Son ouverture d'esprit, sa franchise, sa gentillesse sont autant d'éléments qui m'ont permis d'atteindre les objectifs de l'entreprise dans le cadre du doctorat. J'ai pris un grand plaisir à travailler avec lui.

Je désire en outre remercier Eric Adjakossa pour sa contribution à l'article sur le generalized Elastic Net, notamment son aide dans les développements théoriques. Sa générosité et sa gentillesse m'ont beaucoup touchée.

Je remercie Liliane Bel d'avoir accepté de présider mon jury de thèse. Je remercie également grandement Anne-Laure Boulesteix et Pierre Neuvial d'avoir accepté d'être rapporteurs de ma thèse. Je les remercie pour tous leurs précieux commentaires et suggestions qui m'ont permis d'améliorer la qualité de ce manuscrit. Enfin, ma gratitude s'adresse également à Riccardo De Bin pour avoir accepté d'examiner ce travail.

Je remercie Sanofi R&D et l'ANRT d'avoir financé ce travail de thèse.

Bien évidemment, je remercie l'ensemble du service de biostatistique et programming de Sanofi pour leur aide. Je remercie Caroline Paccard, mon ancienne manager, pour m'avoir accueillie au sein de son équipe. J'exprime en particulier ma gratitude à Inoncent pour son soutien, à Emilie pour son support sur les projet, à Annick pour ses connaissances en bioinformatique, à Vinh pour les discussion sur les algorithmes. Enfin, je remercie mon co-bureau Antoine qui m'a fourni toujours tout ce que je cherchais : informatique, labo médicale, restaurants, plongée etc.

J'exprime en particulier ma gratitude à Françoise, notre assistante, pour sa sympathie et son efficacité dans l'organisation et la résolution des problèmes administratifs, et aussi pour les bons moments hors du travail : repas de Noël, Flamenco, café etc. J'adore ses confitures faites maison.

Je voudrais également remercier mes amis à Sanofi : Emma, Rana, Jihane et Gabriel. J'apprécie leur organisation des repas, soirées, sorties qui m'ont permis d'oublier momentanément le travail. Les rapports humains dont j'ai profité à leur côté ont fait naître de réels liens d'amitiés qui a mes yeux n'ont pas de prix.

Je tiens également à remercier tous les personnes de l'équipe SOLsTis de l'UMR MIA Paris-Saclay (Julie, Julien, Christelle...) pour leur aide, leur soutien et leur gentillesse.

Je remercie tous les thésards de l'Agro pour la bonne ambiance de travail mais également pour les nombreux bons moments passés ensemble. Entre autres : Mary, Marina, Jérémy, Tâm, Saint-Clair, Bastien, Armand...

Mes remerciements vont aussi à ma famille et mes amis pour leur soutien au cours de ces trois années et sans lesquels je n'en serai pas là aujourd'hui.

Contents

Chapter 1 - Introduction

1.1. Biological context

Traditional medicine follows the one-size-fits-all approach to assess drugs and other therapies to cure large groups of people with a similar illness, and the treatment is based on what is most likely to work for patients with similar symptoms. But people do not respond to a treatment in the same way. Some drugs work very well for some people, while do not help for others or even cause side effects (Alberti and Cavaletti, 2014). Compared to traditional medicine, precision medicine is much more targeted; it aims at matching the right treatment to the right patients (Figure 1.1). According to the US National Cancer Institute's definition,

"Precision medicine is a form of medicine that uses information about a person's genes, proteins and environment to prevent, diagnose and treat disease"

With the revolution of biotechniques and the new era of precision medicine, understanding the cause of a disease, finding out drug response mechanisms, and predicting the response to treatment for therapeutic decision-making are becoming increasingly important in medical research. The availability of growing amounts of data from population studies is enabling researchers to have access to patient omics (e.g. genomes) and clinical data, and therefore give access to ever more detailed molecular outlines of the human body (Krassowski et al., 2020). These biological molecules are called the biomarkers; examples include genes in RNAseq data, proteins in proteomics data, and SNPs in GWAS data. However, how to make the best use of these data requires advanced research (Ginsburg and Phillips, 2018).

An important feature of biomarker data is the high dimensionality: the number of biomarkers is usually larger than the sample size. For example, RNAseq data provides over 20,000 gene expressions. Meanwhile, the sample size is generally limited ($10 \sim 100$) due to the cost. Moreover, it is commonly assumed by the scientific community that only a small subset of biomarkers is sufficient to explain the outcome (e.g. clinical endpoint). The identification of such biomarkers is therefore of fundamental and practical interest. It can help to make discovery on novel signaling pathways, predict patients' responses to the therapy and eventually optimize the therapeutical strategy for a given patient. However, separating relevant biomarkers from the background is difficult due to the noise associated with sample collection and assay variability. The presence of complicated interactions between biomarkers also increases the difficulty (Yamada et al., 2020). Novel techniques are needed to discover active biomarkers and better understand diseases at the molecular level.

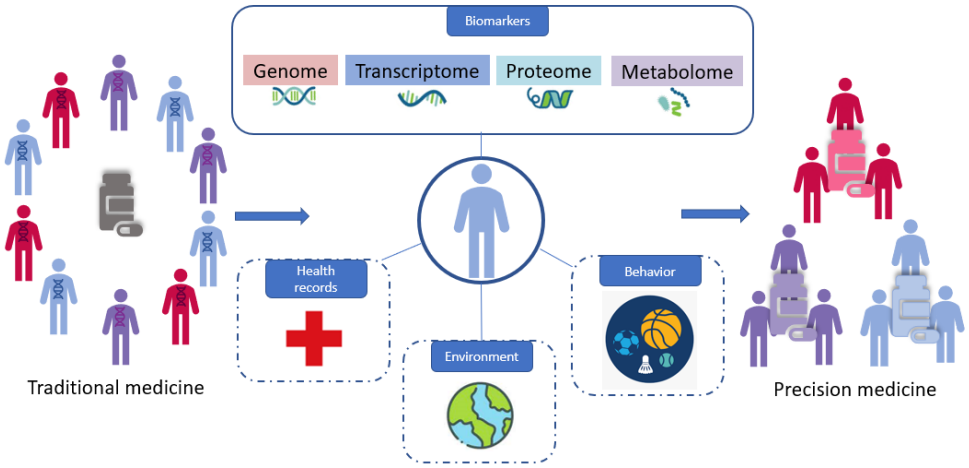

Figure 1.1: Traditional medicine v.s. Precision medicine

1.2. Examples in precision medicine

Over years of research, scientists have learned more about the biological mechanism that controls how diseases start and how they behave. Knowing how biomarkers and illnesses interact has helped to fine-tune treatments to make them work better. According to the glossary of FDA (Food and Drug Administration) and NIH (National Institutes of Health), different types of biomarkers have been defined in the framework of medical development. Here we only focus on two of them: prognostic biomarker and predictive biomarker. A prognostic biomarker informs about a likely disease outcome (e.g. disease recurrence, disease progression, death) independent of treatment received. As illustrated in Figure 1.2, we assume that the biomarker has only two status: positive (B+) and negative (B-), and an experiment treatment (Exp) is compared to a placebo (Pbo)/standard treatment. The clinical endpoint is measured under different treatments and different biomarker status. If a biomarker is prognostic, the response is different according to the biomarker status (positive v.s. negative) but irrelevant to the treatment (placebo v.s. experiment). An example of a prognostic biomarker is BRCA1/2 mutations. For women with breast cancer, BRCA1/2 mutations suggest a higher risk of developing a contralateral breast cancer (Ahmed et al., 2009). To reduce this risk, some women with BRCA1/2 mutation-associated breast cancer undergo contralateral prophylactic mastectomy, and this procedure has been reported to reduce the risk of future contralateral breast cancer by at least 90% (Sprundel et al., 2005; Domchek et al., 2010). Another example is the MammaPrint signature developed in breast cancer (Sotiriou and Pusztai, 2009). MammaPrint uses microarray to measure the expression of 70 genes that are key hallmarks of cancer, based on which patients are separated into two groups, i.e., low or high risk for disease recurrence. In the prospective randomized clinical trial MINDACT study

(Rutgers et al., 2020), patients classified as being at low risk had an excellent outcome of disease-free survival. According to the EGTM (European Group on Tumour Markers) guidelines, the MammaPrint test "may be used for determining prognosis and guiding decision making with respect to the administration of adjuvant chemotherapy in patients with newly diagnosed invasive breast cancer." Prognostic biomarkers indicate the potential risk of the progression of the disease, therefore contribute to decisions about whether or how aggressively to intervene with the treatment.

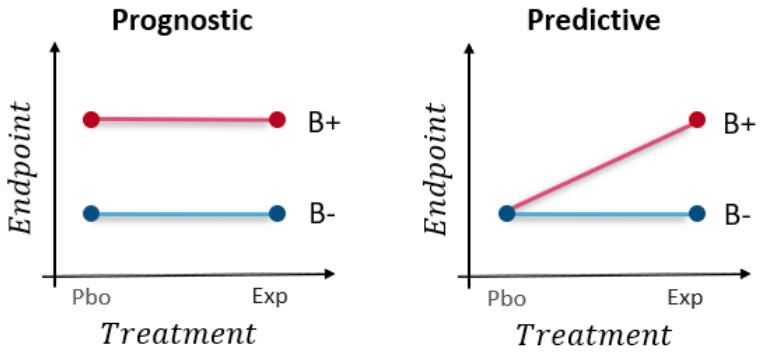

Figure 1.2: Prognostic biomarker (left) v.s. Predictive biomarker (right)

Unlike prognostic biomarkers, a biomarker is considered as predictive if on a given clinical endpoint, the treatment effect (experimental compared with control/standard therapy) is different for biomarker-positive patients compared with biomarker-negative patients. In Figure 1.2, only the B+ group showed an effect from the experiment compared to placebo on the clinical endpoint, while no treatment effect was observed in the B- group. Predictive biomarkers can therefore help to determine which patients might be more likely to respond or be resistant to specific therapies. One example is erlotinib maintenance treatment for advanced non–small cell lung cancer. Patients with tumors harboring an EGFR mutation had a higher survival rate when assigned with erlotinib than with a placebo. In contrast, the EGFR wild-type group showed no clear benefit from erlotinib (Ballman, 2015). Another example is the measurement of HER2 in breast cancer therapy decisions. Overexpression of HER2 leads to tumor growth and enhances cell proliferation and invasion (Rimawi et al., 2015). According to this mechanism, four forms of anti-HER2 therapy are available (Martin and López-Tarruella, 2016), and HER2 gene overexpression appears to be necessary for patients to respond to these treatments. Therefore, only HER2-positive patients can receive anti-HER2 therapies. Apart from therapeutical decision-making, predictive biomarkers are also important in clinical development. For example, in a randomized controlled clinical trial of an investigational therapy, a biomarker can be used to select patients

for enrollment in a clinical trial or to stratify patients into biomarker positive and biomarker negative groups. If the biomarker is predictive of a favorable outcome, then the effect of the investigational therapy compared to a control therapy will be greater in patients with the biomarker. Distinguishing prognostic and predictive biomarkers can be difficult in some cases (Mishina, 2020). Especially with only one treatment presented, the effect on a given clinical endpoint can come from either prognostic or predictive effect or both of them.

The development of precision medicine is changing the way that patients are treated. Vargas and Harris (2016) summarized the lung cancer treatments with precision medicine based on different types of biomarkers (genomic, transcriptomic, epigenomic, proteomic, etc.) and listed promising biomarkers in lung cancer. From a more global point of view, Tsimberidou et al. (2020) revisited the history of the development in precision medicine and indicated that new strategies, including gene-directed therapies, will enable optimization of treatment for individual patients and expedite drug discovery and approval.

The objective of my thesis is therefore to develop novel methods that can correctly select relevant biomarkers in high-dimensional settings, especially when the biomarkers are correlated. We first considered the linear regression case to identify prognostic biomarkers when the endpoint is continuous. With the presence of two comparative treatments, we then developed a novel method to simultaneously identify both prognostic and predictive models by using an ANCOVA type linear model. At last, the linear regression model was extended to the logistic regression model with the purpose of classification when the endpoint is binary.

1.3. Variable selection for high-dimensional data in the linear regression model

1.3.1. Background

We first consider the linear regression model. We denote the continuous responses (endpoint) $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ of n samples, where A^T denotes the transpose of A. Then we consider the following model:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{1.1}$$

where $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is the design matrix containing the expression of p biomarkers ($p \gg n$) and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a sparse vector to estimate, namely with a majority of null coefficients. In Model (1.1), $\boldsymbol{\epsilon}$ is the error term. Variable selection aims at identifying all important variables whose regression coefficients are estimated as non-null. Several reviews can be found on the topic of variable selection (Saeys et al. (2007) and Heinze et al. (2018) for example). To summarize,

following are three classes of methods mainly used for high-dimensional variable selection for omics data.

Univariate test

The univariate approach consists in independently studying each biomarker by evaluating its strength of association with the response in a regression model (McDonald, 2009). However, the multiplicity of the statistical tests can make this approach less powerful (Lee and Lee, 2020). To address this issue, multiple testing corrections have been proposed to make statistical tests more stringent. The best-known adjustment is the Bonferroni correction, and other less conservative ones include Bonferroni-Holm (Holm, 1979) and Hochberg (Hochberg, 1988) techniques. To control the false discovery rate, commonly used adjustments in omics data analysis include Benjamini and Hochberg (1995) and Benjamini and Yekutieli (2001). Although very simple to implement, this approach does not take into account the correlation between biomarkers, which may be an important limitation in the context of genomic data.

Wrapper approaches

Wrapper approaches are other methods to select a subset of variables. The most popular ones are forward, backward and stepwise selection. The choice of variables is carried out by an automatic procedure. In each step, a variable is considered for addition to or subtraction from the set of explanatory variables based on some prespecified criterion (Hocking, 1976). Application of this approach to biomarker selection can be found in Xiong et al. (2001) and Lu et al. (2020). However, these approaches often show a high risk of overfitting and are computationally expensive for high-dimensional data (Smith, 2018).

Penalized regression

Penalized regression is often used in the context of high-dimensional data. It consists in adding a penalty term to the likelihood of Model (1.1). A widely used penalty is a ℓ_1 norm of coefficients, namely the Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996a). It consists in minimizing the following criterion:

$$L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{1.2}$$

where $\|\mu\|_2^2 = \sum_{i=1}^{n} \mu_i^2$ and $\|\mu\|_1 = \sum_{i=1}^{n} |\mu_i|$ for $\mu = (\mu_i, \ldots, \mu_n)^T$.

Other penalty forms include the Elastic Net (Zou and Hastie, 2005), SCAD (Fan and Li, 2001), the Dantzig selector (Candes and Tao, 2007), Nonnegative Garrote (Breiman, 1995), adaptive Lasso (Zou, 2006), grouped Lasso (Yuan and Lin, 2006) as examples. A more general review on the topic of variable selection in high dimensional settings can be found in Fan and Lv (2009). Due to their

attractive ability to perform variable selection and coefficient estimation simultaneously (Fan and Li, 2006), penalized approaches have been widely applied to genomic analysis (Ogutu et al., 2012; Li and Sillanpää, 2012; Desta and Ortiz, 2014). Therefore, this thesis mainly focuses on this type of approach.

1.3.2. High Correlations between biomarkers

A notorious difficulty of model selection in high dimensional frameworks comes from the correlation between the covariates. The correlation can easily be spurious in high-dimensional genomic data, which can lead to the selection of wrong models. Figure 1.3 presents empirical correlations between gene expression data (after preprocessing) in Prostate dataset (Singh et al., 2002) (2135 genes, 102 samples) and Breast cancer dataset (Sotiriou et al., 2006) (1111 genes, 189 samples). We can clearly observe strong correlations in blocks, which means that genes in each block are correlated.

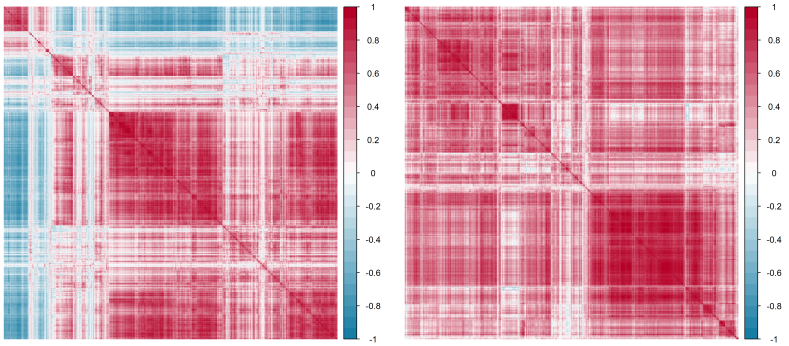

Figure 1.3: Heatmap of correlations between expressions of genes. Prostate cancer (left) and Breast cancer (right) datasets.

Under such high correlations, the Lasso is known to be inconsistent in variable selection, despite its various advantages. Sign consistency ensures that the active variables (non-null coefficients) of $\boldsymbol{\beta}$ are estimated by non-null coefficients with the same sign and that the non-active variables (null coefficients) are estimated by null coefficients. More precisely, an estimator of $\boldsymbol{\beta}$ is sign consistent if

$$\mathbb{P}\left(sign(\widehat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta})\right) \xrightarrow[n\to+\infty]{} 1 \tag{1.3}$$

where $sign(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$.

The consistency of subset selection received considerable attention, and various investigation has been devoted to the model selection consistency of Lasso: Zhao and Yu (2006); Meinshausen and Bühlmann (2006); Lounici (2008); Meinshausen and Yu (2009). During my PhD, I mainly focused on the Irrepresentable Condition

established by Zhao and Yu (2006). The authors 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.

Irrepresentable condition (IC): Let $S = \{j,\ \beta_j \neq 0\}$ be the set of active variables, S^c the set of non-active variables and $\boldsymbol{X}_A$ the submatrix of $\boldsymbol{X}$ containing only the indices of columns which are in the set A. Hence, the empirical covariance matrix of the covariates, $C_n = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, can be rewritten as follows:

$$C_n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix},$$

where $C_{11}^n = n^{-1}\boldsymbol{X}_S^T\boldsymbol{X}_S$, $C_{12}^n = n^{-1}\boldsymbol{X}_S^T\boldsymbol{X}_{S^c}$, $C_{21}^n = n^{-1}\boldsymbol{X}_{S^c}^T\boldsymbol{X}_S$, $C_{22}^n = n^{-1}\boldsymbol{X}_{S^c}^T\boldsymbol{X}_{S^c}$. Then, the design matrix $\boldsymbol{X}$ satisfies the **Irrepresentable Condition** if for some constant $\alpha \in (0, 1)$,

$$\left|\left(C_{21}^n (C_{11}^n)^{-1}\mathrm{sign}(\beta_S)\right)_j\right| \leq 1 - \alpha, \text{ for all } j. \tag{1.4}$$

Intuitively, this condition means that the correlation between the active and non-active explanatory variables is smaller than 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). This phenomenon is typically observed in omics data. 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.

1.3.3. The main idea: whitening

To deal with the issue of high correlations between the biomarkers, several strategies have been proposed. The Elastic Net introduced by Zou and Hastie (2005) combines both the ℓ_1 penalty ($\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\beta_i|$) of the Lasso and the ℓ_2 penalty ($\|\boldsymbol{\beta}\|_2 = \sum_{i=1}^{p} \beta_i^2$). Elastic Net has the grouping effect of selecting groups of correlated variables. Preconditioning is another type of methods to deal with correlation. Jia and Rohe (2015) and Wang and Leng (2016) proposed to left-multiply $\boldsymbol{X}$, $\boldsymbol{y}$ and $\boldsymbol{\epsilon}$ in Model (1.1) by specific matrices to remove the correlations between the columns of $\boldsymbol{X}$. Another recently published method, named Precision Lasso (Wang et al., 2019), proposes to handle the correlation issue by assigning similar weights to correlated variables.

In my thesis, I proposed an alternative and novel technique to remove the correlations that may exist between the predictors (biomarkers). Suppose the n rows $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ of $\boldsymbol{X}$ are assumed to be independent Gaussian random vectors with a covariance matrix equal to Σ. 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 $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$.

Therefore, (1.1) can be rewritten as follows:

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \tag{1.5}$$

where $\widetilde{\boldsymbol{\beta}} = \Sigma^{1/2}\boldsymbol{\beta} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^T\boldsymbol{\beta}$.

With such a transformation, the covariance matrix of the rows of $\widetilde{\boldsymbol{X}}$ is equal to identity and the columns of $\widetilde{\boldsymbol{X}}$ are thus uncorrelated. Figure 1.4 presents the heatmap of the correlations (same datasets as presented in Figure 1.3) after the whitening transformation. The covariance matrix Σ was estimated by package cvCovEst (Boileau et al., 2021).

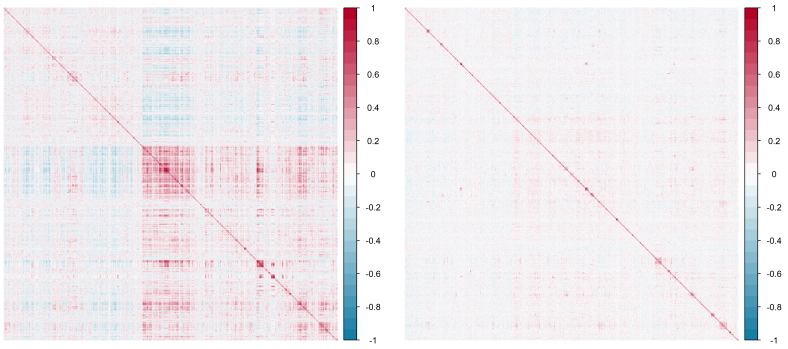

Figure 1.4: Heatmap of correlations after whitening. Prostate cancer (left) and Breast cancer (right) datasets.

1.3.4. Contribution of Chapter 2

This section summarizes the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2021), A variable selection approach for highly correlated predictors in high-dimensional genomic data, *Bioinformatics*, 37(16), 2238–2244.

The proposed method is implemented in the `WLasso` R package available from the CRAN.

We propose a novel variable selection approach, WLasso (Whitening Lasso), with the previously introduced idea of whitening. After transformation of Model

(1.5), we propose to minimize the following criterion with respect to $\widetilde{\boldsymbol{\beta}}$:

$$L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1, \tag{1.6}$$

which guarantees a sparsity enforcing constraint on $\boldsymbol{\beta}$ thanks to the ℓ_1 penalty. Note that note that $L_\lambda(\boldsymbol{\beta}) = L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}})$. We thus obtain

$$\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda) = \arg\min_{\widetilde{\boldsymbol{\beta}}} L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}}).$$

To estimate $\widetilde{\boldsymbol{\beta}}$, we will use the following modified estimator which can be seen as a thresholding of the components of $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda)$. For K in $\{1, \ldots, p\}$, let Top_K be the set of indices corresponding to the K largest values of the components of $|\widehat{\widetilde{\boldsymbol{\beta}}}_0|$, then the estimator of $\widetilde{\boldsymbol{\beta}}$ is $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(\widehat{K})})_{1\leq j\leq p}$ where $\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}$ is defined by:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}(\lambda) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{0j}(\lambda), & j \in \text{Top}_K \\ K\text{th largest value of } |\widehat{\widetilde{\boldsymbol{\beta}}}_{0j}|, & j \notin \text{Top}_K. \end{cases} \tag{1.7}$$

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

$$\widehat{\boldsymbol{\beta}}_j^{(M)}(\lambda) = \begin{cases} \widehat{\boldsymbol{\beta}}_{0j}(\lambda), & j \in \text{Top}_M \\ 0, & j \notin \text{Top}_M. \end{cases} \tag{1.8}$$

Variables with non-null coefficients in $\widehat{\boldsymbol{\beta}}$ are considered as associated with the response variable. In Chapter 2, we showed in various numeric experiments that when the biomarkers are highly correlated, WLasso outperforms the compared approaches in sparse high-dimensional frameworks.

1.3.5. Contribution of Chapter 3

This section summarizes the article:
Zhu, W., Adjakossa, E., Lévy-Leduc, C., and Ternès, N. (2021). Sign Consistency of the Generalized Elastic Net Estimator. Submitted and also available on *arXiv preprint* (arXiv:2106.05454).

In this section, aside from the transformation in Model (1.5), we propose combining an ℓ_1 and ℓ_2 penalty and consider the following criterion:

$$L_{\lambda,\eta}^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1 + \eta \left\|\widetilde{\boldsymbol{\beta}}\right\|_2^2, \text{ with } \lambda, \eta > 0. \tag{1.9}$$

Since it consists in adding a ℓ_2 penalty part to the Generalized Lasso as in the Elastic Net, we will call it generalized Elastic Net (gEN). The gEN estimator is defined by

$$\widehat{\boldsymbol{\beta}} = \Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}, \tag{1.10}$$

with

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right). \tag{1.11}$$

With this new estimator, we defined a corresponding irrepresentable condition called **Generalized Irrepresentable Condition (GIC)**:
There exist $\lambda, \eta, \alpha, \delta_4 > 0$ such that for all j,

$$\left|\left((C_{21}^n + \frac{\eta}{n}\Sigma_{21})(C_{11}^n + \frac{\eta}{n}\Sigma_{11})^{-1}\left(\text{sign}(\boldsymbol{\beta}_1) + \frac{2\eta}{\lambda}\boldsymbol{\beta}_1\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1\right)_j\right| \leq 1-\alpha, \text{ for all } j, \tag{1.12}$$

where $\boldsymbol{\beta}_1$ denotes the non-null components in the vector $\boldsymbol{\beta}$. We proved that this condition is sufficient for the gEN estimator to be sign consistent under certain conditions. Moreover, we compared **GIC** with **EIC** (Elastic Net Irrepresentable Condition, Jia and Yu (2010)) and **IC**, and demonstrate that there exist cases where **GIC** is satisfied but **EIC** and **IC** are not.

1.4. Identification of prognostic and predictive biomarkers in high-dimensional linear models with PPLasso

1.4.1. Identification of predictive biomarkers

With the advancement of precision medicine, there has been an increasing interest in identifying prognostic or predictive biomarkers. Previously introduced WLasso and generalized Elastic Net are developed for the purpose of selecting prognostic biomarkers. However, the discovery of predictive biomarkers has seen much less attention. Limited to binary endpoint, Foster et al. (2011) 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. (2012) proposed a new method to detect the interaction between the treatment and the biomarkers by modifying the covariates. This method can be implemented on a continuous/binary/time-to-event endpoint. Lipkovich et al. (2011) proposed a method called SIDES, which adopts a recursive partitioning algorithm for screening treatment-by-biomarker interactions. This method was further improved in Lipkovich and Dmitrienko (2014) by adding another step of preselection on predictive biomarkers based on variable importance. The method was demonstrated with a continuous endpoint. More recently, Sechidis et al. (2018) 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 endpoints. Moreover, most of these methods were assessed in a 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 problems of identifying predictive biomarkers. However, the identification of prognostic

biomarkers is also key in this context. The clinical impact of 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 in treatment counseling (Windeler, 2000). We proposed a novel approach called PPLasso (Predictive Prognostic Lasso) to simultaneously identify prognostic and predictive biomarkers in a high-dimensional setting with ANCOVA-type linear models.

1.4.2. Contribution of Chapter 4

This section summarizes the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2022). Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso. Submitted and also available on *arXiv preprint* (arXiv:2202.01970).

The proposed method is implemented in the PPLasso R package available from the CRAN.

Let $\boldsymbol{y}$ be a continuous response and t_1, t_2 two treatments. Let also $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) 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:

$$\boldsymbol{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \dots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \dots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix}. \tag{1.13}$$

To take into account the potential correlation that may exist between the biomarkers in the different treatments, we shall assume that the rows of $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) are independent centered Gaussian random vectors with a covariance matrix equal to Σ_1 (resp. Σ_2).

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

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \boldsymbol{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1p} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2p} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}, \tag{1.14}$$

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

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

α_1 (resp. α_2) corresponding to the effects of treatment t_1 (resp. t_2). Moreover, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \ldots, \beta_{1p})^T$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \ldots, \beta_{2p})^T$) are the coefficients associated to each of the p biomarkers in treatment t_1 (resp. t_2) group, and $\epsilon_{11}, \ldots, \epsilon_{2n_2}$ are standard independent Gaussian random variables independent of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. When t_1 stands for the standard treatment or placebo, prognostic (resp. predictive) biomarkers are defined as those having non-null coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$) and non prognostic (resp. non predictive) biomarkers correspond to the indices having null coefficients in $\boldsymbol{\beta}_1$ (resp. in $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$).

Model (1.14) can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{1.15}$$

with $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$.

To estimate $\boldsymbol{\gamma}$ using a sparsity enforcing constraint, we consider a first estimator of $\boldsymbol{\gamma}$ obtained by minimizing the following criterion with respect to $\boldsymbol{\gamma}$:

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \boldsymbol{\gamma} \right\|_1, \tag{1.16}$$

where $D_1 = [\mathrm{Id}_p, \mathbf{0}_{p,p}]$ and $D_2 = [-\mathrm{Id}_p, \mathrm{Id}_p]$, 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.

Since the inconsistency of Lasso variable selection originates from the correlations between the variables, we propose to remove the correlation by "whitening" the matrix $\boldsymbol{X}$. More precisely, we consider $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$, where

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \Sigma_1 & 0 \\ 0 & 0 & 0 & \Sigma_2 \end{bmatrix} \tag{1.17}$$

and define $\Sigma^{-1/2}$ by replacing in (1.17) Σ_i by $\Sigma_i^{-1/2}$, where $\Sigma_i^{-1/2} = \mathbf{U}_i \mathbf{D}_i^{-1/2} \mathbf{U}_i^T$, $\mathbf{U}_i$ and $\mathbf{D}_i$ being the matrices involved in the spectral decomposition of Σ_i for $i = 1$

or 2. With such a transformation the columns of $\widetilde{\boldsymbol{X}}$ are decorrelated and Model (1.15) can be rewritten as follows:

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon} \tag{1.18}$$

where $\widetilde{\boldsymbol{\gamma}} = \Sigma^{1/2}\boldsymbol{\gamma}$. The objective function (1.16) thus becomes:

$$L^{\text{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}) = \frac{1}{2}\left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}}\right\|_2^2 + \lambda_1 \left\|\begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \Sigma^{-1/2}\widetilde{\boldsymbol{\gamma}}\right\|_1 . \tag{1.19}$$

Similar thresholding was then imposed as previously explained in Section 1.3.4 to obtain the final estimation $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$. The biomarkers with non-null coefficients in $\widehat{\boldsymbol{\beta}}_1$ (resp. $\widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1$) are considered as prognostic (resp. predictive) biomarkers.

We compared PPLasso to other methods that also deal with the correlations and we showed that PPLasso outperforms them on both prognostic and predictive biomarker identification in various scenarios.

1.5. Variable selection in high-dimensional logistic regression models using a whitening approach

1.5.1. Biomarker selection for binary responses

Previously, we assumed that the response $\boldsymbol{y}$ was continuous. This section focuses on binary responses, which can be seen as a classification problem. Classification is an important topic in biomedical research. For example, following the RECIST (Response Evaluation Criteria in Solid Tumours) guidance (Watanabe et al., 2009) on the oncology research, patients are usually defined as complete response, partial response, stable disease, and progression according to the response to the treatment. Patients in the first two categories (complete response and partial response) are considered as responders to the treatment, while the others are considered as non-responders. For the disease of Rheumatoid Arthritis, ACR (American College of Rheumatology) criteria are used to assess the treatment response and discriminate efficient treatment from placebo treatment in a clinical trial setting. ACR response is scored as a percentage improvement. For example, ACR50 is a binary outcome indicating whether the improvement is greater than 50%. Another example is tumor classification. With the development of bioinformatics, cancer classification from omics data has become an important topic in genome research (Ramaswamy et al., 2001; Tibshirani et al., 2002; Menyhárt and Győrffy, 2021).

Compared to other classifiers such as decision tree (Utgoff, 1989) and SVM (Support Vector Machine) (Cortes and Vapnik, 1995), logistic regression (Walker

and Duncan, 1967) is a popular classification method with an explicit statistical interpretation and can provide classification probabilities for a binary response (Menard, 2002). However, as previously explained, with the high dimensional omics data, it is essential to obtain a small number of key genes and improve the classification accuracy, which leads us to consider the problem of variable selection in high dimensional logistic regression model (Park and Hastie, 2007). Recently, regularization approaches have been widely applied to biomarker discovery and disease classification (Zhu and Hastie, 2004; Wu, 2006; Ma and Huang, 2008; Liu et al., 2020). Besides the high dimensionality, the correlations between biomarkers should also be taken into account. To deal with the correlations, several methods have been proposed. Adapted to logistic regression, the most well-known ones include Elastic Net (Zou and Hastie, 2005) and Adaptive Lasso (Zou, 2006) as previously mentioned in Section 1.3.1. Several filter approaches were also proposed to take into consideration the correlations in the classification framework. Relief (Kira and Rendell, 1992) is sensitive to feature interactions and has inspired a family of Relief-based feature selection algorithms, notably the ReliefF (Kononenko et al., 1997). It was widely used in biomedical research (Urbanowicz et al., 2018). Fast Correlation Based Filter (FCBF) (Yu and Liu, 2003) is another approach in high-dimensional feature selection that evaluates feature relevance and redundancy based on correlation measures. In this chapter, we propose a novel method that can identify active biomarkers in high-dimensional data and provide classification on collected samples.

1.5.2. Contribution of Chapter 5

This section summarizes the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2022). Variable selection in high-dimensional logistic regression models using a whitening approach. Submitted and also available on *arXiv preprint* (arXiv:2206.14850)

The proposed method is implemented in the `WLogit` R package which will soon be available from the CRAN.

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

$$\pi_{\boldsymbol{\beta}}(X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}. \tag{1.20}$$

The logistic regression with ℓ_1 regularization solves the feature selection problem

by adding a penalty function to the log-likelihood of the logistic regression model:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \tag{1.21}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$, and the log-likelihood $l(\boldsymbol{\beta})$ is defined by:

$$l(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \cdot X^{(i)} \boldsymbol{\beta} - \log(1 + e^{X^{(i)} \boldsymbol{\beta}}) \right], \tag{1.22}$$

with $X^{(i)}$ the ith row of $\boldsymbol{X}$. With the penalty function and properly chosen parameter λ, some components of $\widehat{\boldsymbol{\beta}}$ are set to zero.

As previously mentioned, the Lasso criterion can fail to select the true subset of active biomarkers when the correlation between active and non-active biomarkers is large, which is stated in the irrepresentable condition for linear regression models in Equation (1.4). A similar condition was obtained by Ravikumar et al. (2010) and Bunea (2008) in the logistic regression case. Let Q be defined by:

$$Q = \boldsymbol{X}^T \boldsymbol{H} \boldsymbol{X}, \tag{1.23}$$

where $\boldsymbol{H}$ is a diagonal matrix with

$$H_{ii} = \pi_{\boldsymbol{\beta}}(X^{(i)})/(1 - \pi_{\boldsymbol{\beta}}(X^{(i)})), 1 \leq i \leq n. \tag{1.24}$$

Let $S = \{j,\ \beta_j \neq 0\}$ be the set of active variables with size d, S^c the set of non-active variables, Q_{SS} denotes the $d \times d$ sub-matrix of Q indexed by S. With this notation, the condition states:
There exists $\alpha \in (0, 1]$ such that:

$$\left| Q_{S^c S} (Q_{SS})^{-1} \right|_\infty \leq 1 - \alpha, \tag{1.25}$$

where $|A|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^{p} |A_{jk}|$ for any real matrix having p rows and p columns.

We then propose to remove the correlation by 'whitening' the matrix $\boldsymbol{X}$. More precisely, we consider $\widetilde{\boldsymbol{X}} = \boldsymbol{X} \check{\Sigma}^{-1/2}$, where $\check{\Sigma}$ is a covariance estimator obtained from $\boldsymbol{H}^{1/2} \boldsymbol{X}$, where $\boldsymbol{H}$ is defined in Equation (1.24). With this transformation, $\widetilde{\boldsymbol{X}}^T \boldsymbol{H} \widetilde{\boldsymbol{X}}$ should be close to the identity matrix I_p, thus the irrepresentable condition should be satisfied. After the whitening step, Model (1.20) can be rewritten as:

$$\pi_{\widetilde{\boldsymbol{\beta}}}(\widetilde{X}^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)} \right)}{1 + \exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)} \right)}, \tag{1.26}$$

where $\widetilde{\boldsymbol{\beta}} = \check{\Sigma}^{1/2}\boldsymbol{\beta}$. The log-likelihood after the transformation can be written as:

$$l^{wt}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i \cdot \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}} - \log\left(1 + e^{\widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}}}\right)\right\}. \tag{1.27}$$

Then an estimator of $\widetilde{\boldsymbol{\beta}}$ is obtained by solving the following problem:

$$\underset{\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^p}{\arg\min}\, l^{wt}(\widetilde{\boldsymbol{\beta}}) + \lambda\left\|\check{\Sigma}^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1. \tag{1.28}$$

To solve this optimization problem, we usually form a quadratic approximation of the log-likelihood (1.27) by using a Taylor expansion at the current estimates (Friedman et al., 2010):

$$l_Q^{wt}(\widetilde{\boldsymbol{\beta}}) = -\frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}})^2 + C(\widetilde{\boldsymbol{\beta}}^o)^2 \tag{1.29}$$

$$= -\frac{1}{2n}\sum_{i=1}^{n}(\sqrt{w_i}z_i - \sqrt{w_i}\widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}})^2 + C(\widetilde{\boldsymbol{\beta}}^o)^2 \tag{1.30}$$

with

$$z_i = \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}} + \frac{y_i - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})}{\pi_{\widetilde{\boldsymbol{\beta}}^o}(X^{(i)})(1 - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)}))}, \text{(working response)}$$

$$w_i = \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})(1 - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})), \text{(weights)} \tag{1.31}$$

where $\pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})$ is the evaluation of $\pi_{\widetilde{\boldsymbol{\beta}}}$ (defined in Model (1.26)) at the current parameters $\widetilde{\boldsymbol{\beta}}^o$. The final estimator can then be derived by the IRLS (Iterative Re-weighted Least Square) algorithm (Daubechies et al., 2010).

After obtaining the estimation of $\widetilde{\boldsymbol{\beta}}$, similar thresholding was then imposed as previously explained in Section 1.3.4 to obtain the final estimation $\widehat{\boldsymbol{\beta}}$. The biomarkers with non-null coefficients are considered as active ones. A classifier is also obtained with $\widehat{\boldsymbol{\beta}}$.

The performance of WLogit is assessed using synthetic data in several scenarios and compared with other approaches. The results suggest that WLogit can identify almost all active biomarkers even in the cases where the biomarkers are highly correlated, while the other methods fail, which consequently leads to higher classification accuracy. The performance is also evaluated on the classification of two Lymphoma subtypes, and the obtained classifier also outperformed other methods.

Chapter 2 - A variable selection approach for highly correlated predictors in high-dimensional linear regression models

Publication

The content of this chapter is in the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2021), A variable selection approach for highly correlated predictors in high-dimensional genomic data, *Bioinformatics*, 37(16), 2238–2244.

The proposed method is implemented in the `WLasso` R package available from the CRAN.

Abstract

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

Contents

2.1. 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 (Saeys et al. (2007) and Heinze et al. (2018) 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 (1996a)), 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:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ is the variable to explain (clinical endpoint), $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is the design matrix containing the expression of biomarkers such that the correlation matrix of its columns is Σ, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ is a sparse vector to estimate, namely with a majority of null coefficients, and $\boldsymbol{\epsilon}$ 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 (1996a)):

$$L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{2.2}$$

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 $\boldsymbol{X}_A$ the submatrix of $\boldsymbol{X}$ containing only the

indices of columns which are in the set A. Then, the design matrix $\boldsymbol{X}$ satisfies the Irrepresentable Condition (IC) if, for some constant $\alpha \in (0, 1)$,

$$\left| \left(\boldsymbol{X}_{S^c}^T \boldsymbol{X}_S (\boldsymbol{X}_S^T \boldsymbol{X}_S)^{-1} \text{sign}(\boldsymbol{\beta}_S) \right)_j \right| \leq 1 - \alpha, \text{ for all } j, \tag{2.3}$$

where $\text{sign}(x) = 1$ 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 $\boldsymbol{X}$ and $\boldsymbol{y}$ before applying the Lasso criterion. For example, Jia and Rohe (2015) and Wang and Leng (2016) proposed to left-multiply $\boldsymbol{X}$, $\boldsymbol{y}$ and thus $\boldsymbol{\epsilon}$ in Model (2.1) by specific matrices to remove the correlations between the columns of $\boldsymbol{X}$. 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 paper, 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 (2.1) in order to remove the correlations existing between the columns of $\boldsymbol{X}$ and thus to "whiten" them and make the IC valid but without changing the error term $\boldsymbol{\epsilon}$. This prevents us from noise inflation, see (2.4). Then, the variable (biomarker) selection is performed thanks to the generalized Lasso criterion devised by Tibshirani and Taylor (2011). The full details of this method are provided in Section 2.2. An extensive simulation study is presented in Section 2.3 to assess the selection performance of WLasso and to compare it to other methods in different settings. WLasso is also applied to a publicly available dataset in breast cancer in Section 2.4. Finally, we

discuss our findings and give concluding remarks in Section 2.5.

2.2. 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 $\boldsymbol{X}$) and in applying the generalized Lasso criterion proposed by Tibshirani and Taylor (2011) for variable selection purpose.

2.2.1. Model Transformation

Inspired by the literature on preconditioning, we propose to rewrite Model (2.1) in order to remove the correlation existing between the columns of $\boldsymbol{X}$. 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 $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$. Therefore, (2.1) can be rewritten as follows:

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \tag{2.4}$$

where $\widetilde{\boldsymbol{\beta}} = \Sigma^{1/2}\boldsymbol{\beta} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^T\boldsymbol{\beta}$. With such a transformation, since the n rows $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of $\boldsymbol{X}$ are assumed to be independent Gaussian random vectors with a covariance matrix equal to Σ, the covariance matrix of the rows of $\widetilde{\boldsymbol{X}}$ is equal to identity and the columns of $\widetilde{\boldsymbol{X}}$ 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 $\boldsymbol{\epsilon}$ 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 $\boldsymbol{y}$ were generated according to Model (2.1) with $p = 500$, $n = 50$, $\boldsymbol{\beta}$ having 10 non null components which are equal to 2 and with Σ defined by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \tag{2.5}$$

where Σ_{11} is the correlation matrix of active variables with off-diagonal entries equal to α_1, Σ_{22} is the one of non active variables with off-diagonal entries equal to α_3 and Σ_{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 2.1 displays the percentage of components j for which the Irrepresentable Condition (2.3) 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 2.2 are obtained from observations $\boldsymbol{y}$ generated according to the previous scenario.

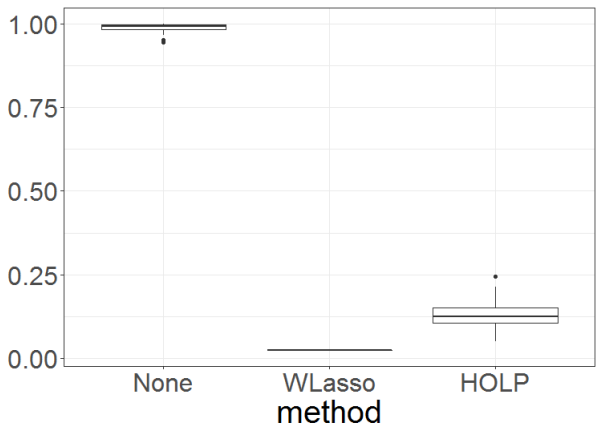

Figure 2.1: Proportion of components j such that (2.3) is violated. These results were obtained from 100 replications.

2.2.2. Estimation of $\widetilde{\beta}$

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

$$\|y - X\beta\|_2^2 + \lambda \|D_0\beta\|_1 ,$$

where D_0 is a specific matrix. Note that this criterion boils down to the classical Lasso criterion if D_0 is the identity matrix. In Model (2.4), we thus propose to minimize the following criterion with respect to $\widetilde{\beta}$:

$$L_\lambda^{\text{gen}}(\widetilde{\beta}) = \left\| y - \widetilde{X}\widetilde{\beta} \right\|_2^2 + \lambda \left\| \Sigma^{-1/2}\widetilde{\beta} \right\|_1 , \tag{2.6}$$

which guarantees a sparsity enforcing constraint on β thanks to the ℓ_1 penalty. We thus obtain

$$\widehat{\widetilde{\beta}}_0(\lambda) = \arg\min_{\widetilde{\beta}} L_\lambda^{\text{gen}}(\widetilde{\beta}).$$

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

$$\widehat{\widetilde{\beta}}_j^{(K)}(\lambda) = \begin{cases} \widehat{\widetilde{\beta}}_{0j}(\lambda), & j \in \text{Top}_K \\ K\text{th largest value of } |\widehat{\widetilde{\beta}}_{0j}|, & j \notin \text{Top}_K . \end{cases} \tag{2.7}$$

The choice of $\widehat{K} = \widehat{K}(\lambda)$ is explained in Section 2.2.4. Figure 2.2 displays the average of $(|\widehat{\widetilde{\beta}}_{0j}(\lambda) - \widetilde{\beta}_j(\lambda)|)_{1\le j\le p}$ and $(|\widehat{\widetilde{\beta}}_j^{(\widehat{K})}(\lambda) - \widetilde{\beta}_j(\lambda)|)_{1\le j\le p}$ for all the values of

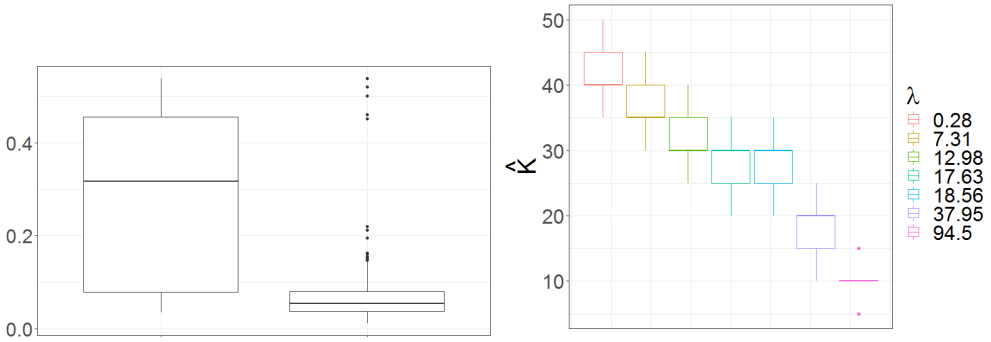

Figure 2.2: Left: Boxplots of the average of $(|\widehat{\widetilde{\beta}}_{0j}(\lambda) - \widetilde{\beta}_j(\lambda)|)_{1\leq j\leq p}$ (left) and $(|\widehat{\widetilde{\beta}}_j^{(\widehat{K})}(\lambda) - \widetilde{\beta}_j(\lambda)|)_{1\leq j\leq p}$ (right) for all λ obtained from 100 replications. Right: Boxplots of $\widehat{K}(\lambda)$ for different values of λ obtained from 100 replications.

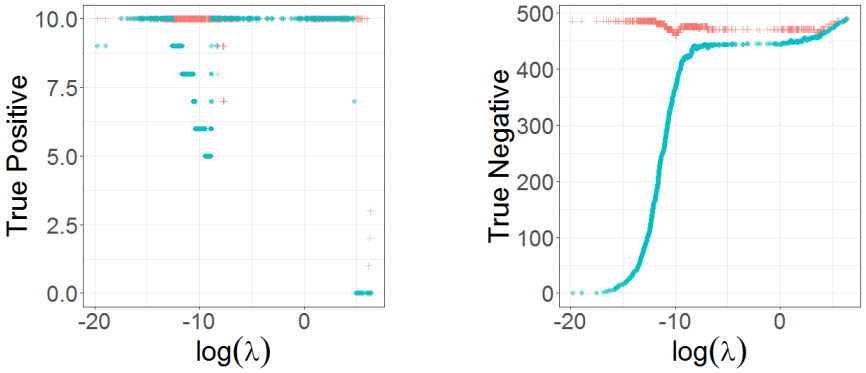

Figure 2.3: Number of True Positive and True Negative for $\widehat{\beta}$ in red and $\widehat{\beta}_0$ in blue for a given vector of observations y.

λ that are considered and boxplots of $\widehat{K}(\lambda)$ for some λ. We can see from this figure that the thresholding improves the estimation of $\widetilde{\beta}$.

2.2.3. Estimation of β

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

$$\widehat{\beta}_j^{(M)}(\lambda) = \begin{cases} \widehat{\beta}_{0j}(\lambda), & j \in \mathrm{Top}_M \\ 0, & j \notin \mathrm{Top}_M. \end{cases} \tag{2.8}$$

The choice of $\widehat{M}(\lambda)$ is explained in Section 2.2.4.

As we can see from Figure 2.3, more true non null (active) components of β (true positive) and more true null (non active) components of β (true negative)

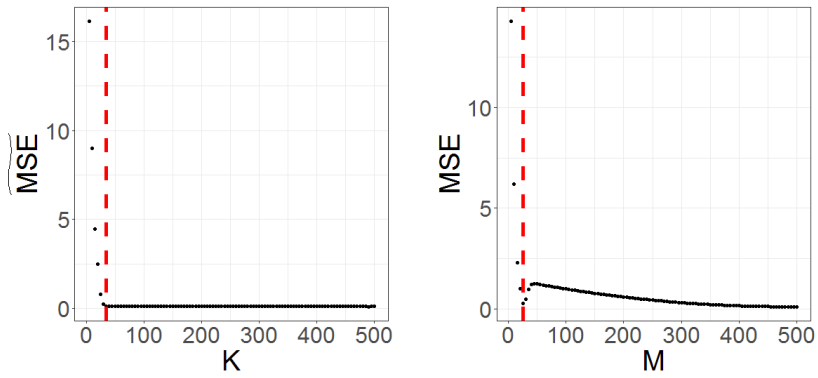

Figure 2.4: $\widetilde{\mathrm{MSE}}_K(\lambda)$ (left) and $\mathrm{MSE}_M(\lambda)$ (right) for λ chosen thanks to the strategy explained in Section 2.3.2 for a given vector of observations $\boldsymbol{y}$ and $\gamma = 0.95$. The vertical dotted lines correspond to $\widehat{K}(\lambda)$ and $\widehat{M}(\lambda)$, respectively.

can be retrieved with $\widehat{\boldsymbol{\beta}}$ than with $\widehat{\boldsymbol{\beta}}_0$. We can thus conclude from Figures 2.2 and 2.3 that both thresholdings improve the variable selection.

2.2.4. Choice of the parameters

To choose the parameters K and M in (2.7) and (2.8) 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 $\widehat{K}$. Let

$$\widetilde{\mathrm{MSE}}_K(\lambda) = \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\widehat{\boldsymbol{\beta}}}^{(K)}(\lambda)\|_2^2,$$

where $\boldsymbol{y}$, $\widetilde{\boldsymbol{X}}$ and $\widetilde{\widehat{\boldsymbol{\beta}}}^{(K)}(\lambda)$ are defined in (2.1), (2.4) and (2.7), respectively and

$$\widehat{K}(\lambda) = \arg\min\left\{K \geq 1 \text{ s.t. } \frac{\widetilde{\mathrm{MSE}}_{K+1}(\lambda)}{\widetilde{\mathrm{MSE}}_K(\lambda)} \geq \gamma\right\}, \text{ where } \gamma \in (0,1).$$

Large values of γ will lead to large values of $\widehat{K}(\lambda)$ and thus to a weak thresholding of the estimator of $\widetilde{\boldsymbol{\beta}}$. In practice, as it is shown in Section 2.3, taking γ in (0.9,0.99) provides satisfactory and almost similar results.

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

$$\mathrm{MSE}_M(\lambda) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}^{(M)}(\lambda)\|_2^2, \tag{2.9}$$

where $\boldsymbol{y}$, $\boldsymbol{X}$ and $\widehat{\boldsymbol{\beta}}^{(M)}(\lambda)$ are defined in (2.1) and (2.8), respectively.
Both criteria are displayed in Figure 2.4 for a value of λ which is chosen according to the strategy explained in Section 2.3.2.

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 2.5 displays the TPR and FPR for different values of K and M. We can see from this

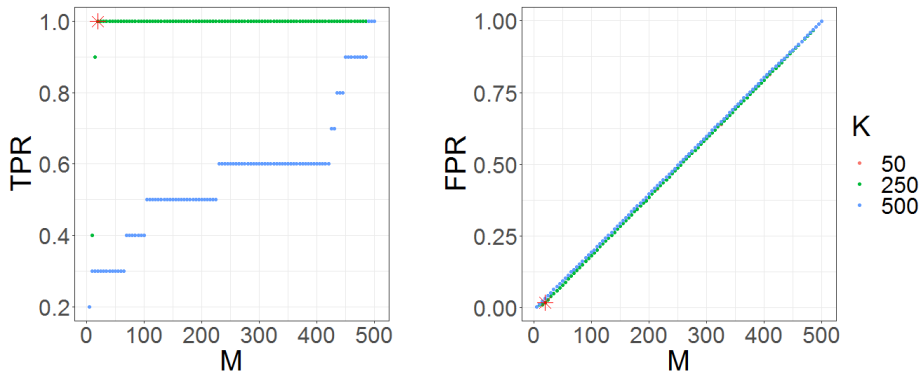

Figure 2.5: TPR (left) and FPR (right) for different values of M and K. The TPR and FPR obtained for $M = \hat{M}$ and $K = \hat{K}$ are displayed with a red star ('*'). (red dots and green ones are overlaped)

figure that our choice of M and K, displayed with a red star, guarantees a good trade-off between the TPR and FPR.

2.2.5. Estimation of Σ

Since the matrix Σ is unknown in practice, it has to be estimated. In the particular situation where Σ has the block structure described in (2.5), 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 (2.1). The corresponding $p \times p$ sample correlation matrix $\boldsymbol{R} = (R_{i,j})$ is defined by:

$$\boldsymbol{R}_{i,j} = \frac{S_{i,j}}{\sigma_i \sigma_j}, \ \forall 1 \leq i, j \leq p, \tag{2.10}$$

where

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

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 $\widehat{\Sigma}$ 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.

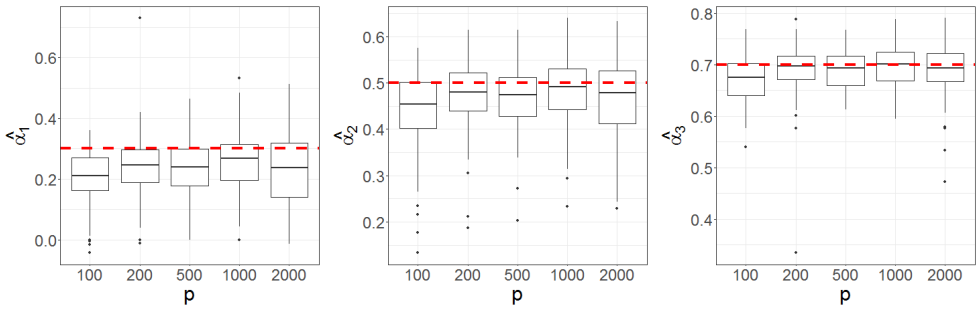

Figure 2.6: Estimation of the parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. The horizontal dotted lines correspond to the true values of the parameters. These results are obtained from 100 replications for each value of p.

Then, for a given clustering C:

$$\rho_{i,j} = \begin{cases} \dfrac{1}{\#C(i)\#C(j)} \displaystyle\sum_{k \in C(i), \ell \in C(j)} R_{k,\ell}, & \text{if } C(i) \neq C(j) \\ \dfrac{1}{\#C(i)(\#C(i)-1)} \displaystyle\sum_{k \in C(i), \ell \in C(i), k \neq \ell} R_{k,\ell}, & \text{if } C(i) = C(j) \end{cases}, \tag{2.11}$$

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 (2.10).

We illustrate the performance of our method in Figure 2.6 in the case where Σ has the structure (2.5) 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 $\widehat{\Sigma}$ is efficient.

2.2.6. Summary of the WLasso method

The WLasso method can be summarized as follows:

- First step: Estimation of the matrix Σ by $\widehat{\Sigma}$, see Section 2.2.5.
- Second step: Transformation of Model (2.1) into Model (2.4) to remove the correlation existing between the columns of $\boldsymbol{X}$, see Section 2.2.1 where Σ is replaced by $\widehat{\Sigma}$.
- Third step: Estimation of $\widetilde{\boldsymbol{\beta}}$ defined in (2.4), see Section 2.2.2.
- Fourth step: Estimation of $\boldsymbol{\beta}$ defined in (2.1), see Section 2.2.3 and identification of its null and non null components.

2.3. 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 (2.1) where the n rows of $\boldsymbol{X}$ are assumed to be independent Gaussian random vectors with a covariance matrix equal to Σ and $\boldsymbol{\epsilon}$ is a standard Gaussian random vector independent of $\boldsymbol{X}$. 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 Σ which contains the correlation values between the biomarkers, namely the correlations between the columns of the design matrix $\boldsymbol{X}$, several structures were considered:

- Block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$ and $(0.5, 0.7, 0.9)$;
- Independent setting where Σ is the identity matrix.

The results that are presented hereafter are obtained from 100 replications. Note that other correlation structures are considered in the Supplementary Material. Since the objective of WLasso is to work with structures where the IC is violated, it should not be surprising that WLasso performs similarly to Lasso if this condition is satisfied.

2.3.1. Estimation of Σ

To evaluate the impact of the estimation of Σ, simulations were performed to compare the performance of WLasso when Σ is known and when it is estimated. The results are displayed in Figure 2.7 for several values of γ (0.9, 0.95, 0.97) which is a parameter appearing in Section 2.2.4. 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 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 Σ is known. In the case where Σ is estimated

by using the approach described in Section 2.2.5, 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 2.7 are obtained when $(b, n) = (0.5, 50)$ but we obtained similar conclusions for $(b, n) = (1, 50)$, $(b, n) = (0.5, 100)$ and $(1, 100)$.

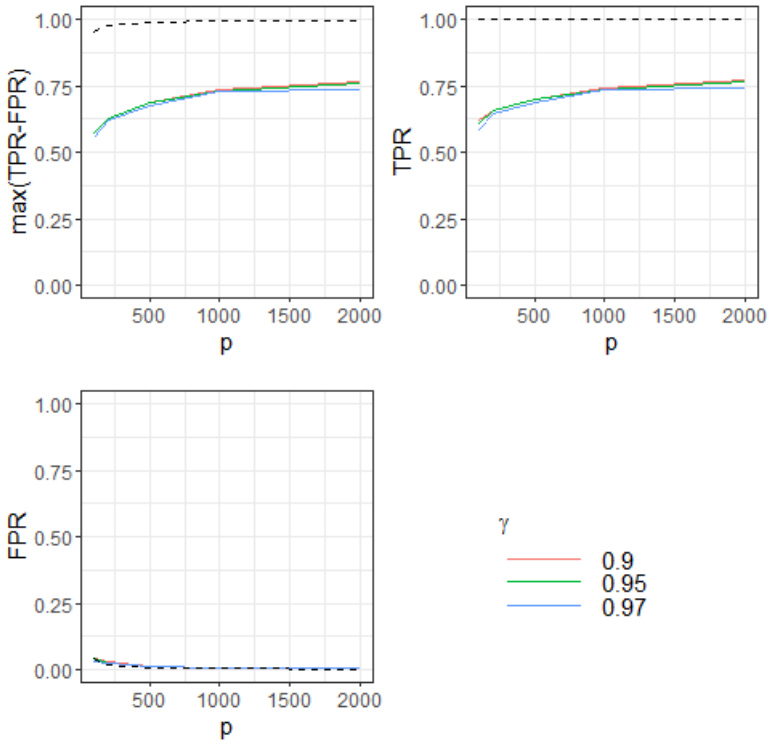

Figure 2.7: Average over the replications of max(TPR-FPR) and of the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 50$. Dotted line: Σ, solid line: $\widehat{\Sigma}$.

2.3.2. Choice of λ

For tuning the parameter λ involved in WLasso, we propose choosing the value which minimizes $\mathrm{MSE}_{\widehat{M}(\lambda)}(\lambda)$ defined in (2.9). In Figure 2.14 of the Supplementary material, we compare the performance of WLasso with this choice of λ (dotted line, also noted as WLasso in the simulation) to the optimal one, called WLasso optimal, obtained when λ is chosen to yield the largest difference between the TPR and the FPR. 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 $b = 1$.

2.3.3. Comparison with other methods

In this section, we compare our methodology with other approaches: the classical Lasso described in Tibshirani (1996a) and two recently proposed methods aiming at handling the correlations between the columns of the design matrix $\boldsymbol{X}$: 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 WLasso 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 $\hat{\boldsymbol{\beta}}_{\text{HOLP}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^{-1}\boldsymbol{y}$. Then, for each s in $\{1, \dots, p\}$, the components of $\boldsymbol{\beta}$ which are estimated as non null are the s largest among the $|\hat{\boldsymbol{\beta}}_{\text{HOLP},j}|$, where $\hat{\boldsymbol{\beta}}_{\text{HOLP},j}$ denotes the jth components of $\hat{\boldsymbol{\beta}}_{\text{HOLP}}$. 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 2.8 and 2.9 in the case where $n = 50$ and $b = 0.5$ and Σ has the block-wise correlation structure defined in (2.5) 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 2.3.2, $\gamma = 0.95$ and Σ is estimated using the procedure explained in Section 2.2.5. 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 2.8 and 2.9 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 2.3.2 and Σ was estimated.

Figure 2.10 displays the results when the sample size n is increased and equal to

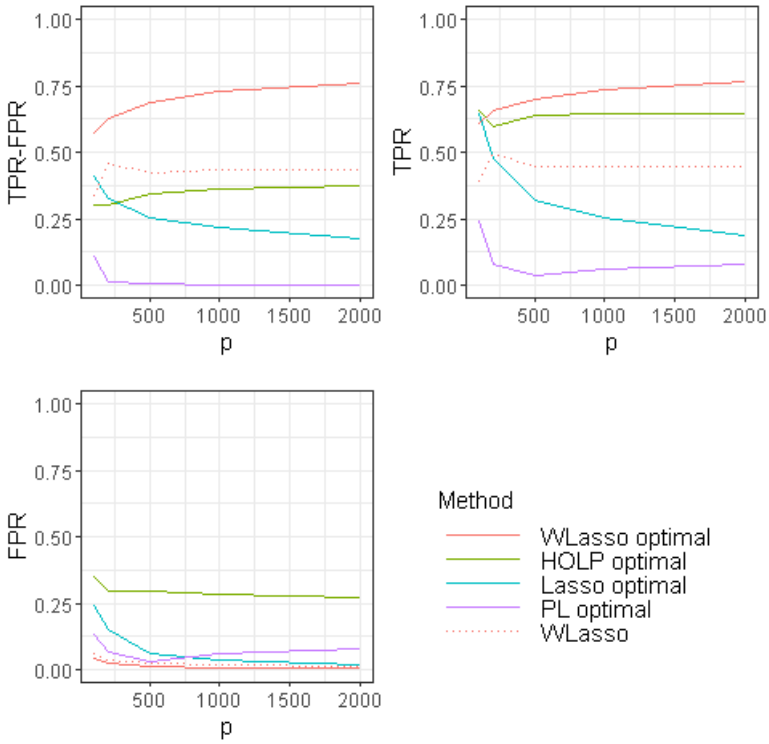

Figure 2.8: Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 50$.

100. We observe from this figure that the overall performance has been improved and that both WLasso optimal and WLasso 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 2.11 displays the performance of the different methodologies in the case where $\Sigma = \mathrm{Id}$, $n = 50$ and $b = 0.5$, that is in the case where there is no correlation between the biomarkers (columns of $\boldsymbol{X}$). We can see from this figure that even in this case, WLasso, 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 $n = 100$, WLasso obtains the best results in terms of TPR-FPR for p larger than 250, see the Supplementary material.

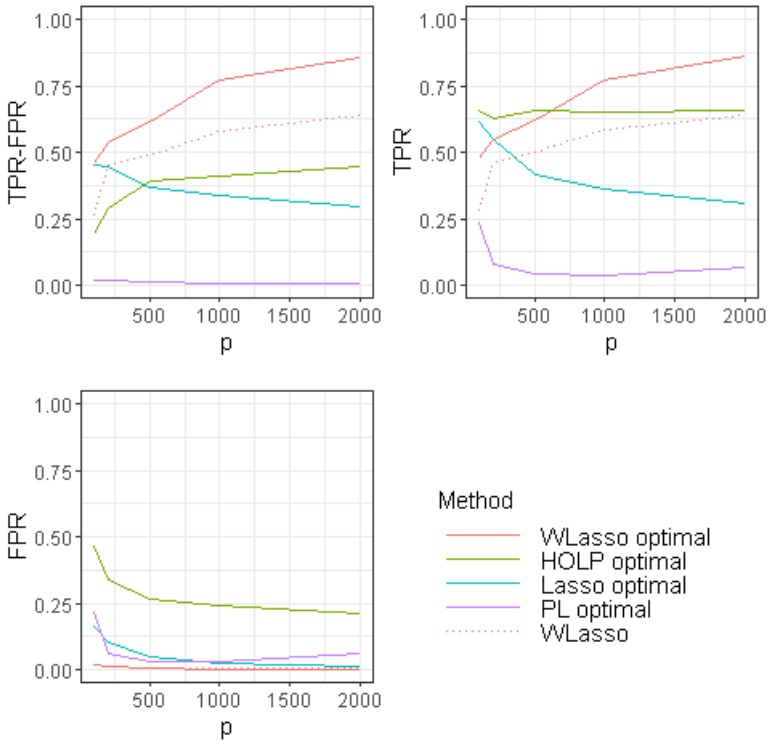

Figure 2.9: Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$, $b = 0.5$ and $n = 50$.

2.3.4. Numerical performance

Figure 2.12 displays the computational times of WLasso 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 (2.4GHz) CPU. We can see from this figure that it takes only 6 minutes for processing data when $p = 2000$.

Moreover, we can observe from Figure 2.13 that the most time consuming step of WLasso is the one where the generalized Lasso criterion is used (blue part in Figure 2.13). 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).

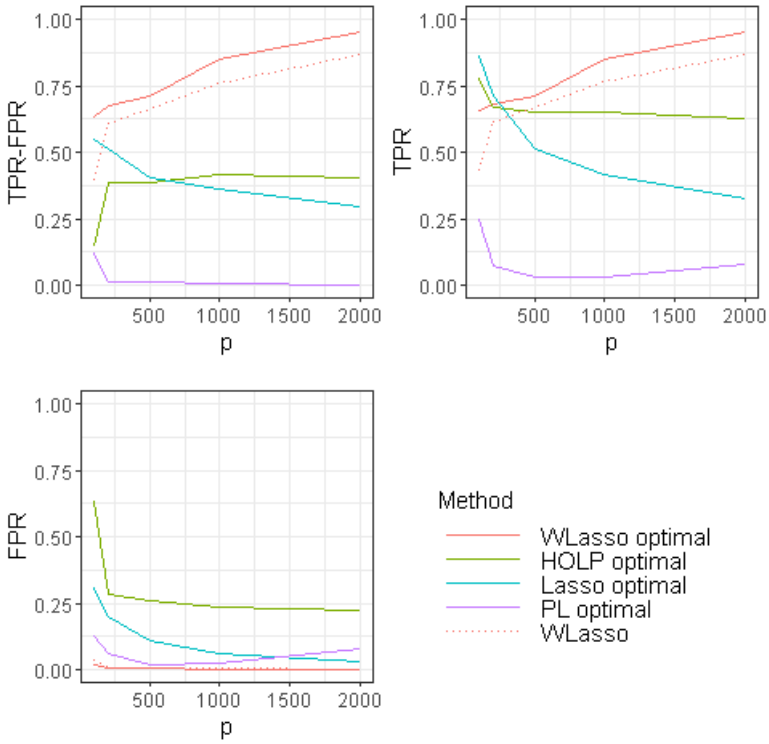

Figure 2.10: Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 100$.

2.4. 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 $\boldsymbol{y}$ to explain (response variable) as it is well known to be associated with breast cancer prognosis as recently explained by Wu et al. (2020).As all patients in the application are ER+ breast cancer patients, the choice

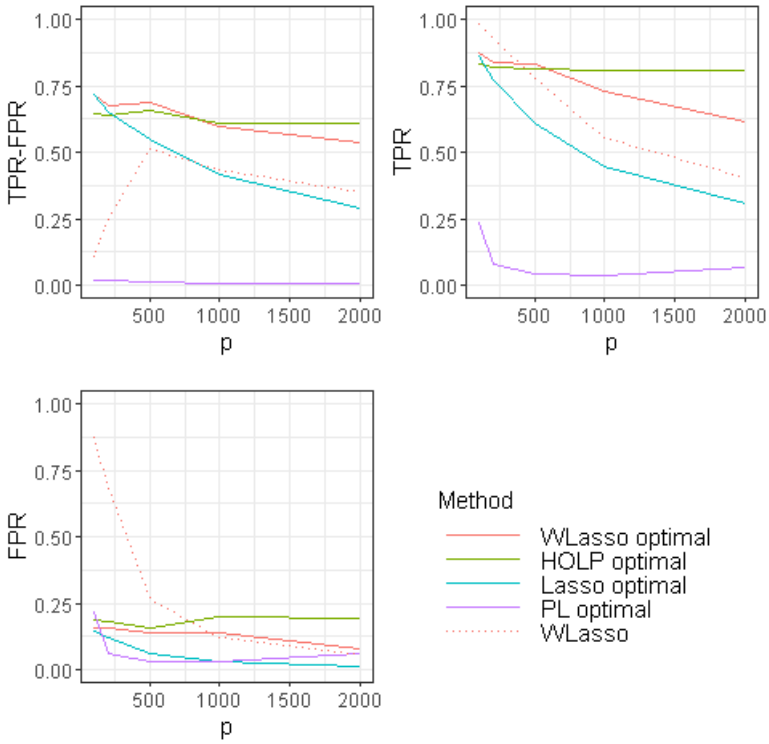

Figure 2.11: Top left: Average of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and average of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\Sigma = \mathrm{Id}$, $b = 0.5$ and $n = 50$.

of this gene expression is justified. The standardized 1,111 probes were considered as explanatory variables.

We implemented the approaches investigated in Section 2.3 and illustrated their genes selection. For WLasso, we used the methodology described in Section 2.2.5 for estimating the correlation matrix Σ. The heatmap of the correlation between probes is provided in Figure 2.28 of the 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. Table 2.1 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 $\boldsymbol{X}\boldsymbol{X}^T$ which is not invertible in this case. The matrix $\boldsymbol{X}^T$ is indeed not full rank in this dataset. WLasso and Lasso selected almost the same number of genes, *i.e.* 63 and 66

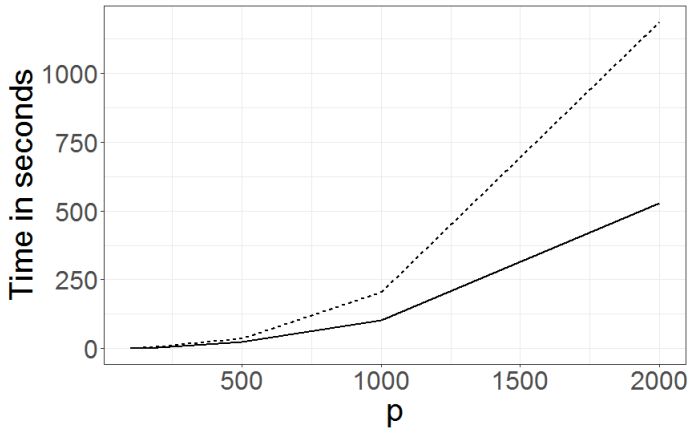

Figure 2.12: Computational time of WLasso when $n = 50$, "maxsteps" has the default value, namely 2000 (dotted line) and maxsteps=500 (solid line).

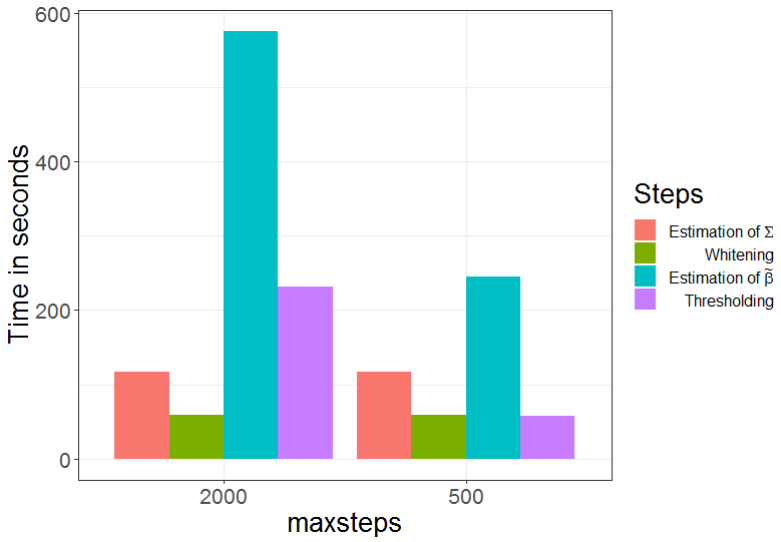

Figure 2.13: Time allocation for each part of WLasso for $p = 2000$ and two values of the parameter "maxsteps".

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.

2.5. Conclusion

In this paper, 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 Σ could be improved by using more sophisticated approaches such as Perrot-Dockès et al. (2020). Secondly, the 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 WLasso 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.

Appendix

This supplementary material provides additional numerical experiments, figures and a table for Chapter 2: "A variable selection approach for highly correlated predictors in high-dimensional genomic data".

Figure 2.14 illustrates Section 2.3.2.

Figures 2.15, 2.16, 2.17, 2.18, 2.19, 2.20, 2.21 and 2.22 provide similar results as those displayed in Figure 2.8 of the paper in the following cases:

- Block-wise correlation structure for Σ with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$ with
  - $b = 1$ and $n = 50$
  - $b = 1$ and $n = 100$
- Block-wise correlation structure for Σ with $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$ with
  - $b = 1$ and $n = 50$
  - $b = 0.5$ and $n = 100$
  - $b = 1$ and $n = 100$
- Σ is equal to identity with
  - $b = 1$ and $n = 50$
  - $b = 0.5$ and $n = 100$
  - $b = 1$ and $n = 100$

Figures 2.23, 2.24, 2.25, 2.26 and 2.27 provide similar results as those displayed in Figure 2.8 of the paper for other correlation structures.

Firstly, for testing the robustness of our approach, we considered the case where Σ has the block-wise structure (2.5) of the paper but α_1 (resp. α_2, α_3) is randomly chosen in an interval of length 0.01 or 0.05 around 0.3 (resp. 0.5, 0.7). The results for $n = 50$ and $b = 0.5$ are displayed in Figures 2.23 and 2.24. We can see from these figures that the performance of our approach are not very altered by this additional jitter except when it is too large (0.05).

Secondly, to extend the case where Σ is equal to identity, we considered the case where Σ has a particular block-wise structure where $\alpha_1 = \alpha_2 = \alpha_3 = \rho$ for $\rho = 0.1$, 0.3 and 0.5. The results for $n = 100$ and $b = 0.5$ are displayed in Figures 2.25, 2.26 and 2.27. We can see from these figures that in this case the performance of our approach are similar to the ones of the Lasso except for the largest value of ρ (0.5) where the True Positive Rate for the Lasso is 3% larger than WLasso for the same False Positive Rate.

Figures 2.28 and Table 2.1 give additional information for the Application Section.

Table 2.1: Selected genes from each method.

	selected genes
WLasso	DSP, TOP2A, FHL1, CD55, SLC39A6, SERPINA3, USP1, MYC, MARCH7, ITM2A, SLPI, TLE1, HLA-DQA1, BAMBI, GALNT3, LPL, ADIRF, SMN1, CLDN3, SYT1, MAOB, CKS2, PMAIP1, S100P, REEP1, ABCA8, MYB, TFF1, DNALI1, CXCL13, AGTR1, DACH1, ZNF415, BLNK, PITX1, LOC101928189, ABI1, RGS5, CLNS1A, ABAT, GATA3, CXCL12, SPP1, CCL19, NQO1, RHOB, ELN, RND3, HLA-DQB1, NFIB, COBL, TMEM158, SLC1A1, ERBB4, HBA1, SELENBP1, NAT1, CXCL14, C6orf211, PSD3, SPAG16, IL20RA, KLF4
Lasso	CCL5, CDH1, TOP2A, ZFP36L2, TPD52, ALCAM, RIOK3, RGS2, YES1, FH, ATXN1, BAMBI, SNRPE, LPL, RAB4A, BCL2, FRZB, HEPH, CA12, PEX3, ALOX5AP, SORBS2, FLRT2, FGFR3, MYB, MELK, ADORA2A, KIT, ACAP1, INHBB, ACOX2, PLAU, MAPT, SCGB1D2, NFAT5, H2BFS, ABI1, CADM1, HLA-DQB1, IGHM, ABAT, HLA-DQB1, NR2F1, ADH1B, RRM2, CTSZ, GOLGA8A, CALM1, PBX1, SYNM, ACKR3, COBL, TP73-AS1, KRT6B, BGN, DHRS2, URI1, NAT1, TFAP2B, SLC7A8, COL10A1, PRC1, MLPH, FAM134B, COL5A2, ACTB
PL	CDH1, TIMP3, JUN, FHL1, CALD1, SDHD, MX1, WWTR1, RANBP9, MED14, PHKB, FBLN1, MYO6, RBP1, FRZB, GSTT1, MMP11, PPP1R3C, PMAIP1, CCL5, MYB, CXCL13, NPY1R, CPB1, PROS1, ACTG1P4, IFI16, KRT7, ABAT, PDZD2, MXRA5, GATA3, MKRN1, TGOLN2, CERS6, SPARC, MFAP4, AHSA2, NFIB, SHANK2, ELOVL2, MTMR6, CALU, HYMAI, TAF1B, C3, TRAPPC11, C6orf211, MST4, FAM134B, S100A14, EXOC5, TPRKB, CKLF, CECR1, ARMC9, WAC, PLA2G12A, SERINC3, PYCARD, WIZ
HOLP	

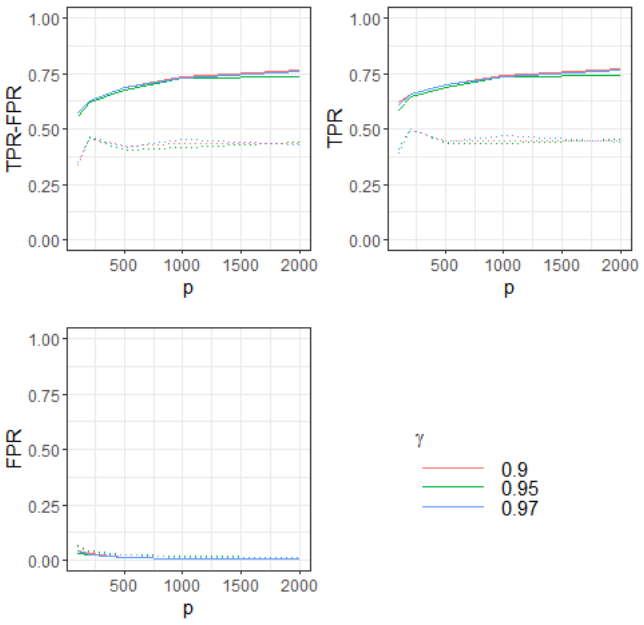

Figure 2.14: Top left: Average over the replications of max(TPR-FPR) and of the corresponding True Positive Rate (top right) and False Positive Rate (bottom left) for $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 0.5$ and $n = 50$. Solid line: optimal choice of λ, dotted line: choice of λ explained in Section 2.3.2 of the paper.

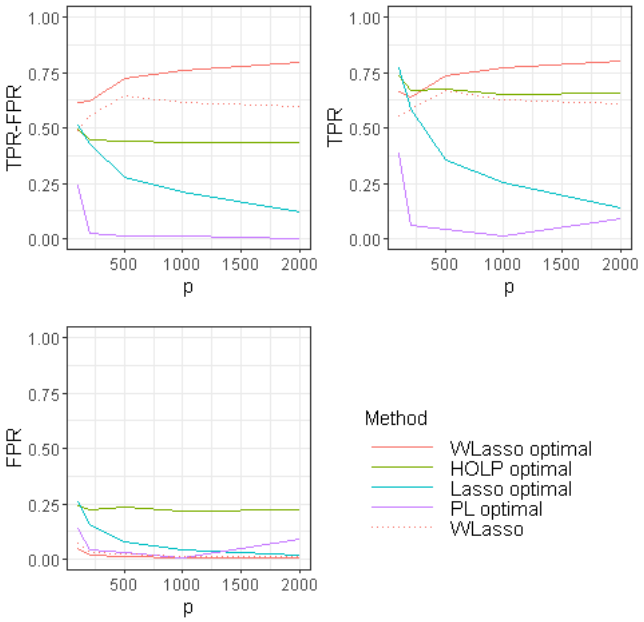

Figure 2.15: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) of the paper with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 1$ and $n = 50$.

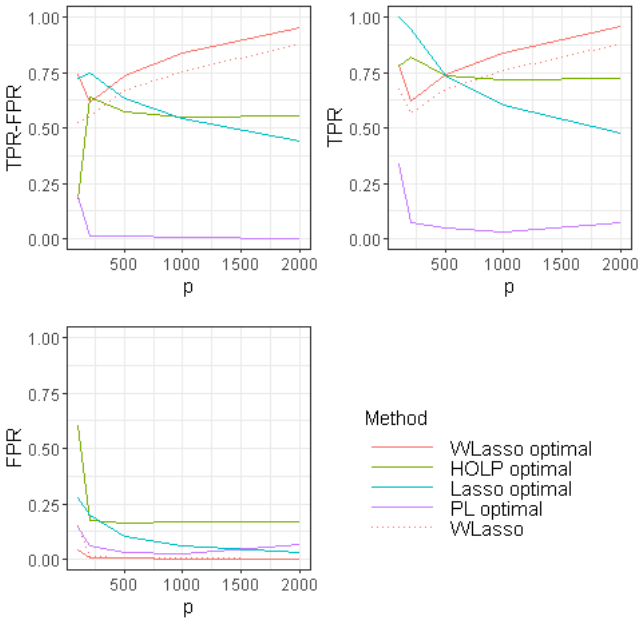

Figure 2.16: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, $b = 1$ and $n = 100$.

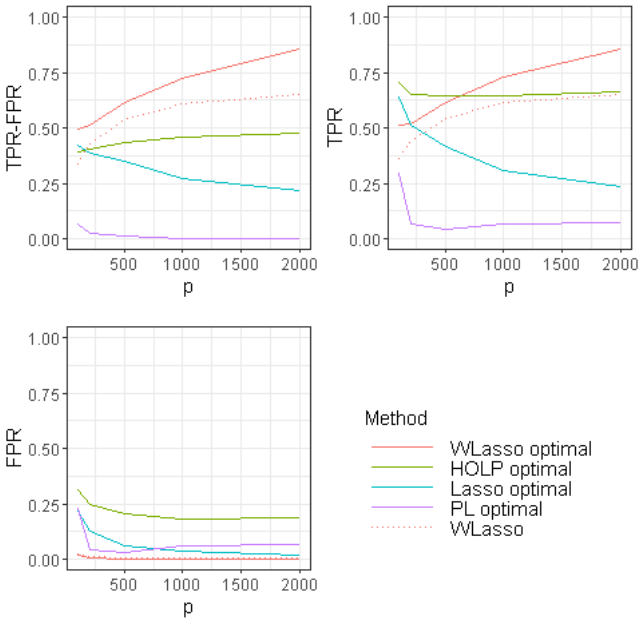

Figure 2.17: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$, $b = 1$ and $n = 50$.

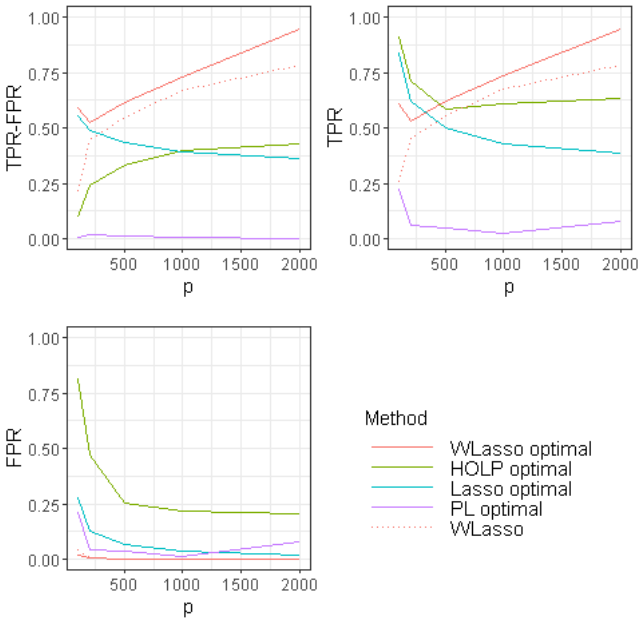

Figure 2.18: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) of the paper with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$, $b = 0.5$ and $n = 100$.

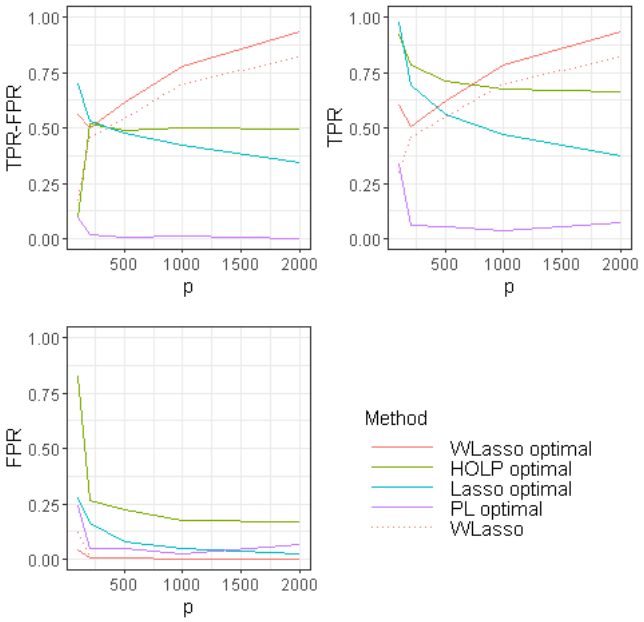

Figure 2.19: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when Σ has the block-wise correlation structure defined in (2.5) of the paper with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.5, 0.7, 0.9)$, $b = 1$ and $n = 100$.

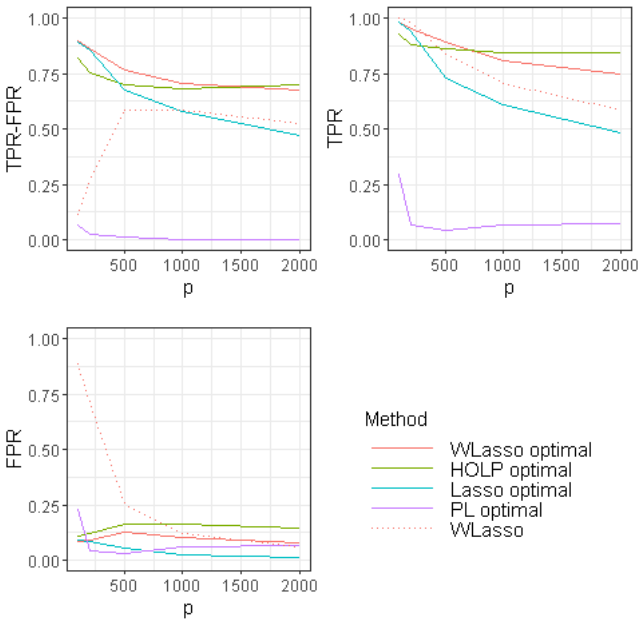

Figure 2.20: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\Sigma = \mathrm{Id}$, $b = 1$ and $n = 50$.

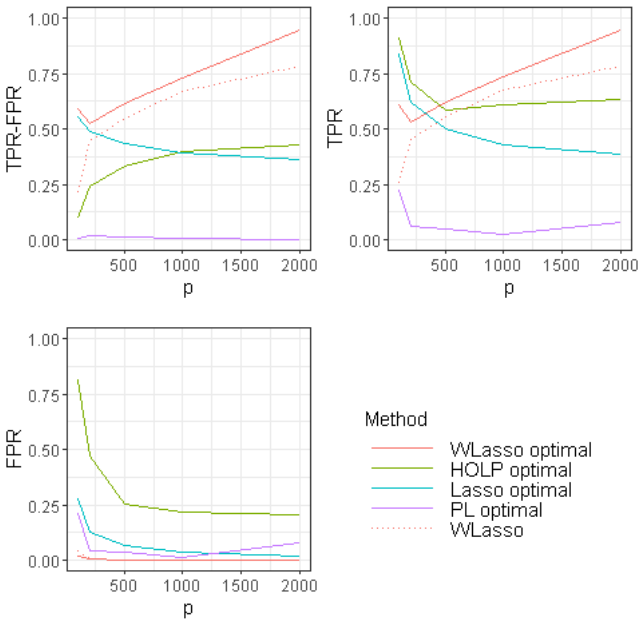

Figure 2.21: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\Sigma = \mathrm{Id}$, $b = 0.5$ and $n = 100$.

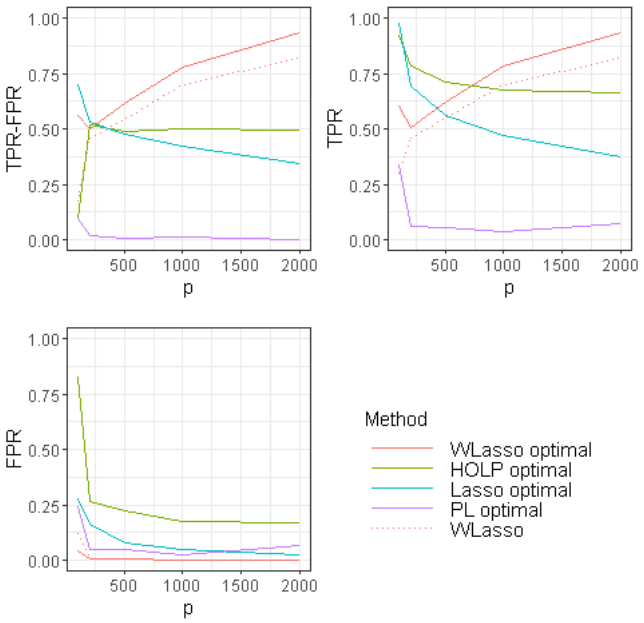

Figure 2.22: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, Precision Lasso (PL), WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $\Sigma = \mathrm{Id}$, $b = 1$ and $n = 100$.

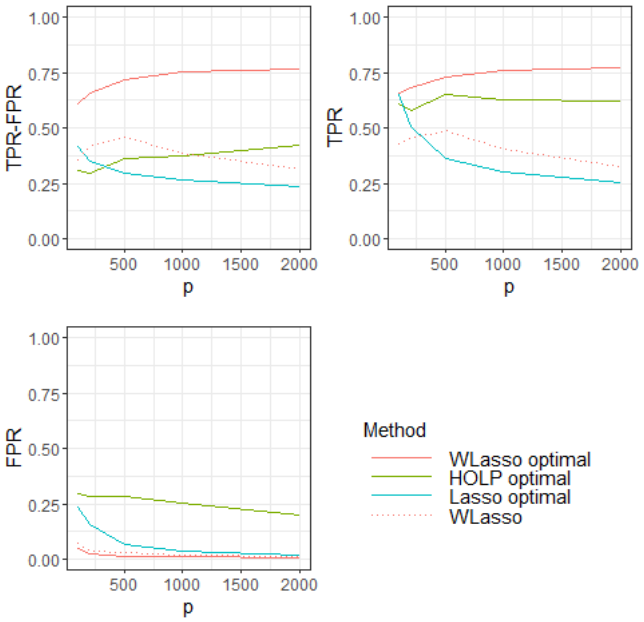

Figure 2.23: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $b = 0.5$, $n = 50$, Σ has the block-wise structure (2.5) of the paper but α_1 (resp. α_2, α_3) is randomly chosen in an interval of length 0.01 around 0.3 (resp. 0.5, 0.7).

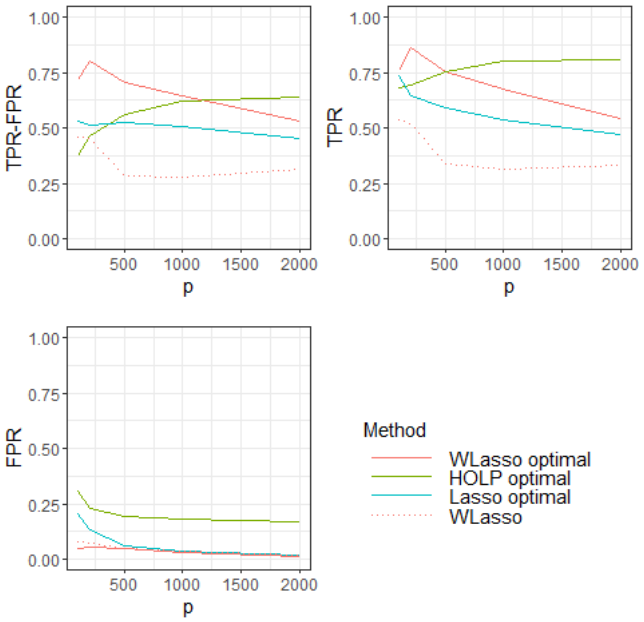

Figure 2.24: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when when $b = 0.5$, $n = 50$, Σ has the block-wise structure (2.5) of the paper but α_1 (resp. α_2, α_3) is randomly chosen in an interval of length 0.05 around 0.3 (resp. 0.5, 0.7).

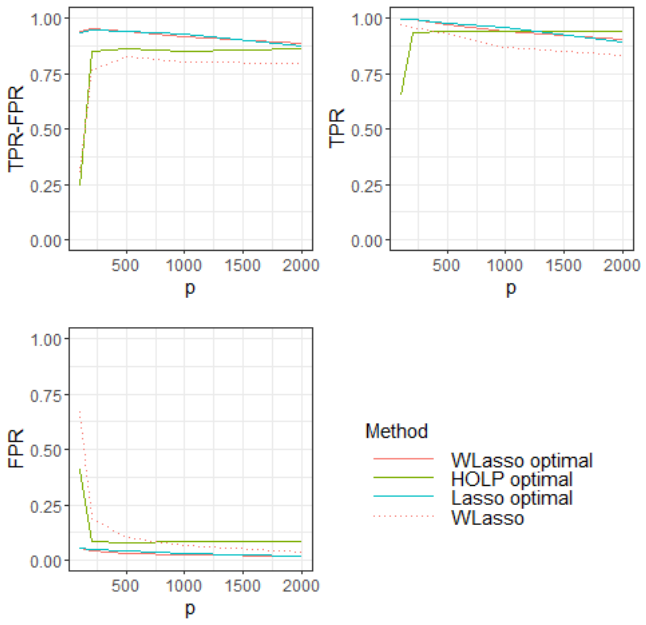

Figure 2.25: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when $b = 0.5$, $n = 100$, Σ has the block-wise structure (2.5) of the paper with $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$.

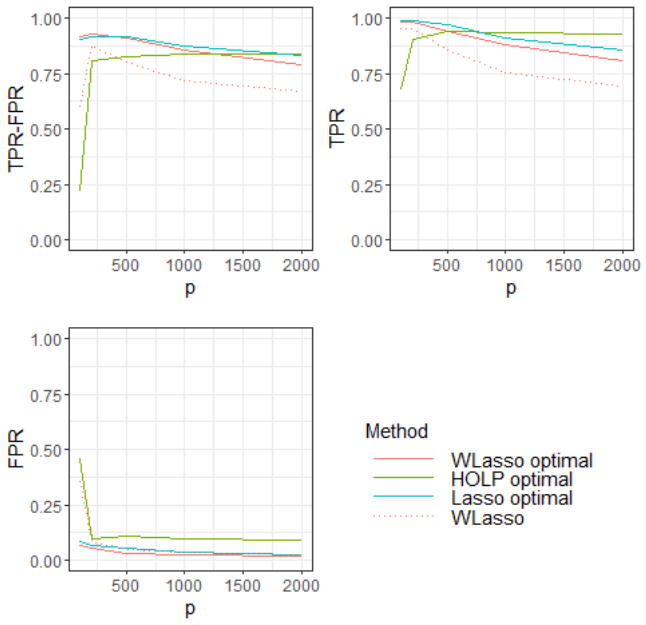

Figure 2.26: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when when $b = 0.5$, $n = 100$, Σ has the block-wise structure (2.5) of the paper with $\alpha_1 = \alpha_2 = \alpha_3 = 0.3$.

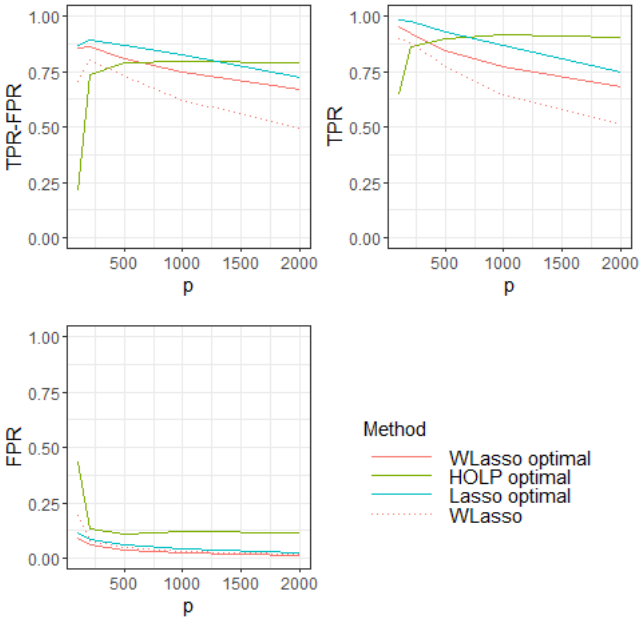

Figure 2.27: Top left: Average over the replications of max(TPR-FPR) for Lasso, HOLP, WLasso (solid line) and of (TPR-FPR) for WLasso obtained for the λ chosen by the strategy proposed in Section 2.3.2 of the paper (dotted line). Average of the corresponding TPR (top right) and FPR (bottom) when when $b = 0.5$, $n = 100$, Σ has the block-wise structure (2.5) of the paper with $\alpha_1 = \alpha_2 = \alpha_3 = 0.5$.

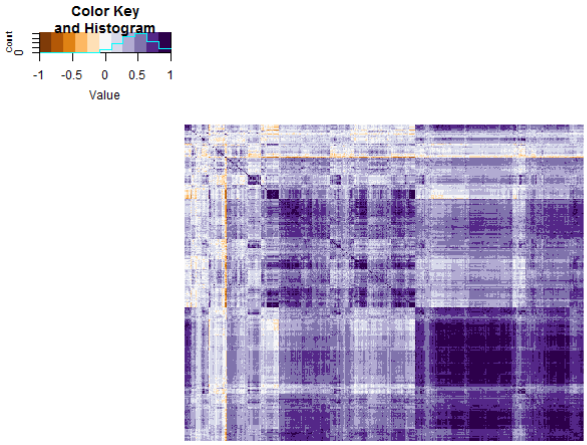

Figure 2.28: Heatmap of correlations between the probes.

Chapter 3 - Sign Consistency of the Generalized Elastic Net Estimator

Publication

The content of this chapter is in the article:
Zhu, W., Adjakossa, E., Lévy-Leduc, C., and Ternès, N. (2021). Sign Consistency of the Generalized Elastic Net Estimator. Submitted and also available on *arXiv preprint* (arXiv:2106.05454).

Abstract

In this paper, we propose a novel variable selection approach in the framework of high-dimensional linear models where the columns of the design matrix are highly correlated. It consists in rewriting the initial high-dimensional linear model to remove the correlation between the columns of the design matrix and in applying a generalized Elastic Net criterion since it can be seen as an extension of the generalized Lasso. The properties of our approach called gEN (generalized Elastic Net) are investigated both from a theoretical and a numerical point of view. More precisely, we provide a new condition called GIC (Generalized Irrepresentable Condition) which generalizes the EIC (Elastic Net Irrepresentable Condition) of Jia and Yu (2010) under which we prove that our estimator can recover the positions of the null and non null entries of the coefficients when the sample size tends to infinity. We also assess the performance of our methodology using synthetic data and compare it with alternative approaches. Our numerical experiments show that our approach improves the variable selection performance in many cases.

Contents

3.1. Introduction

Variable selection has become an important and actively used task for understanding or predicting an outcome of interest in many fields such as medicine (Lu et al., 2020; Gunter et al., 2011; Gu et al., 2013; Zhu et al., 2021), social media (Tufekci, 2014; Lin et al., 2016; Tomeny et al., 2017), or finance (Sermpinis et al., 2018; Amendola et al., 2017; Uniejewski et al., 2019). Through decades, numerous variable selection methods have been developed such as subset selection (Draper and Smith, 1998) or regularization techniques (Bickel et al., 2006). Subset selection methods achieve sparsity by selecting the best subset of relevant variables using the Akaike information criterion (Akaike, 1998) or the Bayesian information criterion (Schwarz et al., 1978) but are shown to be NP-hard and could be unstable in practice (Welch, 1982; Breiman et al., 1996). The regularized variable selection techniques have become popular for their capability to overcome the above difficulties (Tibshirani, 1996a; Hastie et al., 2015; Zou and Hastie, 2005; Wu et al., 2009). Among them, the Lasso approach (Tibshirani, 1996a) is one of the most popular and can be defined as follows. Let $\boldsymbol{y}$ satisfy the following linear model

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^{\star} + \boldsymbol{\epsilon}, \tag{3.1}$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ is the response variable, T denoting the transposition, $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ is the design matrix with n rows of observations on p covariates, $\boldsymbol{\beta}^{\star} = (\beta_1^{\star}, \ldots, \beta_p^{\star})^T \in \mathbb{R}^p$ is a sparse vector, namely contains a lot of null components, and $\boldsymbol{\epsilon}$ is a Gaussian vector with zero-mean and a covariance matrix equal to $\sigma^2\mathbb{I}_n$, $\mathbb{I}_n$ denoting the identity matrix in $\mathbb{R}^n$. The Lasso approach estimates $\boldsymbol{\beta}^{\star}$ with a sparsity enforcing constraint by minimizing the following penalized least-squares criterion:

$$L_{\lambda}^{Lasso}(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1, \tag{3.2}$$

where $\|a\|_1 = \sum_{k=1}^{p} |a_k|$ denotes the ℓ_1 norm of the vector $(a_1, \ldots, a_p)^T$, $\|b\|_2^2 = \sum_{k=1}^{n} b_k^2$ denotes the ℓ_2 norm of the vector $(b_1, \ldots, b_n)^T$, and λ is a positive constant corresponding to the regularization parameter. The Lasso popularity largely comes from the fact that the resulting estimator

$$\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, L_{\lambda}^{Lasso}(\boldsymbol{\beta})$$

is sparse (has only a few nonzero entries), and sparse models are often preferred for their interpretability (Zhao and Yu, 2006). Moreover, $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ can be proved to be sign consistent under some assumptions, namely there exists λ such that

$$\lim_{n\to\infty} \mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)\right) = sign(\boldsymbol{\beta}^{\star})\right) = 1,$$

where $\text{sign}(x) = 1$ if $x > 0$, -1 if $x < 0$ and 0 if $x = 0$. Before giving the conditions under which (Zhao and Yu, 2006) prove the sign consistency of $\widehat{\boldsymbol{\beta}}^{Lasso}$, we first introduce some notations. Without loss of generality, we shall assume as in (Zhao and Yu, 2006) that the first q components of $\boldsymbol{\beta}^\star$ are non null (*i.e.* the components that are associated to the active variables, and denoted as $\boldsymbol{\beta}_1^\star$) and the last $p-q$ components of $\boldsymbol{\beta}^\star$ are null (*i.e.* the components that are associated to the non active variables, and denoted as $\boldsymbol{\beta}_2^\star$). Moreover, we shall denote by $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) the first q (resp. the last $p-q$) columns of $\boldsymbol{X}$. Hence, $C_n = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, which is the empirical covariance matrix of the covariates, can be rewritten as follows:

$$C_n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix},$$

with $C_{11}^n = n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_1$, $C_{12}^n = n^{-1}\boldsymbol{X}_1^T\boldsymbol{X}_2$, $C_{21}^n = n^{-1}\boldsymbol{X}_2^T\boldsymbol{X}_1$, $C_{22}^n = n^{-1}\boldsymbol{X}_2^T\boldsymbol{X}_2$. It is proved by Zhao and Yu in (Zhao and Yu, 2006) that $\widehat{\boldsymbol{\beta}}^{Lasso}(\lambda)$ is sign consistent when the following Irrepresentable Condition (IC) is satisfied:

$$\left|\left(C_{21}^n (C_{11}^n)^{-1}\text{sign}(\boldsymbol{\beta}_1^\star)\right)_j\right| \leq 1-\alpha, \text{ for all } j, \tag{3.3}$$

where α is a positive constant. In the case where $p \gg n$, Wainwright develops in (Wainwright, 2009) the necessary and sufficient conditions, for both deterministic and random designs, on p, q, and n for which it is possible to recover the positions of the null and non null components of $\boldsymbol{\beta}^\star$, namely its support, using the Lasso.

When there are high correlations between covariates, especially the active ones, the C_{11}^n matrix may not be invertible, and the Lasso estimator fails to be sign consistent. To circumvent this issue, Zou and Hastie (Zou and Hastie, 2005) introduced the Elastic Net estimator defined by:

$$\widehat{\boldsymbol{\beta}}^{EN}(\lambda,\eta) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\arg\min}\, L_{\lambda,\eta}^{EN}(\boldsymbol{\beta}), \tag{3.4}$$

where

$$L_{\lambda,\eta}^{EN}(\boldsymbol{\beta}) = \|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 + \eta\|\boldsymbol{\beta}\|_2 \text{ with } \lambda,\eta > 0.$$

Yuan and Lin prove in (Yuan and Lin, 2007) that when the following Elastic Net Condition (EIC) is satisfied the Elastic Net estimator defined by (3.4) is sign consistent when p and q are fixed: there exist positive λ and η such that

$$\left|\left(C_{21}^n\left(C_{11}^n + \frac{\eta}{n}\mathbb{I}_q\right)^{-1}\left(\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\boldsymbol{\beta}_1^\star\right)\right)_j\right| \leq 1-\alpha, \text{ for all } j. \tag{3.5}$$

Moreover, when p, q, and n go to infinity with $p \gg n$, Jia and Yu prove in (Jia and Yu, 2010) that the sign consistency of the Elastic Net estimator holds if additionally to Condition (3.5) n goes to infinity at a rate faster than $q\log(p-q)$.

In the case where the active and non active covariates are highly correlated, IC (3.3) and EIC (3.5) may be violated. To overcome this issue several approaches were proposed: the Standard PArtial Covariance (SPAC) method (Xue and Qu, 2017) and preconditioning approaches among others. Xue and Qu (Xue and Qu, 2017) developed the so-called SPAC-Lasso which enjoys strong sign consistency in both finite-dimensional ($p < n$) and high-dimensional ($p \gg n$) settings. However, the authors mentioned that the SPAC-Lasso method only selects the active variables that are not highly correlated to the non active ones, which may be a weakness of this approach. The preconditioning approaches consist in transforming the given data $\boldsymbol{X}$ and $\boldsymbol{y}$ before applying the Lasso criterion. For example, (Jia and Rohe, 2015) and (Wang and Leng, 2016) proposed to left-multiply $\boldsymbol{X}$, $\boldsymbol{y}$ and thus $\boldsymbol{\epsilon}$ in Model (3.1) by specific matrices to remove the correlations between the columns of $\boldsymbol{X}$. A major drawback of the latter approach, 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.

Recently, Zhu et al. (2021) proposed another strategy under the following assumption:

1. $\boldsymbol{X}$ is assumed to be a random design matrix such that its rows $(\boldsymbol{x}_i)_{1\leq i\leq n}$ are i.i.d. zero-mean Gaussian random vectors having a covariance matrix equal to Σ.

More precisely, they propose to rewrite Model (3.1) in order to remove the correlation existing between the columns of $\boldsymbol{X}$. 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: $\Sigma = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$, then, denoting $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$, (3.1) can be rewritten as follows:

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}^{\star} + \boldsymbol{\epsilon}, \tag{3.6}$$

where $\widetilde{\boldsymbol{\beta}}^{\star} = \Sigma^{1/2}\boldsymbol{\beta}^{\star} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^T\boldsymbol{\beta}^{\star}$. With such a transformation, the covariance matrix of the n rows of $\widetilde{\boldsymbol{X}}$ is equal to identity and the columns of $\widetilde{\boldsymbol{X}}$ 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 $\boldsymbol{\epsilon}$ is not modified thus avoiding an increase of the noise which can overwhelm the benefits of a well conditioned design matrix. Their approach then consists in minimizing the following criterion with respect to $\widetilde{\boldsymbol{\beta}}$:

$$\left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda\left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1, \tag{3.7}$$

where $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$ in order to ensure a sparse estimation of $\boldsymbol{\beta}^{\star}$ thanks to the penalization by the ℓ_1 norm.

This criterion actually boils down to the Generalized Lasso proposed by (Tibshirani and Taylor, 2011):

$$L_{\lambda}^{genlasso}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|D\widetilde{\boldsymbol{\beta}}\right\|_1, \text{ with } \lambda > 0 \tag{3.8}$$

and $D = \Sigma^{-1/2}$.

Since, as explained in (Tibshirani and Taylor, 2011), some problems may occur when the rank of the design matrix is not full, we will consider in this paper the following criterion:

$$L_{\lambda,\eta}^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1 + \eta \left\|\widetilde{\boldsymbol{\beta}}\right\|_2^2, \text{ with } \lambda, \eta > 0. \tag{3.9}$$

Since it consists in adding an L_2 penalty part to the Generalized Lasso as in the Elastic Net, we will call it generalized Elastic Net (gEN). We prove in Section 3.2 that under Assumption 1 and the Generalized Irrepresentable Condition (GIC) (3.12) given below among others, $\widehat{\boldsymbol{\beta}}$ is a sign-consistent estimator of $\boldsymbol{\beta}^\star$ where $\widehat{\boldsymbol{\beta}}$ is defined by

$$\widehat{\boldsymbol{\beta}} = \Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}, \tag{3.10}$$

with

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right), \tag{3.11}$$

$L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right)$ being defined in Equation (3.9). The Generalized Irrepresentable Condition (GIC) can be stated as follows: There exist $\lambda, \eta, \alpha, \delta_4 > 0$ such that for all j,

$$\mathbb{P}\left(\left|\left((C_{21}^n + \frac{\eta}{n}\Sigma_{21})(C_{11}^n + \frac{\eta}{n}\Sigma_{11})^{-1}\left(\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda}\boldsymbol{\beta}_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^\star\right)_j\right| \leq 1 - \alpha\right) = 1 - o\left(e^{-n^{\delta_4}}\right). \tag{3.12}$$

Note that GIC coincides with EIC when $\boldsymbol{X}$ is not random and $\Sigma = \mathbb{I}_p$. Moreover, GIC does not require C_{11}^n to be invertible. Since EIC and IC are both particular cases of GIC, if the IC or EIC holds, then there exist λ or η such that the GIC holds.

The rest of the paper is organized as follows. Section 3.2 is devoted to the theoretical results of the paper. More precisely, we prove that under some mild conditions $\widehat{\boldsymbol{\beta}}$ defined in (3.10) is a sign-consistent estimator of $\boldsymbol{\beta}^\star$. To support our theoretical results, some numerical experiments are presented in Section 3.3. The proofs of our theoretical results can be found in Section 3.5.

3.2. Theoretical results

The goal of this section is to establish the sign consistency of the Generalized Elastic Net estimator defined in (3.10). To prove this result, we shall use the following lemma.

Lemma 3.2.1. *Let $\boldsymbol{y}$ satisfying Model (3.1) under Assumption 1 and $\widehat{\boldsymbol{\beta}}$ be defined in (3.10). Then,*

$$\mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}\right) = sign(\boldsymbol{\beta}^{\star})\right) \geq \mathbb{P}\left(A_n \cap B_n\right), \tag{3.13}$$

where

$$A_n := \left\{\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1)\right| < \sqrt{n}\left(\left|\boldsymbol{\beta}_1^{\star}\right| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^{\star})\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right|\right)\right\},$$

$$B_n := \left\{\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2)\right| \leq \frac{\lambda}{2\sqrt{n}} - \frac{\lambda}{2\sqrt{n}}\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\boldsymbol{\beta}_1^{\star}) + \frac{2\eta}{\lambda}\Sigma_{11}\boldsymbol{\beta}_1^{\star}\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1^{\star}\right|\right\},$$

and

$$\mathscr{C}_{11}^{n,\Sigma} = C_{11}^{n} + \frac{\eta}{n}\Sigma_{11},\ \mathscr{C}_{21}^{n,\Sigma} = C_{21}^{n} + \frac{\eta}{n}\Sigma_{21},\ W_n = \frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\epsilon} = \begin{bmatrix} W_n(1) \\ W_n(2) \end{bmatrix}, \tag{3.14}$$

with

$$W_n(1) = \frac{1}{\sqrt{n}}\boldsymbol{X}_1^T\boldsymbol{\epsilon}\ and\ W_n(2) = \frac{1}{\sqrt{n}}\boldsymbol{X}_2^T\boldsymbol{\epsilon}.$$

The proof of Lemma 3.2.1 is given in Section 3.5.

The following theorem gives the conditions under which the sign consistency of the generalized Elastic Net estimator $\widehat{\boldsymbol{\beta}}$ defined in (3.10) holds.

Theorem 3.2.2. *Assume that $\boldsymbol{y}$ satisfies Model (3.1) under Assumption 1 with $p = p_n$ such that $p_n \exp\left(n^{-\delta}\right)$ tends to 0 as n tends to infinity for all positive δ. Assume also that there exist some positive constants M_1, M_2, M_3 and α satisfying*

$$M_1 < \frac{\boldsymbol{\beta}_{\min}^2}{9\sigma^2} \quad and \quad \frac{\sqrt{2+\sqrt{2}}\sqrt{M_3}\sigma}{\alpha} < \frac{\boldsymbol{\beta}_{\min}}{3M_2\sqrt{q}}, \tag{3.15}$$

and that there exist $\lambda > 0$ and $\eta > 0$ such that (3.12) and

$$\frac{\lambda}{n} < \frac{2\boldsymbol{\beta}_{\min}}{3M_2\sqrt{q}}, \tag{3.16}$$

$$\frac{\lambda}{n} \geq \frac{2\sqrt{2+\sqrt{2}}\sqrt{M_3}\sigma}{\alpha}, \tag{3.17}$$

$$\frac{\eta}{n} < \frac{1}{3M_2\lambda_{\max}\left(\Sigma_{11}\right)} \times \frac{\boldsymbol{\beta}_{\min}}{\left\|\boldsymbol{\beta}_1^{\star}\right\|_2}, \tag{3.18}$$

hold as n tends to infinity, where $\boldsymbol{\beta}_{\min} = \min_{1\leq j\leq q}\left|\left(\boldsymbol{\beta}_1^{\star}\right)_j\right|$. Suppose also that there exist some positive constants δ_1, δ_2, δ_3 such that, as $n \to \infty$,

$$\mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) \leq M_1\right) = 1 - o\left(e^{-n^{\delta_1}}\right), \tag{3.19}$$

$$\mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \leq M_2\right) = 1 - o\left(e^{-n^{\delta_2}}\right), \tag{3.20}$$

$$\mathbb{P}\left(\lambda_{\max}\left(H_B H_B^T\right) \leq M_3\right) = 1 - o\left(e^{-n^{\delta_3}}\right), \tag{3.21}$$

where $\lambda_{\max}(A)$ *denotes the largest eigenvalue of* A*,*

$$H_A = \frac{1}{\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \boldsymbol{X}_1^T \text{ and } H_B = \frac{1}{\sqrt{n}}\left(\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \boldsymbol{X}_1^T - \boldsymbol{X}_2^T\right),$$

$\mathscr{C}_{11}^{n,\Sigma}$ *and* $\mathscr{C}_{21}^{n,\Sigma}$ *being defined in (3.14) and* $\boldsymbol{X}_1$ *(resp.* $\boldsymbol{X}_2$*) denoting the first* q *(resp. the last* $p-q$*) columns of* $\boldsymbol{X}$*. Then,*

$$\mathbb{P}\left(sign\left(\widehat{\boldsymbol{\beta}}\right) = sign(\boldsymbol{\beta}^\star)\right) \to 1, \ as\, n \to \infty,$$

where $\widehat{\boldsymbol{\beta}}$ *is defined in* (3.10)*.*

Note that Conditions (3.16) and (3.17) are consistent thanks to (3.15).

The proof of Theorem 3.2.2 is given in Section 3.5 and a discussion on the assumptions of Theorem 3.2.2 is provided in Section 3.3.

3.3. Numerical experiments

The goal of this section is to discuss the assumptions and illustrate the results of Theorem 3.2.2. For this, we generated datasets from Model (3.1) where the matrix Σ appearing in 1 is defined by

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix}. \tag{3.22}$$

In (3.22), Σ_{11} is the correlation matrix of the active variables having its off-diagonal entries equal to α_1, Σ_{22} is the correlation matrix of the non active variables having its off-diagonal entries equal to α_3 and Σ_{12} is the correlation matrix between the active and the non active variables with entries equal to α_2. In the numerical experiments, $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. Moreover, $\boldsymbol{\beta}^\star$ appearing in Model (3.1) has q non zero components which are equal to b and $\sigma = 1$. The number of predictors p is equal to 200, 400, or 600 and the sample size n takes the same values for each value of p.

3.3.1. Discussion on the assumptions of Theorem 3.2.2

We first show that GIC defined in (3.12) can be satisfied even when EIC and IC, defined in (3.5) and (3.3) respectively, are not fulfilled. For this, we computed for different values of λ and η the following values:

$$\begin{aligned}
\text{IC} &= \max_j \left(\left| \left(C_{21}^n (C_{11}^n)^{-1} (\text{sign}(\boldsymbol{\beta}_1^\star)) \right)_j \right| \right) \\
\text{EIC} &= \min_{\lambda,\eta} \max_j \left(\left| \left(C_{21}^n (C_{11}^n + \frac{\eta}{n} \mathbb{I}_q)^{-1} (\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \boldsymbol{\beta}_1^\star) \right)_j \right| \right) \\
\text{GIC} &= \min_{\lambda,\eta} \max_j \left(\left| \left((C_{21}^n + \frac{\eta}{n} \Sigma_{21})(C_{11}^n + \frac{\eta}{n} \Sigma_{11})^{-1} \left(\text{sign}(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right)_j \right| \right)
\end{aligned} \tag{3.23}$$

and Figure 3.1 displays the boxplots of these criteria obtained from 100 replications. We can see from these figures that in all the considered cases GIC is satisfied (*i.e.* all values are smaller than 1) whereas EIC and IC are not. The values of p and n do not seem to have a big impact on EIC and IC. However, contrary to p, n seems to have an influence on GIC which increases with n when $b = 1$ and decreases when n increases when $b = 10$.

Figures 3.2 and 3.3 show the behavior of $\lambda_{\max}\left(H_A H_A^T\right)$, $\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)$ and $\lambda_{\max}\left(H_B H_B^T\right)$ appearing in (3.19), (3.20) and (3.21) with respect to η for different values of n, p and for $q = 5$ or 10. These plots thus provide lower bounds for M_1, M_2 and M_3 appearing in the previous equations. Observe that (3.18) can be rewritten as:

$$\eta M_2 < \frac{n}{3\lambda_{\max}(\Sigma_{11})} \times \frac{\boldsymbol{\beta}_{\min}}{\left\|\boldsymbol{\beta}_1^\star\right\|_2}. \tag{3.24}$$

Based on the plots at the bottom right of Figures 3.2 and 3.3, we can see that there exist η's satisfying Condition 3.24 and thus (3.18) and that the interval in which the adapted η's lie is larger when $q = 5$ than when $q = 10$.

Based on the average of M_1 previously obtained, the left part of (3.15) is always satisfied as soon as $b > \sqrt{18}$. Based on the average of M_2 and M_3 previously obtained, the average of left-hand side and of the right-hand side of the right part of Equation (3.15) are displayed in Figures 3.4 and 3.5. We can see from these figures that it is only satisfied for large values of b. Moreover, it is more often satisfied when $q = 5$ than for $q = 10$.

We will show in the next section that even if the cases where all the conditions of the theorem are not fulfilled our method is robust enough to outperform the Elastic Net defined in (3.4) even in these cases.

3.3.2. Comparison with other methods

To assess the performance of our approach (gEN) in terms of sign-consistency with respect to other methods and to illustrate the results of Theorem 3.2.2, we

Figure 3.1: Boxplot of values defined in (3.23) and obtained from 100 replications.

computed the True Positive Rate (TPR), namely the proportion of active variables selected, and the False Positive Rate (FPR), namely the proportion of non active variables selected, of the Elastic Net and gEN estimators defined in (3.4) and (3.10), respectively. Note that the values of η are chosen in a regular grid going from 0.0001 (all variables included) to 50 (no variables are chosen), and the values of λ are automatically set by the package `genlasso`. The corresponding results are displayed in Figures 3.6 to 3.8. We can see from these figures that the gEN and the Elastic Net estimators have a TPR equal to 1 but that the FPR of gEN is smaller

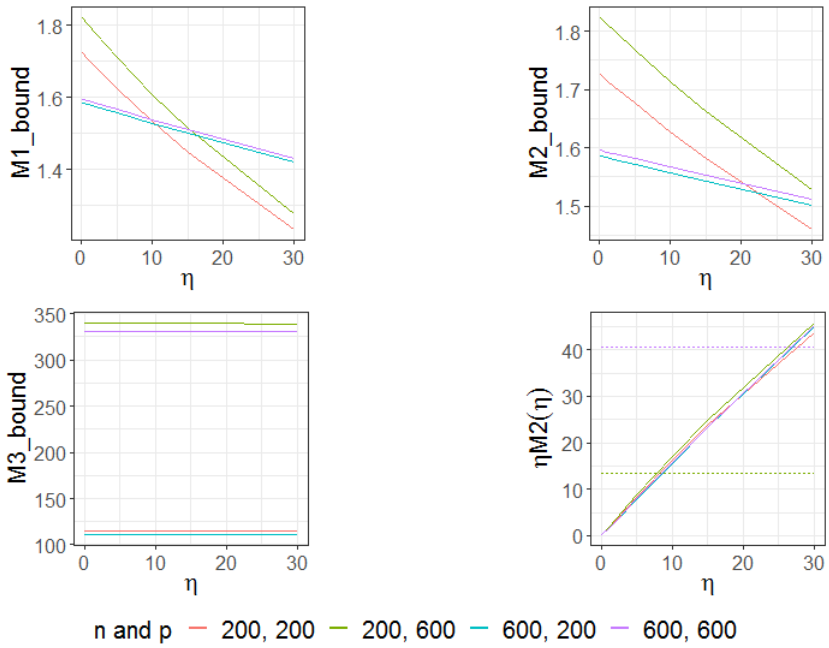

Figure 3.2: Top left: Average of $\lambda_{\max}\left(H_A H_A^T\right)$ in (3.19) as a function of η. Top right: Average of $\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)$ in (3.20) as a function of η. Bottom left: Average of $\lambda_{\max}\left(H_B H_B^T\right)$ in (3.21) as a function of η. Bottom right: Average of the left (resp. right) part of (3.24) in plain (resp. dashed) line. The averages are obtained from 10 replications. Here $q = 5$.

than Elastic Net. We can also see that the difference between the performance of gEN and Elastic Net is larger for high signal-to-noise ratios ($b = 10$). It has to be noticed that when TPR=1 for our approach it also means that the signs of the non null $\beta_i^\star$ are also properly retrieved.

Since Σ is unknown in practice, we also provide in Figure 3.9 the performance of our approach when Σ is estimated by using the R package `cvCovEst`. This latter approach is denoted gEN_est. More precisely, Figure 3.9 displays the empirical mean of the largest difference between the True Positive Rate and the False Positive Rate over the replications. It is obtained by selecting for each replication the value of λ and η achieving the largest difference between the TPR and FPR and by averaging these differences. This figure also displays the corresponding TPR and FPR for gEN with oracle Σ and Elastic Net for different values of n. We observe that all compared methods have TPR=1, which means that all True Positives have been retrieved. However, the False Positives selected by gEN (whether Σ was estimated or not) is smaller. Although this rate is slightly higher for gEN when Σ was estimated, gEN_est still outperforms ELastic Net.

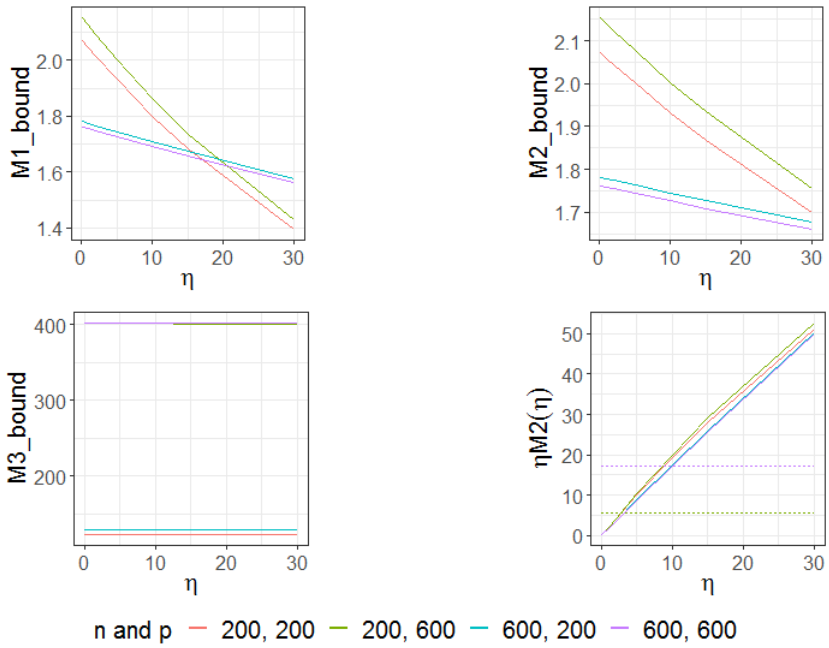

Figure 3.3: Top left: Average of $\lambda_{\max}\left(H_A H_A^T\right)$ in (3.19) as a function of η. Top right: Average of $\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)$ in (3.20) as a function of η. Bottom left: Average of $\lambda_{\max}\left(H_B H_B^T\right)$ in (3.21) as a function of η. Bottom right: Average of the left (resp. right) part of (3.24) in plain (resp. dashed) line. The averages are obtained from 10 replications. Here $q = 10$.

3.4. Discussion

In this paper, we proposed a novel variable selection approach called gEN (generalized Elastic Net) in the framework of linear models where the columns of the design matrix are highly correlated and thus when the standard Lasso criterion usually fails. We proved that under mild conditions, among which the GIC, which is valid when other standard conditions like EIC or IC are not fulfilled, our method provides a sign-consistent estimator of $\beta^\star$. For a more thorough discussion regarding the application of our approach in practical situations, we refer the reader to Zhu et al. (2021).

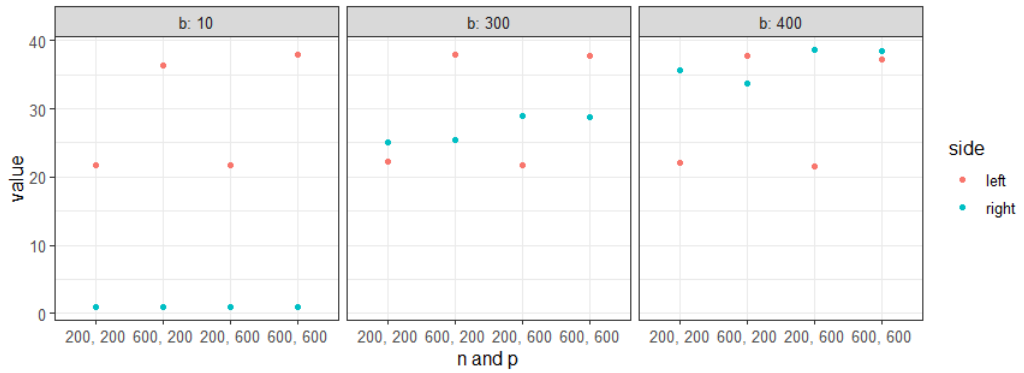

Figure 3.4: Average of the left-hand (resp. right-hand) side of the second part of (3.15) in red (resp. blue) for $q = 5$.

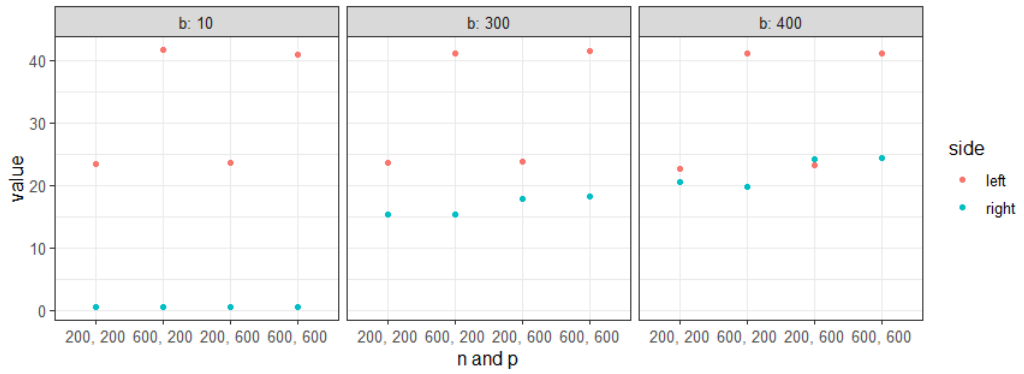

Figure 3.5: Average of the left-hand (resp. right-hand) side of the second part of (3.15) in red (resp. blue) for $q = 10$.

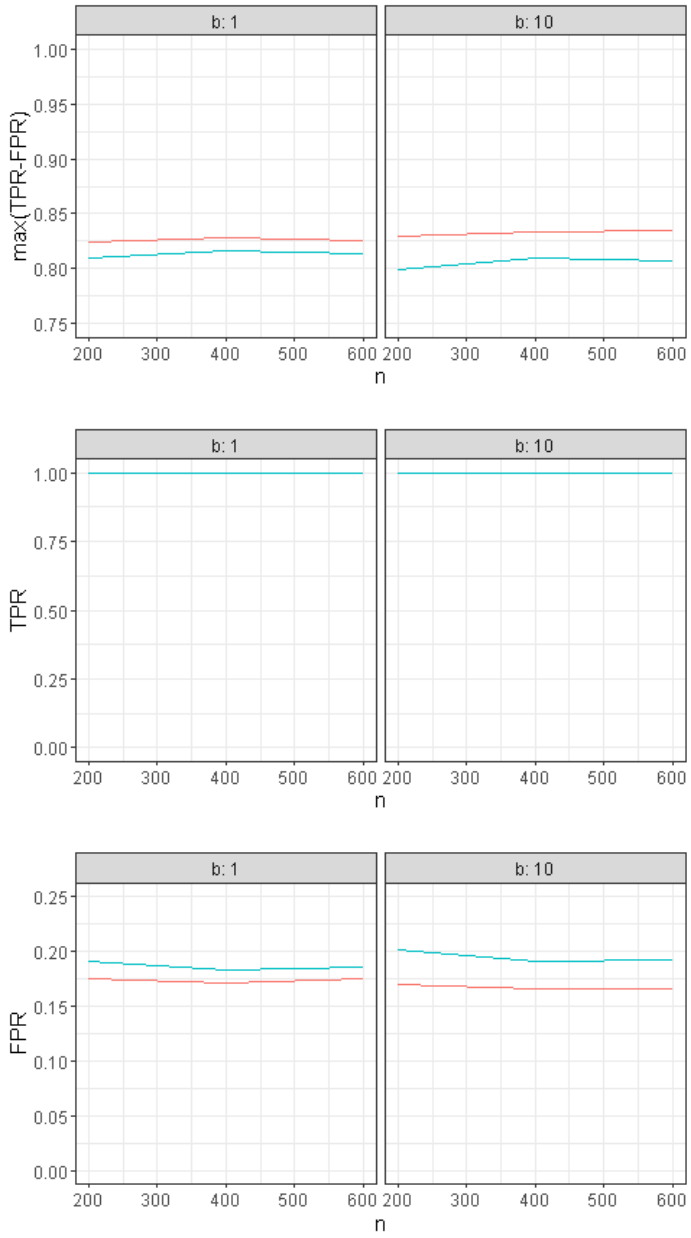

Figure 3.6: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 200$.

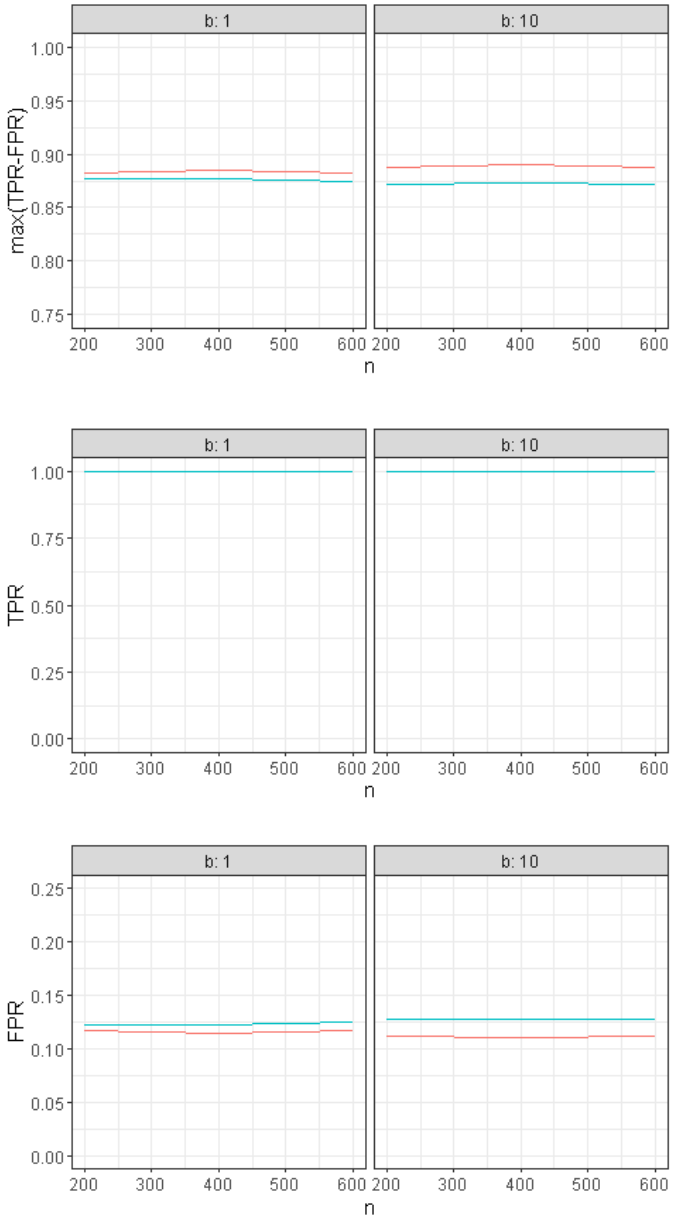

Figure 3.7: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 400$.

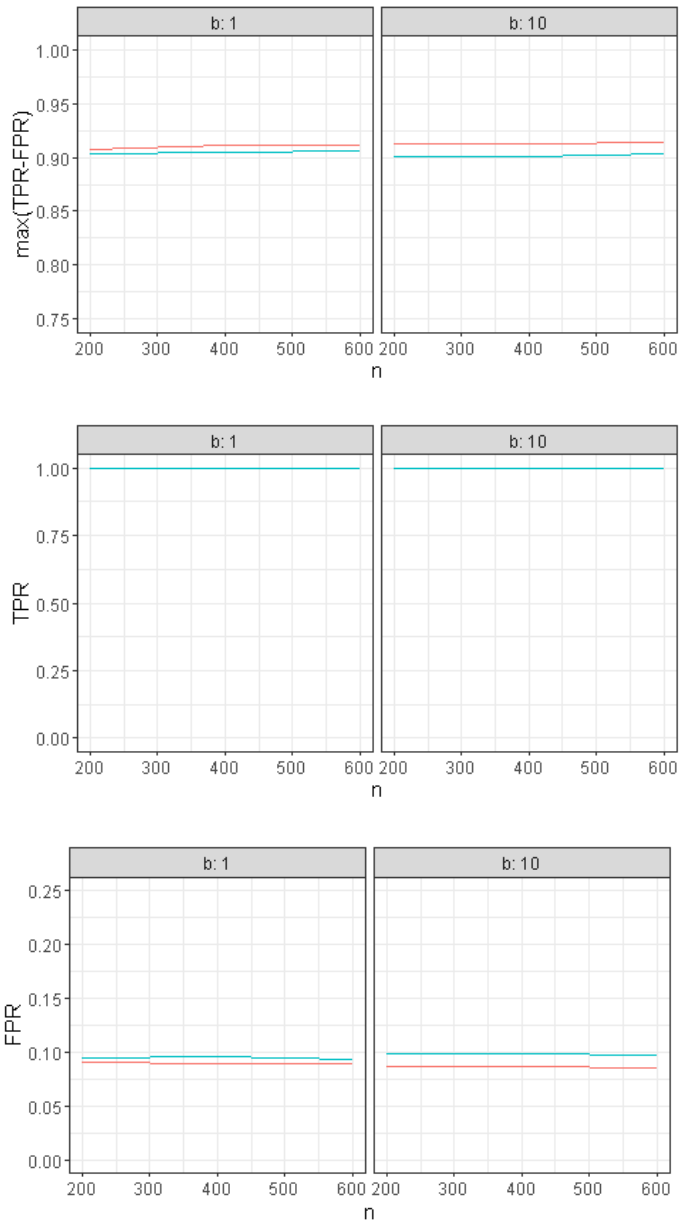

Figure 3.8: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red) and Elastic Net (in blue) with $p = 600$.

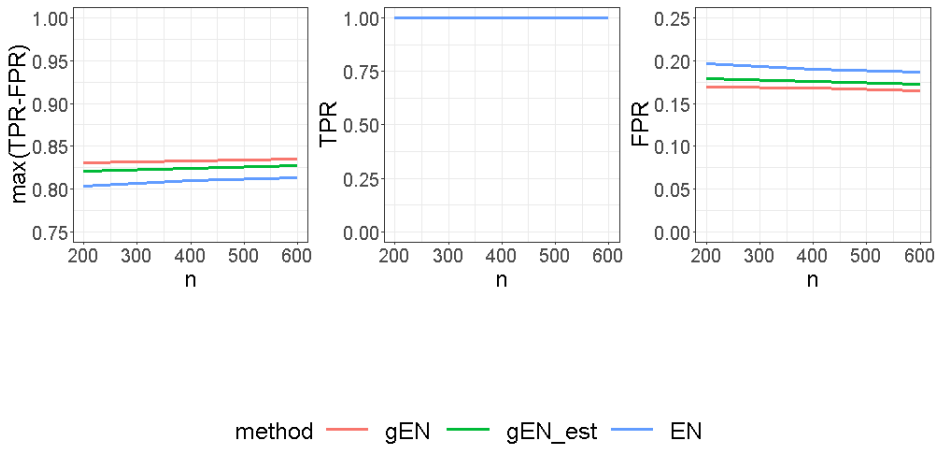

Figure 3.9: Average of max(TPR-FPR) and the corresponding TPR and FPR for gEN (in red), gEN_est (in green) and Elastic Net (in blue) with $p = 200$.

3.5. Proofs

3.5.1. Proof of Lemma 3.2.1

Note that (3.9) given by:

$$L^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1 + \eta \left\|\widetilde{\boldsymbol{\beta}}\right\|_2^2$$

can be rewritten as

$$L^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y}^* - \widetilde{\boldsymbol{X}}^*\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda \left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1,$$

where

$$\boldsymbol{y}^* = \begin{pmatrix} \boldsymbol{y} \\ 0 \end{pmatrix}, \quad \widetilde{\boldsymbol{X}}^* = \begin{pmatrix} \widetilde{\boldsymbol{X}} \\ \sqrt{\eta}\mathbb{I}_p \end{pmatrix}.$$

Then, $\widehat{\widetilde{\boldsymbol{\beta}}}$ satisfies

$$\widetilde{\boldsymbol{X}}_*{}^T \left(\boldsymbol{y}^* - \widetilde{\boldsymbol{X}}^*\widehat{\widetilde{\boldsymbol{\beta}}}\right) = \frac{\lambda}{2}(\Sigma^{-1/2})^T z, \tag{3.25}$$

where A^T denotes the transpose of the matrix A, and

$$\begin{cases} z_j = sign\left((\Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}})_j\right), & \text{if } (\Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}})_j \neq 0 \\ z_j \in [-1, 1], & \text{if } (\Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}})_j = 0 \end{cases}.$$

Equation (3.25) can be rewritten as:

$$\boldsymbol{X}^T\boldsymbol{y} - \left(\boldsymbol{X}^T\boldsymbol{X} + \eta\Sigma\right)\widehat{\boldsymbol{\beta}} = \frac{\lambda}{2}z$$

which leads to

$$\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{\beta}^\star - \widehat{\boldsymbol{\beta}}) + \boldsymbol{X}^T\boldsymbol{\epsilon} - \eta\Sigma\widehat{\boldsymbol{\beta}} = \frac{\lambda}{2}z,$$

by using that $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^\star + \boldsymbol{\epsilon}$. By using the following notations: $\widehat{\mathbf{u}} = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^\star$,

$$C_n = \frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \text{ and } W_n = \frac{1}{\sqrt{n}}\boldsymbol{X}^T\boldsymbol{\epsilon},$$

Equation (3.25) becomes

$$\left(C_n + \frac{\eta}{n}\Sigma\right)\sqrt{n}\widehat{\mathbf{u}} + \frac{\eta}{\sqrt{n}}\Sigma\boldsymbol{\beta}^\star - W_n = -\frac{\lambda}{2\sqrt{n}}z. \tag{3.26}$$

With the following notations:

$$C_n = \begin{pmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{pmatrix},\ \Sigma = \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix},\ \widehat{\mathbf{u}} = \begin{pmatrix} \widehat{\mathbf{u}}_1 \\ \widehat{\mathbf{u}}_2 \end{pmatrix},\ W_n = \begin{pmatrix} W_n(1) \\ W_n(2) \end{pmatrix},\ \boldsymbol{\beta}^\star = \begin{pmatrix} \boldsymbol{\beta}_1^\star \\ \mathbf{0} \end{pmatrix},$$

the first components of Equation (3.26) are:

$$\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \left(C_{12}^n + \frac{\eta}{n}\Sigma_{12}\right)\sqrt{n}\widehat{\mathbf{u}}_2 + \frac{\eta}{\sqrt{n}}\Sigma_{11}\boldsymbol{\beta}_1^\star - W_n(1) = -\frac{\lambda}{2\sqrt{n}}sign(\boldsymbol{\beta}_1^\star). \tag{3.27}$$

If $\widehat{\mathbf{u}} = \begin{pmatrix} \widehat{\mathbf{u}}_1 \\ 0 \end{pmatrix}$, it can be seen as a solution of the generalized Elastic Net criterion where, by Equation (3.27), $\widehat{\mathbf{u}}_1$ is defined by:

$$\sqrt{n}\widehat{\mathbf{u}}_1 = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - \frac{\eta}{\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1{}^\star - \frac{\lambda}{2\sqrt{n}}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1{}^\star), \tag{3.28}$$

where we used (3.14).

Note that the event A_n can be rewritten as follows:

$$\begin{aligned}\sqrt{n}\left(-\left|\boldsymbol{\beta}_1^\star\right| + \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^\star)\right| + \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right|\right) \\ < \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) < \\ \sqrt{n}\left(\left|\boldsymbol{\beta}_1^\star\right| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^\star)\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right|\right)\end{aligned}$$

which implies

$$\begin{aligned}\sqrt{n}\left(-\left|\boldsymbol{\beta}_1^\star\right| + \frac{\lambda}{2n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^\star) + \frac{\eta}{n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right) \\ < \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) < \\ \sqrt{n}\left(\left|\boldsymbol{\beta}_1^\star\right| + \frac{\lambda}{2n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\boldsymbol{\beta}_1^\star) + \frac{\eta}{n}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\boldsymbol{\beta}_1^\star\right),\end{aligned}$$

using that $-|x| \leq x \leq |x|$, $\forall x \in \mathbb{R}$. Then, by using (3.28), we get that $\sqrt{n}|\widehat{\mathbf{u}}_1| < \sqrt{n}|\beta_1^\star|$ and thus $|\widehat{\mathbf{u}}_1| < |\beta_1^\star|$. Notice that $|\widehat{\mathbf{u}}_1| < |\beta_1^\star|$ implies that $\widehat{\beta}_1 \neq 0$ and that $sign(\widehat{\beta}_1) = sign(\beta_1^\star)$. Moreover, since $\widehat{\mathbf{u}}_2 = 0$, we get that $sign(\widehat{\beta}) = sign(\beta^\star)$.

The last components of (3.26) satisfy:

$$\left(C_{21}^n + \frac{\eta}{n}\Sigma_{21}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \left(C_{22}^n + \frac{\eta}{n}\Sigma_{22}\right)\sqrt{n}\widehat{\mathbf{u}}_2 + \frac{\eta}{\sqrt{n}}\Sigma_{21}\beta_1^\star - W_n(2) = -\frac{\lambda}{2\sqrt{n}}z_2,$$

where by (3.25), $|z_2| \leq 1$. Hence,

$$\left|\left(C_{21}^n + \frac{\eta}{n}\Sigma_{21}\right)\sqrt{n}\widehat{\mathbf{u}}_1 + \frac{\eta}{\sqrt{n}}\Sigma_{21}\beta_1^\star - W_n(2)\right| \leq \frac{\lambda}{2\sqrt{n}},$$

which can be rewritten as follows by using (3.28):

$$\left|\mathscr{C}_{21}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(W_n(1) - \frac{\eta}{\sqrt{n}}\Sigma_{11}\beta_1^\star - \frac{\lambda}{2\sqrt{n}}sign(\beta_1^\star)\right) + \frac{\eta}{\sqrt{n}}\Sigma_{21}\beta_1^\star - W_n(2)\right| \leq \frac{\lambda}{2\sqrt{n}}. \tag{3.29}$$

When the event B_n is satisfied:

$$\begin{aligned} -\frac{\lambda}{2\sqrt{n}} + \frac{\lambda}{2\sqrt{n}}\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\beta_1^\star) - \frac{2\eta}{\lambda}\Sigma_{11}\beta_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\beta_1^\star\right| \\ \leq \mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) - W_n(2) \\ \leq \frac{\lambda}{2\sqrt{n}} - \frac{\lambda}{2\sqrt{n}}\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\beta_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\beta_1^\star\right) - \frac{2\eta}{\lambda}\Sigma_{21}\beta_1^\star\right|. \end{aligned} \tag{3.30}$$

By using that $-|x| \leq x \leq |x|$ for all x in $\mathbb{R}$, we get that it implies that

$$\left|\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1) - W_n(2) - \frac{\lambda}{2\sqrt{n}}\mathscr{C}_{21}^{n,\Sigma}\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\left(sign(\beta_1^\star) + \frac{2\eta}{\lambda}\Sigma_{11}\beta_1^\star\right) + \frac{\eta}{\sqrt{n}}\Sigma_{21}\beta_1^\star\right| \leq \frac{\lambda}{2\sqrt{n}},$$

which corresponds to (3.29). Thus, if A_n and B_n are satisfied, we get that $sign\left(\widehat{\beta}\right) = sign\left(\beta\right)$, which concludes the proof.

3.5.2. Proof of Theorem 3.2.2

By Lemma 3.2.1,

$$\mathbb{P}\left(sign\left(\widehat{\beta}\right) = sign(\beta^\star)\right) \geq \mathbb{P}\left(A_n \cap B_n\right) \geq 1 - \mathbb{P}\left(A_n^c\right) - \mathbb{P}\left(B_n^c\right),$$

where A_n^c and B_n^c denote the complementary of A_n and B_n, respectively. Thus, to prove the theorem it is enough to prove that $\mathbb{P}\left(A_n^c\right) \to 0$ and $\mathbb{P}\left(B_n^c\right) \to 0$ as $n \to \infty$.

Recall that

$$A_n := \left\{\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}W_n(1)\right| < \sqrt{n}\left(\left|\beta_1^\star\right| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}sign(\beta_1^\star)\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right|\right)\right\}.$$

Let ζ and τ be defined by

$$\zeta = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) \text{ and } \tau = \sqrt{n}\left(\left|\beta_1^\star\right| - \frac{\lambda}{2n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right| - \frac{\eta}{n}\left|\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right|\right).$$

Then,

$$\mathbb{P}(A_n) = \mathbb{P}\left(\forall j, |\zeta_j| < \tau_j\right).$$

Thus,

$$\mathbb{P}(A_n^c) = \mathbb{P}\left(\exists j, |\zeta_j| \geq \tau_j\right) \leq \sum_{j=1}^{q} \mathbb{P}\left(|\zeta_j| \geq \tau_j\right).$$

Note that

$$\begin{aligned}
\mathbb{P}(|\zeta_j| \geq \tau_j) &= \mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\left(\left|(\beta_1^\star)_j\right| - \frac{\lambda}{2n}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| - \frac{\eta}{n}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right|\right)\right) \\
&= \mathbb{P}\left(|\zeta_j| + \frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| + \frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right| \geq \sqrt{n}\left|(\beta_1^\star)_j\right|\right) \\
&\leq \mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right) + \mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right) \\
&+ \mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right). \qquad (3.31)
\end{aligned}$$

Observe that

$$\zeta = \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) = \frac{1}{\sqrt{n}}\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)^{-1} \boldsymbol{X}_1^T \boldsymbol{\epsilon} = H_A \boldsymbol{\epsilon},$$

where

$$H_A = \frac{1}{\sqrt{n}}\left(C_{11}^n + \frac{\eta}{n}\Sigma_{11}\right)^{-1} \boldsymbol{X}_1^T,$$

$\boldsymbol{X}_1$ denoting the columns of the design matrix $\boldsymbol{X}$ associated to the q active covariates. Thus, for all j in $\{1, \ldots, q\}$,

$$\zeta_j = \sum_{k=1}^{n} (H_A)_{jk}\boldsymbol{\epsilon}_k.$$

By using the Cauchy-Schwarz inequality,

$$\begin{aligned}
|\zeta_j| = \left|\sum_{k=1}^{n} (H_A)_{jk}\boldsymbol{\epsilon}_k\right| &\leq \left(\sum_{k=1}^{n} (H_A)_{jk}^2\right)^{1/2}\left(\sum_{k=1}^{n}\boldsymbol{\epsilon}_k^2\right)^{1/2} \\
&= \sqrt{\left(H_A H_A^T\right)_{jj}} \times \|\boldsymbol{\epsilon}\|_2 \\
&\leq \sqrt{\lambda_{\max}\left(H_A H_A^T\right)} \times \|\boldsymbol{\epsilon}\|_2.
\end{aligned}$$

Hence, the first term in the r.h.s. of (3.31) satisfies the following inequalities:

$$\begin{aligned}\mathbb{P}\left(|\zeta_j| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right) &\leq \mathbb{P}\left(\sqrt{\lambda_{\max}\left(H_A H_A^T\right)} \times \|\epsilon\|_2 \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right) \\ &\leq \mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) \times \|\epsilon\|_2^2 \geq n\frac{(\beta_1^\star)_j^2}{9}\right). \qquad (3.32)\end{aligned}$$

Since by (3.19), there exist $M_1 > 0$ and $\delta_1 > 0$ such that

$$\mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) \leq M_1\right) = 1 - o\left(e^{-n^{\delta_1}}\right), \text{ as } n \to \infty,$$

we have:

$$\begin{aligned}\mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) \times \|\epsilon\|_2^2 \geq n\frac{(\beta_1^\star)_j^2}{9}\right) &\leq \mathbb{P}\left(\|\epsilon\|_2^2 \geq n\frac{(\beta_1^\star)_j^2}{9M_1}\right) + \mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) > M_1\right) \\ &\leq \mathbb{P}\left(\frac{\|\epsilon\|_2^2}{\sigma^2} \geq \frac{n\beta_{\min}^2}{9M_1\sigma^2}\right) + o\left(e^{-n^{\delta_1}}\right).\end{aligned}$$

Using that $\frac{\|\epsilon\|_2^2}{\sigma^2} \sim \chi^2(n)$, we get, by Lemma (1) of Laurent and Massart (2000), that

$$\mathbb{P}\left(\lambda_{\max}\left(H_A H_A^T\right) \times \|\epsilon\|_2^2 \geq n\frac{(\beta_1^\star)_j^2}{9}\right) \leq \exp\left(-\frac{t}{2} + \frac{1}{2}\sqrt{n(2t-n)}\right) + o\left(e^{-n^{\delta_1}}\right), \qquad (3.33)$$

since $t = \frac{n\beta_{\min}^2}{9M_1\sigma^2} > \frac{n}{2}$ using that $\frac{2\beta_{\min}^2}{9\sigma^2} > M_1$ by (3.15).
By putting together Equations (3.32) and (3.33) we get

$$\mathbb{P}\left(|\zeta_j| > \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right) \leq \exp\left(-\frac{t}{2} + \frac{1}{2}\sqrt{n(2t-n)}\right) + o\left(e^{-n^{\delta_1}}\right), \qquad (3.34)$$

with $t = \frac{n\beta_{\min}^2}{9M_1\sigma^2} > \frac{n}{2}$.
Let us now derive an upper bound for the second term in the r.h.s. of (3.31):

$$\mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right).$$

By using the Cauchy-Schwarz inequality, we get that:

$$\begin{aligned}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| &= \left|\sum_{k=1}^{q}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)_{jk}\left(sign(\beta_1^\star)\right)_k\right| \\ &\leq \sqrt{\sum_{k=1}^{q}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)_{jk}^2} \times \|sign(\beta_1^\star)\|_2 \leq \sqrt{\left(\mathscr{C}_{11}^{n,\Sigma}\right)_{jj}^{-2}} \times \sqrt{q} \\ &\leq \lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \times \sqrt{q}.\end{aligned}$$

Then,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\lambda}{2\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} sign(\beta_1^\star)\right)_j\right| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right)\\
\leq\ &\mathbb{P}\left(\frac{\lambda}{2}\sqrt{\frac{q}{n}}\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right)\\
\leq\ &\mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{2n}{3\lambda\sqrt{q}}\left|(\beta_1^\star)_j\right|\right)\\
\leq\ &\mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{2n}{3\lambda\sqrt{q}}\beta_{\min}\right) = o\left(e^{-n^{\delta_2}}\right), \text{ as } n \to \infty, \qquad (3.35)
\end{aligned}$$

since $\frac{2n}{3\lambda\sqrt{q}}\beta_{\min} > M_2$ by (3.16). Let us now derive an upper bound for the third term in the r.h.s. of (3.31):

$$\mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right| > \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right).$$

We have

$$\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right| = \left|\sum_{k=1}^{q}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\right)_{jk}(\beta_1^\star)_k\right| \leq \sqrt{\sum_{k=1}^{q}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\right)_{jk}^2} \times \left\|\beta_1^\star\right\|_2$$

$$\leq \sqrt{\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}^2\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)} \times \left\|\beta_1^\star\right\|_2 \leq \lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)\lambda_{\max}(\Sigma_{11}) \times \left\|\beta_1^\star\right\|_2.$$

Thus,

$$\begin{aligned}
&\mathbb{P}\left(\frac{\eta}{\sqrt{n}}\left|\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\Sigma_{11}\beta_1^\star\right)_j\right| \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right)\\
\leq\ &\mathbb{P}\left(\frac{\eta}{\sqrt{n}}\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right)\lambda_{\max}(\Sigma_{11})\left\|\beta_1^\star\right\|_2 \geq \sqrt{n}\frac{\left|(\beta_1^\star)_j\right|}{3}\right)\\
\leq\ &\mathbb{P}\left(\lambda_{\max}\left(\left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1}\right) \geq \frac{n\beta_{\min}}{3\eta\left\|\beta_1^\star\right\|_2\lambda_{\max}(\Sigma_{11})}\right) = o\left(e^{-n^{\delta_2}}\right), \text{ as } n \to \infty, \qquad (3.36)
\end{aligned}$$

since $\frac{n\beta_{\min}}{3\eta\left\|\beta_1^\star\right\|_2\lambda_{\max}(\Sigma_{11})} > M_2$ by (3.18).
By putting together Equations (3.34), (3.35) and (3.36), we get:

$$\mathbb{P}\left(A_n^c\right) \leq q\exp\left[-\frac{n}{2}\left(\kappa - \sqrt{2\kappa - 1}\right)\right] + q \times o\left(e^{-n^{\delta_1}}\right) + 2q \times o\left(e^{-n^{\delta_2}}\right), \qquad (3.37)$$

with $\kappa = \frac{\beta_{\min}^2}{9M_1\sigma^2}$. Note that $\kappa - \sqrt{2\kappa - 1} > 0$ since $\kappa = \frac{\beta_{\min}^2}{9M_1\sigma^2} > 1$ by (3.15). Equation (3.37) then implies that

$$\mathbb{P}\left(A_n^c\right) \to 0 \text{ as } n \to \infty.$$

Let us now prove that $\mathbb{P}\left(B_n^c\right) \to 0$ as $n \to \infty$.
Recall that

$$B_n := \left\{ \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2) \right| \leq \frac{\lambda}{2\sqrt{n}} \right.$$
$$\left. - \frac{\lambda}{2\sqrt{n}} \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \Sigma_{11} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right| \right\},$$

Let

$$\boldsymbol{\psi} = \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} W_n(1) - W_n(2) = \frac{1}{\sqrt{n}} \left(\mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \boldsymbol{X}_1^T - \boldsymbol{X}_2^T \right) \boldsymbol{\epsilon} =: H_B \boldsymbol{\epsilon}$$

and

$$\boldsymbol{\mu} = \frac{\lambda}{2\sqrt{n}} - \frac{\lambda}{2\sqrt{n}} \left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \Sigma_{11} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right|.$$

Then,

$$\mathbb{P}(B_n^c) = \mathbb{P}(\exists j, |\psi_j| > \mu_j) \leq \sum_{j=q+1}^{p} \mathbb{P}(|\psi_j| > \mu_j).$$

By using the Cauchy-Schwarz inequality, we get that:

$$|\psi_j| = \left| \sum_{k=1}^{n} (H_B)_{jk} \boldsymbol{\epsilon}_k \right| \leq \left(\sum_{k=1}^{n} (H_B)_{jk}^2 \right)^{1/2} \times \|\boldsymbol{\epsilon}\|_2 = \sqrt{\left(H_B H_B^T\right)_{jj}} \times \|\boldsymbol{\epsilon}\|_2 \leq \sqrt{\lambda_{\max}(H_B H_B^T)} \times \|\boldsymbol{\epsilon}\|_2, \tag{3.38}$$

where

$$H_B H_B^T = \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} C_{11}^n \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \mathscr{C}_{12}^{n,\Sigma} - \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} C_{12}^n - C_{21}^n \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \mathscr{C}_{12}^{n,\Sigma} + C_{22}^n.$$

By (3.21), there exist $M_3 > 0$ and $\delta_3 > 0$ such that

$$\mathbb{P}\left(\lambda_{\max}\left(H_B H_B^T \right) \leq M_3 \right) = 1 - o\left(e^{-n^{\delta_3}} \right), \text{ as } n \to \infty.$$

By the GIC condition (3.12), there exist $\alpha > 0$ and $\delta_4 > 0$ such that for all j,

$$\mathbb{P}\left(\left| \mathscr{C}_{21}^{n,\Sigma} \left(\mathscr{C}_{11}^{n,\Sigma}\right)^{-1} \left(sign(\boldsymbol{\beta}_1^\star) + \frac{2\eta}{\lambda} \Sigma_{11} \boldsymbol{\beta}_1^\star \right) - \frac{2\eta}{\lambda} \Sigma_{21} \boldsymbol{\beta}_1^\star \right| \leq 1 - \alpha \right) = 1 - o\left(e^{-n^{\delta_4}} \right).$$

Thus, we get that:

$$
\begin{aligned}
\mathbb{P}(B_n^c) &\leq \sum_{j=q+1}^{p} \mathbb{P}\left(|\psi_j| > \mu_j\right) \\
&\leq \sum_{j=q+1}^{p} \mathbb{P}\left(|\psi_j| > \frac{\lambda\alpha}{2\sqrt{n}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq \sum_{j=q+1}^{p} \mathbb{P}\left(\sqrt{\lambda_{\max}(H_B H_B^T)} \times \|\epsilon\|_2 > \frac{\lambda\alpha}{2\sqrt{n}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right), \text{ using Equation (3.38)} \\
&\leq \sum_{j=q+1}^{p} \mathbb{P}\left(\lambda_{\max}(H_B H_B^T) \times \|\epsilon\|_2^2 > \frac{\lambda^2\alpha^2}{4n}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq \sum_{j=q+1}^{p} \mathbb{P}\left(\frac{\|\epsilon\|_2^2}{\sigma^2} > \frac{\lambda^2\alpha^2}{4nM_3\sigma^2}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq (p-q)\exp\left(-\frac{s}{2} + \frac{1}{2}\sqrt{n(2s-n)}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq (p-q)\exp\left(-\frac{n}{2}\left(\frac{s}{n} - \sqrt{2\frac{s}{n} - 1}\right)\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \\
&\leq (p-q)\exp\left(-\frac{n}{2}\right) + (p-q)o\left(e^{-n^{\delta_3}}\right) + (p-q)o\left(e^{-n^{\delta_4}}\right) \qquad (3.39)
\end{aligned}
$$

with $\frac{s}{n} = \frac{\lambda^2\alpha^2}{4n^2M_3\sigma^2}$ since $\frac{\lambda^2\alpha^2}{4n^2M_3\sigma^2} \geq 2 + \sqrt{2}$ by (3.17).

Finally, Equation (3.39) implies that

$$\mathbb{P}(B_n^c) \to 0, \text{ as } n \to \infty,$$

which concludes the proof.

Chapter 4 - Identification of prognostic and predictive biomarkers in high-dimensional linear models with PPLasso

Publication

The content of this chapter is in the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2022). Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso. Submitted and also available on *arXiv preprint* (arXiv:2202.01970).

The proposed method is implemented in the PPLasso R package available from the CRAN.

Abstract

In clinical development, identification of prognostic and predictive biomarkers is essential to precision medicine. Prognostic biomarkers can be useful for anticipating the prognosis of individual patients, and predictive biomarkers can be used to identify patients more likely to benefit from a given 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 (Prognostic Predictive Lasso) integrating 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 Lasso type approaches on both prognostic and predictive biomarker identification in various scenarios. Finally, our method is applied to publicly available transcriptomic data from clinical trial RV144. Our method is implemented in the PPLasso R package which is available from the Comprehensive R Archive Network (CRAN).

Contents

4.1. Introduction

With the advancement of precision medicine, there has been an increasing interest in identifying prognostic or predictive biomarkers in clinical development. 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. Ballman (2015) and Clark (2008) provided a comprehensive explanation and concrete examples to distinguish prognostic from predictive biomarkers, respectively.

Concerning the biomarker selection, the high dimensionality of genomic data is one of the main challenges as explained in Fan and Li (2006). To identify effective biomarkers in high-dimensional settings, several approaches can be considered including hypothesis-based tests described in McDonald (2009), wrapper approaches proposed in Saeys et al. (2007), and penalized approaches such as Lasso designed by Tibshirani (1996b) 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 Smith (2018). More efforts have been devoted to penalized methods given their ability to automatically perform variable selection and coefficient estimation simultaneously as highlighted in Fan and Lv (2009). However, Lasso showed some potential drawbacks when biomarkers are highly correlated. Particularly, when the Irrepresentable Condition (IC) proposed by Zhao and Yu (2006) 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 Wang et al. (2019). Several methods have been proposed to adress this issue. Elastic Net (Zou and Hastie, 2005) combines the ℓ_1 and ℓ_2 penalties and is particularly effective in tackling correlation issues and can generally outperform Lasso. Adaptive Lasso (Zou, 2006) proposes to assign adaptive weights for penalizing different coefficients in the ℓ_1 penalty, and its oracle property was demonstrated. Wang and Leng (2016) proposed the HOLP approach which consists in removing the correlation between the columns of the design matrix; Wang et al. (2019) proposed to handle the correlation by assigning similar weights to correlated variables in their approach called Precision Lasso; Zhu et al. (2021) proposed to remove the correlations by applying a whitening transformation to the data before using the generalized Lasso criterion designed by Tibshirani and Taylor (2011).

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. (2011) proposed to first predict response probabilities for treatment and use this proba-

bility as the response in a classification problem to find effective biomarkers. Tian et al. (2012) 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. (2011) proposed a method called SIDES, which adopts a recursive partitioning algorithm for screening treatment-by-biomarker interactions. This method was further improved in Lipkovich and Dmitrienko (2014) by adding another step of preselection on predictive biomarkers based on variable importance. The method was demonstrated with continuous endpoint. More recently, Sechidis et al. (2018) 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 identification of 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 (Windeler, 2000). In this paper, we propose a novel approach called PPLasso (Prognostic Predictive Lasso) to simultaneously identify prognostic and predictive biomarkers in a high dimensional setting with continuous endpoints, as presented in Section 4.2. Extensive numerical experiments are given in Section 4.3 to assess the performance of our approach and to compare it to other methods. PPLasso is also applied to the clinical trial RV144 in Section 4.4. Finally, we give concluding remarks in Section 4.5.

4.2. Method

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 Faraway (2002).

4.2.1. Statistical modeling

Let $\boldsymbol{y}$ be a continuous response or endpoint and t_1, t_2 two treatments. Let also $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) 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:

$$\boldsymbol{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \dots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \dots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix}. \tag{4.1}$$

To take into account the potential correlation that may exist between the biomarkers in the different treatments, we shall assume that the rows of $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) are independent centered Gaussian random vectors with a covariance matrice equal to Σ_1 (resp. Σ_2).

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

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \boldsymbol{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1p} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2p} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}, \tag{4.2}$$

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

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

with α_1 (resp. α_2) corresponding to the effects of treatment t_1 (resp. t_2). Moreover, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1p})^T$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2p})^T$) are the coefficients associated to each of the p biomarkers in treatment t_1 (resp. t_2) group, T denoting the matrix transposition and $\epsilon_{11}, \dots, \epsilon_{2n_2}$ are standard independent Gaussian random variables independent of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. When t_1 stands for the standard treatment or placebo, prognostic biomarkers are defined as those having non-zero

coefficients in β_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 $\beta_2 - \beta_1$ because they aim to highlight different effects between two different treatments.

Model (4.2) can be written as:

$$\boldsymbol{y} = \boldsymbol{X}\gamma + \epsilon, \tag{4.3}$$

with $\gamma = (\alpha_1, \alpha_2, \beta_1^T, \beta_2^T)^T$. 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 (Tibshirani (1996b)):

$$\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\gamma\right\|_2^2 + \lambda\left\|\gamma\right\|_1, \tag{4.4}$$

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 β_1 and $\beta_2 - \beta_1$ to allow different sparsity levels. Hence we propose to replace the penalty $\lambda\left\|\gamma\right\|_1$ in (4.4) by

$$\lambda_1\left\|\beta_1\right\|_1 + \lambda_2\left\|\beta_2 - \beta_1\right\|_1. \tag{4.5}$$

Thus, a first estimator of γ could be found by minimizing the following criterion with respect to γ:

$$\frac{1}{2}\left\|\boldsymbol{y} - \boldsymbol{X}\gamma\right\|_2^2 + \lambda_1\left\|\begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1}D_2 \end{bmatrix}\gamma\right\|_1, \tag{4.6}$$

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 $\boldsymbol{X}$. More precisely, we consider $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$, where

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \Sigma_1 & 0 \\ 0 & 0 & 0 & \Sigma_2 \end{bmatrix} \tag{4.7}$$

and define $\Sigma^{-1/2}$ by replacing in (4.7) Σ_i by $\Sigma_i^{-1/2}$, where $\Sigma_i^{-1/2} = \mathbf{U}_i\mathbf{D}_i^{-1/2}\mathbf{U}_i^T$, $\mathbf{U}_i$ and $\mathbf{D}_i$ being the matrices involved in the spectral decomposition of Σ_i for $i = 1$

or 2. With such a transformation the columns of $\widetilde{\boldsymbol{X}}$ are decorrelated and Model (4.3) can be rewritten as follows:

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon} \tag{4.8}$$

where $\widetilde{\boldsymbol{\gamma}} = \Sigma^{1/2}\boldsymbol{\gamma}$. The objective function (4.6) thus becomes:

$$L^{\text{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}) = \frac{1}{2}\left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}}\right\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \Sigma^{-1/2}\widetilde{\boldsymbol{\gamma}} \right\|_1 . \tag{4.9}$$

4.2.2. Estimation of $\widetilde{\gamma}$

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

$$\widehat{\widetilde{\boldsymbol{\gamma}}}_0(\lambda_1,\lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\widetilde{\boldsymbol{\beta}}}_{10}^T, \widehat{\widetilde{\boldsymbol{\beta}}}_{20}^T) = \arg\min_{\widetilde{\boldsymbol{\gamma}}} L^{\text{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}), \tag{4.10}$$

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 $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda_1,\lambda_2) = (\widehat{\widetilde{\boldsymbol{\beta}}}_{10}(\lambda_1,\lambda_2)^T, \widehat{\widetilde{\boldsymbol{\beta}}}_{20}(\lambda_1,\lambda_2)^T)^T$. For K_1 (resp. K_2) in $\{1,\dots,p\}$, 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 $\widetilde{\boldsymbol{\beta}} = (\widetilde{\boldsymbol{\beta}}_1^T, \widetilde{\boldsymbol{\beta}}_2^T)$ 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:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_{ij}^{(K_i)}(\lambda_1,\lambda_2) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1,\lambda_2), & j \in \text{Top}_{K_i} \\ K_1\text{th largest value of } |\widehat{\widetilde{\boldsymbol{\beta}}}_{i0j}(\lambda_1,\lambda_2)|, & j \notin \text{Top}_{K_i}. \end{cases} \tag{4.11}$$

Note that the corrections are only performed on $\widehat{\widetilde{\boldsymbol{\beta}}}_0$, the estimators $\widehat{\widetilde{\alpha}}_1$ and $\widehat{\widetilde{\alpha}}_2$ were not modified.

To illustrate the interest of using a thresholding step, we generated a dataset based on Model 4.3 with parameters described in Section 4.3.1 and $p = 500$. Moreover, to simplify the graphical illustrations, we focus on the case where $\lambda_1 = \lambda_2 = \lambda$. Figure 4.1 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. The choice of K_1 and K_2 will be explained in Section 4.2.4.

4.2.3. Estimation of γ

With $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}_1^T, \widehat{\widetilde{\boldsymbol{\beta}}}_2^T)$, the estimators of $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$ can be obtained by $\widehat{\boldsymbol{\beta}}_{10} = \Sigma_1^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_1$ and $(\widehat{\boldsymbol{\beta}}_{20} - \widehat{\boldsymbol{\beta}}_{10}) = \Sigma_2^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_2 - \Sigma_1^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}_1$. As previously, another thresholding was applied to $\widehat{\boldsymbol{\beta}}_{10}$ and $\widehat{\boldsymbol{\beta}}_{20}$: for $i = 1$ or 2,

$$\widehat{\boldsymbol{\beta}}_{ij}^{(M_i)}(\lambda_1,\lambda_2) = \begin{cases} \widehat{\boldsymbol{\beta}}_{i0j}(\lambda_1,\lambda_2), & j \in \text{Top}_{M_i} \\ 0, & j \notin \text{Top}_{M_i}, \end{cases} \tag{4.12}$$

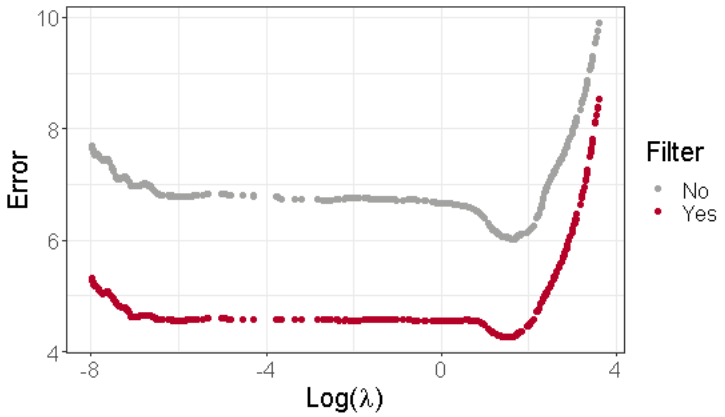

Figure 4.1: Estimation error $\left\|\widehat{\widetilde{\beta}}_0(\lambda) - \widetilde{\beta}\right\|_2$ (gray) and $\left\|\widehat{\widetilde{\beta}}(\lambda) - \widetilde{\beta}\right\|_2$ (red) for all λ.

for each fixed λ_1 and λ_2. The biomarkers with non-zero coefficients in $\widehat{\beta}_1 = \widehat{\beta}_1^{(M_1)}$ (resp. $\widehat{\beta}_2^{(M_2)} - \widehat{\beta}_1^{(M_1)}$) are considered as prognostic (resp. predictive) biomarkers, where the choice of M_1 and M_2 is explained in Section 4.2.4.

To illustrate the benefits of using an additional thresholding step, we used the dataset described in Section 4.2.2. Moreover, to simplify the graphical illustrations, we also focus on the case where $\lambda_1 = \lambda_2 = \lambda$. Figure 4.8 in the Supplementary material displays the number of True Positive (TP) and False Positive (FP) in prognostic and predictive biomarker identification with and without the second thresholding. We 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 (4.10).

4.2.4. 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:

$$\widetilde{\mathrm{MSE}}_{K_1,K_2}(\lambda_1,\lambda_2) = \|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widehat{\widetilde{\gamma}}^{(K1,K2)}(\lambda_1,\lambda_2)\|_2^2, \tag{4.13}$$

where $\widehat{\widetilde{\gamma}}^{(K1,K2)}(\lambda_1,\lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\widetilde{\beta}}_1^{(K_1)^T}, \widehat{\widetilde{\beta}}_2^{(K_2)^T})$ defined in (4.10) and in (4.11). It is displayed in the left part of Figure 4.2.

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:

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

The $\widehat{K_2}$ associated to each K_1 are displayed with '*' in the left part of Figure 4.2.

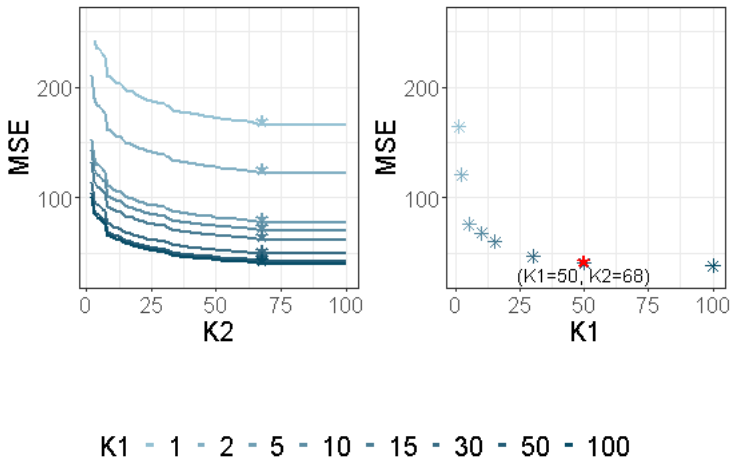

Figure 4.2: Illustration of how to choose K_1 and K_2 ($\delta = 0.95$), final choice is marked with '*'.

Then $\widehat{K_1}$ is chosen by using a similar criterion:

$$\widehat{K_1}(\lambda_1, \lambda_2) = \arg\min \left\{ K_1 \geq 1 \text{ s.t. } \frac{\widetilde{\text{MSE}}_{(K_1+1, \widehat{K}_2)}(\lambda_1, \lambda_2)}{\widetilde{\text{MSE}}_{(K_1, \widehat{K}_2)}(\lambda_1, \lambda_2)} \geq \delta \right\}.$$

The values of $\widetilde{\text{MSE}}_{(K_1, \widehat{K}_2)}(\lambda_1, \lambda_2)$ are displayed in the right part of Figure 4.2 in the particular case where $\lambda_1 = \lambda_2 = \lambda$, $\delta = 0.95$ and with the same dataset as the one used in Section 4.2.2. $\widehat{K}_1$ is displayed with a red star.

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

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

with $\widehat{\gamma}^{(M_1, M_2)}(\lambda_1, \lambda_2) = (\widehat{\widetilde{\alpha}}_1, \widehat{\widetilde{\alpha}}_2, \widehat{\boldsymbol{\beta}}_1^{(M_1)^T}, \widehat{\boldsymbol{\beta}}_2^{(M_2)^T})$ defined in (4.10) and (4.12). In the following, $\widehat{\gamma}(\lambda_1, \lambda_2) = \widehat{\gamma}^{(\widehat{M}_1, \widehat{M}_2)}(\lambda_1, \lambda_2)$.

4.2.5. Estimation of Σ_1 and Σ_2

As the empirical correlation matrix is known to be a non accurate estimator of Σ when p is larger than n, a new estimator has to be used. Thus, for estimating Σ we adopted a cross-validation based method designed by Boileau et al. (2021) and implemented in the `cvCovEst` R package (Boileau et al., 2021). This method chooses the estimator having the smallest estimation error among several compared methods (sample correlation matrix, POET (Fan et al. (2013)) and Tapering (Cai et al. (2010)) as examples). Since the samples in treatments t_1 and t_2 are assumed to be collected from the same population, Σ_1 and Σ_2 are assumed to be equal.

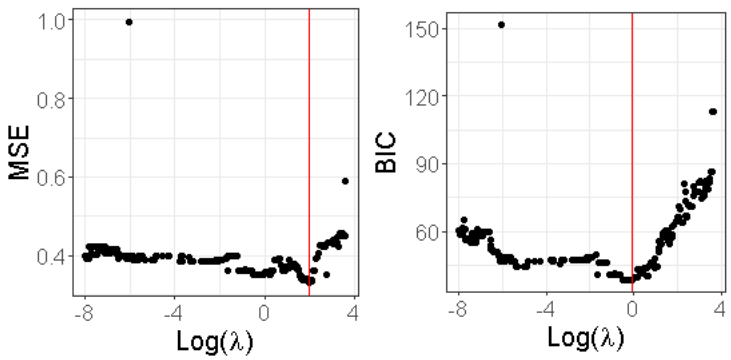

Figure 4.3: MSE and BIC for all λ. The λ minimizing each criterion is displayed with a vertical line.

4.2.6. 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:

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

where n is the total number of samples, $\mathrm{MSE}(\lambda) = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\gamma}}(\lambda)\|_2^2$ 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 Section 4.2.2 are displayed in Figure 4.3.

Table 4.2 in the supplementary material 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.

4.3. 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 and Adaptive Lasso since they also take into account the correlations. For Lasso, Elastic Net and Adaptive Lasso, in order to directly estimate prognostic and

predictive effects, $\boldsymbol{X}$ and γ in Model (4.3) were replaced by

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

and $\gamma^* = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1^*, \boldsymbol{\beta}_2^*)$, respectively, where $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are defined in (4.1), $\mathbf{0}_{i,j}$ (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 Lipkovich et al. (2017), $\boldsymbol{\beta}_1^* = \boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2^* = \boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$. The sparsity enforcing constraint was 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 $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$.

4.3.1. Simulation setting

All simulated datasets were generated from Model (4.3) where the n_1 (n_2) rows of $\boldsymbol{X}_1$ ($\boldsymbol{X}_2$) are assumed to be independent Gaussian random vectors with a covariance matrix $\Sigma_1 = \Sigma_2 = \Sigma_{bm}$, and $\boldsymbol{\epsilon}$ is a standard Gaussian random vector independent of $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. We defined Σ_{bm} as:

$$\Sigma_{bm} = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^T & \Sigma_{22} \end{bmatrix} \tag{4.14}$$

where Σ_{11} (resp. Σ_{22}) are the correlation matrix of prognostic (resp. non-prognostic) biomarkers with off-diagonal entries equal to a_1 (resp. a_3). Morever, Σ_{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 Xue and Qu (2017). We checked that the Irrepresentable Condition (IC) of Zhao and Yu (2006) 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 $\boldsymbol{\beta}_{2j} = b_2 = 2$ $(j = 6, ..., 10)$).

4.3.2. 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 TPR_{all} and minimize the FPR_{all}. All metrics were calculated by averaging the results of 100 replications for each scenario.

4.3.3. Biomarker selection results

For the proposed method, different results were presented. PPLasso_{Σ} (resp. PPLasso) corresponds to the results of the method by considering the true (resp. estimated) matrix Σ_{bm}. For estimating Σ_{bm}, we used the approach explained in Section 4.2.5. Two choices of λ are also presented: "optimal" and "min(bic)". The former gives the optimal selection that maximizes $(\text{TPR}_{\text{all}} - \text{FPR}_{\text{all}})$ which is also the choice used for Lasso, Elastic Net and Adaptive Lasso in these simulations. All these three methods are implemented with the `glmnet` R package, and the parameter α used for Elastic Net varies from 0.1 to 0.9. The choice of "min(bic)" is only applied to our method, which minimizes the BIC criterion defined in Section 4.2.6. For ease of presentation, the abbreviation EN (resp. AdLasso) refers to Elastic Net (resp. Adaptive Lasso) in the following.

Figure 4.4 shows the selection performance of PPLasso and other compared methods in the simulation scenario presented in Section 4.3.1. PPLasso achieved to select all prognostic biomarkers (TPR_{prog} almost 1) even for large p, with limited false positive prognostic biomarkers selected. As compared to the optimal λ maximizing $(\text{TPR}_{\text{all}} - \text{FPR}_{\text{all}})$, the one selected with the BIC tends to select some false positives (average: 33 ($\text{FPR}_{\text{prog}} = 0.17$) for $p = 200$ and 10 ($\text{FPR}_{\text{prog}} = 0.005$) for $p = 2000$). The results obtained from the oracle and estimated Σ_{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 $\text{TPR}_{\text{pred}} = 0.98$ and 0.80 for oracle and estimated Σ_{bm}, respectively, with $p = 2000$). In this scenario where the IC is violated, PPLasso globally outperforms Lasso, Elastic Net and Adaptive Lasso. 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 Σ

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 Σ_{bm} is the identity matrix.

For the scenario with compound symmetry structure displayed in Figure 4.5, all the methods successfully identified the true prognostic biomarkers (TPR_{prog} close

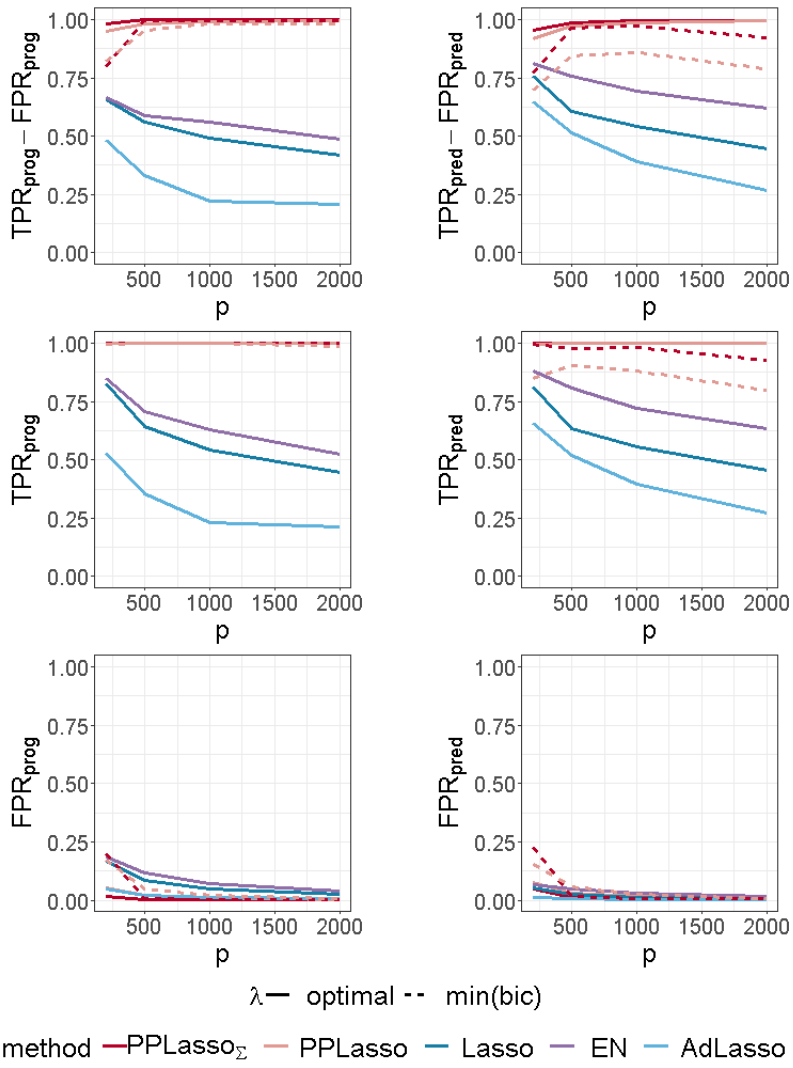

Figure 4.4: Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers.

to 1 even for large p) with limited false positive selection. On the other hand, the compared methods (Lasso, Elastic Net, Adaptive Lasso) 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)), $\text{PPLasso}_{\text{est}}$ 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 Figure 4.6, prognostic biomarkers were globally well identified by all the compared methods with a slightly higher

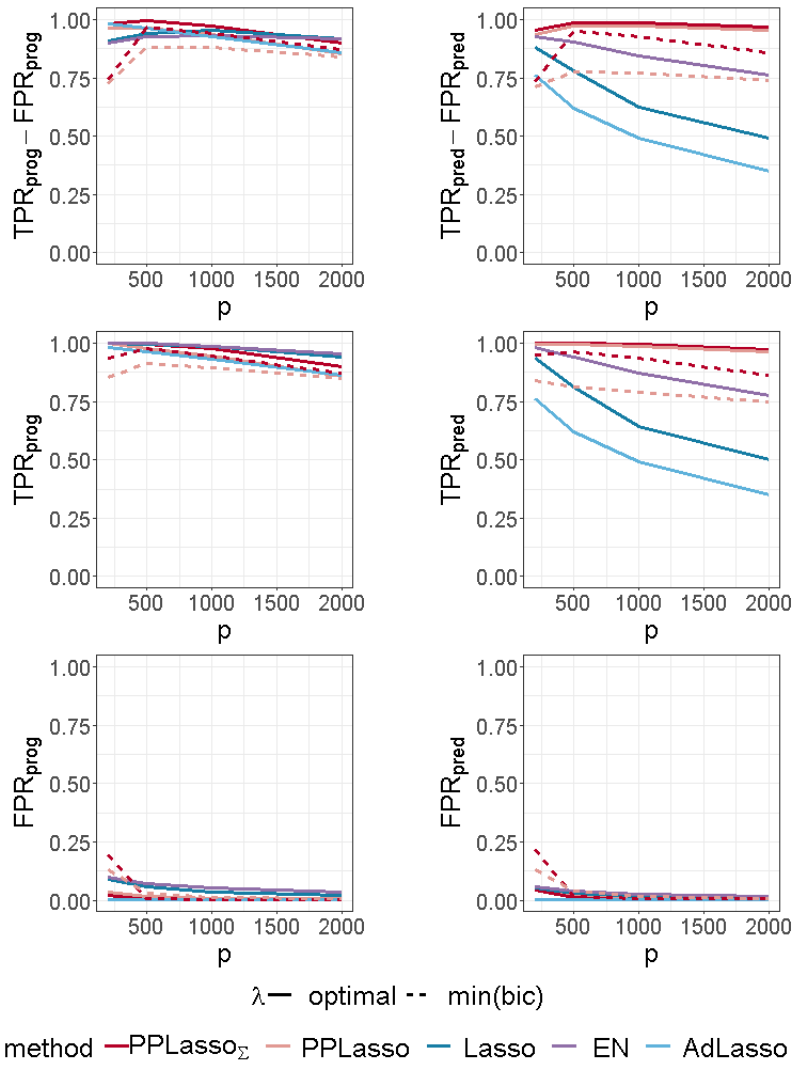

Figure 4.5: Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers for the compound symmetry correlation structure.

TPR_{prog} for Lasso and Elastic Net as compared to PPLasso but also with a slightly higher FPR_{prog}. With regards to predictive biomarkers, PPLasso using Σ_{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.

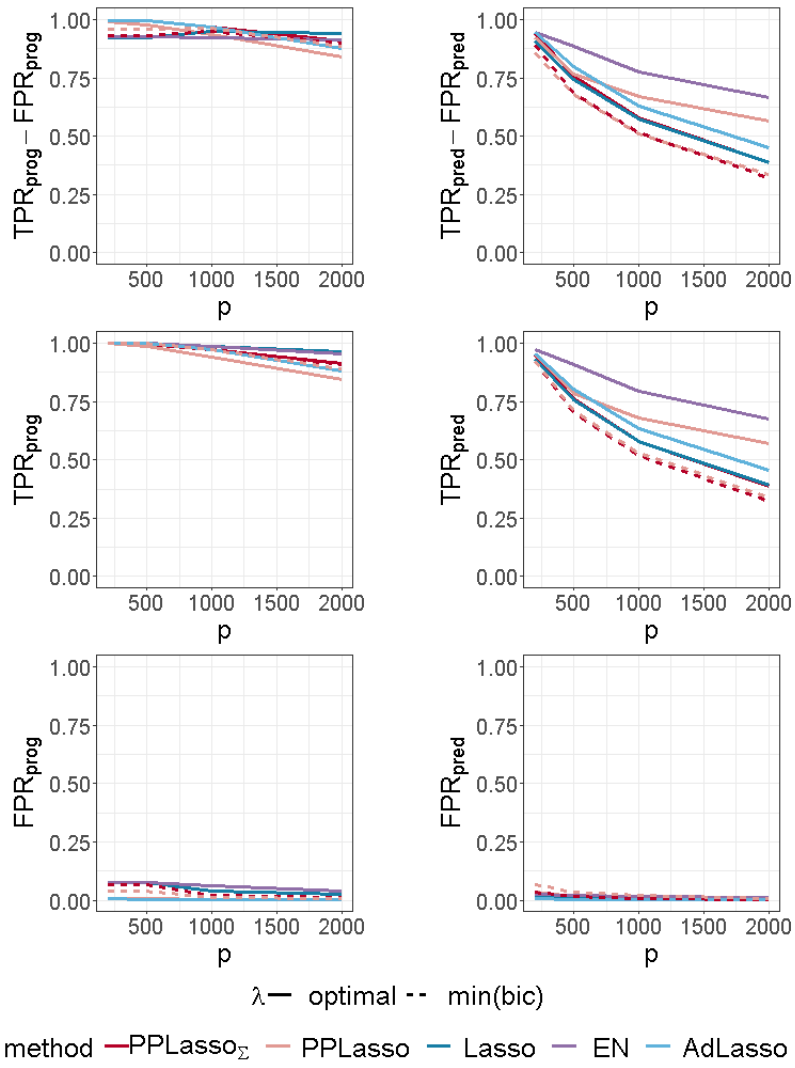

Figure 4.6: Average of (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers (independent setting).

Impact of the effect size of active biomarkers

To evaluate the impact of the effect size on biomarker selection performance, the scenario presented in Section 4.3.1 was considered with different values of b_2: 1.5, 2 and 2.5.

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 Figure 4.4 and Figure 4.9 of the supplementary material. The selection performance of PPLasso when λ is selected with min(bic) is also reduced by decreasing b_2, especially when Σ_{bm} is also estimated. Nevertheless, the selection performance of PPLasso remains better than for the other compared methods for which the performance displayed are associated to the optimal value of λ. 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 Figure 4.10 of the supplementary material. As compared with PPLasso, for which the selection performance remained stable as p increased, Lasso, Elastic Net and Adaptive Lasso 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 Section 4.3.1. 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 Figure 4.11 of the supplementary material that it became even harder to identify predictive biomarkers. TPR_{pred} decreased compared to Figure 4.4, 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 Figure 4.12 of the supplementary material 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 Figure 4.13 of the supplementary material 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}.

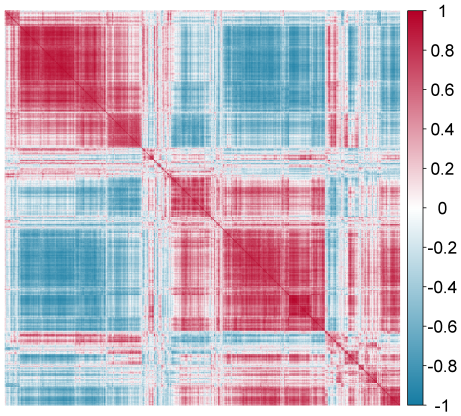

Figure 4.7: Heatmaps of the correlation matrix estimated by the cvCovEst R package.

4.4. Application to real clinical trials

We applied the previously described methods to publicly available transcriptomic data from the RV144 vaccine trial (Rerks-Ngarm et al. (2009)). 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 (Fourati et al. (2019), Gene Expression Omnibus accession code: GSE103671).

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 (Fourati et al. (2019), Akbay et al. (2020)). The gene expression has been normalized as in the original publication of Fourati et al. (2019). 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 Section 4.2.6.

The estimation of Σ 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

Table 4.1: Selected genes from PPLasso, Lasso, Elastic Net and Adaptive Lasso.

	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

matrix and a dense target matrix (*denseLinearShrinkEst*) derived by Ledoit and Wolf (2020) provides the smallest estimation error. Figure 4.7 displays the estimated Σ and highlights the strong correlation between the genes. Table 4.3 of the Supplementary material gives details on the compared estimators.

Prognostic and predictive genes selected by PPLasso, Lasso, Elastic Net and Adaptive Lasso are listed in Table 4.1. 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), 2 predictive genes (YTHDC1 and RPL21)). Interestingly, some genes selected by most methods such as SLAMF7, TNFRSF6B, TNFRSF18 or NUCKS1 have already been discussed in the HIV-1 field. Moreover, among the predictive genes selected by the PPLasso, some are linked to pathways that have been highlighted as possible target for HIV-1 such as BIRC3 and TLR8.

4.5. 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.

Appendix

This supplementary material provides additional numerical experiments, figures and a table for Chapter 4: "Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso".

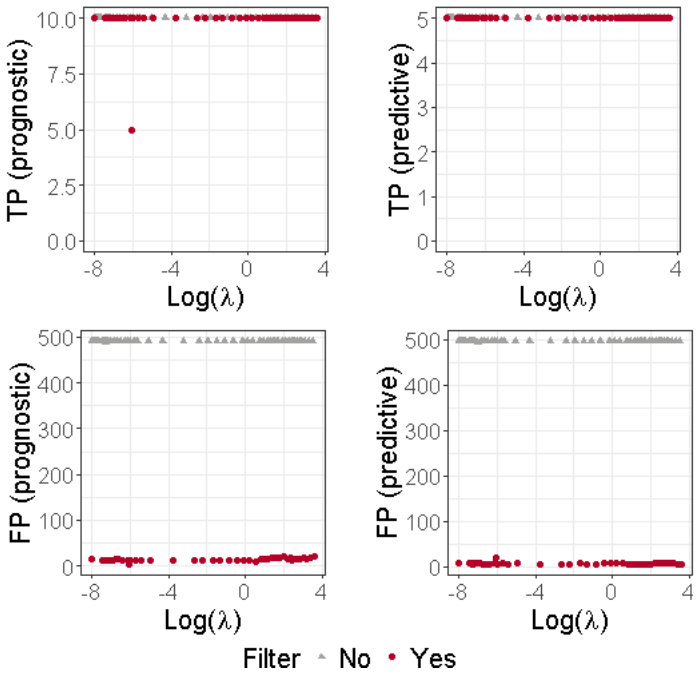

Figure 4.8: Number of True Positives and True Negatives for $\widehat{\beta}$ and $\widehat{\beta}_0$ on prognostic/predictive biomarkers.

	MSE	BIC
TPR(prognostic)	1.000	1.000
FPR(prognostic)	0.038	0.024
TPR(predictive)	1.000	1.000
FPR(predctive)	0.008	0.006

Table 4.2: TPR and FPR associated to prognostic and predictive biomarker identification with the λ chosen in Figure 4.3.

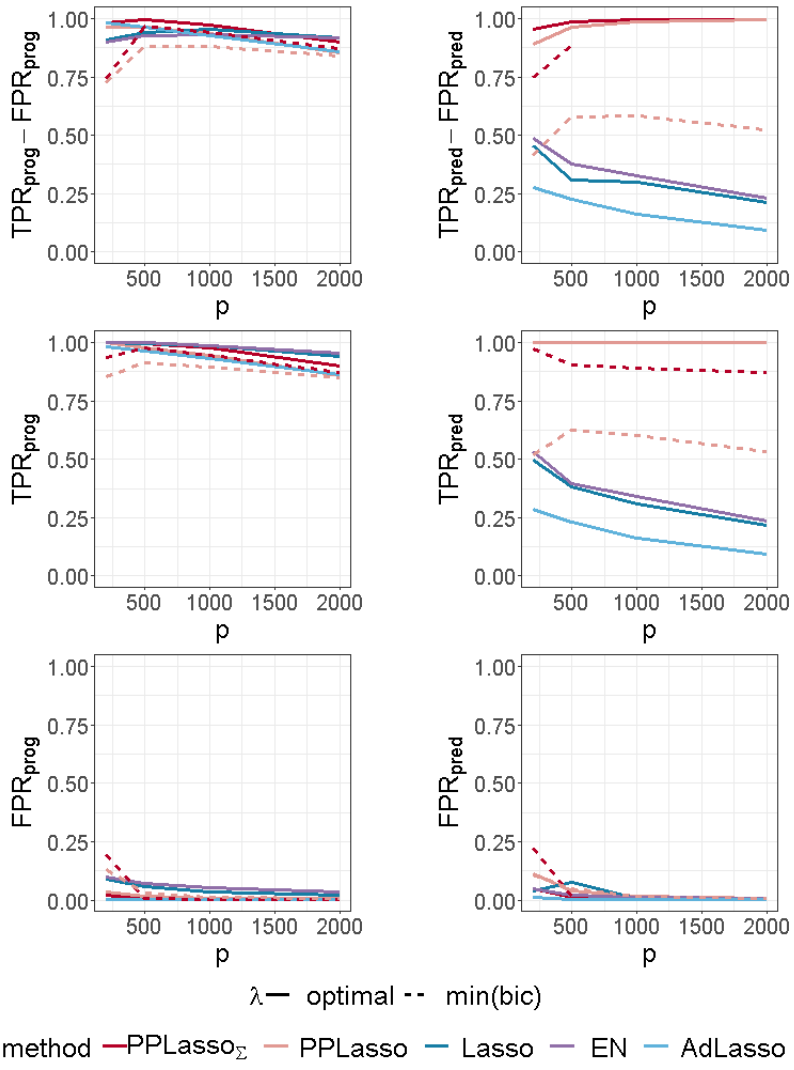

Figure 4.9: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers ($b_2 = 1.5$).

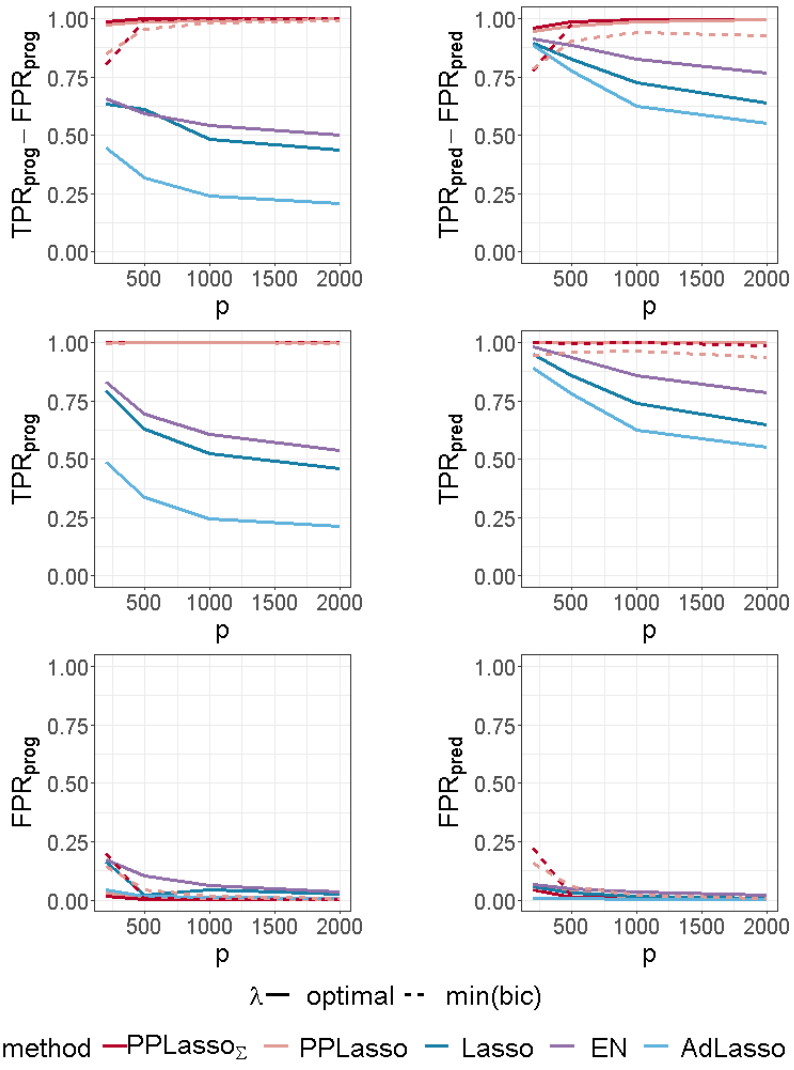

Figure 4.10: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers ($b_2 = 2.5$).

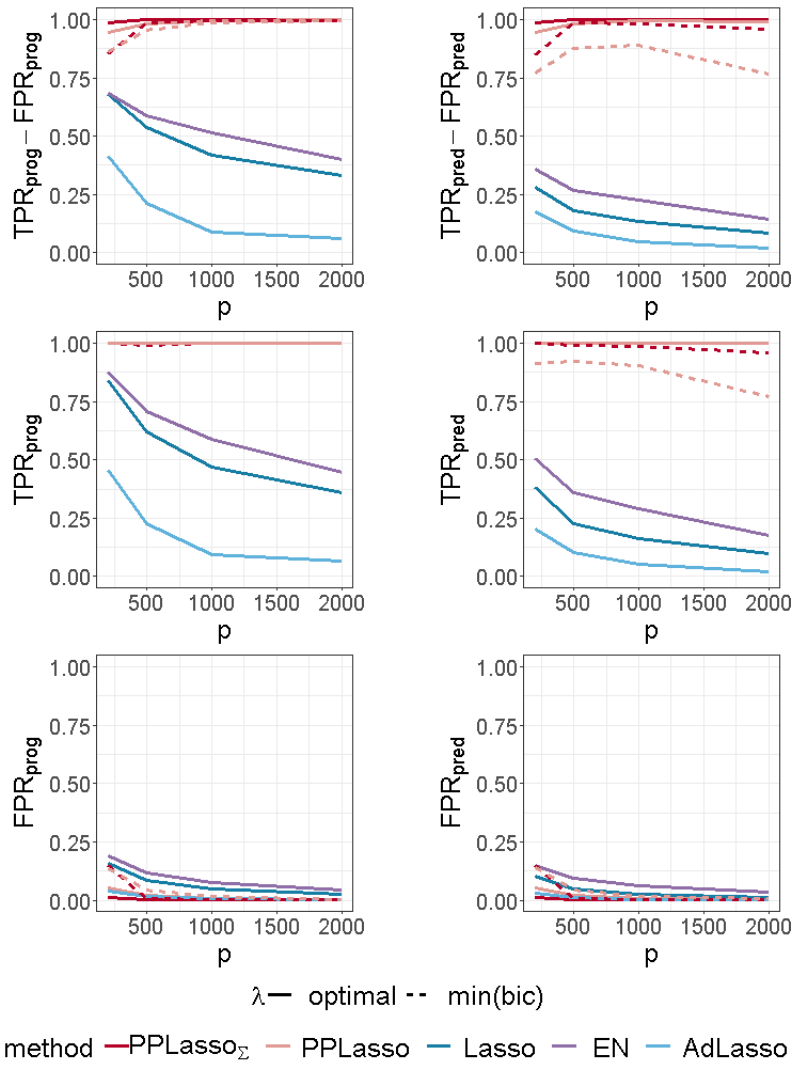

Figure 4.11: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers (10 predictive biomarkers).

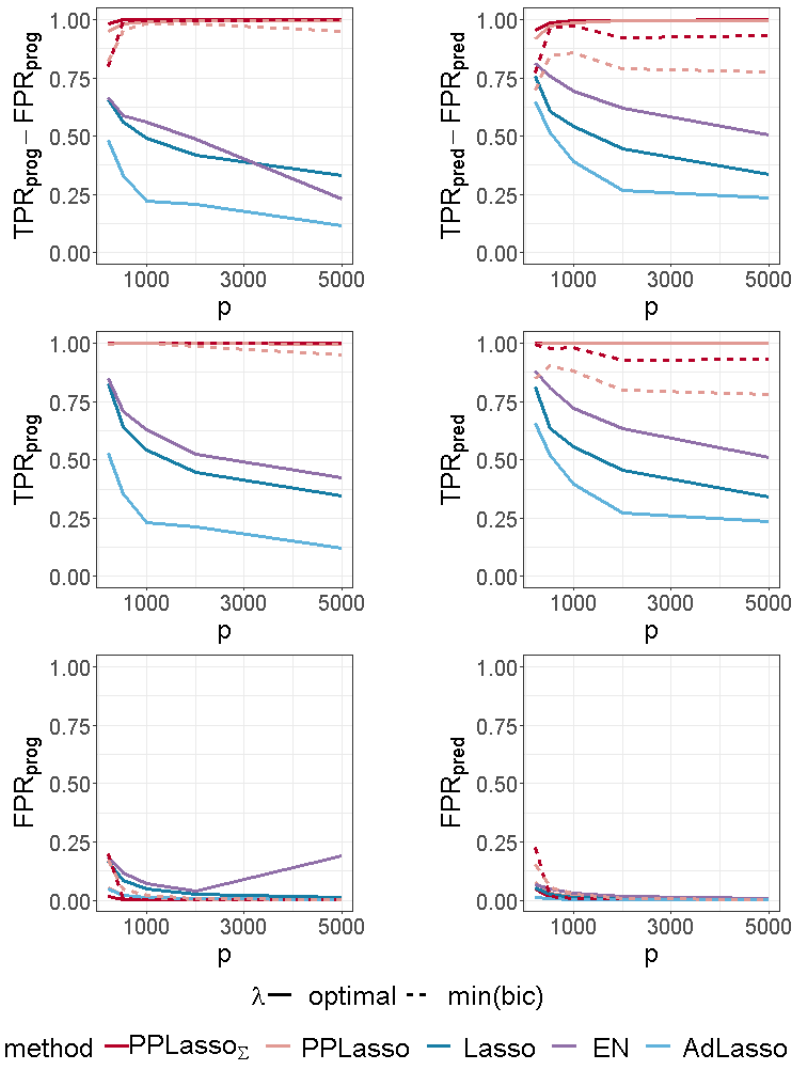

Figure 4.12: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers (with $p = 5000$).

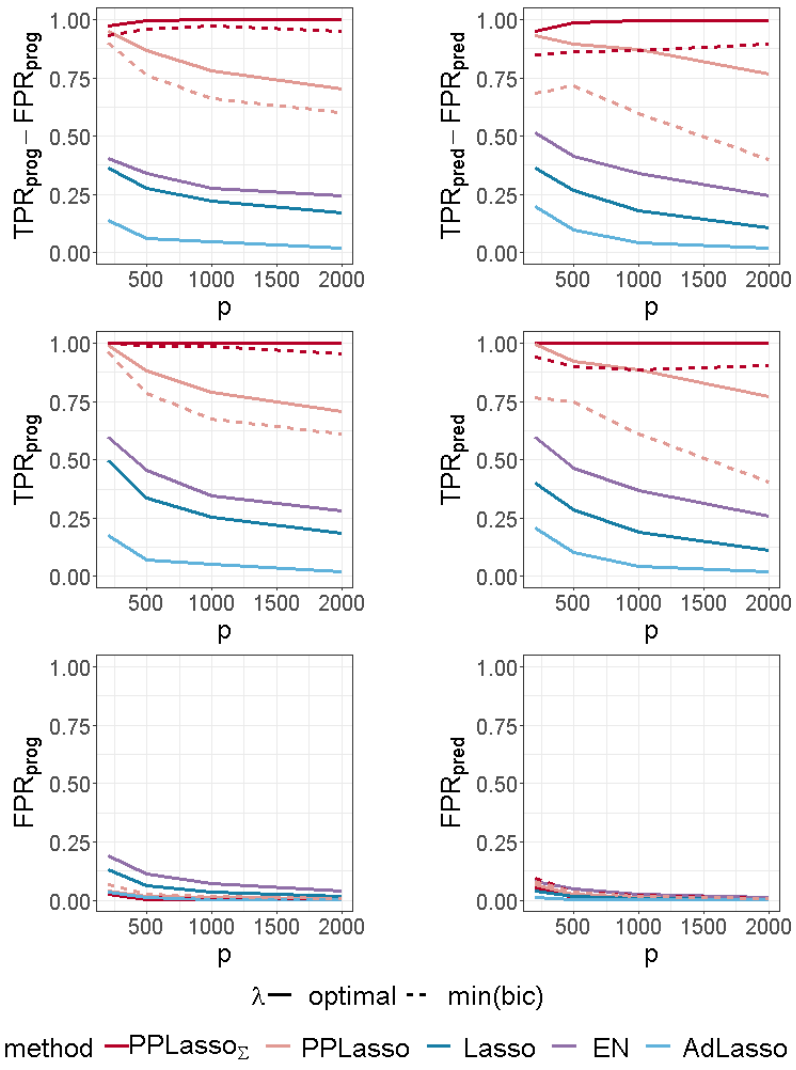

Figure 4.13: (TPR-FPR) and the corresponding True Positive Rate (TPR) and False Positive Rate (FPR) for prognostic (left) and predictive (right) biomarkers ($n_1 = n_2 = 25$).

Estimator	**Hyperparameters**	**Empirical risk**
denseLinearShrinkEst	-	*102546*
sampleCovEst	-	102547
linearShrinkLWEst	-	103496
poetEst	lambda=0.1, k=2	104522
poetEst	lambda=0.2, k=2	105358
poetEst	lambda=0.1, k=1	105972
poetEst	lambda=0.2, k=1	108222
thresholdingEst	gamma=0.2	137798
thresholdingEst	gamma=0.4	186844

Table 4.3: Empirical risk of tested methods with different hyperparameters.

Chapter 5 - Variable selection in high-dimensional logistic regression models using a whitening approach

Publication

The content of this chapter is in the article:
Zhu, W., Lévy-Leduc, C., and Ternès, N. (2022). Variable selection in high-dimensional logistic regression models using a whitening approach. Submitted and also available on *arXiv preprint* (arXiv:2206.14850).

The proposed method is implemented in the `WLogit` R package which will soon be available from the CRAN.

Abstract

In bioinformatics, the rapid development of sequencing technology has enabled us to collect an increasing amount of omics data. Classification based on omics data is one of the central problems in biomedical research. However, omics data usually has a limited sample size but high feature dimensions, and it is assumed that only a few features (biomarkers) are active, i.e. informative to discriminate between two categories (cancer subtypes, responder/non-responder to a treatment for example). Identifying active biomarkers for classification has therefore become fundamental for omics data analysis. Focusing on binary classification, we propose an innovative feature selection method aiming at dealing with the high correlations between the biomarkers. Various research has shown the notorious influence of correlated biomarkers and the difficulty of accurately identifying active ones. Our method, WLogit, consists in whitening the design matrix to remove the correlations between biomarkers, then using a penalized criterion adapted to the logistic regression model to select features. The performance of WLogit is assessed using synthetic data in several scenarios and compared with other approaches. The results suggest that WLogit can identify almost all active biomarkers even in the cases where the biomarkers are highly correlated, while the other methods fail, which consequently leads to higher classification accuracy. The performance is also evaluated on the classification of two Lymphoma subtypes, and the obtained classifier also outperformed other methods. Our method is implemented in the `WLogit` R package available from the Comprehensive R Archive Network (CRAN).

Contents

5.1. Introduction

With the advances in high-throughput molecular techniques, omics technologies can generate large-scale molecular data, such as genomic, transcriptomic, proteomic, and metabolomic data. Classification based on the molecular levels is one of the essential issues in genome research. Examples include tumor classification (Quackenbush, 2006), disease classification (Loscalzo et al., 2007) and distinguishing between responder v.s. non-responder to a treatment (Gustafsson et al., 2014). Different machine learning techniques have been applied to solve this classification problem. Compared to classifiers such as decision tree (Utgoff, 1989) and SVM (Cortes and Vapnik, 1995), logistic regression (Walker and Duncan, 1967) is a popular classification method with an explicit statistical interpretation and can provide classification probabilities for a binary response (Menard, 2002).

However, classification based on omics data is a challenging task. In most omics datasets, the number of biomarkers is much larger than the sample size. Under such a situation, it is generally believed that only a few biomarkers are relevant to disease outcomes, they are called active biomarkers. The presence of irrelevant biomarkers can lead to overparameterized models that increase the risk of overfitting (Sung et al., 2012). Therefore, selecting the active biomarkers can simplify the classifier without the loss of classification accuracy and ease the computational burden. Various methods for feature selection in bioinformatics were developed and reviews can be found in Ang et al. (2015) and Jardillier et al. (2018). To address this issue, regularization via the Lasso (Tibshirani, 1996a) is often implemented to reduce the subset of biomarkers. It adds a penalty equal to the sum of the absolute value of the coefficients that can result in sparse models with few non-zero coefficients and eliminate biomarkers with zero coefficients.

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

$$\pi_{\boldsymbol{\beta}}(X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}. \tag{5.1}$$

The logistic regression with ℓ_1 regularization solves the feature selection problem by adding a penalty function to the log-likelihood of the logistic regression model:

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \tag{5.2}$$

where $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$, and the log-likelihood $l(\boldsymbol{\beta})$ is defined by:

$$l(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \cdot X^{(i)} \boldsymbol{\beta} - \log(1 + e^{X^{(i)} \boldsymbol{\beta}}) \right], \tag{5.3}$$

with $X^{(i)}$ the ith row of $\boldsymbol{X}$. With the penalty function and properly chosen parameter λ, some components of $\widehat{\boldsymbol{\beta}}$ are set to zero. Recently, penalization approaches have been widely applied to biomarker discovery and disease classification (Zhu and Hastie, 2004; Wu, 2006; Ma and Huang, 2008; Liu et al., 2020). A more comprehensive review of different regularizations for analyzing high-dimensional omics data can be found in Vinga (2021).

Despite various advantages, the Lasso criterion can fail to select the true subset of active biomarkers when all biomarkers are highly correlated, especially when the correlation between active and non-active biomarkers is large. This phenomenon was explicitly explained by Zhao and Yu (2006), where a condition is established for Lasso to consistently select the true model in the classical Gaussian regression model. The condition is called the Irrepresentable Condition (IC) (or incoherent condition by Meinshausen and Yu (2009)), and related properties in a Gaussian linear model were reached independently by Zhao and Yu (2006) and Meinshausen and Yu (2009). A similar condition was obtained by Ravikumar et al. (2010) and Bunea (2008) in the logistic regression case. Let Q be defined by:

$$Q = \boldsymbol{X}^T \boldsymbol{H} \boldsymbol{X}, \tag{5.4}$$

where $\boldsymbol{H}$ is a diagonal matrix with

$$H_{ii} = \pi_{\boldsymbol{\beta}}(X^{(i)})/(1 - \pi_{\boldsymbol{\beta}}(X^{(i)})), 1 \leq i \leq n. \tag{5.5}$$

Let $S = \{j,\ \beta_j \neq 0\}$ be the set of active variables with size d, S^c the set of non-active variables. Q_{SS} denotes the $d \times d$ sub-matrix of Q indexed by S. With this notation, the condition states:

There exists $\alpha \in (0, 1]$ such that:

$$\left| Q_{S^c S} (Q_{SS})^{-1} \right|_\infty \leq 1 - \alpha, \tag{5.6}$$

where $|A|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^{p} |A_{jk}|$ for any real symmetric matrix having p rows and p columns.

To deal with the correlations between variables, several methods have been proposed. The most well-known ones include Elastic Net (Zou and Hastie, 2005) and Adaptive Lasso (Zou, 2006). The former combines the ℓ_1 and ℓ_2 penalties,

and the latter assigns weights to each of the parameters in forming the ℓ_1 penalty of Lasso. Several filter approaches were also proposed to take into consideration the correlations in the classification framework. Relief (Kira and Rendell, 1992) is sensitive to feature interactions and has inspired a family of Relief-based feature selection algorithms, notably the ReliefF (Kononenko et al., 1997). It was widely used in biomedical research (Urbanowicz et al., 2018). Fast Correlation Based Filter (FCBF) (Yu and Liu, 2003) is another approach in high-dimensional feature selection that evaluates feature relevance and redundancy based on correlation measures.

In this article, we propose a novel feature selection method to take this issue into account by removing the correlations between biomarkers in the high dimensional logistic regression model. Inspired by the idea of WLasso (Whitening Lasso) proposed by Zhu et al. (2021), we first 'whiten' the columns of $\boldsymbol{X}$. Then, the biomarker selection is performed thanks to a regularized quadratic approximation of the log-likelihood. More details on this method are presented in Section 5.2. In Section 5.3, the performance of the proposed method is assessed via numerical experiments and compared with several methods focusing on the same problem. In Section 5.4, we apply the proposed procedure to a publicly available omic dataset aiming at identifying active biomarkers to classify on two Lymphoma subtypes. Finally, we discuss our findings and give concluding remarks in Section 5.5.

5.2. Method

To solve the optimization problem (5.2), one may directly minimize the penalized log-likelihood (Park and Hastie, 2007; Wang et al., 2019), or use least square approximation as proposed by Friedman et al. (2010), which proposes to form a quadratic approximation of the log-likelihood (5.3) by using a Taylor expansion at the current estimates:

$$l_Q(\boldsymbol{\beta}) = -\frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - X^{(i)}\boldsymbol{\beta})^2 + C(\boldsymbol{\beta}^o)^2 \tag{5.7}$$

$$= -\frac{1}{2n}\sum_{i=1}^{n}(\sqrt{w_i}z_i - \sqrt{w_i}X^{(i)}\boldsymbol{\beta})^2 + C(\boldsymbol{\beta}^o)^2 \tag{5.8}$$

with

$$z_i = X^{(i)}\boldsymbol{\beta} + \frac{y_i - \pi_{\boldsymbol{\beta}^o}(X^{(i)})}{\pi_{\boldsymbol{\beta}^o}(X^{(i)})(1-\pi_{\boldsymbol{\beta}^o}(X^{(i)}))}, \text{(working response)}$$

$$w_i = \pi_{\boldsymbol{\beta}^o}(X^{(i)})(1-\pi_{\boldsymbol{\beta}^o}(X^{(i)})), \text{(weights)} \tag{5.9}$$

where $\pi_{\boldsymbol{\beta}^o}(X^{(i)})$ is the evaluation of $\pi_{\boldsymbol{\beta}}$ (defined in Model (5.1)) at the current parameters $\boldsymbol{\beta}^o$. The final estimator can be derived by the IRLS (Iterative Re-weighted

Least Square) algorithm (Daubechies et al., 2010).

Interestingly, the logistic irrepresentable condition (5.6) coincides with the Irrepresentable condition in linear regression (Zhao and Yu, 2006), when replacing the matrix $\boldsymbol{X}$ by $\sqrt{\boldsymbol{w}}\boldsymbol{X}$, where $\sqrt{\boldsymbol{w}}$ is a diagonal matrix with diagonal entries equal to $(\sqrt{w_1}, \ldots, \sqrt{w_n})$ as defined in (5.9).

5.2.1. Transformation

Since the inconsistency of the Lasso estimator comes from the correlations between the biomarkers, we propose to remove the correlation by "whitening" the matrix $\boldsymbol{X}$. More precisely, we consider $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\check{\Sigma}^{-1/2}$, where $\check{\Sigma}$ is a covariance estimator obtained from $\boldsymbol{H}^{1/2}\boldsymbol{X}$ where $\boldsymbol{H}$ is defined in Equation (5.5). With this transformation, $\widetilde{\boldsymbol{X}}^T\boldsymbol{H}\widetilde{\boldsymbol{X}}$ should be close to the identity matrix I_p, thus the irrepresentable condition should be satisfied. Figure 5.1 shows the percentage of elements on the left-hand side of Equation (5.6) that violated the condition. Data for illustration was generated on one scenario in numerical experiments: the balanced case with blockwise correlation structure when $p = 500$. This dataset will be used in the rest of the section to illustrate different steps in our method. Since in practice we do not know $\pi_\beta(X^{(i)})$, the oracle $\boldsymbol{H}$ with true coefficients and estimated $\boldsymbol{H}$ (see Section 5.2.5 for details) were both presented. We verified through this figure that the violation of the irrepresentable condition had been reduced after the transformation.

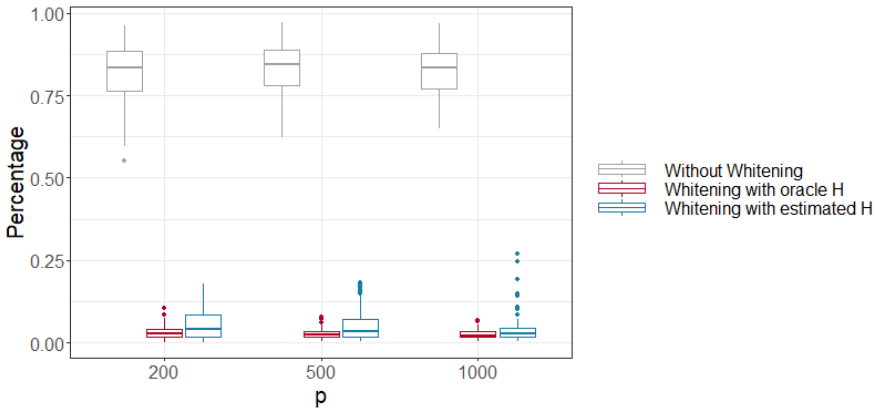

Figure 5.1: Percentage of elements on the left hand-side of Equation (5.6) that violated the IC, before and after transformation with oracle $\boldsymbol{H}$ and estimated $\boldsymbol{H}$.

After the whitening step, Model (5.1) can be rewritten as:

$$\pi_{\widetilde{\boldsymbol{\beta}}}(\widetilde{X}^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)}\right)}, \tag{5.10}$$

where $\widetilde{X}^{(i)}$ denotes the ith row of $\widetilde{X}$, and $\widetilde{\boldsymbol{\beta}} = \check{\Sigma}^{1/2}\boldsymbol{\beta}$. The log-likelihood after the transformation can be written as:

$$l^{wt}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{n}\sum_{i=1}^{n}\left\{y_i \cdot \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}} - \log\left(1 + e^{\widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}}}\right)\right\}. \tag{5.11}$$

Following the same technique of approximation as in (5.7), we can form a quadratic approximation to the transformed (whitened) log-likelihood (5.11), then an estimator of $\widetilde{\boldsymbol{\beta}}$ is obtained by solving the following problem:

$$\underset{\widetilde{\boldsymbol{\beta}}\in\mathbb{R}^p}{\arg\min}\left\{l_Q^{wt}(\widetilde{\boldsymbol{\beta}}) + \lambda\left\|\check{\Sigma}^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1\right\}. \tag{5.12}$$

5.2.2. Estimation of $\widetilde{\beta}$

The estimation is obtained by using an iterative procedure. Let $maxit$ and tol denote the maximum number of iterations and the tolerance. For a fixed λ, the following loops are performed:

- Initialize parameters $\widetilde{\boldsymbol{\beta}}^{(0)}$ by $\widetilde{\boldsymbol{\beta}}^{(0)} = \check{\Sigma}^{1/2}\boldsymbol{\beta}^{(0)}$, where $\boldsymbol{\beta}^{(0)}$ is obtained by ridge regression in the logistic regression model.
- For iteration $j = 1, \ldots, maxit$:
  1. Update working response, weights, weighted response, weighted design matrix in the re-weighted least square regression.
  2. Update coefficients $\widehat{\widetilde{\boldsymbol{\beta}}}^{(j)}$ by solving Equation (5.12).
  3. Calculate $max(|\widehat{\widetilde{\boldsymbol{\beta}}}^{(j)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(j-1)}|)$
  4. For $j > 1$, if $max(|\widehat{\widetilde{\boldsymbol{\beta}}}^{(j)} - \widehat{\widetilde{\boldsymbol{\beta}}}^{(j-1)}|) < tol$, stop and return $\widehat{\widetilde{\boldsymbol{\beta}}}^{(j-1)}$. If $j = maxit$, stop the algorithm and return $\widehat{\widetilde{\boldsymbol{\beta}}}^{(j)}$. If none of these conditions is satisfied, go back to Step 1 until one of the stopping criteria is satisfied.
- Denote the final coefficients by $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda)$.

To estimate $\widetilde{\boldsymbol{\beta}}$, we will not directly use $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda)$ but the following modified estimator which can be seen as a correction of the components of $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda)$. For K in $\{1, \ldots, p\}$, let Top_K be the set of indices corresponding to the K largest values of

the components of $|\widehat{\widetilde{\boldsymbol{\beta}}}_0|$, then the estimator of $\widetilde{\boldsymbol{\beta}}$ is $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(\widehat{K})})_{1\leq j\leq p}$, where $\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}$ is defined by:

$$\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}(\lambda) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{0j}(\lambda), & j \in \mathrm{Top}_K \\ K\text{th largest value of } |\widehat{\widetilde{\boldsymbol{\beta}}}_{0j}|, & j \notin \mathrm{Top}_K. \end{cases} \tag{5.13}$$

To choose the parameter K, we use a strategy based on the log-likelihood of the model. By replacing $\widetilde{\boldsymbol{\beta}}$ in (5.11) by $\widehat{\widetilde{\boldsymbol{\beta}}}^{(K)}(\lambda)$, which is the vector having the $\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}$ for components, we get $l_K^{wt}(\widehat{\widetilde{\boldsymbol{\beta}}}(\lambda))$, and $\widehat{K}$ is chosen as follows

$$\widehat{K}(\lambda) = \arg\min \left\{ K \geq 1 \text{ s.t. } \frac{l_K^{wt}(\widehat{\widetilde{\boldsymbol{\beta}}}(\lambda))}{l_{K+1}^{wt}(\widehat{\widetilde{\boldsymbol{\beta}}}(\lambda))} \geq \gamma \right\}, \text{ where } \gamma \in (0,1).$$

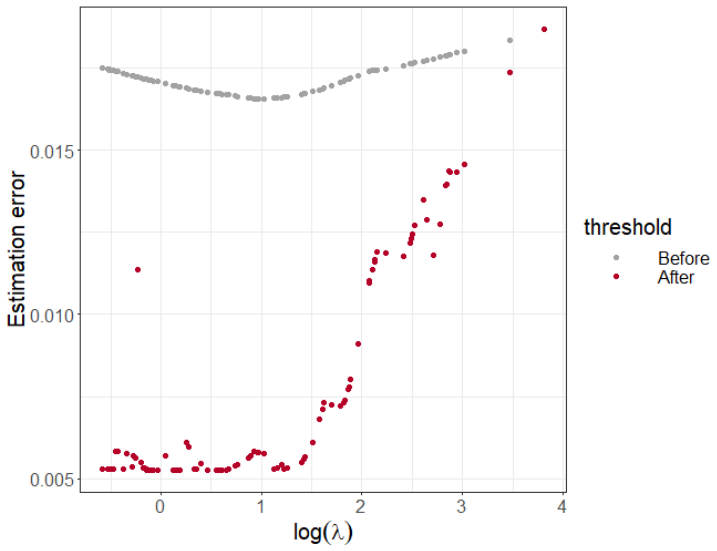

Figure 5.2: Average estimation error for all coefficients before and after the thresholding.

The purpose of this step is to correct the intermediate estimation $\widehat{\widetilde{\boldsymbol{\beta}}}$. Figure 5.2 displays coefficient estimation error of $\widetilde{\boldsymbol{\beta}}$ before and after the thresholding correction. We can see that the correction helps to decrease the coefficient estimation error.

5.2.3. Estimation of β

Resulting from the transformation, a first estimation of $\boldsymbol{\beta}$ is obtained by $\widehat{\boldsymbol{\beta}}_0 = \check{\Sigma}^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}$, and we apply a threshold to get the final estimation $(\widehat{\boldsymbol{\beta}}_j^{(\widehat{M})})_{1\leq j\leq p}$ where

$$\widehat{\boldsymbol{\beta}}_j^{(M)}(\lambda) = \begin{cases} \widehat{\boldsymbol{\beta}}_{0j}(\lambda), & j \in \mathrm{Top}_M \\ 0, & j \notin \mathrm{Top}_M, \end{cases} \tag{5.14}$$

and Top_M is defined in a similar way as previously. The choice of the parameter M was also based on the log-likelihood. By replacing β in (5.3) by $\widehat{\beta}^{(M)}(\lambda)$, which is the vector having the $\widehat{\beta}_j^{(M)}$ for components, we get $l_M(\widehat{\boldsymbol{\beta}}(\lambda))$. Using the same strategy as in Section 5.2.2, M is chosen as follows:

$$\widehat{M}(\lambda) = \arg\min \left\{ K \geq 1 \text{ s.t. } \frac{l_M(\widehat{\boldsymbol{\beta}}(\lambda))}{l_{M+1}(\widehat{\boldsymbol{\beta}}(\lambda))} \geq \gamma \right\}, \text{ where } \gamma \in (0,1).$$

As we can see from Figure 5.10 in Supplementary, the thresholding step successfully removed non active variables while keeping most of the true active ones in the model.

5.2.4. Choice of the parameter λ

Suppose the estimation of $\boldsymbol{\beta}$ was obtained following Section 5.2.2 and Section 5.2.3. For simplicity, we note it as $\widehat{\boldsymbol{\beta}}(\lambda)$ over the sequence of λ, and the corresponding log-likelihood is $l(\widehat{\boldsymbol{\beta}}(\lambda))$. We chose λ by:

$$\hat{\lambda} = \arg\max_{\lambda} l(\widehat{\boldsymbol{\beta}}(\lambda)). \tag{5.15}$$

Notice that if multiple λs maximize the log-likelihood, we chose the one leading to the most parsimonious model.

5.2.5. Estimation of $\check{\Sigma}$

In practice, $\check{\Sigma}$ is calculated by estimating the variance-covariance matrix from $\boldsymbol{H}^{1/2}\boldsymbol{X}$. As the diagonal of $\boldsymbol{H}$ defined in Equation (5.5) is unknown because no information on $\boldsymbol{\beta}$ is available, the latter can be roughly estimated by ridge regression in the logistic regression model when $p > n$. We denote this estimator by $\widehat{\boldsymbol{\beta}}_{ridge}$ and obtain $\widehat{\boldsymbol{H}}$ with $\widehat{H_{ii}} = \pi_{\widehat{\boldsymbol{\beta}}_{ridge}}(X^{(i)})/(1-\pi_{\widehat{\boldsymbol{\beta}}_{ridge}}(X^{(i)}))$ for $i = 1, \ldots, n$. Finally, $\check{\Sigma}$ is calculated by estimating the variance-covariance matrix of $\widehat{\boldsymbol{H}}^{1/2}\boldsymbol{X}$, by using the method implemented in the package cvCovEst of Boileau et al. (2022).

5.2.6. Summary of WLogit algorithm

1. Calculate $\check{\Sigma}$, the empirical variance-covariance matrix of $\boldsymbol{H}^{1/2}\boldsymbol{X}$, as described in Section 5.2.5
2. Compute $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\check{\Sigma}^{-1/2}$
3. For each λ:
   (a) Estimate $\widetilde{\boldsymbol{\beta}}$ as described in Section 5.2.2.
   (b) Estimate $\boldsymbol{\beta}$ as described in Section 5.2.3.
4. Choose λ as described in Section 5.2.4, then perform variable selection and/or prediction of $\boldsymbol{y}$ based on $\widehat{\boldsymbol{\beta}}(\hat{\lambda})$.

5.3. Numerical experiments

This section aims at evaluating WLogit and comparing it with existing ones. We simulated data from Model (5.1), where the rows of $\boldsymbol{X}$ are assumed to be independent Gaussian random vectors with covariance matrix equal to Σ. The response $\boldsymbol{y}$ was generated following Model (5.1), and the vector $\boldsymbol{\beta}$ has 10 non-zero elements with an effect size equal to 1. The sample size is equal to $n = 100$, and we considered the balanced case where there are 50 responses y_i equal to 1 and 50 equal to 0, and an imbalanced case where there are 20 responses y_i equal to 1 and 80 equal to 0. The number of predictors (biomarkers) took its values from 200 to 2000. 100 replications were generated for each scenario.

In our simulations, we mainly considered correlation structures in which the irrepresentable condition was violated. We defined Σ with a blockwise structure:

$$\Sigma = \begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}, \tag{5.16}$$

where Σ_{11} (resp. Σ_{22}) are the correlation matrix of active (resp. non-active) biomarkers with off-diagonal entries equal to α_1 (resp. α_3), Σ_{12} is the correlation matrix between active and non-active variables with entries equal to α_2. In our simulations, we chose $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$, one of the frameworks proposed by Xue and Qu (2017). Although this structure was proposed in the context of linear regression, we checked that the irrepresentable condition for the logistic model was also violated (as displayed in Figure 5.1). Additionally to this special case, we also investigated the case where no correlation exists between predictors, i.e., Σ is the identity matrix, and in this case, the irrepresentable condition is satisfied.

5.3.1. Compared methods

Compared methods include two other penalized approaches: Lasso and Elastic Net adapted to the logistic regression model. Elastic Net is noted as EN in the figures. The parameters in these two algorithms are chosen by 10-fold cross-validation and implemented by the R package `glmnet`. We also compared our method with other approaches not involving the penalized regression family: ReliefF and FCBF. They also take into account the correlations between predictors and are widely used in the identification of biomarkers. ReliefF was implemented by the R package `CORElearn` with parameter `estimator="ReliefFexpRank"`. Since this method only gives the rank of predictors, we selected the same number of predictors as WLogit with the highest rank. FCBF was implemented by the Bioconductor package `FCBF`. We kept the default parameters for these two methods.

5.3.2. Evaluation

The evaluation of the performance of the compared methods was based on two aspects: (1) the accuracy of biomarker selection and (2) the accuracy of sample classification, which can be seen as a prediction task. Figure 5.3 shows different steps in the numerical experiments and the two types of evaluation.

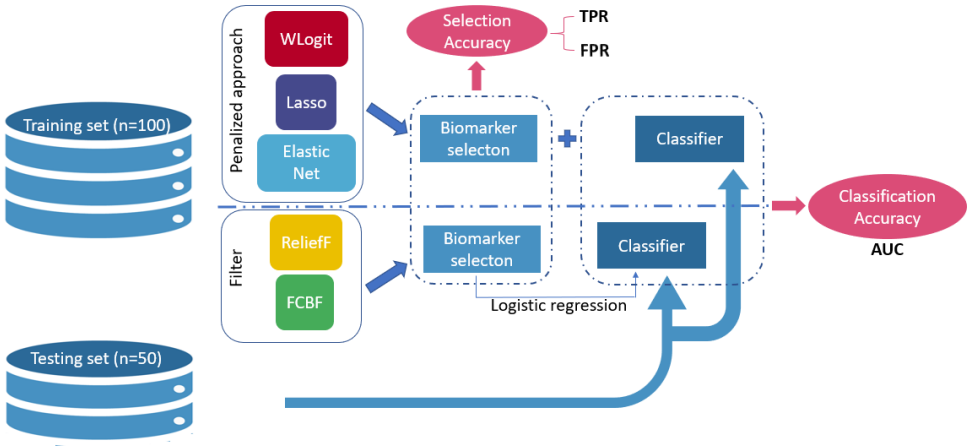

Figure 5.3: Simulation process and evaluation of the compared methods.

Biomarker selection

We generate training sets as described at the beginning of this section. Each method selected a subset of predictors, and we evaluate the selection by True Positive Rate (TPR) and False Positive Rate (FPR). The reported values for TPR and FPR are obtained by averaging these values from 100 replications.

Sample classification

For penalized regression approaches (WLogit, Lasso, and Elastic Net), a classifier was already available with selected predictors since these approaches also give regression coefficients estimation at the same time. For ReliefF and FCBF, when a subset of predictors was chosen, the logistic regression classifier was built with the estimation of coefficients on each chosen predictor. The evaluation was then performed on another simulated testing set with the same settings as the training set, except with only half the sample size (100 (training) v.s. 50 (testing)). The evaluation on the testing test will provide the prediction accuracy of the selected set of predictors, which is presented by the AUC (Area Under the receiver operating characteristic (ROC) curve).

5.3.3. Results

The corresponding results are displayed in Figures 5.4 and 5.5 in the case where Σ has the blockwise correlation structure defined in Model (5.16) with parameters $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$. The corresponding TPR and FPR for each method are displayed. We can see from Figure 5.4 that WLogit largely outperforms the other methods: the TPR is always the largest and close to 1 (0.95 for $p = 200$ and 0.86 for $p = 2000$). Lasso, Elastic Net, and FCBF performed similarly. They can identify a very limited number of active variables (TPR smaller than 0.20). Although the FPR for WLogit was larger when $p = 200$ (FPR= 0.17), it decreased when p increases (FPR= 0.01 for $p = 2000$). When $p = 2000$, the FPR for all the methods is similar. With the same subset size of selected variables as WLogit, ReliefF performed poorly: the TPR is close to 0, and the FPR is the largest when p is not large.

Figure 5.5 presents the average of AUC on the testing set for all methods, based on the classifiers developed on the training set (variable selection evaluated in Figure 5.4). WLogit showed the best classification accuracy stable at a high level (> 0.96) even when the number of predictors increases, which may come from the fact that it has identified more active variables than others. Lasso and Elastic Net performed similarly (AUC= 0.86 and 0.83 for Lasso and Elastic Net, respectively, when $p = 2000$). Although FCBF showed competitive predictor selection accuracy, the classification accuracy (AUC= 0.64 when $p = 2000$) is lower than the one of Lasso and Elastic Net. Moreover, its classification accuracy decreased with the increase of p and was even lower than ReliefF from $p = 1000$ (0.66 for FCBF and 0.68 for ReliefF when $p = 1000$). This may come from the fact that the selected biomarkers from FCBF underwent a re-estimations of coefficients by a logistic regression, while for Lasso and Elastic Net, their coefficients were directly derived from the feature selection step, which provided more accurate prediction.

Figure 5.6 displays the performance of the different approaches in the case where $\Sigma = I_p$, when there is no correlation between the biomarkers. Even if WLogit is designed for handling the correlations when the IC is violated, it still outperformed other methods in terms of biomarker selection. The TPR is the largest among all methods, while the FPR is the smallest (FPR< 0.05). For example when $p = 2000$, the TPRs were 0.43 (WLogit), 0.25 (Lasso), 0.16 (Elastic Net), 0.03 (ReliefF) and 0.08 (FCBF). The FPRs for all the methods were limited. The most performant methods were then: WLogit, Lasso, Elastic Net, FCBF, and ReliefF, in this order. The same conclusion can be reached in sample classification accuracy from Figure 5.7: WLogit always had the highest AUC (0.86 when $p = 200$ and 0.66 when $p = 2000$) compared to other methods. We found that a high accuracy on sample classification is usually given by a high accuracy on predictor selection.

Similar results for the imbalanced case were observed and can be found in Supplementary materials. We noticed that the classification accuracy is slightly lower for all methods compared with balanced cases. However, WLogit always gives the best accuracy on both biomarker selection accuracy and sample classification.

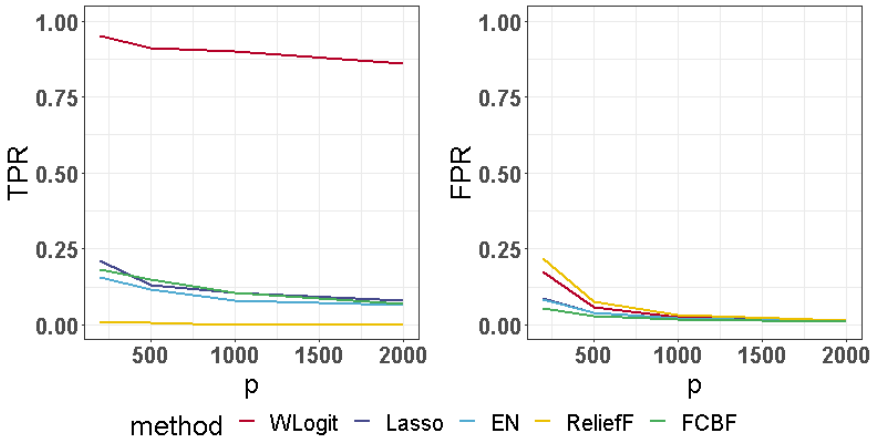

Figure 5.4: True Positive Rate (left) and False Positive Rate (right) for different methods in the balanced case when Σ is defined in (5.16) with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$.

5.4. Application to gene expression data in patients with lymphoma

We applied the previously described approaches to gene expression data from 77 patients with lymphoma first published by Shipp et al. (2002). This dataset contains 58 diffuse large B-cell lymphomas (DLBCL) and 19 follicular lymphomas (FL) samples. The original data contains 7,129 gene expression data. We followed the preprocessing procedures implemented in Glaab et al. (2012) which kept a total of 2648 predictors. The heatmap of the correlations between the expression of the selected genes is displayed in Figure 5.8, where we can observe strong correlations.

We applied different methods to select the genes that distinguish the two lymphoma subtypes (DLBCL v.s. FL). To evaluate the prediction performance of each method, we applied the commonly used 10-fold cross-validation. The dataset was separated into ten folds, and for each fit, the variable selection was conducted on the training set consisting of 90% of the whole set. Then, the classifier was built with the subset of selected variables and used for predicting the lymphoma subtype for the remaining 10% samples in the testing set. Finally, we report the ROC curve on the validation set and the corresponding AUC. Figure 5.15 in the Supplementary material presents the classification accuracy for the different methods. Our method, WLogit, achieved the highest AUC (0.95), followed by FCBF (0.85) and Relief (0.84). Lasso (0.80) and Elastic Net (0.80) both have a lower AUC; this

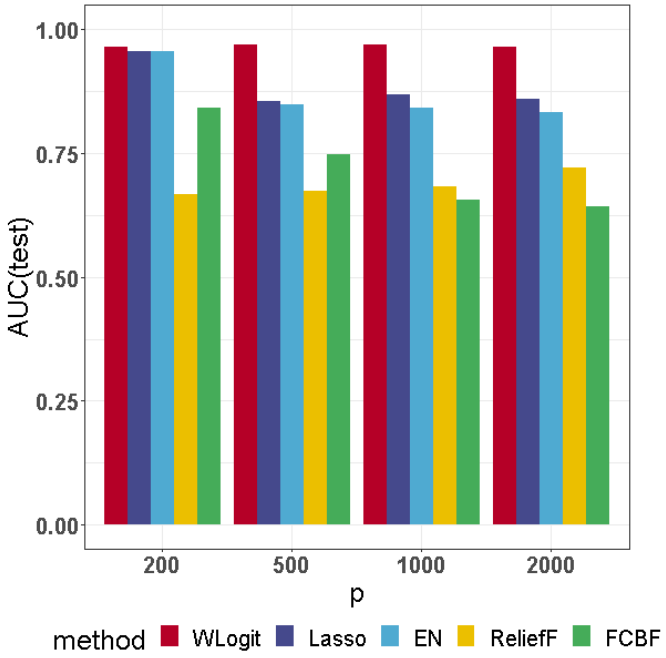

Figure 5.5: AUC on the testing set for different methods in the balanced case when Σ is defined in (5.16) with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$.

result can come from selection failure (no predictor selected) in some folds, which degraded the overall prediction accuracy.

Finally, we used the complete dataset to perform gene selection. Figure 5.9 presents the number of genes selected by each method and the overlap between them. WLogit selected a subset of 18 genes with four genes in common with Elastic Net and one in common with Relief. Lasso selected only one gene that was included in the set of 11 genes selected by Elastic Net. FCBF selected four genes that have no intersection with others. The list of genes selected by each method is given in Supplementary materials, with annotations provided by DAVID database (Sherman et al., 2022).

5.5. Conclusion

This paper proposes a novel biomarker selection method in the high dimensional logistic regression model when the biomarkers are highly correlated. Our approach, called WLogit, consists in using a penalized criterion dedicated to the logistic regression model after having removed the correlations existing between the biomarkers. The numerical experiments showed the strength of our method not only on biomarker selection but also on sample classification.

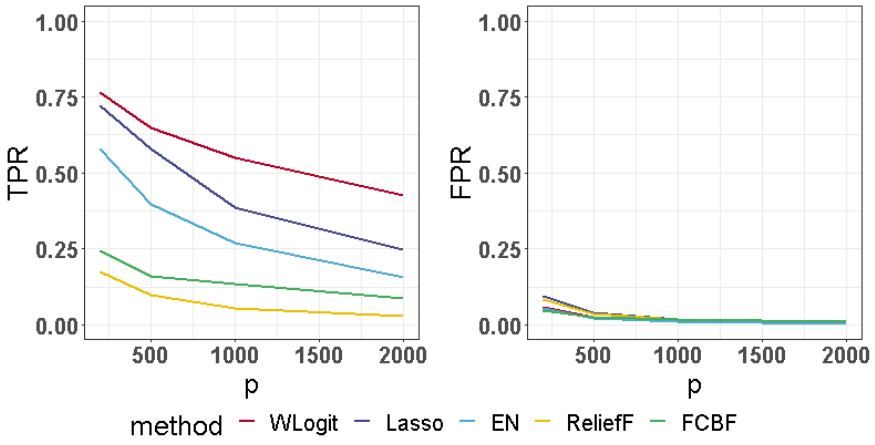

Figure 5.6: True Positive Rate (left) and False Positive Rate (right) for different methods in the balanced case when Σ is the identity matrix.

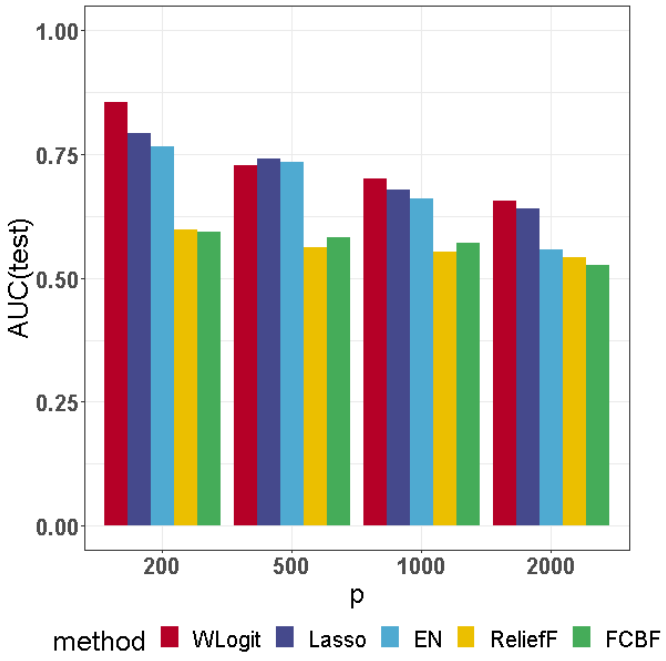

Figure 5.7: AUC on the testing set for different methods in the balanced case when Σ is the identity matrix.

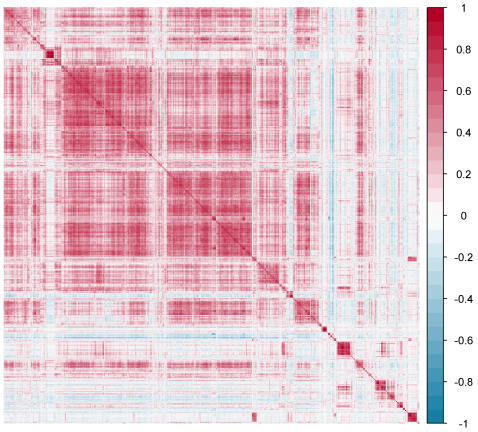

Figure 5.8: Heatmap of correlation of the expression of the genes in the DLBCL dataset.

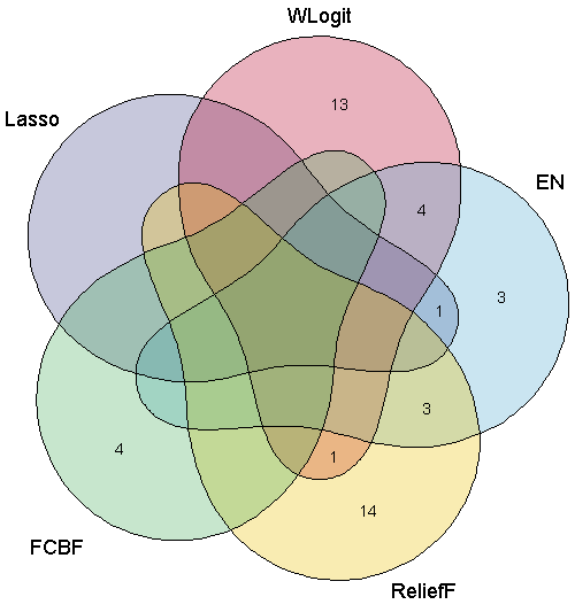

Figure 5.9: Venn plot of selected genes by the different compared methods.

Supplementary material

This supplementary material provides additional numerical experiments, figures and tables for the paper: "Variable selection in high-dimensional logistic regression models using a whitening approach".

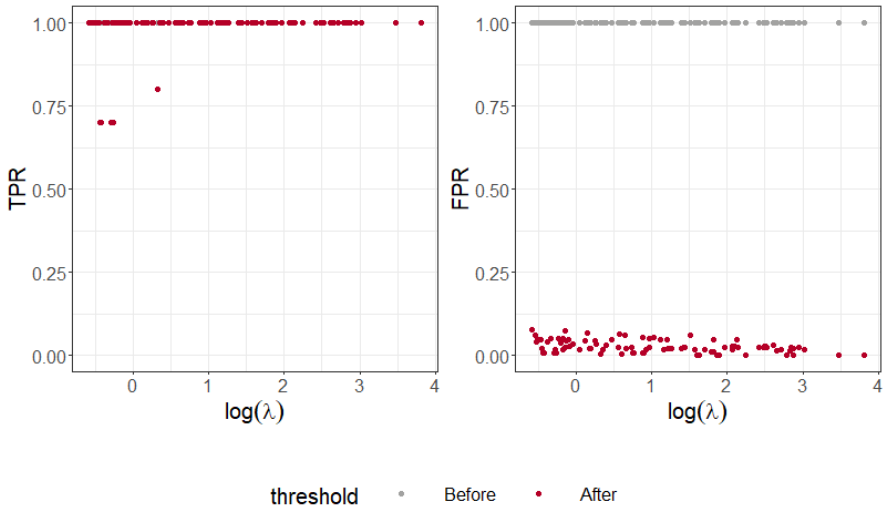

Figure 5.10: True Positive Rate (left) and False Positive Rate (right) before and after the thresholding.

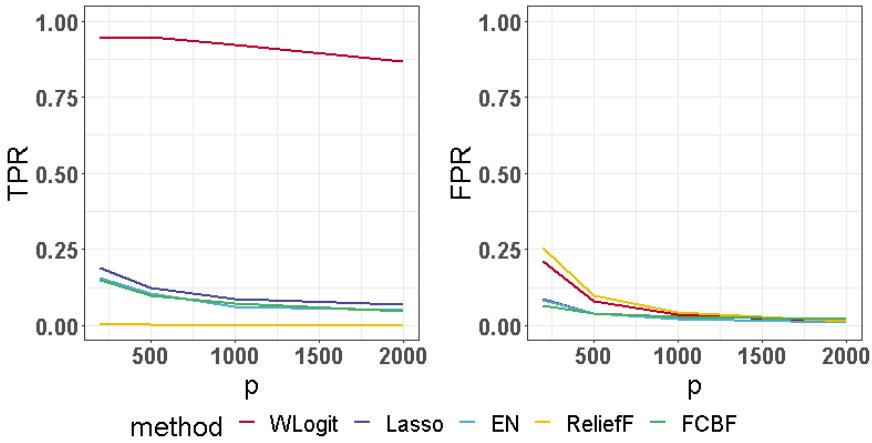

Figure 5.11: True Positive Rate (left) and False Positive Rate (right) for different methods in the imbalanced case when Σ is defined in (5.16) with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$.

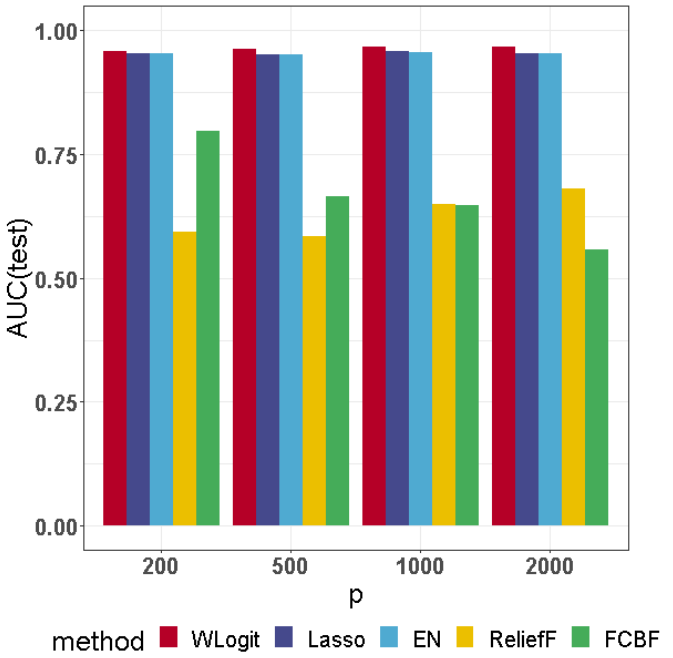

Figure 5.12: AUC on the testing set for different methods in the imbalanced case when Σ is defined in (5.16) with $(\alpha_1, \alpha_2, \alpha_3) = (0.3, 0.5, 0.7)$.

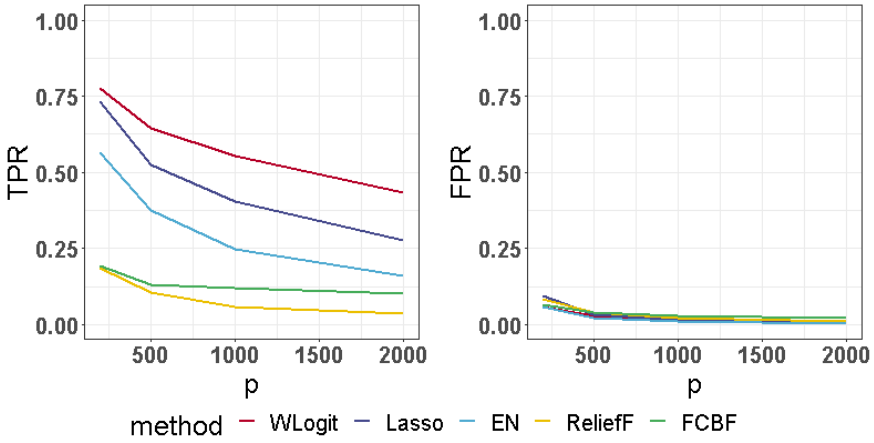

Figure 5.13: True Positive Rate (left) and False Positive Rate (right) for different methods in the imbalanced case when Σ is the identity matrix.

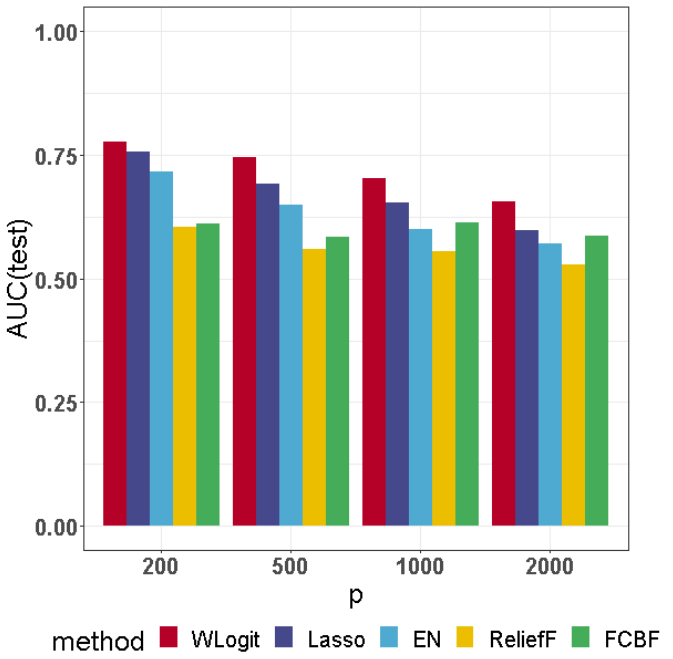

Figure 5.14: AUC on the testing set for different methods in the imbalanced case when Σ is the identity matrix.

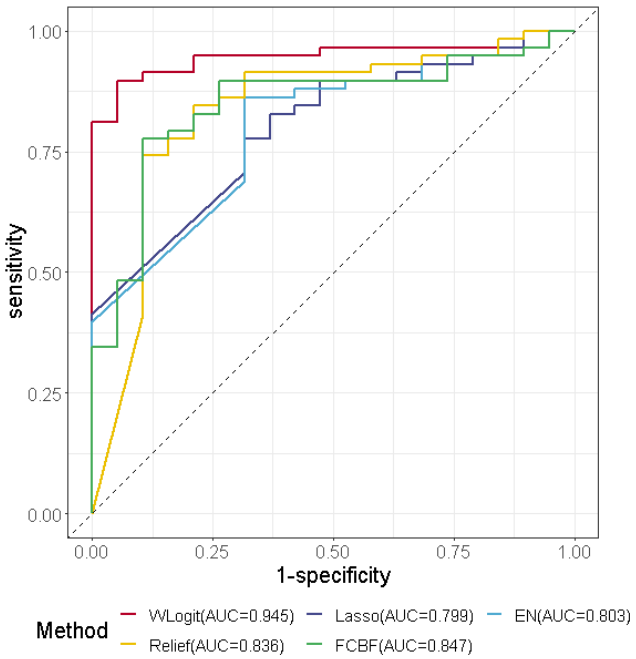

Figure 5.15: ROC curves and AUC for the different compared methods.

Table 5.1: Selected genes and their annotations.

	Selected genes	
	ID	Annotation
WLogit	X52773_AT	retinoid X receptor alpha(RXRA)
	D14662_AT	peroxiredoxin 6(PRDX6)
	V00594_S_AT	metallothionein 2A(MT2A)
	L19686_RNA1_AT	macrophage migration inhibitory factor(MIF)
	AF000562_AT	uroplakin 2(UPK2)
	D87119_AT	tribbles pseudokinase 2(TRIB2)
	S73591_AT	thioredoxin interacting protein(TXNIP)
	X91911_S_AT	GLI pathogenesis related 1(GLIPR1)
	M96684_AT	purine rich element binding protein A(PURA)
	M64925_AT	MAGUK p55 scaffold protein 1(MPP1)
	U49835_S_AT	chitinase 3 like 2(CHI3L2)
	U14187_AT	ephrin A3(EFNA3)
	U63743_at	kinesin family member 2C(KIF2C)
	M63379_AT	clusterin(CLU)
	U36787_AT	holocytochrome c synthase(HCCS)
	M27093_S_AT	dihydrolipoamide branched chain transacylase E2(DBT)
Lasso	U63743_at	kinesin family member 2C(KIF2C)
Elastic Net	AB002409_at	C-C motif chemokine ligand 21(CCL21)
	M23323_s_at	CD3 epsilon subunit of T-cell receptor complex(CD3E)
	U63743_at	kinesin family member 2C(KIF2C)
	V00594_s_at	metallothionein 2A(MT2A)
	X02152_at	lactate dehydrogenase A(LDHA)
	D79987_at	extra spindle pole bodies like 1, separase(ESPL1)
	L19686_rna1_at	macrophage migration inhibitory factor(MIF)
	S73591_at	thioredoxin interacting protein(TXNIP)
	U19495_s_at	C-X-C motif chemokine ligand 12(CXCL12)
Relief	AB002409_at	C-C motif chemokine ligand 21(CCL21)
	D79987_at	extra spindle pole bodies like 1, separase(ESPL1)
	J04031_at	methylenetetrahydrofolate dehydrogenase, cycslohydrolase and formyltetrahydrofolate synthetase 1(MTHFD1)
	L00022_s_at	immunoglobulin heavy constant epsilon(IGHE)
	L42324_at	G protein-coupled receptor 18(GPR18)
	M12963_s_at	alcohol dehydrogenase 1A (class I), alpha polypeptide(ADH1A)
	M15059_at	Fc epsilon receptor II(FCER2)
	M18255_cds2_s_at	protein kinase C beta(PRKCB)
	M64174_at	Janus kinase 1(JAK1)
	M91196_at	interferon regulatory factor 8(IRF8)
	U19495_s_at	C-X-C motif chemokine ligand 12(CXCL12)
	V00594_at	metallothionein 1G(MT1G)
	X01677_f_at	glyceraldehyde-3-phosphate dehydrogenase(GAPDH)
	X52142_at	CTP synthase 1(CTPS1)
	X69433_at	isocitrate dehydrogenase (NADP(+)) 2(IDH2)
	X91911_s_at	GLI pathogenesis related 1(GLIPR1)
	Z11793_at	selenoprotein P(SELENOP)
FCBF	K02777_s_at	T cell receptor delta variable 2(TRDV2)
	M27504_s_at	DNA topoisomerase II beta(TOP2B)
	X52851_rna1_at	peptidylprolyl isomerase A(PPIA)
	X67235_s_at	hematopoietically expressed homeobox(HHEX)

Chapter 6 - Conclusion

6.1. Summary of all developed methods

As described in the previous sections, I developed during my PhD several variable selection methods in different high-dimensional frameworks. We first considered the linear regression model where the response variable is quantitative and only with the presence of prognostic biomarkers. For this purpose, WLasso was designed to perform variable selection in a challenging context where the biomarkers are highly correlated. Then, we proposed another approach based on the Generalized Elastic Net, which combines the ℓ_1 and ℓ_2 penalties and obtains a more generalized estimator. We established under mild conditions the sign consistency of the corresponding estimator. These two methods were mainly developed for prognostic biomarkers identification. For a comparative study (usually a randomized clinical trial comparing two treatments), we proposed the PPLasso approach, which adds biomarker-to-treatment interactions thanks to an ANCOVA-type model. The above three methods were developed for linear models. We finally extended our approach to the classification problem when the response is a binary variable, which led to the development of WLogit. Following the same procedure as PPLasso, WLogit could be further extended to predictive biomarker identification (named by PPLogit, see Section 6.2.2).

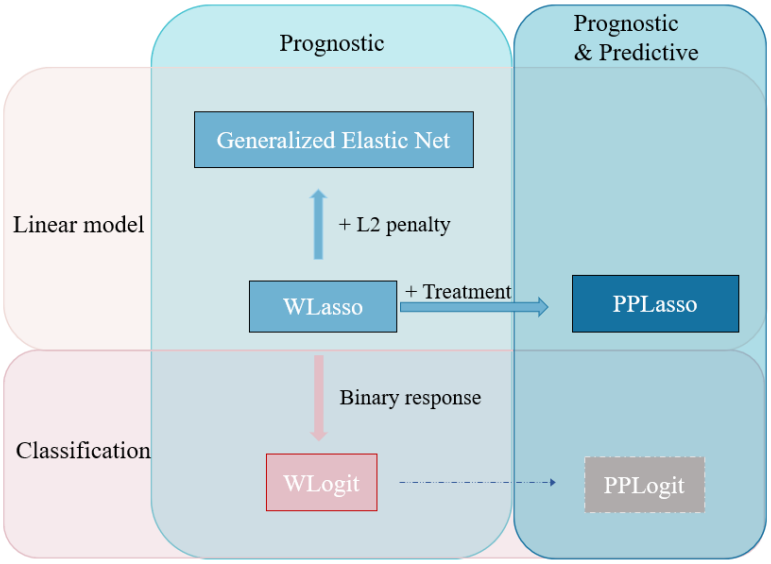

Figure 6.1: Summary of the developed methods

6.2. Future works

6.2.1. Survival data analysis

Our predictor selection methods are only available for continuous and binary responses, but could be extended to survival data. The data available are of the form (time, predictors, event): $(t_1, \boldsymbol{X}^{(1)}, \delta_1), \ldots, (t_N, \boldsymbol{X}^{(N)}, \delta_N)$. The time t_i is either survival or censoring time, and δ_i is the censoring status. If a patient has a survival event of interest (e.g., disease progression, release, death), then t_i will be the time for the survival event and δ_i will be equal to 1. On the other hand, if the patient does not have the survival event, then t_i will be the censoring time, and δ_i will be equal to 0. $\boldsymbol{X}^{(i)}$ denotes the vector of predictors $\boldsymbol{X}_1^{(i)}, \ldots, \boldsymbol{X}_p^{(i)}$ for the ith individual on p covariates. Different models are available to model survival data. The proportional-hazard model, also known as the Cox model, is the most widely used in clinical research. It assumes that:

$$\lambda(t|\boldsymbol{X}^{(i)}) = \lambda_0(t) \exp(\sum_{j=1}^{p} \boldsymbol{X}_j^{(i)} \boldsymbol{\beta}_j) \tag{6.1}$$

where $\lambda(t|\boldsymbol{X}^{(i)})$ is the hazard at time t given predictor values $\boldsymbol{X}^{(i)}$, and $\lambda_0(t)$ is an arbitrary baseline hazard function. $\boldsymbol{\beta}$ is the effect size of each variable. One usually estimates the parameter $\boldsymbol{\beta}$ in the proportional-hazards model (6.1) through maximization of the partial likelihood (without specification of $\lambda_0(t)$):

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{X}^{(i)}\boldsymbol{\beta})}{\sum_{k \in R(t_i)} \exp(\boldsymbol{X}^{(k)}\boldsymbol{\beta})} \tag{6.2}$$

where $R(t) = \{\text{subject } k | t_k \leq t\}$ is the risk set at time t, i.e. the indices of individuals that are still alive at time t. The penalized Cox model proposes to estimate $\boldsymbol{\beta}$ by:

$$\hat{\beta} = \arg\min \{L(\boldsymbol{\beta}) + \lambda \|\boldsymbol{\beta}\|_1\} . \tag{6.3}$$

The direct optimization problem is difficult to adapt to our method, two possible alternatives provide other possibilities to solve this issue.

Survival stacking: transform to a classification problem

As an alternative to Cox model, Craig et al. (2021) proposed a method called "survival stacking" which reshapes survival data, so that survival problems can be addressed as classification problems, thereby enabling the use of classification methods in a survival setting. As a simple example of right-censored dataset

illustrated by Figure 6.2 with the following observations:

$$\mathbf{X} = \begin{array}{c} \begin{array}{cc} \text{predictor 1} & \text{predictor 2} \end{array} \\ \begin{pmatrix} X_1^{(1)} & X_2^{(1)} \\ X_1^{(2)} & X_2^{(2)} \\ X_1^{(3)} & X_2^{(3)} \end{pmatrix} \end{array}, \mathbf{y} = \begin{array}{c} \begin{array}{cc} \text{survival time}(t) & \text{event}(\delta) \end{array} \\ \begin{pmatrix} t_1 & 1 \\ t_2 & 0 \\ t_3 & 1 \end{pmatrix} \end{array},$$

we observe a total of two events at time t_1 and t_3 respectively among the three

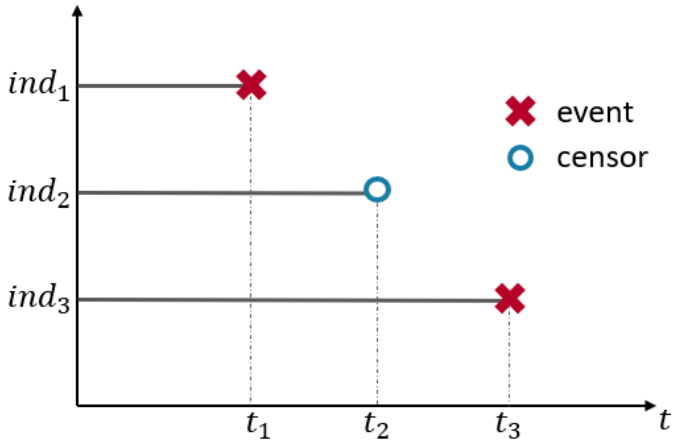

Figure 6.2: Example to illustrate the "stacking" in survival analysis.

individuals, and we defined $t_1 < t_2 < t_3$. For the first event at time point t_1, the corresponding risk set (individuals who still survive or have not been censored) is $\{1, 2, 3\}$, so we use all three observations and created the following predictor matrix and binary outcome vector:

$$\mathbf{X}_{R(t_1)} = \begin{array}{c} \begin{array}{cccc} \text{predictor 1} & \text{predictor 2} & \text{risk set 1} & \text{risk set 2} \end{array} \\ \begin{pmatrix} X_1^{(1)} & X_2^{(1)} & 1 & 0 \\ X_1^{(2)} & X_2^{(2)} & 1 & 0 \\ X_1^{(3)} & X_2^{(3)} & 1 & 0 \end{pmatrix} \end{array}, \tilde{\mathbf{y}}_{R(t_1)} = \begin{array}{c} \text{outcome} \\ \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \end{array}.$$

The newly added variable "risk set 1" corresponds to the risk indicator for individuals at risk at time t_1 (time point of the first event). We repeat for the second observed time t_3 for the second event, when the first two individuals are no longer in the study. The corresponding risk set is $\{3\}$, so our predictor matrix and binary outcome vector are:

$$\mathbf{X}_{R(t_3)} = \begin{array}{c} \begin{array}{cccc} \text{predictor 1} & \text{predictor 2} & \text{risk set 1} & \text{risk set 2} \end{array} \\ \begin{pmatrix} X_1^{(3)} & X_2^{(3)} & 0 & 1 \end{pmatrix} \end{array}, \tilde{\mathbf{y}}_{R(t_3)} = \begin{array}{c} \text{outcome} \\ \begin{pmatrix} 1 \end{pmatrix} \end{array}.$$

The variable "risk set 2" is the indicator for individuals at risk at time t_3 (time point of the second event). Finally, we vertically stack all predictor matrices and outcome vectors to form a single dataset:

$$\mathbf{X} = \begin{array}{c} \begin{array}{cccc} \text{predictor 1} & \text{predictor 2} & \text{risk set 1} & \text{risk set 2} \end{array} \\ \begin{pmatrix} X_1^{(1)} & X_2^{(1)} & 1 & 0 \\ X_1^{(2)} & X_2^{(2)} & 1 & 0 \\ X_1^{(3)} & X_2^{(3)} & 1 & 0 \\ X_1^{(3)} & X_2^{(3)} & 0 & 1 \end{pmatrix} \end{array}, \tilde{\mathbf{y}} = \begin{array}{c} \text{outcome} \\ \begin{pmatrix} 1 \\ 0 \\ 0 \\ 1. \end{pmatrix} \end{array}$$

By reframing survival problems as classification problems, we can now leverage the previously developed WLogit in a survival context. Craig et al. (2021) showed that the survival stacking has a deep relationship to the Cox model: they verified that the coefficients used in the logistic regression model with reshaped data were a close approximation of those from the Cox model. This technique has already been discussed in D'Agostino et al. (1990) and Ingram and Kleinman (1989), but underused in practice.

Restricted mean survival time: transform to a regression problem

The Cox proportional hazard model is a linear model that assumes the relationship between the predictors and the hazard is constant through time, which is not always appropriate. Another alternative tool, the restricted mean survival time (RMST), was proposed to analyze survival data in a more reliable way (Royston and Parmar (2011), Royston and Parmar (2013)). The restricted mean is a measurement of average survival from time 0 to a specified (restricted) time point, and may be estimated as the area under the survival curve up to that point. Suppose $S(t)$ is the survival curve (Kaplan-Meier curve for example). The restricted mean survival time μ to some horizon $t*$ equals the area under the survival curve $S(t)$ from $t = 0$ to $t = t*$:

$$\mu = \int_0^{t*} S(t)\, dt. \tag{6.4}$$

For example, when t is years to death, we may think of μ as the $t*$-year life expectancy. To take into account the covariables, Ambrogi et al. (2022) further proposed a pseudo-value approach that can model RMST with covariables through a linear model. With the ith pseudo-value at time $t*$ defined as:

$$\theta_{t*,i} = n \int_0^{t*} S(t)\, dt - (n-1) \int_0^{t*} S^{-i}(t)\, dt, \tag{6.5}$$

where $S^{-i}(t)$ is the Kaplan-Meier estimator excluding subject i, $\theta_{t*,i}$ therefore presents the contribution of the ith individual to the overall mean time survival at time $t*$. Taking this as the response value of individual i, a linear regression model can be assumed as:

$$\theta_{t*,i} = \alpha + \boldsymbol{X}^{(i)}\boldsymbol{\beta}, \tag{6.6}$$

where α determines the baseline restricted mean time and $\boldsymbol{X}^{(i)}$ are time-independent covariates. The basic model can be the linear regression where we assume a linear relationship between the baseline predictor and the survival time, therefore we can now use the previously developed WLasso method in this context.

6.2.2. Identification of predictive biomarkers

In chapter 4, we developed a method, PPLasso, that can simultaneously identify prognostic and predictive biomarkers in the linear model. Although this method is currently only limited to continuous quantitative response, the idea of using the ANCOVA model for predictive effect can be easily generalized to other types of statistical modeling, including logistic regression. Following the idea of PPLasso, we can construct the same design matrix with biomarker-treatment interactions and treatment effects. Then for logistic regression, we use the re-weighted least square approach as in WLogit. Finally, the thresholding and penalty of PPLasso could be used to identify the two kinds of biomarkers. This corresponds to PPLogit in Figure 6.1. Furthermore, with the idea of adapting to survival analysis as described in Section 6.2.1, we can also develop new approaches for predictive biomarker identification based on time-to-event endpoints: by converting the survival data to binary (survival stacking) or continuous (restricted survival mean time) response, we will be able to apply PPLogit or PPLasso on converted dataset respectively. Thus, one perspective of this work is to develop and make available to users a more sophisticated and complete R package encompassing all the methods for different response types and different types of biomarkers.

6.2.3. Testing covariance matrices in high dimension

During my PhD research, I have been focusing on developing novel methods that take into account the correlations that may exist between covariables. All the developed methods aim to solve the inconsistency of variable selection in the presence of high correlations. However, not all datasets show a strong correlation between covariables. Testing the presence of correlation is therefore useful before implementing our methods. In the absence of correlations, traditional Lasso may be preferred even though the whitening approaches show similar results. Covariance matrix testing in high-dimensional settings requires new approaches since many classical multivariate methods fail when the number of covariables exceeds the sample size. Classical tests are based on the unbiased modified likelihood ratio criterion, which leads to unstable tests that have poor statistical performance when $p > n$. Among the newly developed techniques, (Fisher, 2012) proposed to test the hypothesis $H_0 : \Sigma = I_p$, i.e. the covariance matrix is identity (decorrelated covariates) by proposing two new statistics for high-dimensional data. With the same objective, another method based on posterior Bayes factor (Wang and Xu, 2021) was proposed. The test of block-diagonal covariance matrix was also proposed by (Hyodo et al., 2015), which can be implemented to test pathway structures. Extended to any prefixed structure, CRAMP (Ayyala et al., 2021) per-

forms hypothesis testing for a given covariance structure with the hypothesis test $H_0 : \Sigma = \Sigma_0$ for some $p \times p$ matrix Σ_0. It consists in projecting the high dimensional data randomly into lower-dimensional subspaces, therefore allowing for the use of traditional multivariate tests.

Chapter 7 - En bref

7.1. Contexte biologique

La médecine traditionnelle traite tous les patients de la même manière sans tenir compte de leurs spécifités propres mais parfois les gens ne réagissent pas de la même manière à un traitement. Certains médicaments fonctionnent très bien pour certaines personnes, tandis que d'autres peuvent s'avérer inefficaces ou même provoquer des effets secondaires (Alberti and Cavaletti, 2014). Contrairement à la médecine traditionnelle, la médecine de précision est beaucoup plus ciblée : elle vise à donner un traitement adapté à chaque patient (figure 7.1). Selon la définition de l'Institut national du cancer des États-Unis,

"La médecine de précision est une forme de médecine qui utilise des informations sur les gènes, les protéines et l'environnement d'une personne pour prévenir, diagnostiquer et traiter les maladies."

Avec l'arrivée des nouvelles techniques de séquençage et de la médecine de précision, la compréhension de la cause d'une maladie, la découverte des mécanismes de réponse aux médicaments et la prédiction de la réponse au traitement pour la prise de décision thérapeutique deviennent de plus en plus importantes dans la recherche médicale. La disponibilité de quantités croissantes de données issues d'études de population permet aux chercheurs d'avoir accès aux données "-omiques" (métabolomique, protéomique, transcriptomique,...) et cliniques des patients, et donc d'accéder à des informations au niveau moléculaire (Krassowski et al., 2020). Ces molécules biologiques sont appelées les biomarqueurs, ils incluent : les gènes dans les données RNAseq, les protéines dans les données protéomiques et les SNP dans les données GWAS. Cependant, pour exploiter au mieux ces données de nouvelles méthodes doivent être mises en place (Ginsburg and Phillips, 2018).

Une caractéristique importante des données "-omiques" est le fait que le nombre de variables (ici les biomarqueurs) est généralement supérieur à la taille de l'échantillon. Par exemple, dans le cadre des données RNAseq on peut avoir plus de 20 000 données d'expression de gènes alors que la taille de l'échantillon est généralement limitée ($10 \sim 100$) en raison du coût des expériences. Par ailleurs, étant donné que seul un petit sous-ensemble de biomarqueurs suffit généralement pour expliquer la variable réponse (un critère clinique, par exemple), arriver à identifier ces biomarqueurs est fondamental et difficile. La présence d'interactions complexes entre les biomarqueurs augmente encore la difficulté de leur identification (Yamada et al., 2020). De nouvelles techniques sont donc nécessaires pour

identifier les biomarqueurs véritablement explicatifs d'une variable réponse et ainsi mieux comprendre les maladies au niveau moléculaire.

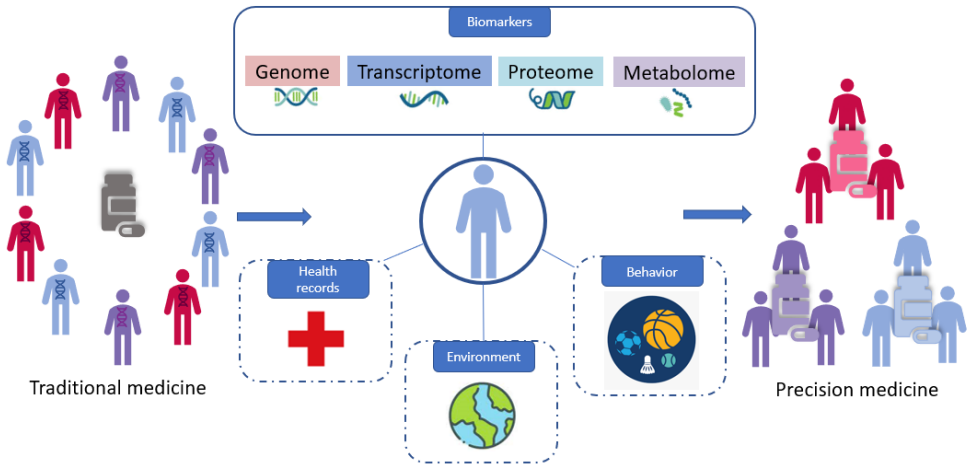

Figure 7.1: médecine traditionnelle v.s. médecine de précision

7.2. Exemples en médecine de précision

Au fil des années, les scientifiques en ont appris davantage sur les mécanismes biologiques qui contrôlent l'apparition et le comportement des maladies. Le fait de savoir comment les biomarqueurs et les maladies interagissent a permis d'affiner les traitements pour les rendre plus efficaces. Selon le glossaire de la FDA (Food and Drug Administration) et du NIH (National Institutes of Health), différents types de biomarqueurs ont été définis. Nous nous focaliserons ici sur deux d'entre eux : les biomarqueurs pronostiques et prédictifs. Un biomarqueur pronostique informe sur l'issue probable d'une maladie (par exemple, la récidive de la maladie, sa progression ou le décès du patient) indépendamment du traitement reçu. Comme l'illustre la figure 7.2, nous supposons que le biomarqueur n'a que deux statuts : positif (B+) et négatif (B-), et qu'un traitement expérimental (Exp) est comparé à un placebo (Pbo)/traitement standard. Le critère d'évaluation clinique est mesuré pour différents traitements et différents statuts du biomarqueur. Si un biomarqueur est considéré comme pronostique, la réponse est différente selon le statut du biomarqueur (positif ou négatif) mais n'est pas liée au traitement (placebo ou expérience). Un exemple de biomarqueur pronostique est celui des mutations BRCA1/2. Chez les femmes atteintes d'un cancer du sein, les mutations BRCA1/2 suggèrent un risque plus élevé de développer un cancer du sein controlatéral. Pour réduire ce risque, certaines femmes atteintes d'un cancer du sein associé à une mutation de BRCA1/2 subissent une mastectomie prophylactique contralatérale, et il a été rapporté que cette procédure réduisait le risque d'un futur cancer du sein contralatéral d'au moins 90 % (Sprundel et al., 2005; Domchek et al., 2010).

Un autre exemple est la signature MammaPrint développée dans le cancer du sein (Sotiriou and Pusztai, 2009). MammaPrint utilise des microréseaux pour mesurer l'expression de 70 gènes qui sont des caractéristiques clés du cancer, sur la base desquels les patients sont séparés en deux groupes : ceux qui présentent un risque faible ou élevé de récidive de la maladie. Dans l'essai clinique prospectif randomisé MINDACT, les patients classés comme étant à faible risque ont obtenu un excellent résultat en termes de survie sans maladie. Selon les lignes directrices de l'EGTM (European Group on Tumour Markers), le test MammaPrint "peut être utilisé pour déterminer le pronostic et guider la prise de décision concernant l'administration d'une chimiothérapie adjuvante chez les patientes atteintes d'un cancer du sein invasif récemment diagnostiqué." Les biomarqueurs pronostiques indiquent le risque potentiel de progression de la maladie et contribuent donc à la prise de décision quant à la nécessité ou à l'agressivité d'un traitement.

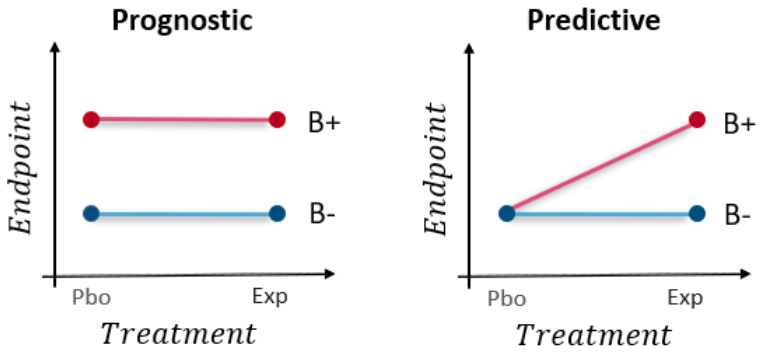

Figure 7.2: Biomarqueurs pronostiques (gauche) v.s. biomarqueurs prédictifs (droite)

Contrairement aux biomarqueurs pronostiques, un biomarqueur est considéré comme prédictif si l'effet du traitement (expérimental par rapport au traitement de contrôle/standard) est différent selon le niveau du biomarqueur (par exemple dans le cas d'un biomarqueur binaire : entre les patients positifs au biomarqueur et les patients négatifs au biomarqueur). Dans la figure 7.2, seul le groupe B+ a montré un effet du traitement expérimental par rapport au placebo sur le critère clinique, alors qu'aucun effet de traitement n'a été observé dans le groupe B-. Les biomarqueurs prédictifs peuvent donc aider à déterminer quels patients sont les plus susceptibles de répondre ou d'être résistants à des thérapies spécifiques. Un exemple est le traitement à base d'erlotinib pour soigner le cancer du poumon. Les patients dont les tumeurs présentent une mutation de l'EGFR ont un taux de survie plus élevé lorsqu'ils sont traités avec de l'erlotinib par rapport à ceux qui ont eu un placebo. En revanche, le groupe des patients de type sauvage EGFR n'a pas montré de bénéfice clair de l'erlotinib (Ballman, 2015). Un autre exemple est la mesure de l'expression du gène HER2 dans les décisions de traitement du

cancer du sein. La surexpression de HER2 entraîne la croissance de la tumeur et favorise la prolifération et l'invasion des cellules (Rimawi et al., 2015). Selon ce mécanisme, quatre formes de thérapie anti-HER2 sont disponibles (Martin and López-Tarruella, 2016) et la surexpression du gène HER2 semble être nécessaire pour que les patients répondent à ces traitements. Par conséquent, seules les patientes HER2-positives peuvent recevoir des thérapies anti-HER2. Outre la prise de décision thérapeutique, les biomarqueurs prédictifs sont également importants pour la recherche clinique. Par exemple, dans un essai clinique contrôlé randomisé d'une thérapie expérimentale, un biomarqueur peut être utilisé pour sélectionner les patients à inclure dans un essai clinique ou pour stratifier les patients en groupes de biomarqueurs positifs et négatifs. Si le biomarqueur est prédictif d'une issue favorable, l'effet du traitement expérimental par rapport à un traitement de référence sera plus important chez les patients porteurs du biomarqueur. Distinguer les biomarqueurs pronostiques et prédictifs peut être difficile dans certains cas en particulier lorsqu'un seul traitement est présent, l'effet sur un critère d'évaluation clinique donné pouvant provenir d'un effet pronostique ou prédictif, voire des deux.

Le développement de la médecine de précision change la façon dont les patients sont traités. Vargas and Harris (2016) a résumé les traitements du cancer du poumon par la médecine de précision en se basant sur différents types de biomarqueurs (génomique, transcriptomique, épigénomique, protéomique, etc.) et a listé les biomarqueurs prometteurs dans ce dernier. D'un point de vue plus global, Tsimberidou et al. (2020) a revisité l'histoire du développement de la médecine de précision et a indiqué que de nouvelles stratégies, notamment les thérapies utilisant les expressions génétiques, permettront d'optimiser le traitement de chaque patient et d'accélérer la découverte de nouveaux médicaments.

L'objectif de ma thèse est donc de développer de nouvelles méthodes qui peuvent sélectionner correctement les biomarqueurs pertinents dans des contextes de grande dimension, en particulier lorsque les biomarqueurs sont corrélés. Nous avons d'abord considéré le cas de la régression linéaire multiple pour identifier les biomarqueurs pronostiques lorsque la variable réponse est continue. Avec la présence de deux traitements potentiels, nous avons ensuite développé une nouvelle méthode pour identifier simultanément les marqueurs pronostiques et prédictifs en utilisant un modèle linéaire de type ANCOVA. Enfin, nous avons étendu nos méthodes au cas de la régression logistique lorsque la variable à expliquer est binaire.

7.3. Sélection de variables dans le modèle de régression linéaire multiple en grande dimension

7.3.1. Contexte

Soit $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ la réponse de n échantillons, où A^T désigne la transposée de A. Nous considérons alors le modèle suivant :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{7.1}$$

où $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_p)$ est la matrice de design contenant l'expression de p biomarqueurs ($p \gg n$) et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ est un vecteur parcimonieux à estimer, c'est-à-dire avec une majorité de coefficients nuls. Dans le modèle (7.1), $\boldsymbol{\epsilon}$ est le terme d'erreur. La sélection de variables vise à identifier toutes les variables dont les coefficients de régression sont estimés comme non nuls. Plusieurs articles de revue ont été écrits sur le sujet comme Saeys et al. (2007) et Heinze et al. (2018) par exemple et indiquent que trois types de méthodes sont principalement utilisées pour faire de la sélection de variables dans des données "-omiques" de grande dimension.

Tests univariés

L'approche univariée consiste à étudier indépendamment chaque biomarqueur en évaluant sa force d'association avec la réponse dans un modèle de régression linéaire (McDonald, 2009). Cependant, la multiplicité des tests statistiques peut rendre cette approche moins puissante (Lee and Lee, 2020). Pour résoudre ce problème, des corrections de tests multiples ont été proposées. La correction la plus connue est la correction de Bonferroni, et d'autres moins conservatrices incluent les techniques de Bonferroni-Holm (Holm, 1979) et de Hochberg (Hochberg, 1988). Pour contrôler le taux de fausse découverte, les ajustements couramment utilisés dans l'analyse des données "-omiques" sont les suivants : Benjamini and Hochberg (1995) et Benjamini and Yekutieli (2001). Bien que très simple à mettre en œuvre, cette approche ne tient pas compte de la corrélation entre les biomarqueurs, ce qui peut constituer une limitation importante dans le contexte des données "-omiques".

Approches forward, backward et stepwise

Dans le cadre de ces approches le choix des variables est effectué par une procédure qui, à chaque étape, envisage d'ajouter ou de soustraire une variable à l'ensemble des variables explicatives en fonction d'un critère pré-spécifié (Hocking (1976)). On trouve des applications de cette approche à la sélection de biomarqueurs dans Xiong et al. (2001) et Lu et al. (2020). Cependant, ces approches présentent souvent un risque élevé de surajustement et sont coûteuses en temps de calcul pour les données de grande dimension (Smith, 2018).

Approches régularisées

Les approches régularisées sont souvent utilisées dans le contexte de données de grande dimension. Elles consistent à ajouter une pénalité de type ℓ_1 à la vraisemblance du modèle (7.1) comme c'est le cas par exemple pour le critère Lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996a). Il consiste à minimiser le critère suivant :

$$L_\lambda(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 , \tag{7.2}$$

où $\|\mu\|_2^2 = \sum_{i=1}^{n} \mu_i^2$ et $\|\mu\|_1 = \sum_{i=1}^{n} |\mu_i|$ lorsque $\mu = (\mu_i, \ldots, \mu_n)^T$.

D'autres types de pénalités ont été proposées comme celles qui sont par exemple utiliser dans l'Elastic Net (Zou and Hastie, 2005), le SCAD (Fan and Li, 2001), le Dantzig selector (Candes and Tao, 2007), le nonnegative Garrote (Breiman, 1995), le Lasso adaptatif (Zou, 2006) ou le group Lasso (Yuan and Lin, 2006). Une revue plus générale sur le sujet de la sélection de variables dans des contextes de grande dimension peut être trouvée dans (Fan and Lv, 2009). En raison de leur capacité à effectuer la sélection de variables et l'estimation de coefficients simultanément (Fan and Li, 2006), les approches régularisées ont été largement appliquées à l'analyse génomique (Ogutu et al., 2012; Li and Sillanpää, 2012; Desta and Ortiz, 2014). Par conséquent, nous nous concentrerons dans cette thèse principalement sur ce type d'approches.

7.3.2. Fortes corrélations entre les biomarqueurs

Une difficulté notoire de la sélection de variables dans les contextes de grande dimension provient de la corrélation entre les variables explicatives (covariables, biomarqueurs). La corrélation peut facilement conduire la sélection de variables (biomarqueurs) non pertinentes en particulier dans le cadre des données génomiques de grande dimension. La figure 7.3 présente les corrélations empiriques entre les données d'expression des gènes (après prétraitement) dans un ensemble de données sur le cancer de la prostate (Singh et al., 2002) (2135 gènes, 102 échantillons) et un ensemble de données sur le cancer du sein (Sotiriou et al., 2006) (1111 gènes, 189 échantillons). Nous pouvons clairement observer de fortes corrélations par blocs, ce qui signifie que les gènes de chaque bloc sont corrélés.

En présence de telles corrélations, le Lasso est connu pour ne pas être consistant en signe. La consistance en signe garantit que les coefficients non nuls de $\boldsymbol{\beta}$ (associés aux variables actives) sont estimés par des coefficients non nuls de même signe et que les coefficients nuls (associés aux variables non actives) sont estimés par des coefficients nuls. Plus précisément, un estimateur de $\boldsymbol{\beta}$ est consistant en signe si

$$\mathbb{P}\left(sign(\widehat{\boldsymbol{\beta}}) = sign(\boldsymbol{\beta})\right) \xrightarrow[n\to+\infty]{} 1 \tag{7.3}$$

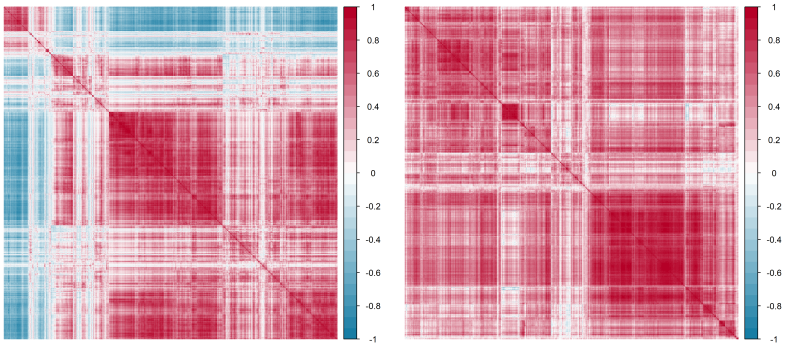

Figure 7.3: Heatmap des corrélations entre les expressions des gènes (prostate à gauche et cancer du sein à droite).

où $sign(x) = 1$ si $x > 0$, -1 si $x < 0$ et 0 si $x = 0$.

La consistance de la sélection de variables a fait l'objet d'une attention considérable, et divers travaux ont été consacrés à l'étude de la consistance de la sélection de modèles obtenus par le Lasso : Zhao and Yu (2006); Meinshausen and Bühlmann (2006); Lounici (2008); Meinshausen and Yu (2009). Pendant ma thèse, je me suis principalement concentrée sur la Condition d'Irreprésentabilité établie par Zhao and Yu (2006). Les auteurs ont prouvé que cette condition est nécessaire et suffisante pour retrouver le support de $\boldsymbol{\beta}$, c'est-à-dire pour retrouver les composantes nulles et non nulles dans le vecteur $\boldsymbol{\beta}$ et ainsi fournir un estimateur consistant en signe. Cette condition est définie comme suit.

Condition d'Irreprésentabilité (IC) : Soit $S = \{j,\ \beta_j \neq 0\}$ l'ensemble des variables actives, S^c l'ensemble des variables non actives et $\boldsymbol{X}_A$ la sous-matrice de $\boldsymbol{X}$ contenant uniquement les indices des colonnes qui sont dans l'ensemble A. Par conséquent, la matrice de covariance empirique des covariables, $C_n = n^{-1}\boldsymbol{X}^T\boldsymbol{X}$, peut être réécrite comme suit :

$$C_n = \begin{bmatrix} C_{11}^n & C_{12}^n \\ C_{21}^n & C_{22}^n \end{bmatrix},$$

where $C_{11}^n = n^{-1}\boldsymbol{X}_S^T\boldsymbol{X}_S$, $C_{12}^n = n^{-1}\boldsymbol{X}_S^T\boldsymbol{X}_{S^c}$, $C_{21}^n = n^{-1}\boldsymbol{X}_{S^c}^T\boldsymbol{X}_S$, $C_{22}^n = n^{-1}\boldsymbol{X}_{S^c}^T\boldsymbol{X}_{S^c}$. Alors, la matrice de design $\boldsymbol{X}$ satisfait à la condition d'irreprésentabilité si pour une certaine constante $\alpha \in (0, 1)$,

$$\left|\left(C_{21}^n (C_{11}^n)^{-1}\text{sign}(\boldsymbol{\beta}_S)\right)_j\right| \leq 1 - \alpha,\ \text{pout tout } j. \tag{7.4}$$

Intuitivement, cette condition signifie que la corrélation entre les variables actives et non actives est inférieure à la corrélation entre les variables actives. Par conséquent, cette condition est plus susceptible d'être violée lorsque les corrélations entre les variables actives et non actives sont importantes.

Dans les données génomiques en grande dimension, cette condition est difficile à garantir car la corrélation entre les biomarqueurs est généralement élevée (Michalopoulos et al., 2012). Ce phénomène est typiquement observé dans les données omiques. (Wang et al., 2019) a testé la condition d'irreprésentabilité sur plusieurs données génomiques et a mis en évidence que la condition est violée dans presque tous les jeux de données étudiés.

7.3.3. Idée pour supprimer la présence de corrélations : utiliser le blanchiment

Plusieurs stratégies ont été proposées pour résoudre le problème des corrélations élevées entre les biomarqueurs. L'Elastic Net introduit par Zou and Hastie (2005) combine à la fois la pénalité ℓ_1 ($\|\boldsymbol{\beta}\|_1 = \sum_{i=1}^{p} |\boldsymbol{\beta}_i|$) du Lasso et la pénalité ℓ_2 ($\|\boldsymbol{\beta}\|_2 = \sum_{i=1}^{p} \boldsymbol{\beta}_i^2$). L'Elastic Net a pour effet de sélectionner des groupes de variables corrélées. Le préconditionnement est un autre type de méthodes permettant de traiter la corrélation. Jia and Rohe (2015) et Wang and Leng (2016) ont proposé de multiplier à gauche $\boldsymbol{X}$, $\boldsymbol{y}$ et $\boldsymbol{\epsilon}$ dans le modèle (7.1) par des matrices spécifiques pour supprimer les corrélations entre les colonnes de $\boldsymbol{X}$. Une autre méthode publiée récemment, nommée Precision Lasso (Wang et al., 2019), propose de traiter le problème de corrélation en attribuant des poids similaires aux variables corrélées.

Dans ma thèse, j'ai proposé une nouvelle approche pour éliminer les corrélations qui peuvent exister entre les covariables (biomarqueurs). Supposons que les n lignes $\boldsymbol{X}^{(1)}, \ldots, \boldsymbol{X}^{(n)}$ de $\boldsymbol{X}$ sont supposées être des vecteurs aléatoires gaussiens indépendants avec une matrice de covariance égale à Σ. Soit $\Sigma^{-1/2} := \boldsymbol{U}\boldsymbol{D}^{-1/2}\boldsymbol{U}^T$ où $\boldsymbol{U}$ et $\boldsymbol{D}$ sont les matrices impliquées dans la décomposition spectrale de la matrice symétrique Σ donnée par : $\Sigma = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^T$. Nous notons alors $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$.

Le modèle (7.1) peut ainsi être réécrit comme suit :

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} + \boldsymbol{\epsilon}, \tag{7.5}$$

où $\widetilde{\boldsymbol{\beta}} = \Sigma^{1/2}\boldsymbol{\beta} := \boldsymbol{U}\boldsymbol{D}^{1/2}\boldsymbol{U}^T\boldsymbol{\beta}$.

Avec une telle transformation, la matrice de covariance des lignes de $\widetilde{\boldsymbol{X}}$ est égale à l'identité et les colonnes de $\widetilde{\boldsymbol{X}}$ sont donc non corrélées. La figure 7.4 présente la heatmap des corrélations (mêmes ensembles de données que ceux présentés dans la figure 7.3) après la transformation de blanchiment. La matrice de covariance Σ a été estimée par le package cvCovEst (Boileau et al., 2021).

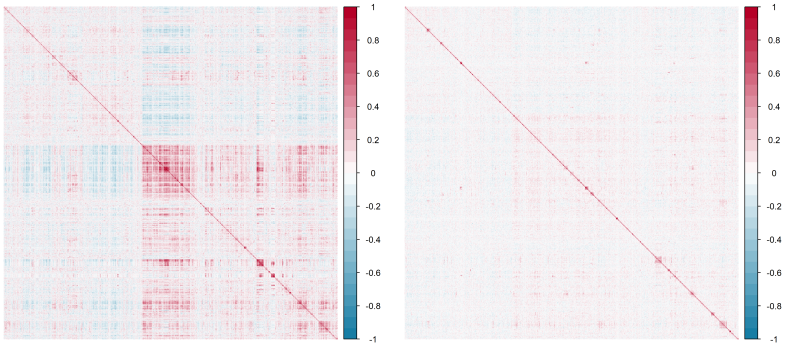

Figure 7.4: Heatmap des corrélations après le blanchiment (prostate à gauche et cancer du sein à droite).

7.3.4. Contribution du chapitre 2

Cette section résume l'article suivant :
Zhu, W., Lévy-Leduc, C., et Ternès, N. (2021), A variable selection approach for highly correlated predictors in high-dimensional genomic data, *Bioinformatics*, 37(16), 2238-2244.

La méthode proposée est implémentée dans le package R `WLasso` disponible sur le CRAN.

Nous proposons une nouvelle approche de sélection de variables, WLasso (Whitening Lasso), avec l'idée de blanchiment introduite précédemment. Après transformation du modèle (7.5), nous proposons de minimiser le critère suivant par rapport à $\widetilde{\boldsymbol{\beta}}$:

$$L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}}) = \left\| \boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}} \right\|_2^2 + \lambda \left\| \Sigma^{-1/2}\widetilde{\boldsymbol{\beta}} \right\|_1, \tag{7.6}$$

qui garantit une contrainte de parcimonie sur $\boldsymbol{\beta}$ grâce à la pénalité ℓ_1. Nous obtenons donc

$$\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda) = \arg\min_{\widetilde{\boldsymbol{\beta}}} L_\lambda^{\text{gen}}(\widetilde{\boldsymbol{\beta}}).$$

Pour estimer $\widetilde{\boldsymbol{\beta}}$, nous utiliserons l'estimateur modifié suivant qui peut être vu comme un seuillage des composantes de $\widehat{\widetilde{\boldsymbol{\beta}}}_0(\lambda)$. Pour K dans $\{1, \ldots, p\}$, soit Top_K l'ensemble des indices correspondant aux K plus grandes valeurs des composantes de $\widehat{\widetilde{\boldsymbol{\beta}}}_0$, alors l'estimateur de $\widetilde{\boldsymbol{\beta}}$ est $\widehat{\widetilde{\boldsymbol{\beta}}} = (\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(\widehat{K})})_{1\le j\le p}$ où $\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}$ est défini par :

$$\widehat{\widetilde{\boldsymbol{\beta}}}_j^{(K)}(\lambda) = \begin{cases} \widehat{\widetilde{\boldsymbol{\beta}}}_{0j}(\lambda), & j \in \text{Top}_K \\ K\text{ième plus grande valeur de } |\widehat{\widetilde{\boldsymbol{\beta}}}_{0j}|, & j \notin \text{Top}_K. \end{cases} \tag{7.7}$$

Pour estimer $\boldsymbol{\beta}$, nous allons d'abord considérer $\widehat{\boldsymbol{\beta}}_0 = \Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}$ et ensuite appliquer

une stratégie de seuillage. Ainsi, nous proposons d'estimer $\boldsymbol{\beta}$ par $\widehat{\boldsymbol{\beta}} = (\widehat{\boldsymbol{\beta}}_j^{(\widehat{M})})_{1\leq j\leq p}$ où $\widehat{\boldsymbol{\beta}}_j^{(M)}$ est défini par :

$$\widehat{\boldsymbol{\beta}}_j^{(M)}(\lambda) = \begin{cases} \widehat{\boldsymbol{\beta}}_{0j}(\lambda), & j \in \mathrm{Top}_M \\ 0, & j \notin \mathrm{Top}_M. \end{cases} \tag{7.8}$$

Les variables dont les coefficients ne sont pas nuls dans $\widehat{\boldsymbol{\beta}}$ sont considérées comme associées à la variable de réponse. Dans le chapitre 2, nous avons montré dans diverses expériences numériques que, lorsque les biomarqueurs sont fortement corrélés, notre approche `WLasso` obtient de meilleures performances statistiques que les approches auxquelles nous l'avons comparée.

7.3.5. Contribution du chapitre 3

Cette section résume l'article suivant :
Zhu, W., Adjakossa, E., Lévy-Leduc, C., et Ternès, N. (2021). Sign Consistency of the Generalized Elastic Net Estimator. Soumis avec *arXiv preprint* (arXiv:2106.05454).

Dans cette section, outre la transformation du modèle (7.5), nous proposons de combiner une pénalité ℓ_1 et ℓ_2, et de considérer le critère suivant :

$$L_{\lambda,\eta}^{gEN}(\widetilde{\boldsymbol{\beta}}) = \left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\beta}}\right\|_2^2 + \lambda\left\|\Sigma^{-1/2}\widetilde{\boldsymbol{\beta}}\right\|_1 + \eta\left\|\widetilde{\boldsymbol{\beta}}\right\|_2^2, \text{ avec } \lambda,\eta > 0. \tag{7.9}$$

Puisqu'il consiste à ajouter une partie pénalité de ℓ_2 au Lasso généralisé comme dans l'Elastic Net, nous l'appellerons Elastic Net généralisé (gEN). L'estimateur gEN est défini par :

$$\widehat{\boldsymbol{\beta}} = \Sigma^{-1/2}\widehat{\widetilde{\boldsymbol{\beta}}}, \tag{7.10}$$

avec

$$\widehat{\widetilde{\boldsymbol{\beta}}} = \arg\min_{\widetilde{\boldsymbol{\beta}}} L_{\lambda,\eta}^{gEN}\left(\widetilde{\boldsymbol{\beta}}\right). \tag{7.11}$$

Avec ce nouvel estimateur, nous avons défini une condition d'irreprésentabilité appelée **Generalized Irrepresentable Condition (GIC)**:
Il existe $\lambda, \eta, \alpha, \delta_4 > 0$ tels que pour tout j,

$$\left|\left((C_{21}^n + \frac{\eta}{n}\Sigma_{21})(C_{11}^n + \frac{\eta}{n}\Sigma_{11})^{-1}\left(\mathrm{sign}(\boldsymbol{\beta}_1) + \frac{2\eta}{\lambda}\boldsymbol{\beta}_1\right) - \frac{2\eta}{\lambda}\Sigma_{21}\boldsymbol{\beta}_1\right)_j\right| \leq 1-\alpha, \text{ pour tout } j, \tag{7.12}$$

où $\boldsymbol{\beta}_1$ désigne les composantes non nulles du vecteur $\boldsymbol{\beta}$. Nous avons prouvé que cette condition est suffisante pour que l'estimateur gEN soit consistant en signe sous certaines conditions. De plus, nous avons comparé **GIC** avec **EIC** (Elastic Net Irrepresentable Condition, Jia and Yu (2010)) et **IC**, et démontré qu'il existe des cas où **GIC** est satisfaite alors qu'**EIC** et **IC** ne le sont pas.

7.4. Identification de biomarqueurs pronostiques et prédictifs dans des modèles linéaires en grande dimension avec PPLasso

7.4.1. Identification de biomarqueurs prédictifs

Avec les progrès de la médecine de précision, l'identification de biomarqueurs pronostiques ou prédictifs suscite un intérêt croissant. Les méthodes WLasso et Elastic Net généralisés introduites précédemment ont été développées dans le but de sélectionner des biomarqueurs pronostiques. Cependant, peu d'approches existent pour identifier des biomarqueurs prédictifs. Pour des variables réponses binaires, Foster et al. (2011) a proposé de prédire d'abord les probabilités de réponse au traitement et d'utiliser cette probabilité comme réponse dans un problème de classification pour trouver des biomarqueurs efficaces. Tian et al. (2012) a proposé une nouvelle méthode pour détecter l'interaction entre le traitement et les biomarqueurs en modifiant les covariables. Cette méthode peut être mise en œuvre pour des réponses continues/binaires/de survie. Lipkovich et al. (2011) a proposé une méthode appelée SIDES, qui adopte un algorithme de partitionnement récursif pour le dépistage des interactions traitement-biomarqueur. Cette méthode a été améliorée dans Lipkovich and Dmitrienko (2014) en ajoutant une autre étape de présélection sur les biomarqueurs prédictifs fondée sur l'importance des variables. Leur approche fonctionne pour des variables réponses continues. Plus récemment, Sechidis et al. (2018) a appliqué des approches issues de la théorie de l'information pour classer les biomarqueurs selon leur effet pronostique/prédictif. Leur méthode n'est applicable que pour des réponses binaires ou de survie. De plus, la plupart de ces méthodes ont été évaluées dans une situation où la taille de l'échantillon est relativement importante et le nombre de biomarqueurs est limité, ce qui n'est pas le cas pour les données génomiques.

Dans la littérature mentionnée ci-dessus, les auteurs se sont concentrés sur l'identification de biomarqueurs prédictifs. Cependant, l'effet pronostique est également essentiel dans la recherche biomédicale. Nous avons proposé une nouvelle approche appelée PPLasso (Predictive Prognostic Lasso) pour identifier simultanément des biomarqueurs pronostiques et prédictifs dans un cadre de grande dimension en utilisant un modèle linéaire de type ANCOVA.

7.4.2. Contribution du chapitre 4

Cette section résume l'article suivant :
Zhu, W., Lévy-Leduc, C., et Ternès, N. (2022). Identification of prognostic and predictive biomarkers in high-dimensional data with PPLasso. Soumis avec *arXiv preprint* (arXiv:2202.01970).

La méthode proposée est implementée dans le package R `PPLasso` disponible sur le CRAN.

Soient $\boldsymbol{y}$ une réponse continue et t_1, t_2 deux traitements. Soit $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) la matrice de design pour les n_1 (resp. n_2) patients recevant le traitement t_1 (resp. t_2), chacune contenant des mesures sur p biomarqueurs candidats :

$$\boldsymbol{X}_1 = \begin{bmatrix} X_{11}^1 & X_{11}^2 & \dots & X_{11}^p \\ X_{12}^1 & X_{12}^2 & \dots & X_{12}^p \\ \dots & & & \\ X_{1n_1}^1 & X_{1n_1}^2 & \dots & X_{1n_1}^p \end{bmatrix}, \boldsymbol{X}_2 = \begin{bmatrix} X_{21}^1 & X_{21}^2 & \dots & X_{21}^p \\ X_{22}^1 & X_{22}^2 & \dots & X_{22}^p \\ \dots & & & \\ X_{2n_2}^1 & X_{2n_2}^2 & \dots & X_{2n_2}^p \end{bmatrix}. \tag{7.13}$$

Pour prendre en compte la corrélation potentielle qui peut exister entre les biomarqueurs des différents traitements, nous supposerons que les lignes de $\boldsymbol{X}_1$ (resp. $\boldsymbol{X}_2$) sont des vecteurs aléatoires gaussiens centrés indépendants dont la matrice de covariance est égale à $\boldsymbol{\Sigma_1}$ (resp. $\boldsymbol{\Sigma_2}$).

Pour modéliser le lien qui existe entre $\boldsymbol{y}$ et les différents types de biomarqueurs, nous proposons d'utiliser le modèle suivant :

$$\boldsymbol{y} = \begin{pmatrix} y_{11} \\ y_{12} \\ \vdots \\ y_{1n_1} \\ y_{21} \\ y_{22} \\ \vdots \\ y_{2n_2} \end{pmatrix} = \boldsymbol{X} \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \beta_{11} \\ \beta_{12} \\ \vdots \\ \beta_{1p} \\ \beta_{21} \\ \beta_{22} \\ \vdots \\ \beta_{2p} \end{pmatrix} + \begin{pmatrix} \epsilon_{11} \\ \epsilon_{12} \\ \vdots \\ \epsilon_{1n_1} \\ \epsilon_{21} \\ \epsilon_{22} \\ \vdots \\ \epsilon_{2n_2} \end{pmatrix}, \tag{7.14}$$

où $(y_{i1}, \dots, y_{in_i})$ correspond à la réponse des patients recevant le traitement t_i, i étant égal à 1 ou 2,

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

α_1 (resp. α_2) correspondant aux effets du traitement t_1 (resp. t_2). En outre, $\boldsymbol{\beta}_1 = (\beta_{11}, \beta_{12}, \dots, \beta_{1p})^T$ (resp. $\boldsymbol{\beta}_2 = (\beta_{21}, \beta_{22}, \dots, \beta_{2p})^T$) sont les coefficients associés à chacun des p biomarqueurs dans le groupe t_1 (resp. t_2), et $\epsilon_{11}, \dots, \epsilon_{2n_2}$ sont des variables aléatoires gaussiennes centrées indépendantes de $\boldsymbol{X}_1$ et $\boldsymbol{X}_2$. Lorsque t_1 représente le traitement standard ou le placebo, les biomarqueurs pronostiques

(resp. prédictifs) sont définis comme ceux ayant des coefficients non nuls dans $\boldsymbol{\beta}_1$ (resp. dans $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$) et les biomarqueurs non pronostiques (resp. non prédictifs) correspondent aux indices ayant des coefficients nuls dans $\boldsymbol{\beta}_1$ (resp. dans $\boldsymbol{\beta}_2 - \boldsymbol{\beta}_1$).

Le modèle (7.14) peut être écrit comme suit :

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\gamma} + \boldsymbol{\epsilon}, \tag{7.15}$$

avec $\boldsymbol{\gamma} = (\alpha_1, \alpha_2, \boldsymbol{\beta}_1^T, \boldsymbol{\beta}_2^T)^T$.

Pour estimer $\boldsymbol{\gamma}$ de façon parcimonieuse, nous considérons un premier estimateur de $\boldsymbol{\gamma}$ obtenu en minimisant le critère suivant par rapport à $\boldsymbol{\gamma}$:

$$\frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\gamma}\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \boldsymbol{\gamma} \right\|_1, \tag{7.16}$$

où $D_1 = [\mathrm{Id}_p, \mathbf{0}_{p,p}]$ et $D_2 = [-\mathrm{Id}_p, \mathrm{Id}_p]$, Id_p désignant la matrice identité de taille p et $\mathbf{0}_{i,j}$ désignant une matrice ayant i lignes et j colonnes et ne contenant que des zéros.

Cependant, vu que la non consistance en signe du Lasso provient en général des fortes corrélations existant entre les variables, nous proposons de supprimer les corrélations en "blanchissant" la matrice $\boldsymbol{X}$. Plus précisément, nous considérons $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\Sigma^{-1/2}$, où

$$\Sigma = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & \Sigma_1 & 0 \\ 0 & 0 & 0 & \Sigma_2 \end{bmatrix} \tag{7.17}$$

et définissons $\Sigma^{-1/2}$ en remplaçant dans (7.17) Σ_i par $\Sigma_i^{-1/2}$, où $\Sigma_i^{-1/2} = \mathbf{U}_i \mathbf{D}_i^{-1/2} \mathbf{U}_i^T$, $\mathbf{U}_i$ et $\mathbf{D}_i$ étant les matrices impliquées dans la décomposition spectrale de Σ_i pour $i = 1$ ou 2. Avec une telle transformation, les colonnes de $\widetilde{\boldsymbol{X}}$ sont décorrélées et le modèle (7.15) peut être réécrit comme suit :

$$\boldsymbol{y} = \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}} + \boldsymbol{\epsilon} \tag{7.18}$$

où $\widetilde{\boldsymbol{\gamma}} = \Sigma^{1/2}\boldsymbol{\gamma}$. La fonction objectif (7.16) devient donc :

$$L^{\mathrm{PPLasso}}_{\lambda_1,\lambda_2}(\widetilde{\boldsymbol{\gamma}}) = \frac{1}{2}\left\|\boldsymbol{y} - \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{\gamma}}\right\|_2^2 + \lambda_1 \left\| \begin{bmatrix} \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & D_1 \\ \mathbf{0}_{p,1} & \mathbf{0}_{p,1} & \frac{\lambda_2}{\lambda_1} D_2 \end{bmatrix} \Sigma^{-1/2} \widetilde{\boldsymbol{\gamma}} \right\|_1. \tag{7.19}$$

Un seuillage similaire a ensuite été imposé comme expliqué précédemment dans la section 7.3.4 pour obtenir l'estimation finale de $(\widehat{\boldsymbol{\beta}}_1, \widehat{\boldsymbol{\beta}}_2)$. Les biomarqueurs dont les coefficients ne sont pas nuls dans $\widehat{\boldsymbol{\beta}}_1$ (resp. $\widehat{\boldsymbol{\beta}}_2 - \widehat{\boldsymbol{\beta}}_1$) sont considérés comme des biomarqueurs pronostiques (resp. prédictifs).

7.5. Sélection de variable dans le modèle de régression logistique en grande dimension

7.5.1. Sélection de biomarqueurs pour les réponses binaires

Auparavant, nous avons supposé que la réponse $\boldsymbol{y}$ était continue. Cette section se concentre sur les réponses binaires dont la modélisation peut être vue comme un problème de classification. Cette dernière est un sujet important dans la recherche biomédicale. Par exemple, selon le guide RECIST (Response Evaluation Criteria in Solid Tumours) (Watanabe et al., 2009) sur la recherche en oncologie, l'évaluation de la réponse des patients est généralement définis comme une réponse complète, une réponse partielle, une maladie stable et une progression en fonction de la réponse au traitement. Les patients des deux premières catégories (réponse complète et réponse partielle) sont considérés comme des répondeurs au traitement, tandis que les autres sont considérés comme des non-répondeurs. Pour la maladie de la polyarthrite rhumatoïde, les critères ACR (American College of Rheumatology) sont utilisés pour évaluer la réponse au traitement dans le cadre d'un essai clinique. La réponse ACR est notée en pourcentage d'amélioration. Par exemple, ACR50 est un résultat binaire indiquant si l'amélioration est supérieure à 50 %. Un autre exemple est la classification des tumeurs. Avec le développement de la bioinformatique, la classification des cancers à partir de données omiques est devenue un sujet important dans la recherche sur le génome (Ramaswamy et al., 2001; Tibshirani et al., 2002; Menyhárt and Győrffy, 2021).

Comparée à d'autres classifieurs tels que les arbres de décision (Utgoff, 1989) ou les SVM (Support Vector Machine) (Cortes and Vapnik, 1995), la régression logistique (Walker and Duncan, 1967) est une méthode de classification très utilisée qui permet une interprétation statistique explicite et qui peut fournir des probabilités de classification pour une réponse binaire (Menard, 2002). Cependant, comme expliqué précédemment, avec les données omiques de grande dimension, il est essentiel d'obtenir un petit nombre de gènes clés et d'améliorer la précision de la classification, ce qui nous amène à considérer le problème de la sélection des variables dans le modèle de régression logistique de grande dimension en utilisant des approches régularisées (Park and Hastie, 2007). Récemment, les méthodes régularisées ont été largement appliquées à la découverte de biomarqueurs et à la classification des maladies (Zhu and Hastie, 2004; Wu, 2006; Ma and Huang, 2008; Liu et al., 2020). Outre le grand nombre de variables, les corrélations entre les biomarqueurs doivent également être prises en compte. Pour traiter ces corrélations, plusieurs méthodes ont été proposées. Les plus connues sont l'Elastic Net (Zou and Hastie, 2005) et l'Adaptive Lasso (Zou, 2006) mentionnés dans la section 7.3 et adaptés au cadre de la régression logistique. Plusieurs approches de filtrage ont également été proposées pour prendre en compte les corrélations dans le cadre de la classification. Relief (Kira and Rendell, 1992) est sensible aux

interactions entre les variables et a inspiré une famille d'algorithmes de sélection fondée sur Relief, notamment ReliefF (Kononenko et al., 1997). Ce dernier a été largement utilisé dans la recherche biomédicale (Urbanowicz et al., 2018). FCBF (Yu and Liu, 2003) est une autre approche de sélection de variables en grande dimension qui évalue la pertinence et la redondance des variables sur la base de mesures de corrélation. Dans ce chapitre, nous proposons une nouvelle méthode qui peut identifier les biomarqueurs actifs dans les données de grande dimension et fournir une classification sur les échantillons collectés.

7.5.2. Contribution du Chapter 5

Cette section résume l'article suivant qui va être prochainement soumis : Zhu, W., Lévy-Leduc, C., and Ternès, N. (2022). Variable selection in high-dimensional logistic regression models using a whitening approach.

La méthode proposée est mise en œuvre dans le package R `WLogit` prochainement disponible sur le CRAN.

Soient $\boldsymbol{X}$ une matrice de design de taille $n \times p$ où $X_j^{(i)}$ correspond à la mesure du $j$ème biomarqueur pour le $i$ème échantillon, et $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T$ le vecteur de la taille de l'effet pour chaque biomarqueur où $\boldsymbol{\beta}$ contient un grand nombre de composantes égales à zéro. Nous supposons que les réponses binaires $y_1, y_2, ..., y_n$ sont des variables aléatoires indépendantes ayant une distribution de Bernoulli avec le paramètre $\pi_{\boldsymbol{\beta}}(X^{(i)})$ ($y_i \sim Bernoulli(\pi_{\boldsymbol{\beta}}(X^{(i)}))$), où pour tous les i dans $\{1, \ldots, n\}$,

$$\pi_{\boldsymbol{\beta}}(X^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \beta_j X_j^{(i)}\right)}. \tag{7.20}$$

Les approches régularisées avec une pénalité ℓ_1 dans le contexte de la régression logistique résolvent le problème de la sélection des variables en ajoutant une pénalité ℓ_1 à la log-vraisemblance du modèle de régression logistique :

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left\{ l(\boldsymbol{\beta}) + \lambda \left\| \boldsymbol{\beta} \right\|_1 \right\}, \tag{7.21}$$

où $\|\boldsymbol{\beta}\|_1 = \sum_{k=1}^{p} |\beta_k|$, et la log-vraisemblance $l(\boldsymbol{\beta})$ est definie par:

$$l(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^{n} \left[y_i \cdot X^{(i)} \boldsymbol{\beta} - \log(1 + e^{X^{(i)} \boldsymbol{\beta}}) \right], \tag{7.22}$$

avec $X^{(i)}$ la i-ième ligne de $\boldsymbol{X}$. Avec la fonction de pénalité et le paramètre λ correctement choisi, certaines composantes de $\widehat{\boldsymbol{\beta}}$ sont mises à zéro.

Comme nous l'avons mentionné précédemment, le critère Lasso peut échouer à sélectionner le véritable sous-ensemble de biomarqueurs actifs lorsque la corrélation

entre les biomarqueurs actifs et non actifs est importante, ce qui est énoncé dans la condition d'irreprésentabilité pour les modèles de régression linéaire dans l'équation (7.4). Une condition similaire a été obtenue par Ravikumar et al. (2010) et Bunea (2008) dans le cas de la régression logistique. Soit Q défini par :

$$Q = \boldsymbol{X}^T \boldsymbol{H} \boldsymbol{X}, \tag{7.23}$$

où $\boldsymbol{H}$ est une matrice diagonale avec

$$H_{ii} = \pi_{\boldsymbol{\beta}}(X^{(i)})/(1 - \pi_{\boldsymbol{\beta}}(X^{(i)})), 1 \leq i \leq n. \tag{7.24}$$

Soit $S = \{j,\ \beta_j \neq 0\}$ l'ensemble des variables actives de taille d, S^c l'ensemble des variables non actives, Q_{SS} désigne la sous-matrice $d \times d$ de Q indexée par S. Avec cette notation, la condition s'énonce de la façon suivante :
Il existe $\alpha \in (0, 1]$ tel que :

$$\left|Q_{S^cS}(Q_{SS})^{-1}\right|_\infty \leq 1 - \alpha, \tag{7.25}$$

où $|A|_\infty = \max_{j=1,\ldots,p} \sum_{k=1}^{p} |A_{jk}|$ pour toute matrice symétrique réelle ayant p lignes et p colonnes.

Nous proposons de supprimer la corrélation en "blanchissant" la matrice $\boldsymbol{X}$. Plus précisément, nous considérons $\widetilde{\boldsymbol{X}} = \boldsymbol{X}\check{\Sigma}^{-1/2}$, où $\check{\Sigma}$ est un estimateur de covariance obtenu à partir de $\boldsymbol{H}^{1/2}\boldsymbol{X}$, où $\boldsymbol{H}$ est défini dans l'équation (7.24). Avec cette transformation, $\widetilde{\boldsymbol{X}}^T \boldsymbol{H} \widetilde{\boldsymbol{X}}$ devrait être proche de la matrice identité I_p, donc la condition d'irreprésentabilité devrait être satisfaite. Après l'étape de blanchiment, le modèle (7.20) peut être réécrit comme suit :

$$\pi_{\widetilde{\boldsymbol{\beta}}}(\widetilde{X}^{(i)}) = \frac{\exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)}\right)}{1 + \exp\left(\sum_{j=1}^{p} \widetilde{\beta}_j \widetilde{X}_j^{(i)}\right)}, \tag{7.26}$$

où $\widetilde{\boldsymbol{\beta}} = \check{\Sigma}^{1/2}\boldsymbol{\beta}$. La log-vraisemblance après la transformation peut donc être réécrite comme suit :

$$l^{wt}(\widetilde{\boldsymbol{\beta}}) = \frac{1}{n} \sum_{i=1}^{n} \left\{ y_i \cdot \widetilde{X}^{(i)} \widetilde{\boldsymbol{\beta}} - \log\left(1 + e^{\widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}}}\right) \right\}. \tag{7.27}$$

On obtient alors un estimateur de $\widetilde{\boldsymbol{\beta}}$ en résolvant le problème suivant :

$$\arg\min_{\widetilde{\boldsymbol{\beta}} \in \mathbb{R}^p} l^{wt}(\widetilde{\boldsymbol{\beta}}) + \lambda \left\| \check{\Sigma}^{-1/2} \widetilde{\boldsymbol{\beta}} \right\|_1. \tag{7.28}$$

Pour résoudre ce problème d'optimisation, nous proposons d'utiliser une approximation quadratique de la log-vraisemblance (7.27) en utilisant un développement

de Taylor à l'ordre 2 évalué aux valeurs courantes des estimateurs (Friedman et al., 2010) :

$$l_Q^{wt}(\widetilde{\boldsymbol{\beta}}) = -\frac{1}{2n}\sum_{i=1}^{n} w_i(z_i - \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}})^2 + C(\widetilde{\boldsymbol{\beta}}^o)^2 \tag{7.29}$$

$$= -\frac{1}{2n}\sum_{i=1}^{n}(\sqrt{w_i}z_i - \sqrt{w_i}\widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}})^2 + C(\widetilde{\boldsymbol{\beta}}^o)^2 \tag{7.30}$$

avec

$$z_i = \widetilde{X}^{(i)}\widetilde{\boldsymbol{\beta}} + \frac{y_i - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})}{\pi_{\widetilde{\boldsymbol{\beta}}^o}(X^{(i)})(1 - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)}))}, \text{(working response)}$$

$$w_i = \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})(1 - \pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})), \text{(weights)} \tag{7.31}$$

où $\pi_{\widetilde{\boldsymbol{\beta}}^o}(\widetilde{X}^{(i)})$ est l'évaluation de $\pi_{\widetilde{\boldsymbol{\beta}}^o}$ (définie dans le modèle (7.26)) aux paramètres courants $\widetilde{\boldsymbol{\beta}}^o$. L'estimateur final peut alors être obtenu grâce à l'algorithme IRLS (Iterative Re-weighted Least Square) (Daubechies et al., 2010).
Après avoir obtenu l'estimation de $\widetilde{\boldsymbol{\beta}}$, un seuillage similaire à celui proposé pour les modèles précédents a ensuite été imposé pour obtenir l'estimation finale $\widehat{\boldsymbol{\beta}}$. Les biomarqueurs dont les coefficients ne sont pas nuls sont considérés comme actifs. Un classifieur peut également être obtenu à l'aide de $\widehat{\boldsymbol{\beta}}$.

Les performances de WLogit sont évaluées à l'aide de données synthétiques dans plusieurs scénarios et comparées à d'autres approches. Les résultats suggèrent que WLogit peut identifier presque tous les biomarqueurs actifs, même dans les cas où les biomarqueurs sont fortement corrélés, alors que les autres méthodes échouent, ce qui conduit par conséquent à une précision de classification plus élevée. Les performances de la méthode ont également été évaluées pour la classification de deux sous-types de lymphome, et le classifieur obtenu surpasse également les autres méthodes.

Bibliography

Ahmed, M., F. Lalloo, A. Howell, and D. Evans (2009, 11). Risks of contralateral breast cancer in brca1 and brca2 mutation carriers. *Breast Cancer Research : BCR 12*.

Akaike, H. (1998). Information theory and an extension of the maximum likelihood principle. In *Selected papers of hirotugu akaike*, pp. 199–213. Springer.

Akbay, B., A. Shmakova, Y. Vassetzky, and S. Dokudovskaya (2020, 04). Modulation of mtorc1 signaling pathway by hiv-1. *Cells 9*, 1090.

Alberti, P. and G. Cavaletti (2014, 08). *Management of Side Effects in the Personalized Medicine Era: Chemotherapy-Induced Peripheral Neuropathy*, Volume 1175.

Ambrogi, F., S. Iacobelli, and P. Andersen (2022, 03). Analyzing differences between restricted mean survival time curves using pseudo-values. *BMC Medical Research Methodology 22*.

Amendola, A., F. Giordano, M. Parrella, and M. Restaino (2017, 02). Variable selection in high-dimensional regression: A nonparametric procedure for business failure prediction. *Applied Stochastic Models in Business and Industry 33*.

Ang, J. C., A. Mirzal, H. Haron, and H. N. A. Hamed (2015). Supervised, unsupervised, and semi-supervised feature selection: a review on gene selection. *IEEE/ACM transactions on computational biology and bioinformatics 13*(5), 971–989.

Ayyala, D. N., S. Ghosh, and D. F. Linder (2021). Covariance matrix testing in high dimension using random projections. *Computational Statistics*, 1–31.

Ballman, K. (2015, 09). Biomarker: Predictive or prognostic? *Journal of Clinical Oncology 33*.

Benjamini, Y. and Y. Hochberg (1995, 11). Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B 57*, 289 – 300.

Benjamini, Y. and D. Yekutieli (2001, 08). The control of the false discovery rate in multiple testing under dependency. *Ann. Stat. 29*.

Bickel, P. J., B. Li, A. B. Tsybakov, S. A. van de Geer, B. Yu, T. Valdés, C. Rivero, J. Fan, and A. van der Vaart (2006). Regularization in statistics. *Test 15*(2), 271–344.

Boileau, P., N. Hejazi, and B. Collica (2022). *cvCovEst: Cross-Validated Covariance Matrix Estimation*. R package version 1.1.0.

Boileau, P., N. S. Hejazi, B. Collica, M. J. van der Laan, and S. Dudoit (2021). 'cvcovest': Cross-validated covariance matrix estimator selection and evaluation in 'r'. *Journal of Open Source Software 6*(63), 3273.

Boileau, P., N. S. Hejazi, M. J. van der Laan, and S. Dudoit (2021). Cross-validated loss-based covariance matrix estimator selection in high dimensions.

Breiman, L. (1995). Better subset regression using the nonnegative garrote. *Technometrics 37*(4), 373–384.

Breiman, L. et al. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics 24*(6), 2350–2383.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization. *Electronic Journal of Statistics 2*, 1153 – 1194.

Cai, T., C.-H. Zhang, and H. Zhou (2010, 10). Optimal rates of convergence for covariance matrix estimation. *The Annals of Statistics 38*.

Candes, E. and T. Tao (2007). The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics 35*(6), 2313 – 2351.

Clark, G. (2008, 05). Prognostic factors versus predictive factors: Examples from a clinical trial of erlotinib. *Molecular oncology 1*, 406–12.

Cortes, C. and V. Vapnik (1995). Support-vector networks. In *Machine Learning*, pp. 273–297.

Craig, E., C. Zhong, and R. Tibshirani (2021). Survival stacking: casting survival analysis as a classification problem. *arXiv preprint arXiv:2107.13480*.

D'Agostino, R. B., M.-L. Lee, A. J. Belanger, L. A. Cupples, K. Anderson, and W. B. Kannel (1990). Relation of pooled logistic regression to time dependent cox regression analysis: the framingham heart study. *Statistics in medicine 9*(12), 1501–1515.

Daubechies, I., R. DeVore, M. Fornasier, and C. S. Güntürk (2010). Iteratively reweighted least squares minimization for sparse recovery. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences 63*(1), 1–38.

Desta, Z. and R. Ortiz (2014, 06). Genomic selection: Genome-wide prediction in plant improvement. *Trends in plant science 19*.

Domchek, S., T. Friebel, C. Singer, D. Evans, H. Lynch, C. Isaacs, J. Garber, S. Neuhausen, E. Matloff, R. Eeles, G. Pichert, L. t'veer, N. Tung, J. Weitzel, F. Couch, W. Rubinstein, P. Ganz, M. Daly, O. Olopade, and T. Rebbeck (2010, 09). Association of risk-reducing surgery in brca1 or brca2 mutation carriers with cancer risk and mortality. *JAMA : the journal of the American Medical Association 304*, 967–75.

Draper, N. R. and H. Smith (1998). *Applied regression analysis*, Volume 326. John Wiley & Sons.

Fan, J. and R. Li (2001, 02). Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc 96*, 1348–1360.

Fan, J. and R. Li (2006, 03). Statistical challenges with high dimensionality: Feature selection in knowledge discovery. *Proc. Madrid Int. Congress of Mathematicians 3*, 595–622.

Fan, J., Y. Liao, and M. Mincheva (2013, 09). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society. Series B, Statistical methodology 75*.

Fan, J. and J. Lv (2009). A selective overview of variable selection in high dimensional feature space. *Stat. Sinica 20 1*, 101–148.

Faraway, J. J. (2002). *Practical regression and ANOVA using R*. University of Bath.

Fisher, T. J. (2012). On testing for an identity covariance matrix when the dimensionality equals or exceeds the sample size. *Journal of Statistical Planning and Inference 142*(1), 312–326.

Foster, J., J. Taylor, and S. Ruberg (2011, 10). Subgroup identification from randomized clinical trial data. *Statistics in medicine 30*, 2867–80.

Fourati, S., S. Ribeiro, F. Blasco Lopes, A. Talla, F. Lefebvre, M. Cameron, J. Kaewkungwal, P. Pitisuttithum, S. Nitayaphan, S. Rerks-Ngarm, J. Kim, R. Thomas, P. Gilbert, G. Tomaras, R. Koup, N. Michael, M. McElrath, R. Gottardo, and R. Sékaly (2019, 02). Integrated systems approach defines the antiviral pathways conferring protection by the rv144 hiv vaccine. *Nature Communications 10*.

Friedman, J. H., T. Hastie, and R. Tibshirani (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software 33*(1), 1–22.

Gentleman, R., V. Carey, W. Huber, R. Irizarry, and S. Dudoit (2005). *Bioinformatics and Computational Biology Solutions Using R and Bioconductor (Statistics for Biology and Health)*. Berlin, Heidelberg: Springer-Verlag.

Ginsburg, G. and K. Phillips (2018, 05). Precision medicine: From science to value. *Health Affairs 37*, 694–701.

Glaab, E., J. Bacardit, J. M. Garibaldi, and N. Krasnogor (2012, July). Using rule-based machine learning for candidate disease gene prioritization and sample classification of cancer gene expression data. *PLoS ONE 7*(7), e39932.

Gu, X., G. Yin, and J. J. Lee (2013). Bayesian two-step lasso strategy for biomarker selection in personalized medicine development for time-to-event endpoints. *Contemporary clinical trials 36*(2), 642–650.

Gunter, L., J. Zhu, and S. Murphy (2011). Variable selection for qualitative interactions in personalized medicine while controlling the family-wise error rate. *Journal of biopharmaceutical statistics 21*(6), 1063–1078.

Gustafsson, M., M. Edström, D. Gawel, C. E. Nestor, H. Wang, H. Zhang, F. Barrenäs, J. Tojo, I. Kockum, T. Olsson, et al. (2014). Integrated genomic and prospective clinical studies show the importance of modular pleiotropy for disease susceptibility, diagnosis and treatment. *Genome medicine 6*(2), 1–12.

Hastie, T., R. Tibshirani, and M. Wainwright (2015, 05). *Statistical Learning with Sparsity: The Lasso and Generalizations*. Boca Raton: CRC Press.

Heinze, G., C. Wallisch, and D. Dunkler (2018, 01). Variable selection - a review and recommendations for the practicing statistician. *Biometrical J. 60*(3), 1–19.

Hochberg, Y. (1988). A sharper bonferroni procedure for multiple tests of significance. *Biometrika 75*(4), 800–802.

Hocking, R. R. (1976). A biometrics invited paper. the analysis and selection of variables in linear regression. *Biometrics 32*(1), 1–49.

Holm, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian journal of statistics*, 65–70.

Hyodo, M., N. Shutoh, T. Nishiyama, and T. Pavlenko (2015). Testing block-diagonal covariance structure for high-dimensional data. *Statistica Neerlandica 69*(4), 460–482.

Ingram, D. D. and J. C. Kleinman (1989). Empirical comparisons of proportional hazards and logistic regression models. *Statistics in medicine 8*(5), 525–538.

Jardillier, R., F. Chatelain, and L. Guyon (2018, 11). Bioinformatics methods to select prognostic biomarker genes from large scale datasets: A review. *Biotechnology Journal 13*.

Jia, J. and K. Rohe (2015). Preconditioning the lasso for sign consistency. *Electron. J. Statist. 9*(1), 1150–1172.

Jia, J. and B. Yu (2010). On model selection consistency of the elastic net when $p > n$. *Statistica Sinica*, 595–611.

Kalia, M. (2015). Biomarkers for personalized oncology: recent advances and future challenges. *Metabolism 64*(3), S16 – S21. Biomarkers: Current Status and Future Trends.

Kira, K. and L. A. Rendell (1992). A practical approach to feature selection. In *Machine learning proceedings 1992*, pp. 249–256. Elsevier.

Kononenko, I., E. Šimec, and M. Robnik-Šikonja (1997). Overcoming the myopia of inductive learning algorithms with relieff. *Applied Intelligence 7*(1), 39–55.

Krassowski, M., V. Das, S. Sahu, and B. Misra (2020, 12). State of the field in multi-omics research: From computational needs to data mining and sharing. *Frontiers in Genetics*.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Ledoit, O. and M. Wolf (2020, 06). The Power of (Non-)Linear Shrinking: A Review and Guide to Covariance Matrix Estimation. *Journal of Financial Econometrics*.

Lee, S. and D. Lee (2020, 12). What is the proper way to apply the multiple comparison test? *Korean Journal of Anesthesiology 73*, 572–572.

Li, Z. and M. Sillanpää (2012, 05). Overview of lasso-related penalized regression methods for quantitative trait mapping and genomic selection. *TAG. Theoretical and applied genetics. Theoretische und angewandte Genetik 125*, 419–35.

Lin, H., J. Jia, L. Nie, G. Shen, and T.-S. Chua (2016). What does social media say about your stress?. In *IJCAI*, pp. 3775–3781.

Lipkovich, I. and A. Dmitrienko (2014, 01). Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. *Journal of biopharmaceutical statistics 24*, 130–53.

Lipkovich, I., A. Dmitrienko, and R. B. D'Agostino Sr. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine 36*(1), 136–196.

Lipkovich, I., A. Dmitrienko, J. Denne, and G. Enas (2011, 07). Subgroup identification based on differential effect search (sides) – a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in medicine 30*, 2601–21.

Liu, X.-Y., S.-B. Wu, W.-Q. Zeng, Z.-J. Yuan, and H.-B. Xu (2020). Logsum+ l2 penalized logistic regression model for biomarker selection and cancer classification. *Scientific reports 10*(1), 1–16.

Loscalzo, J., I. Kohane, and A.-L. Barabasi (2007). Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular systems biology 3*(1), 124.

Lounici, K. (2008). Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics 2*(none), 90 – 102.

Lu, S., L. Dong, C. Fang, S. Liu, L. Kong, Q. Cheng, L. Chen, T. Su, H. Nan, D. Zhang, et al. (2020). Stepwise selection on homeologous prr genes controlling flowering and maturity during soybean domestication. *Nature Genetics 52*(4), 428–436.

Ma, S. and J. Huang (2008). Penalized feature selection and classification in bioinformatics. *Briefings in bioinformatics 9*(5), 392–403.

Martin, M. and S. López-Tarruella (2016, 05). Emerging therapeutic options for her2-positive breast cancer. *American Society of Clinical Oncology Educational Book*, e64–e70.

McDonald, J. (2009, 01). *Handbook of Biological Statistics 2nd Edition*. Sparky House Publishing Baltimore.

Meinshausen, N. and P. Bühlmann (2006, 09). High dimensional graphs and variable selection with the lasso. *Ann. Stat. 34*(3), 1436–1462.

Meinshausen, N. and B. Yu (2009). Lasso-type recovery of sparse representations for high-dimensional data. *The Annals of Statistics 37*(1), 246 – 270.

Menard, S. (2002). *Applied logistic regression analysis*, Volume 106. Sage.

Menyhárt, O. and B. Győrffy (2021). Multi-omics approaches in cancer research with applications in tumor subtyping, prognosis, and diagnosis. *Computational and Structural Biotechnology Journal 19*, 949–960.

Michalopoulos, I., G. Pavlopoulos, A. Malatras, A. Karelas, M. Kostadima, R. Schneider, and S. Kossida (2012, 06). Human gene correlation analysis (hgca): A tool for the identification of transcriptionally co-expressed genes. *BMC research notes 5*, 265.

Mishina, E. (2020, 11). *FDA-NIH Biomarker Working Group. BEST (Biomarkers, EndpointS, and other Tools) Resource [Internet]. Monitoring Biomarker. Silver Spring (MD): Food and Drug Administration (US); 2016-. Available from: https://www.ncbi.nlm.nih.gov/books/NBK326791/ Co-published by National Institutes of Health (US), Bethesda (MD).*

Ogutu, J., T. Schulz-Streeck, and H.-P. Piepho (2012, 05). Genomic selection using regularized linear regression models: Ridge regression, lasso, elastic net and their extensions. *BMC proceedings 6 Suppl 2*, S10.

Park, M. Y. and T. Hastie (2007). L1-regularization path algorithm for generalized linear models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 69*(4), 659–677.

Perrot-Dockès, M., C. Lévy-Leduc, and L. Rajjou (2020). Estimation of large block structured covariance matrices: Application to "multi-omic" approaches to study seed quality. arXiv:1806.10093.

Quackenbush, J. (2006). Microarray analysis and tumor classification. *New England Journal of Medicine 354*(23), 2463–2472.

Ramaswamy, S., P. Tamayo, R. Rifkin, S. Mukherjee, C.-H. Yeang, M. Angelo, C. Ladd, M. Reich, E. Latulippe, J. P. Mesirov, et al. (2001). Multiclass cancer diagnosis using tumor gene expression signatures. *Proceedings of the National Academy of Sciences 98*(26), 15149–15154.

Ravikumar, P., M. J. Wainwright, and J. D. Lafferty (2010). High-dimensional Ising model selection using l1-regularized logistic regression. *The Annals of Statistics 38*(3), 1287 – 1319.

Rerks-Ngarm, S., P. Pitisuttithum, S. Nitayaphan, J. Kaewkungwal, J. Chiu, R. Paris, N. Premsri, C. Namwat, M. De Souza, M. Benenson, S. Gurunathan, J. Tartaglia, J. McNeil, D. Francis, D. Stablein, D. Birx, S. Chunsuttiwat, C. Khamboonruang, and J. Kim (2009, 11). Vaccination with alvac and aidsvax to prevent hiv-1 infection in thailand. *The New England journal of medicine 361*, 2209–20.

Rimawi, M., R. Schiff, and C. Osborne (2015, 01). Targeting her2 for the treatment of breast cancer. *Annual review of medicine 66*, 111–28.

Royston, P. and M. Parmar (2011, 08). The use of restricted mean survival time to estimate the treatment effect in randomized clinical trials when the proportional hazards assumption is in doubt. *Statistics in medicine 30*, 2409–21.

Royston, P. and M. Parmar (2013, 12). Restricted mean survival time: An alternative to the hazard ratio for the design and analysis of randomized trials with a time-to-event outcome. *BMC medical research methodology 13*, 152.

Rutgers, E., L. van 't Veer, C. Poncet, J. Cardozo, S. Delaloge, J.-Y. Pierga, P. Vuylsteke, E. Brain, G. Viale, S. Kümmel, I. Rubio, Z. Gabriele, A. Thompson, K. Zaman, S. Knox, F. Hilbers, A. Peric, B. Meulemans, M. Piccart, and F. Cardoso (2020, 10). Updated results of the mindact trial: 70-gene signature to guide de-escalation of chemotherapy in early breast cancer. *European Journal of Cancer 138*, S14–S15.

Saeys, Y., I. Inza, and P. Larranaga (2007). A review of feature selection techniques in bioinformatics. *Bioinformatics 23*(19), 2507–2517.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *Annals of statistics 6*(2), 461–464.

Sechidis, K., K. Papangelou, P. D. Metcalfe, D. Svensson, J. Weatherall, and G. Brown (2018, 05). Distinguishing prognostic and predictive biomarkers: an information theoretic approach. *Bioinformatics 34*(19), 3365–3376.

Sermpinis, G., S. Tsoukas, and P. Zhang (2018). Modelling market implied ratings using lasso variable selection techniques. *Journal of Empirical Finance 48*, 19–35.

Sherman, B. T., M. Hao, J. Qiu, X. Jiao, M. W. Baseler, H. C. Lane, T. Imamichi, and W. Chang (2022). David: a web server for functional enrichment analysis and functional annotation of gene lists (2021 update). *Nucleic Acids Res*, gkac194.

Shipp, M., K. Ross, P. Tamayo, A. Weng, J. Kutok, T. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G. Pinkus, T. Ray, M. Koval, K. Last, A. Norton, T. Lister, J. Mesirov, D. Neuberg, E. Lander, J. Aster, and T. Golub (2002, 02). Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. *Nature medicine 8*, 68–74.

Singh, D., P. Febbo, K. Ross, D. Jackson, J. Manola, C. Ladd, P. Tamayo, A. Renshaw, A. D'Amico, J. Richie, E. Lander, M. Loda, P. Kantoff, T. Golub, and W. Sellers (2002, 04). Gene expression correlates of clinical prostate cancer behavior. *Cancer cell 1*, 203–9.

Smith, G. (2018, 09). Step away from stepwise. *J. Big Data 5*(32), 1–12.

Sotiriou, C. and L. Pusztai (2009, 03). Gene-expression signatures in breast cancer. *The New England journal of medicine 360*, 790–800.

Sotiriou, C., P. Wirapati, S. Loi, A. Harris, S. Fox, J. Smeds, H. Nordgren, P. Farmer, V. Praz, B. Haibe-Kains, C. Desmedt, D. Larsimont, F. Cardoso, H. Peterse, D. Nuyten, M. Buyse, M. J. Van de Vijver, J. Bergh, M. Piccart, and M. Delorenzi (2006, 02). Gene Expression Profiling in Breast Cancer: Understanding the Molecular Basis of Histologic Grade To Improve Prognosis. *JNCI: Journal of the National Cancer Institute 98*(4), 262–272.

Sprundel, T., M. Schmidt, M. Rookus, R. Brohet, C. Asperen, E. Rutgers, L. van 't Veer, and R. Tollenaar (2005, 09). van sprundel tc, schmidt mk, rookus ma, brohet r, van asperen cj, rutgers ej, van't veer lj, tollenaar rarisk reduction of contralateral breast cancer and survival after contralateral prophylactic mastectomy in brca1 or brca2 mutation carriers. br j cancer 93(3): 287-292. *British journal of cancer 93*, 287–92.

Sung, J., Y. Wang, S. Chandrasekaran, D. M. Witten, and N. D. Price (2012). Molecular signatures from omics data: from chaos to consensus. *Biotechnology journal 7*(8), 946–957.

Tian, L., A. Alizadeh, A. Gentles, and R. Tibshirani (2012, 12). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association 109*.

Tibshirani, R. (1996a). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Stat. Methodol.) 58*(1), 267–288.

Tibshirani, R. (1996b). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B (Stat. Methodol.) 58*(1), 267–288.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences 99*(10), 6567–6572.

Tibshirani, R. J. and J. Taylor (2011, 06). The solution path of the generalized lasso. *Ann. Statist. 39*(3), 1335–1371.

Tomeny, T. S., C. J. Vargo, and S. El-Toukhy (2017). Geographic and demographic correlates of autism-related anti-vaccine beliefs on twitter, 2009-15. *Social science & medicine 191*, 168–175.

Tsimberidou, A., E. Fountzilas, M. Nikanjam, and R. Kurzrock (2020, 03). Review of precision cancer medicine: Evolution of the treatment paradigm. *Cancer Treatment Reviews 86*, 102019.

Tufekci, Z. (2014, 03). Big questions for social media big data: Representativeness, validity and other methodological pitfalls. *Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014*.

Uniejewski, B., G. Marcjasz, and R. Weron (2019). Understanding intraday electricity markets: Variable selection and very short-term price forecasting using lasso. *International Journal of Forecasting 35*(4), 1533–1547.

Urbanowicz, R. J., M. Meeker, W. La Cava, R. S. Olson, and J. H. Moore (2018). Relief-based feature selection: Introduction and review. *Journal of biomedical informatics 85*, 189–203.

Utgoff, P. (1989, 11). Incremental induction of decision trees. *Machine Learning 4*, 161–186.

Vargas, A. and C. Harris (2016, 07). Biomarker development in the precision medicine era: lung cancer as a case study. *Nature reviews. Cancer 16*, 525–537.

Vinga, S. (2021). Structured sparsity regularization for analyzing high-dimensional omics data. *Briefings in Bioinformatics 22*(1), 77–87.

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using l1-constrained quadratic programming (lasso). *IEEE transactions on information theory 55*(5), 2183–2202.

Walker, S. and D. Duncan (1967, 07). Estimation of the probability of an event as a function of several independent variables. *Biometrika 54*, 167–79.

Wang, H., B. Lengerich, B. Aragam, and E. Xing (2019, 09). Precision lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics 35*(7), 1181–1187.

Wang, X. and C. Leng (2016). High dimensional ordinary least squares projection for screening variables. *J. R. Stat. Soc. Ser. B (Stat. Methodol.) 78*(3), 589–611.

Wang, Z. and X. Xu (2021). Testing high dimensional covariance matrices via posterior bayes factor. *Journal of Multivariate Analysis 181*, 104674.

Watanabe, H., M. Okada, Y. Kaji, M. Satouchi, Y. Sato, Y. Yamabe, H. Onaya, M. Endo, M. Sone, and Y. Arai (2009, 12). New response evaluation criteria in solid tumours revised recist guideline (version 1.1). *Gan to kagaku ryoho. Cancer & chemotherapy 36*, 2495–501.

Welch, W. J. (1982). Algorithmic complexity: three np-hard problems in computational statistics. *Journal of Statistical Computation and Simulation 15*(1), 17–25.

Windeler, J. (2000, 03). Prognosis - what does the clinician associate with this notion? *Statistics in medicine 19*, 425–30.

Wu, B. (2006). Differential gene expression detection and sample classification using penalized linear regression models. *Bioinformatics 22*(4), 472–476.

Wu, J. R., Y. Zhao, X. P. Zhou, and X. Qin (2020). Estrogen receptor 1 and progesterone receptor are distinct biomarkers and prognostic factors in estrogen receptor-positive breast cancer: Evidence from a bioinformatic analysis. *Biomedicine & Pharmacotherapy 121*, 109647.

Wu, T. T., Y. F. Chen, T. Hastie, E. Sobel, and K. Lange (2009, 01). Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics 25*(6), 714–721.

Xiong, M., W. Li, J. Zhao, L. Jin, and E. Boerwinkle (2001). Feature (gene) selection in gene expression-based tumor classification. *Molecular Genetics and Metabolism 73*(3), 239–247.

Xue, F. and A. Qu (2017). Variable selection for highly correlated predictors. *arXiv preprint arXiv:1709.04840*.

Yamada, R., D. Okada, J. Wang, T. Basak, and S. Koyama (2020, 05). Interpretation of omics data analyses. *Journal of Human Genetics 66*, 1–10.

Yu, L. and H. Liu (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the 20th international conference on machine learning (ICML-03)*, pp. 856–863.

Yuan, M. and Y. Lin (2006, 02). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B 68*, 49–67.

Yuan, M. and Y. Lin (2007, 04). On the nonnegative garrote estimator. *Journal of the Royal Statistical Society Series B 69*, 143–161.

Zhao, P. and B. Yu (2006, 12). On model selection consistency of lasso. *J. Machine Learn. Res. 7*, 2541–2563.

Zhu, J. and T. Hastie (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics 5*(3), 427–443.

Zhu, W., C. Lévy-Leduc, and N. Ternès (2021). A variable selection approach for highly correlated predictors in high-dimensional genomic data. *Bioinformatics 37*(16), 2238–2244.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association 101*(476), 1418–1429.

Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology) 67*(2), 301–320.