A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles associated with outcome of a non-small-cell lung cancer treatment

Bérangère Bastien, Hafid Chakir, Anne Gégout-Petit, Aurélie Muller-Gueudin, Yaojie Shi

A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles associated with outcome of a non-small-cell lung cancer treatment.

Bérangère Bastien[a], Hafid Chakir[b], Anne Gégout-Petit[b], Aurélie Muller-Gueudin[b], Yaojie Shi[b]

[a] *Transgene S.A.*
Boulevard Gonthier d'Andernach, Parc d'innovation, CS80166,
67405 Illkirch-Graffenstaden Cedex - FRANCE

[b] *Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France,*
anne.gegout-petit@univ-lorraine.fr, aurelie.gueudin@univ-lorraine.fr,tel: +33 3 72 74 53 78 , fax: +33 3 72 74 53 90

Abstract

We propose a new methodology to select and rank covariates associated to a variable of interest in a context of high-dimensional data under dependence but few observations. The methodology imbricates successively clustering of covariates, decorrelation of covariates using Factor Latent Analysis, selection using aggregation of adapted methods and finally ranking. Simulations study shows the interest of the decorrelation inside the different clusters of covariates. The objective of our method is to determine profiles of patients linked with the outcome of a treatment. We apply our method on transcriptomic data of $n = 37$ patients with advanced non-small-cell lung cancer, who have received chemotherapy. The survival time of these patients being known, we apply our method to select the covariates that are the most linked with the outcome treatment among a set of more than 50 000 transcriptomic covariates. We obtain different transcriptomic profiles for the patients whose survival time was short, versus the other patients with longer survival time.

Keywords: Variable selection, genetic profiles, High dimension, Multiple testing procedures, Aggregated methods, Correlated covariates selection,

Ranking

2010 MSC: 62H15, 62H20, 62P10, 62F40

1. Introduction

The purpose of personalised medicine is to select appropriate and optimal therapies based on the context of a patientâĂŹs genetic content or other molecular or cellular analysis. For this, the first step is to select among a set of more than tens of thousands covariates the ones that are linked with the outcome of a given therapy. For instance, if we consider a transcriptomic dataset of patients with advanced non-small-cell lung cancer who have received a treatment, we want to select the covariates associated with the effect of this treatment. The survival time of these patients being known, the question is to find a relation between the treatment outcome (i.e. the survival time) and the transcriptomic profiles of the patients. We propose a methodology, that, firstly selects and ranks the transcriptomic covariates that are the most linked with the outcome treatment, and secondly, that visualises the profiles of the selected transcriptomic covariates, for all the patients of the study.

More generally, the problem to detect association between a variable of interest and many covariates has been tackled by many biologists and statisticians [1, 2, 3, 4, 5]. A common example, coming from biology, is testing which of p genes' expression levels given in a dataset $\mathbf{X}$ is linked significantly with a variable Y, which we will call the variable of interest. The variable of interest may be a binary variable like an outcome of treatment or it may be a quantitative variable such as a phenotype or physiological parameter. Sometimes, the aim of the biologist is not necessarily to detect exhaustively all the genes involved in his problem but to have a list of the most important of them in order to study their biological function. For this purpose, it is interesting to rank the genes according to the strength of their link with the variable of interest. We will use the gene expression example for concreteness, but our aim is to propose

a general methodology in a context of high dimensional data (the number p of covariates is in the order of thousands) while the total number n of samples could be small (for instance between 25 and 100).

In this context, the covariates are high dimensional and correlated. This correlation between covariates, in a high-dimensional context, has to be taken into account in the statistical analysis. Moreover, we are in a context of small sample size ($n \ll p$). Thus, robustness of the statistical analysis has to be quantified.

We cite here some statistical methods that have been developed to select covariates in high-dimensional contexts. The state of the art about the control of false discoveries in multiple testing procedures is very extensive. The famous correction proposed by Bonferroni [6] to control the Family Wise Error Rate (FWER) has been emulated and we can find a review about these methods in [7]. Alternative methods focused on the control of the False Discovery Rate (FDR) [8, 9] or of the local FDR [10] or the q-value [11, 12, 13]. For a review (in french) of the methods, see Bar-Hen *et al.* [3]. Regarding regression in the framework of high dimensional data ($n \ll p$), many methods are available. For exemple, PLS approach [14] is a kind of principal component regression. The lasso regression [15], which performs both variable selection and regularization in penalizing the sums of squares by the L_1-norm of the coefficients. This method has been derived for many kinds of problem like logistic-regression in the case of binary data or network inference [16, 17]. Another versatile tool to select covariates in different non parametric contexts is given by the random forests, with the concept of importance of covariates (see for instance Genuer *et al.*, [18]).

Another important characteristic of the data that has to be taken into account in the analysis of the association is the structure of covariance of the covariates. Most of the multiple testing corrections make the assumption of the independence between the covariates. However it is well-known that omics data for instance are correlated by clusters. In the context of multiple testing, it

has been shown, that covariance between the covariates could bias the uniform repartition of the p-values under the null hypothesis and also inflates the variance of the estimation of the FDR [19, 20]. In [18] it is also shown that even if the method of random forests is robust, importance of covariates calculated by random forests is perturbed by adding other correlated covariates. One of the ways to deal with dependance is to model it by latent factors; it is a way to reduce the information in supposing that the common information of the p covariates is given by $q \ll p$ latent factors as Friguet *et al.* in [19, 20]. More precisely, they propose a way to correct the data according to a regression link with the variable of interest Y in such a way that covariates are independent conditionally to Y. After this correction, they propose a multiple testing procedure based on the Benjamini-Hochberg method [8, 9]. This method of correction will be called FAMT correction (for Factor Analysis for Multiple Testing) in the sequel.

However, the framework of FAMT is to consider the data $\mathbf{X}$ as an only one block of correlated covariates and has to be adapted if $\mathbf{X}$ is structured in several independent clusters of correlated covariates. As we will see in Section 3, the FAMT does not give good results if it is applied directly on the whole set of data $\mathbf{X}$, without taking into account its decomposition in independent clusters. Then, we propose to identify the clusters of correlated covariates before performing FAMT correction on each of the clusters. The clustering of covariates as proposed by Chavent *et al.* in [21] is a good way to arrange covariates into homogeneous clusters, i.e., groups inside of which covariates are strongly related to each other.

Our purpose in this paper is to propose a method adapted to the selection (and ranking) of correlated quantitative covariates associated with a variable of interest. For this, we propose a methodology that takes into account (1) the structure of correlation by clusters of covariates; (2) the correlation inside each cluster of correlated covariates.

Our methodology is divided in two steps: a pretreatment of the covariates (step

1) and a procedure of selection of the pretreated covariates (step 2). The pretreatment consists of (step 1.1) detecting the independent clusters of covariates by using the clustering of covariates proposed by Chavent *et al.* [21], and (step 1.2) applying a "decorrelation" between the covariates inside each cluster using the analysis in factors proposed by Friguet *et al.* [19, 20, 22]. The method of Friguet *et al.* performs a decorrelation of the covariates and compute corrected covariates that are suitable for testing and/or regression.

After that pretreatment, we propose a procedure to select and rank the covariates, by combining different selection methods that take into account the nature of the outcome Y (qualitative or quantitative) and the high dimensional context (multiple testing procedures for the tests, penalised regression, ...). We define a score for each covariate, which is defined by the number of selections among all the selection methods involved in this step. This score can be used to classify the covariates like in [23].

The paper is organized as follows. In Section 2, we detail the model and explain the principle of the main steps of our methodology: the pretreatment of the covariates and the construction of a score of selection. Section 3 is dedicated to simulations studies in order to assess the interest of the proposed pretreatment on one hand and the good working of the whole selection strategy on the other hand. The simulations are performed in two different designs in the case where the variable of interest is binary. Section 4 is dedicated to real data analysis, the purpose is to select covariates that are linked with the outcome of a treatment. Section 5 gives some conclusions and perspectives. An appendix gives a simulation study in the case where the variable of interest is a quantitative continuous variable.

2. Methodology

2.1. Framework and model

We suppose that we have n i.i.d replications of $(Y, \mathbf{X})$ where Y is the variable of interest, and $\mathbf{X} = (X_1, X_2, \ldots, X_p)$ is the vector of covariates, taking its values in $\mathbb{R}^p$. We make the assumption that, conditionally to Y, the covariates are decomposed into K independent clusters:

$$\mathbf{X} = (X_1^{(1)}, \ldots, X_{p_1}^{(1)}, \ldots X_i^{(k)}, \ldots, X_{p_K}^{(K)}) = (\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(K)}),$$

where $p_1 + \ldots + p_K = p$.

More precisely, on one hand, we use the framework of Friguet *et al.* in [19] to model each of the K different parts of the vector of covariates. Inside each cluster $\mathbf{X}^{(k)}$, the common information between the p_k covariates is modeled by regression on q_k latent factors:

$$X_i^{(k)} = m_i^{(k)}(Y) + b_i^{(k)}\mathbf{Z}^{(k)} + \varepsilon_i^{(k)}, \qquad \text{for } i = 1, \ldots, p_k, \tag{1}$$

where $\mathbf{Z}^{(k)}$ is a random q_k-vector such that $\mathbb{E}(\mathbf{Z}^{(k)}\mathbf{Z}^{(k)'}) = I_{q_k}$, $b_i^{(k)}$ is a q_k-vector, and $\varepsilon^{(\mathbf{k})} = (\varepsilon_1^{(k)}, \ldots, , \varepsilon_{p_k}^{(k)})$ is a random centered p_k-vector with independent components, and independent of $\mathbf{Z}^{(k)}$. The common information contained in $\mathbf{X}^{(k)}$ is then concentrated in a small dimension space by q_k latent factors $\mathbf{Z}^{(k)}$. Under the model (1), the correlation between the components of $\mathbf{X}^{(k)}$, conditionally to Y, is given by:

$$\Sigma^{(k)} = B^{(k)}(B^{(k)})' + \Psi^{(k)} \tag{2}$$

where $\Sigma^{(k)}$ is the covariance matrix of the data $\mathbf{X}^{(k)}$, $\Psi^{(k)}$ is a diagonal $p_k \times p_k$ matrix (the covariance matrix of $\varepsilon^{(k)}$) and $B^{(k)}$ is a $p_k \times q_k$ matrix of factor loadings (in Equation (1), $b_i^{(k)}$ represents the ith row of $B^{(k)}$). In the above decomposition, the diagonal elements $\Psi_i^{(k)}$ are also referred to as the specific

variances of the responses $X_i^{(k)}$. Therefore, $B^{(k)}(B^{(k)})'$ appears as the shared variance in the common factor structure, and [19] define the common variance by

$$\mathrm{cv}_k = \frac{\mathrm{trace}(B^{(k)}(B^{(k)})')}{\mathrm{trace}(\Sigma^{(k)})}. \tag{3}$$

On the other hand, we suppose that the informations specific at each cluster (that is vectors $(\mathbf{Z}^{(k)}, \varepsilon^{(k)})_{1\leq k\leq K}$) are independent, then the covariance matrix conditional to Y of the whole vector of covariates has the form given by the Figure 1.

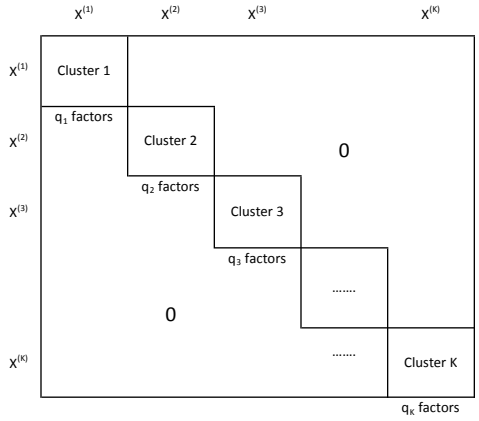

Figure 1: Conditional covariance structure of covariates

2.2. Main prodecure

The procedure is decomposed in a pretreatment of the covariates (step 1) and in a selection method of the covariates (step 2).

2.2.1. Step 1: pretreatment of data (clustering of covariates and decorrelation inside clusters)

The aim of this pretreatment is to perform a decorrelation of the covariates, to obtain corrected covariates that are suitable for testing and/or regression. Indeed, the correlation between covariates has an impact on all the classical

selection procedures: the conventional methods, namely the multiple testing procedures (p-value adjustment methods such as Bonferroni [6], Benjamini-Hochberg procedure [8, 9], q-value [11, 12, 13], or local FDR [2], [3]) are all built on the assumption that covariates are independent. As a results, they are no longer promising. A very detailed discussion can be found in the Friguet's thesis [24].

In estimating together the latent factors $\mathbf{Z}^{(k)}$ and the coefficients of regressions $(B^{(k)}, \Psi^{(k)})$ by an E.M. algorithm in model (1), Friguet *et al.* [19] can correct the covariates such that they are almost independent and as a result, suitable for multiple testing procedures or selection by regression or random forests. More precisely, the corrected data, noted $X_i^{(k)\star} = X_i^{(k)} - b_i^{(k)}\mathbf{Z}^{(k)} = m_i^{(k)}(Y) + \varepsilon_i^{(k)}$, $i = 1, \ldots, p_k$, lead to a standard multiple regression problem where the errors $\varepsilon_i^{(k)}$ are independent. Note that this correction of the data $\mathbf{X}$ is done conditionally on the variable of interest Y.

Of course, the whole vector $\mathbf{X}$ satisfies assumption of Equation (1), and [19] applies this decorrelation procedure on the whole set of covariates $\mathbf{X}$. But instead of applying Friguet's procedure on the whole set of covariates $\mathbf{X}$, we propose to first detect the different clusters $(\mathbf{X}^{(k)})_{1\leq k\leq K}$ and then to apply the decorrelation method on each cluster. Indeed, some simulation studies [25] have shown that the decorrelation was degraded by the dimension of the vector of covariates, whereas it was better after the detection of the independent clusters. By this way, the covariates selection procedure can be highly improved by clustering of covariates (step 1.1) before applying factor analysis to correct the correlation within each cluster (step 1.2), as it is shown in Section 3.

Step 1.1: clustering of covariates. We apply a clustering of covariates in the purpose to find clusters as we assumed in Section 2.1.

We propose to use the Chavent *et al.* [21] algorithm to cluster covariates into homogeneous clusters and thus to reveal structures. This algorithm maximizes an homogeneity criterion, where the homogeneity of a cluster is defined by the

sum of squared Pearson correlations between the covariates present in the cluster and the first principal component of this cluster. This algorithm is expected to roughly find the highly correlated clusters of covariates as we assumed in the Section 2.1. The procedure proposes also a method (based on bootstrap resampling) to find the number K of clusters if it is unknown.

Step 1.2: Factor analysis to correct dependent structure in each cluster. As already explained in the beginning of this section, clustering is followed by decorrelation inside each cluster using the Friguet's procedure.

At the end of this pretreatment procedure, we obtain corrected data, noted $\mathbf{X}^*$ in the sequel.

2.2.2. Step 2: Aggregation of statistical methods applied on the resulting dataset

The statistical methods proposed in this part are not fixed and can be adapted by the practitioner according to its preferred selection methods and the characteristics of the data (nature of variable of interest Y, samples' sizes and so on...).

The idea is the following: we choose several methods to select the pretreated covariates $\mathbf{X}^\star$. Suppose that you perform L methods, then for each covariate $X_j^\star$, you obtain a score $S_j \in \{0, 1, \ldots, L\}$ that is the number of selections among the L methods. By this way, you can rank the covariates according to their link with the outcome Y.

For instance, in the examples developed in our simulation studies and in real data, Y is binary, the size of the samples are low and we choose eight different methods of selection: five different multiple testing procedures applied to the Wilcoxon test (Bonferroni, Benjamin-Hochberg, q-values, local FDR, FAMT), logistic regression penalised by Lasso, and two selections by random forests (threshold step and interpret step, see [18]). The outcomes of this procedure are the scores $S_i, i = 1, \ldots, p$ which are integers included in $[0, 8]$. For example,

if $S_i = 8$, then the corresponding variable has been selected by all the eight methods, whereas if $S_i = 0$, the corresponding variable has been selected by none of them. The scores can be used to rank the covariates according to the strength of their link with the variable of interest.

In the sequel, we call our procedure `ARMADA` for AggRegated Methods for covAriates selection under Dependence.

3. Simulations

We first explain the two different simulation designs in Section 3.1. We then describe the effect of the pretreatment in Section 3.2 and finally, we study the selection procedure itself in Section 3.3.

3.1. Simulation designs

We propose a simulation study with $p = 1600$ covariates and sample size $n = 60$. We first describe how to create dependance in the covariates $\mathbf{X}$, then we present two simulation designs in a classification study. One design in a regression case is given in Appendix A.

The covariates $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4}$ are clustered into four clusters, which are independent conditionally to Y, each of them containing $p_k = 400$ covariates. For this, before to model the dependence with the outcome Y, we generate for each cluster k, a preliminary vector $\tilde{\mathbf{X}}^{(k)}$ that is a gaussian 400-vector, with mean 0 and non-diagonal variance-covariance matrix $\Sigma^{(k)}$. The correlation between the covariates $\tilde{X}_j^{(k)}$ inside the cluster k is designed by a factor analysis model described in Equation (2). More precisions on the simulation procedure of data with covariance design defined by (2) can be found in Friguet's thesis [24]. We simulate data with common variances $\mathrm{cv}^{(k)}$ equal to 0.8 in each cluster (recall that the common variance is defined in Equation (3)). Moreover, the numbers of latent factors in each cluster are $(q^{(1)}, \dots, q^{(4)}) = (4, 6, 8, 10)$.

Now, we create the dependence with outcome Y in perturbing some component of $\tilde{\mathbf{X}}$. The two following simulation designs consider an equiprobable two-class problem, $Y \in \{0, 1\}$ (i.e. $Y = 1$ for $\frac{n}{2}$ subjects, and $Y = 0$ for $\frac{n}{2}$ subjects). In the two designs, there are either 160 (for design 1) or 240 (for design 2) influential covariates, whose links with the response variable Y have different intensities. More precisely, in both cases, the response variable Y is the most strongly linked with the 10 first covariates of each cluster, and the strength of the link is decreasing in the successive clusters of 10 influential covariates. The link between the influential covariates and Y is described for each design in the two following sections.

3.1.1. Design 1

This simulation design is inspired from the simulation design of Friguet *et al.* [19]. Y is linked with 160 influential covariates in $\mathbf{X}$, the others being noise covariates. More precisely,

- for the $m_1 = 40$ first covariates of each cluster, we had dependence with Y to $\tilde{X}_j^{(k)}$ by setting $X_j^{(k)} = \tilde{X}_j^{(k)} + \delta_j \mathbf{1}_{Y=0}$ where:
  - $\delta_j = 1.5$ for $j = 1, \ldots, 10$,
  - $\delta_j = 1$ for $j = 11, \ldots, 20$,
  - $\delta_j = 0.75$ for $j = 21, \ldots, 30$,
  - $\delta_j = 0.5$ for $j = 31, \ldots, 40$.
- $X_j^{(k)} = \tilde{X}_j^{(k)}$ for the $m_0 = 360$ remaining covariates of each cluster, such that they are independent of Y.

3.1.2. Design 2

This simulation design is inspired from the toys-data of Genuer *et al.* [18]. Y is linked with 240 influential covariates in $\mathbf{X}$, the others being noise covariates.

Let us define the simulation model by giving the conditional distribution of X_i for $Y = y$:

- For the $m_1 = 60$ first covariates of each cluster, $X_j^{(k)} = \tilde{X}_j^{(k)} + \delta_j$ where δ_j is a random variable modelled according to the value j:
  - for $j = 1, \ldots, 10$, with probability 0.7, $\delta_j \sim \mathcal{N}(3y, 1)$, and with probability 0.3, $\delta_j \sim \mathcal{N}(0, 1)$;
  - for $j = 11, \ldots, 20$, with probability 0.7, $\delta_j \sim \mathcal{N}(2y, 1)$, and with probability 0.3, $\delta_j \sim \mathcal{N}(0, 1)$;
  - for $j = 21, \ldots, 30$, with probability 0.7, $\delta_j \sim \mathcal{N}(y, 1)$, and with probability 0.3, $\delta_j \sim \mathcal{N}(0, 1)$;
  - for $j = 31, \ldots, 40$, with probability 0.3, $\delta_j \sim \mathcal{N}(3y, 1)$, and with probability 0.7, $\delta_j \sim \mathcal{N}(0, 1)$;
  - for $j = 41, \ldots, 50$, with probability 0.3, $\delta_j \sim \mathcal{N}(2y, 1)$, and with probability 0.7, $\delta_j \sim \mathcal{N}(0, 1)$;
  - for $j = 51, \ldots, 60$, with probability 0.3, $\delta_j \sim \mathcal{N}(y, 1)$, and with probability 0.7, $\delta_j \sim \mathcal{N}(0, 1)$.
- $X_j^{(k)} = \tilde{X}_j^{(k)}$ for the $m_0 = 340$ remaining covariates of each cluster, such that they are independent of Y.

We can remark that these two designs respect the conditional covariance matrix given in Figure 1. The design 1 is exactly in the scope of our model given by Equation (1). Design 2 differs a little bit from the model of Equation (1) because of the term of regression on Y. Note that in real data analysis, we don't know the model from which they are generated. It is why it is interesting to analyse the performance of our method on different kinds of simulated data.

3.2. Interest of our data pretreatment

In order to emphasize the interest of our data pretreatment, we compare the results of a Wilcoxon test after three different data pretreatments:

Procedure 1: nothing is done on the dataset $\mathbf{X}$.

Procedure 2: the covariates $\mathbf{X}$ are decorrelated with the factor analysis procedure `FAMT` [19, 22], taking Y into account, given a new dataset $\mathbf{X}^*_Y$.

Procedure 3: the 4 clusters are estimated with the procedure of Chavent *et al.* [21], implemented in the R package `ClustOfVar`; then the covariates are decorrelated in each cluster, taking Y into account, with the factor analysis procedure of Friguet *et al.* [19, 22], implemented in the R package `FAMT`. This gives a new dataset ${\mathbf{X}^\dagger}_Y$ obtained by the concatenation of the decorrelated clusters.

Remark: *our data pretreatment is the Procedure 3. We have supposed that the number of clusters is known. If that is not the case, the user can choose its own number of clusters by using the graphical tools of the* `ClustOfVar` *procedure (plots of the dendrogram).*

Our objective is to find out the differently expressed covariates in the two groups (groups $Y = 0$ and $Y = 1$) with sample sizes $\frac{n}{2} = 30$. For this, we perform Wilcoxon tests on each of the p pretreated covariates of the dataset (that is $\mathbf{X}$ for Procedure 1, $\mathbf{X}^*_Y$ for Procedure 2, ${\mathbf{X}^\dagger}_Y$ for Procedure 3), given a three sets of p p-values. For each of these procedures, the selected covariates are those with p-values lower than 0.05. We compare these procedures on $N = 100$ runs of $(\mathbf{X}, Y)$. For the comparison, we count the number of influential covariates that are correctly detected (this number is noted TP, for True Positive), this indicator gives an idea of the sensibility of the test after the procedure. To assess the specificity, we count the number of non-influential detected covariates (this number is noted FP, for False Positive). Note that the perfect method would detect all the influential covariates (that is 160 in design 1 and 240 in design 2) and no False Positive. However, according to the detection threshold chosen

for the p-value, the expected number of FP is 72 for design 1 and 68 for design 2. The results are shown in Figure 2 for the design 1 and in Figure 3 for the design 2.

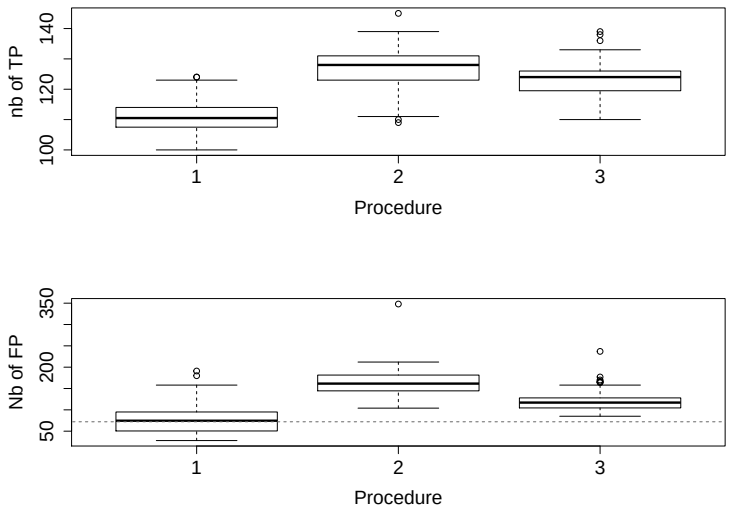

Figure 2: Number of true positive tests (top), false positive tests (bottom) in the design 1 according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 100$ runs.

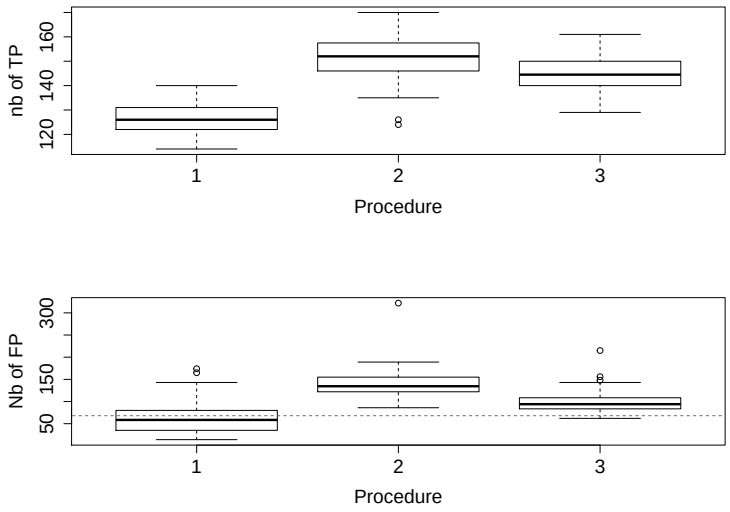

Figure 3: Number of true positive tests (top), false positive tests (bottom) in the design 2 according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 100$ runs.

If we analyse the results given by Figures 2 and 3, we can see that Procedure 1 is in fact the one that has the lowest rate of FP but its power is also the poorest whatever the design. Our Procedure is the one that reduces the mean and the variability of the distributions of the false positive rates. The power of our Procedure is comparable with Procedure 2 with a little bit better performance in design 1 and a little bit worse one in design 2. This results show the interest of our proposed pretreatment before performing selection.

3.3. Results of the whole method (pretreatment and selection)

In order to describe the performances of our method, we show in Figures 4 and 5 the mean `ARMADA` scores obtained on the $N = 100$ runs of $(\mathbf{X}, Y)$ for each design. The scores are given for all the covariates individually, and also by group of influential and noise covariates (the groups of influential covariates are noted by "1.5", "1", "0.75", "0.5" in the design 1, and by "(0.7,3)", "(0.7,2)",

"(0.7,1)", etc. in the design 2 (see Section 3.1); and the group of noise covariates is noted by "-").

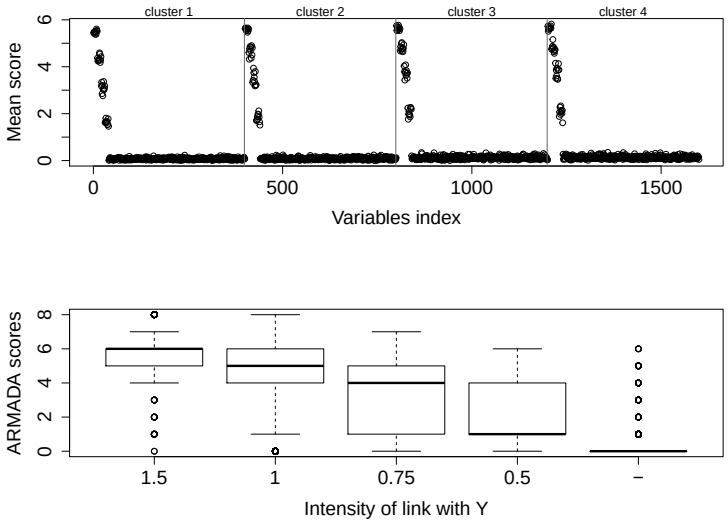

Figure 4: Top: mean of the `ARMADA` scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs. Simulation in the design 1.

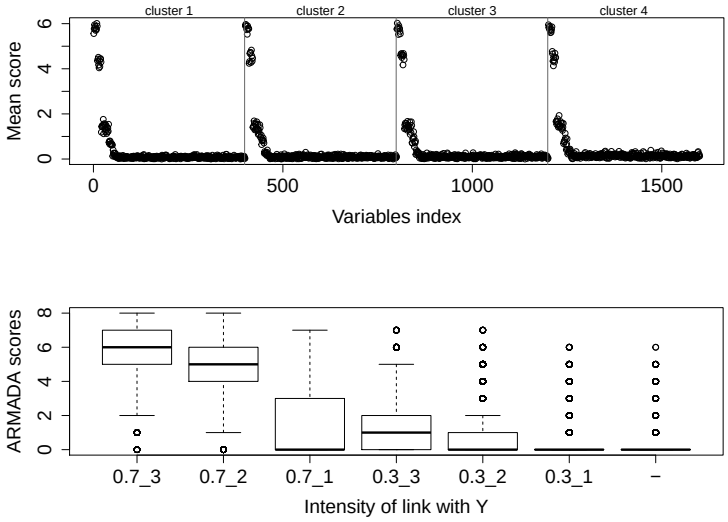

Figure 5: Top: mean of the `ARMADA` scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs. Simulation in the design 2.

We can see on the Figures 4 and 5 that the scores give a clear ranking of the covariates, according to the strength of their link with the response variable Y. The highest scores are obtained by the covariates which are the most strongly linked with the response variable Y. The `ARMADA` method is particularly performant in the design 1, the mean score clearly distinguishes the five groups of covariates according to their link with Y. The distribution of the individual scores inside each group is given by the boxplots. For design 1, scores clearly separate the influential covariates from the others; and inside the influential covariates the two first groups are clearly separated of the last one. The method is not so performant for the design 2 probably because we are note exactly in the model of the study (Equation (1)) but also because the strength of the link with Y is low excepted for the two first groups of covariates that have scores which are well separated from the others by the selection method. Whatever the design, we can precise that around 95% of the noise covariates obtained a `ARMADA`-score that was exactly 0.

	ARMADA	Wilcoxon	FAMT
1.5	0.99 (0.04)	0.99 (0.07)	0.99 (0.02)
1	0.97 (0.15)	0.85 (0.35)	0.95 (0.20)
0.75	0.91 (0.27)	0.62 (0.48)	0.82 (0.38)
0.5	0.79 (0.40)	0.33 (0.47)	0.52 (0.49)
-	0.05 (0.23)	0.05 (0.22)	0.10 (0.30)

Table 1: Results of the $N = 100$ runs in the design 1: rates of selection of the different groups of influential and noise covariates by the method ARMADA, the Wilcoxon test and the FAMT procedure. The corresponding standard deviations are given in brackets.

	ARMADA	Wilcoxon	FAMT
(0.7-3)	0.99 (0.08)	0.99 (0.07)	0.99 (0.04)
(0.7-2)	0.92 (0.27)	0.92 (0.26)	0.96 (0.17)
(0.7-1)	0.44 (0.49)	0.43 (0.49)	0.58 (0.49)
(0.3-3)	0.54 (0.49)	0.41 (0.49)	0.61 (0.48)
(0.3-2)	0.32 (0.46)	0.28 (0.45)	0.41 (0.49)
(0.3-1)	0.12 (0.32)	0.12 (0.33)	0.19 (0.39)
-	0.05 (0.23)	0.05 (0.22)	0.09 (0.29)

Table 2: Results of the $N = 100$ runs in the design 2: rates of selection of the different groups of influential and noise covariates by the method ARMADA, the Wilcoxon test and the FAMT procedure. The corresponding standard deviations are given in brackets.

3.4. Comparison with other selection methods

We propose the following selection criterion in our procedure: the selected covariates are those with scores greater or equal to 1.

We compare this selection procedure with two other selection methods:

- the Wilcoxon test: the selected covariates are those with raw-pvalues (i.e. p-values without any correction) lower than 0.05,
- the FAMT procedure [22]: the selected covariates are those with adjusted p-values lower than 0.05.

To compare the three selection methods, Tables 1 and 2 give the rates of selection for each group of influential covariates, and for the group of noise covariates.

- We can see that our method respect the expected rate of false positives that is not the case for the FAMT method which exhibits a greater rate

of 10 %.

- Moreover, our method gives the best results in the design 1. The rate of selection of the influential covariates is very good compared with the other methods even if the strength of the link is poor.

- In the design 2, our method is competitive with the `FAMT` procedure for the influential covariates, but again `FAMT` procedure has more false positives than ours and consequently more than expected.

Finally, we can conclude with the ROC curves given in Figure 6 that our method outperforms the two others selection methods (the ordinates of the points of the `ARMADA` ROC curve are all higher than the ordinates of the points of the two other ROC curves). Note that the ROC curves of the design 2 give the impression that our method is not competitive with the two others, but this is only caused by the fact that we have traced a solid line between the points (1-specificity, sensibility)$_{\texttt{ARMADA score=0}}$ and (1-specificity, sensibility)$_{\texttt{ARMADA score=1}}$. The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of $(\mathbf{X}, Y)$.

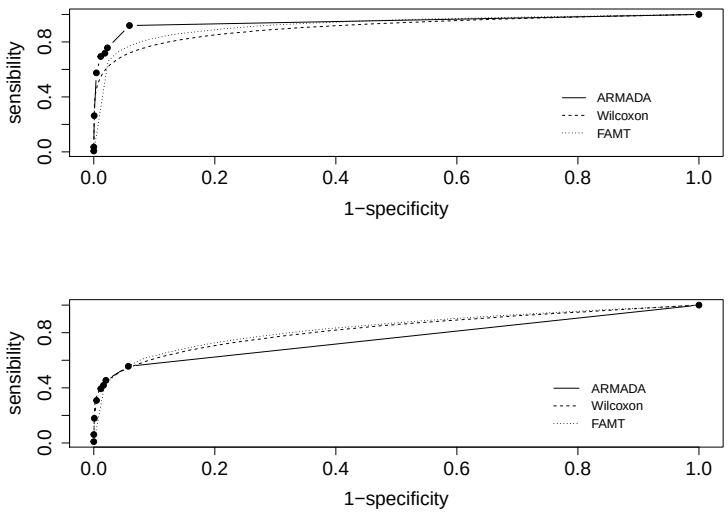

Figure 6: ROC curves for the three selection methods, in the case of design 1 (top), design 2 (bottom).

4. Application to real data

We apply our method on transcriptomic data of $n = 37$ patients with advanced non-small-cell lung cancer, who have received chemotherapy. Even if we are aware of the fact that chemotherapy is not a target therapy, the problematic is really to select suitable transcriptomic covariates in the purpose to detect profiles associated with the effect of a treatment. For each patient, we have 51 336 transcriptomic covariates, and its survival status: 24 patients whose death occurred before 12 months and 13 patients whose death occurred after 12 months, this criteria of death before one year is very common in clinical trials. We applied a first filtering of the covariates, where we decided to ignore the covariates for which the Wilcoxon test does not detect a difference between the 24 patients whose survival time is lower than 12 months and the 13 other ones (we eliminate covariates with Wilcoxon-pvalue greater than 0.05). After this filtering we obtained a dataset with $n = 37$ patients and $p = 6810$ covariates.

4.1. Classification study

In a first time, the biological question was to find the genes which can explain a survival time greater or lower than 12 months. We then consider a binary response variable: $Y = 1$ for the 24 patients whose survival time is lower than 12 months and $Y = 0$ for the 13 patients whose survival time is greater than 12 months. The results are shown in Table 3: 10 covariates are particularly important, with a score equal to 7, whereas 2827 covariates have a score equal to 0, and 3983 covariates have a score greater or equal to 1.

Score	0	1	2	3	4	5	6	7	8
Number of covariates	2827	553	460	596	1170	888	306	10	0

Table 3: Distribution of the covariates scores in the transcriptomic dataset: classification study.

It is clear that, the biologist will not focus on the 3983 covariates with a positive score. But the method clearly gives a hierarchy between the genes and it is sure that the the function of the 10 genes with a score at 7 has to be studied to understand its link with the "success" of the treatment.

4.2. Regression study

As the survival time was known for all the 37 patients, we also apply our method on the same dataset (6810 covariates) but here, Y is the survival time. We then have a regression problem. We have used eight selection methods in the Step 2 of our method: five different multiple testing procedures applied to the Pearson correlation test (Bonferroni, Benjamin-Hochberg, q-values, local FDR, FAMT), regression penalised by Lasso, and two selections by random forests (threshold step and interpret step, see [18]). The results are given in Table 4.

Score	0	1	2	3	4	5	6	7	8
Number of covariates	3988	89	456	509	984	692	86	5	1

Table 4: Repartition of the covariates scores in the transcriptomic dataset: regression study.

4.3. Comparison between regression and classification studies

The comparison between the classification and regression studies is given in Table 5. This table is a little disappointing, because regression and classification do not select the same covariates. Whatever, among the covariates with a C-score equal to 7, there is only one with a R-score lower than 4 (equal to 0 !). But this two analyses are not looking the same kind of link with the covariates. w Moreover, these two approaches give two tools to detect influential covariates. We can combine these two approaches and consider the covariates that are selected by at least one approach, or consider the covariates that are selected by both of them. In the Figure 7, we show the heatmap of the selected covariates which have a classification score **and** a regression score greater than some threshold.

The visualisation of the co-clustering of the selected genes and the survival leads to the distinction of three different groups of patients (noted P_1, P_2, P_3 in Figure 7) of respective sizes 7, 8, 22 from the left to the right of the x-axis. The co-clustering identifies also two clusters of genes (noted G_1 and G_2 for simplicity).

All the people except 2 of the two first group P_1 have a life status $Y = 1$ (among the two exceptions, one is at the threshold with a survival of 11.5 months), all of the people of the third group P_3 have a life status $Y = 0$. The selected covariates clearly separates groups P_1 and P_3 : the patients of the group P_1 have a low expression of the covariates in G_1 and a high expression of the covariates in G_2 and the inverse for group P_3. Patients of group P_2 have intermediate expressions according the two others groups.

	Classification score							
Regression score	0	1	2	3	4	5	6	7
0	2227	328	273	337	531	257	34	1
1	41	7	3	9	17	10	2	0
2	131	35	39	52	119	71	9	0
3	119	48	44	50	117	114	17	0
4	174	65	56	86	256	241	102	4
5	119	64	40	57	116	176	116	4
6	15	4	4	5	12	19	26	1
7	1	2	1	0	1	0	0	0
8	0	0	0	0	1	0	0	0

Table 5: Repartition of the covariates scores in the transcriptomic dataset. The R-scores are given in the 9 rows, the C-scores are given in the 8 columns. For example, 41 covariates have a R-score equal to 1, and a C-score equal to 0.

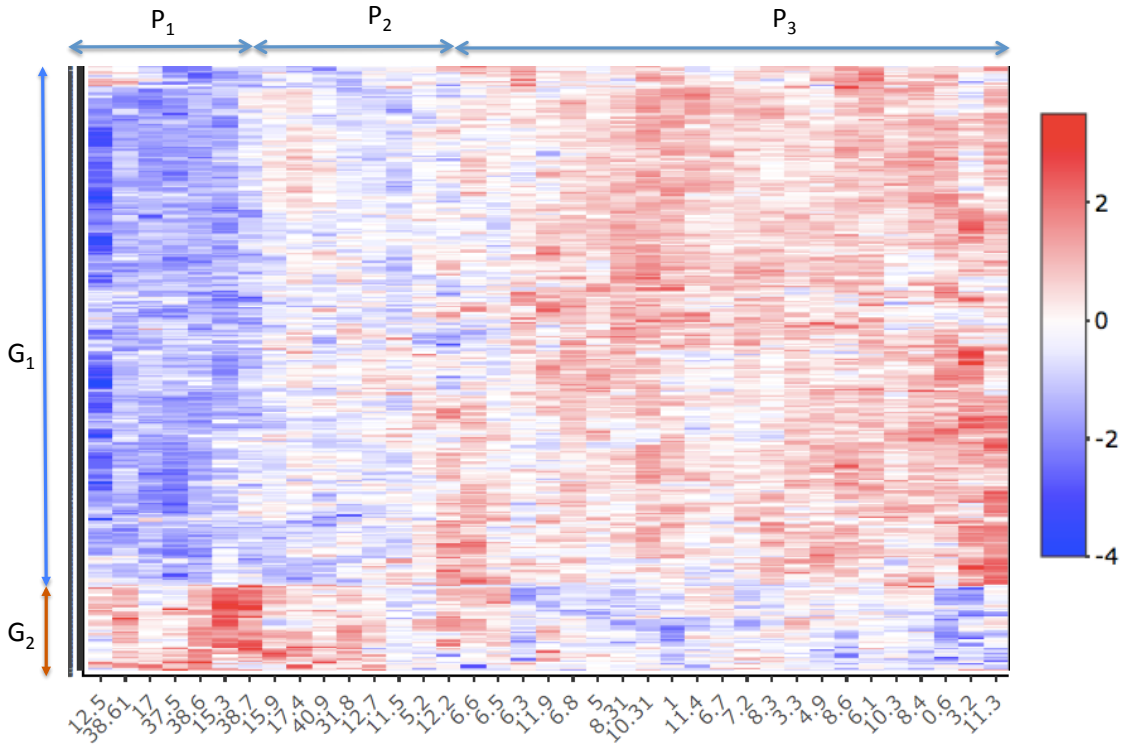

Figure 7: Heatmaps of the 342 covariates which have `ARMADA` scores greater or equal to 5 in **both** classification **and** regression studies. Each column corresponds to one patient. The x-axis represents the patients (marked with their survival time) and the y-axis represents the covariates. The heatmap has been obtained thanks to the R package `heatmaply` after co-clustering of the survival times (on the x-axis) and of the covariates (on the y-axis) with the function `hclust`.

As the number of patients $n = 37$ is small compared to the number of covariates even after filtering ($p = 6810$), we have checked our results with a bootstrap study. The results, for instance in the case of classification, are shown in Figure 8. We can see that the distributions of the bootstrap means of the C-scores have a quite small dispersion. If we consider the covariates which have a non null C-score, they all have a bootstrap mean C-score greater than 1. And if we consider the most important covariates (the 10 covariates that have a C-score equal to 7) : their corresponding bootstrap means of C-scores are all greater than 5.84 and lower than 7.

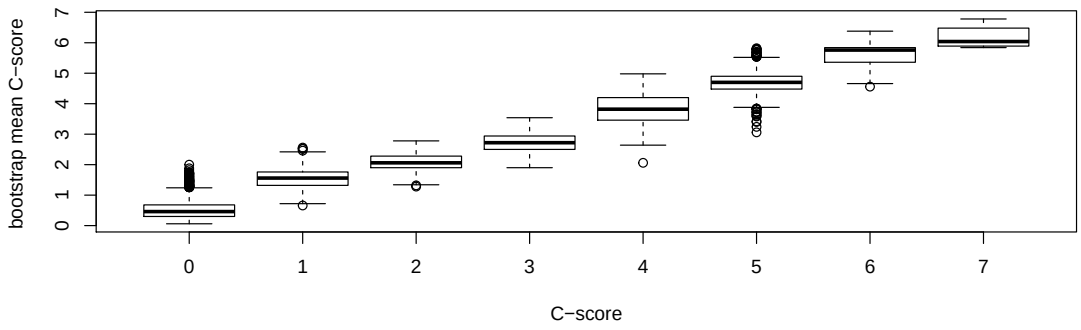

Figure 8: Distributions of the bootstrap mean of C-scores (i.e. means of C-scores obtained on $B = 50$ bootstrap samples), versus their respective C-score, for all the $p = 6810$ covariates.

5. Conclusion and perspectives

We have proposed a new methodology which is able to select the covariates (here the genes) that are linked with a variable of interest (here the treatment of an outcome). After this selection obtained with our method, it is then easy to visualise the selected genes for all the patients, and to classify the genetic profiles of patients with respect to their treatment outcome. In the case of the treatment by chemotherapy in the advanced non-small-cell lung cancer, we have identify three types of genetic profiles defined with two clusters of genes (noted

G_1 and G_2 for simplicity):

- the patients who have an high level of expression for the genes of group G_1 and a low level of expression for the genes of group G_2 have all a low survival time (lower than 1 year);
- on the contrary, the patients who have a low level of expression for the genes of group G_1 and an high level of expression for the genes of group G_2 have all a better survival time (greater than 1 year);
- between these two opposite groups, we have a transitional group of patients who have intermediate expressions according to the two others groups. Apart from two patients, the survival time in this group is greater than 1 year (among the two exceptions, one patients has a survival time of 11.5 months).

This kind of results is very promising for the development of the personalized medicine.

We are developing an R-package, called `ARMADA`, in order to propose our methods to the users who want to do covariates selection in high-dimensional datasets. The package proposes also a graphical representation of the selected covariates, through heatmaps, as we have presented in Figure 7.

Appendix A. Design 3: regression

In this section, we give results of simulations in the purpose to study the behaviour of our algorithm to select covariates linked with a continuous variable of interest (like survival time here). We simulate $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}^{(k)})_{k=1,\dots,4}$ as in Section 3.1 excepted that the number of latent factors $q^{(k)}$ are respectively $1, 4, 6$ and 8, and the common variances $\text{cv}^{(k)}$ are respectively equal to $0, 0.3, 0.5$ and 0.8 in each of the four clusters. We do not transform the covariate so that $\mathbf{X} = \tilde{\mathbf{X}}$ but the quantitative outcome Y is dependent of some covariates by a linear regression link. More precisely, Y is linked with 5 covariates of the three first

clusters, by the following equation:

$$\begin{aligned} Y = & 50X_1^{(1)} + 40X_2^{(1)} + 30X_3^{(1)} + 20X_4^{(1)} + 10X_5^{(1)} \\ & + 50X_1^{(2)} + 40X_2^{(2)} + 30X_3^{(2)} + 20X_4^{(2)} + 10X_5^{(2)} \\ & + 50X_1^{(3)} + 40X_2^{(3)} + 30X_3^{(3)} + 20X_4^{(3)} + 10X_5^{(3)} + \epsilon \end{aligned}$$

where ϵ is a standard gaussian distribution, independent of $\mathbf{X}$. Y is then linked with $5 \times 3 = 15$ covariates in $\mathbf{X}$. Note that, because of the dependence between the covariates, Y is indirectly linked with others covariates of the three first clusters.

Similarly as in Section 3.2, we study the interest of the pretreatment, using the three procedures detailed in 3.2. We see the effect on the Pearson correlation test (instead of the Wilcoxon test in Section 3.2). We produce $N = 100$ runs of $(\mathbf{X}, Y)$ and counted the number of false and true positive (shown in Figure A.10), as well as the `ARMADA` scores. As we can in Figure A.10), compared to the two other procedures, our pretreatment procedure gives the lowest rate of false positives, and with the lowest variability. The rate and variability of the true positives is quite similar with the three procedures.

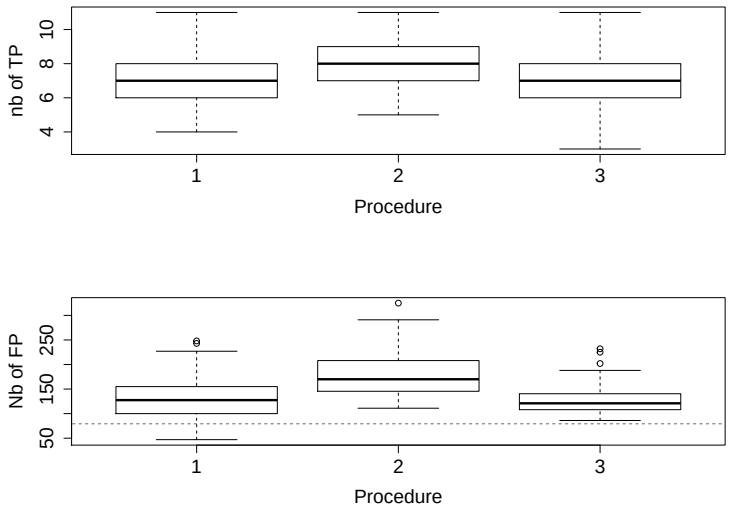

Figure A.9: Number of: true positive tests (top), false positive tests (bottom) in the design 3. Boxplots are calculated on $N = 100$ runs.

As in Section 3.3, the Figure A.10 shows the `ARMADA` scores obtained on these $N = 100$ runs of $(\mathbf{X}, Y)$. The scores give a ranking of the covariates, according to the intensity of their link with respect to the response variable Y. The highest scores are obtained by the covariates which are the most strongly linked with the response variable Y. The boxplots in the bottom of the Figure A.10 show that the median scores of the covariates of groups "50" and "40" are non nul. The mean scores visible in Figure A.10, are greater than 1 for the covariates of groups "50", "40" and "30". We can also precise that around 93% of the noise covariates obtained a score that was 0.

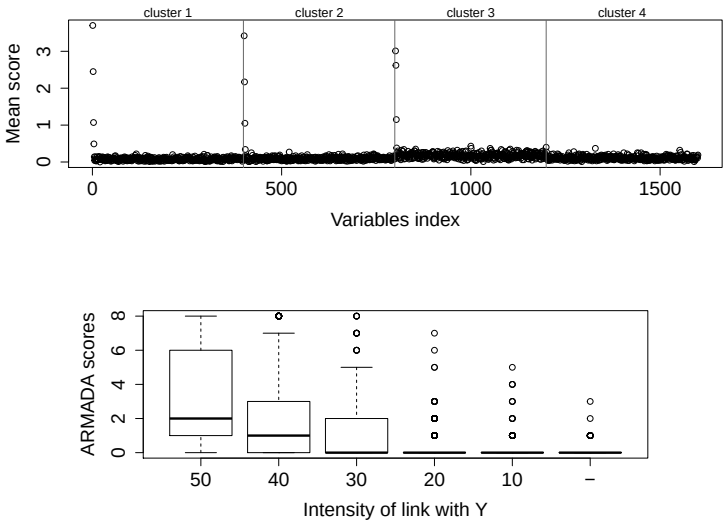

Figure A.10: Top: mean of the `ARMADA` scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs. Simulation in the design 3.

As in Section 3.4, the Table A.6 allows us to compare our method with the Pearson test and the FAMT procedure. We can see that

- all excepted the covariates of groups "20" and "10" in the design 3, have a mean `ARMADA` score greater than 1. Moreover the noise covariates have a mean `ARMADA` score lower than 1.
- only the most strongly linked covariates (those of group "50") have mean pvalues lower than 0.05 in the Pearson correlation test and in the `FAMT` procedure.
- we can see that our method obtains 7% of false positives in the regression case, whereas the two other methods have 8% and 11% of false positives.
- our method is competitive with the `FAMT` procedure for the selection of influential covariates.

As in design 2, the ROC curve of the design 3 gives the impression that our

	ARMADA	Pearson	FAMT
50	0.83 (0.37)	0.82 (0.38)	0.87 (0.33)
40	0.71 (0.45)	0.64 (0.48)	0.75 (0.43)
30	0.45 (0.50)	0.46 (0.50)	0.55 (0.50)
20	0.22 (0.41)	0.25 (0.43)	0.31 (0.46)
10	0.11 (0.32)	0.15 (0.36)	0.17 (0.38)
-	0.07 (0.26)	0.08 (0.27)	0.11 (0.31)

Table A.6: Results of the $N = 100$ runs in the design 3: rates of selection of the different groups of influential and noise covariates by the method ARMADA, the Pearson correlation test and the FAMT procedure. The corresponding standard deviations are given in brackets.

method is not competitive with the two others, but this is only an impression caused by the fact that we have traced a solid line between the points (1-specificity, sensibility)$_{\texttt{ARMADA score=0}}$ and (1-specificity, sensibility)$_{\texttt{ARMADA score=1}}$. The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of $(\mathbf{X}, Y)$. As in the classification study, the ordinates of the points of the ARMADA ROC curve are all higher than the ordinates of the points of the two other ROC curves.

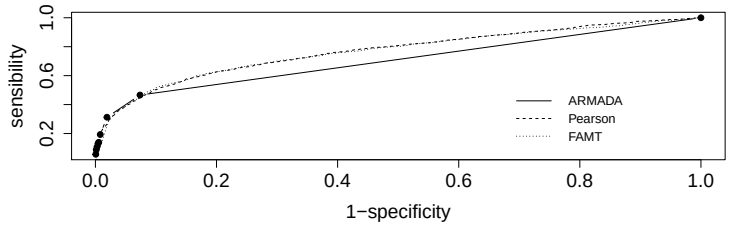

Figure A.11: ROC curves for the three selection methods, in the case of design 3.

Acknowledgements *The real data have been obtained by the Transgene team thanks to Bérangère Bastien. The statistical methodology has been developed by Anne Gégout-Petit and Aurélie Muller-Gueudin, and thanks to Yaojie Shi and Hafid Chakir during their Master internships.*

References

[1] S. Dudoit, Y. H. Yang, M. J. Callow, T. P. Speed, Statistical methods for identifying differentially expressed genes in replicated cdna microarray experiments, Statistica sinica (2002) 111–139.

[2] J. Aubert, A. Bar-Hen, J.-J. Daudin, S. Robin, Determination of the differentially expressed genes in microarray experiments using local fdr, BMC bioinformatics 5 (1) (2004) 1.

[3] A. Bar-Hen, J.-J. Daudin, S. Robin, Comparaisons multiples pour les microarrays, Journal de la Société Française de Statistique 146 (1-2) (2005) 45–62.

[4] K.-A. Lê Cao, S. Boitard, P. Besse, Sparse PLS discriminant analysis: biologically relevant feature selection and graphical displays for multiclass problems, BMC bioinformatics 12 (1) (2011) 1.

[5] O. P. Günther, H. Shin, R. T. Ng, W. R. McMaster, B. M. McManus, P. A. Keown, S. J. Tebbutt, K.-A. Lê Cao, Novel multivariate methods for integration of genomics and proteomics data: applications in a kidney transplant rejection study, Omics: a journal of integrative biology 18 (11) (2014) 682–695.

[6] C. E. Bonferroni, Teoria statistica delle classi e calcolo delle probabilita, Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8 (1936) 3–62.

[7] S. Dudoit, M. J. van der Laan, K. S. Pollard, Multiple testing. part i. single-step procedures for control of general type i error rates, Statistical Applications in Genetics and Molecular Biology 3 (1) (2004) 1–69.

[8] Y. Benjamini, Y. Hochberg, Controlling the false discovery rate: a practical and powerful approach to multiple testing, Journal of the Royal Statistical Society. Series B (Methodological) (1995) 289–300.

[9] Y. Benjamini, D. Yekutieli, The control of the false discovery rate in multiple testing under dependency, The Annals of Statistics (2001) 1165–1188.

[10] B. Efron, Local false discovery rates, Division of Biostatistics, Stanford University, 2005.

[11] J. D. Storey, A direct approach to false discovery rates, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 64 (3) (2002) 479–498.

[12] J. D. Storey, R. Tibshirani, Statistical significance for genomewide studies, Proceedings of the National Academy of Sciences 100 (16) (2003) 9440–9445.

[13] J. D. Storey, The positive false discovery rate: a bayesian interpretation and the q-value, The Annals of Statistics (2003) 2013–2035.

[14] M. Tenenhaus, V. E. Vinzi, Y.-M. Chatelin, C. Lauro, Pls path modeling, Computational Statistics & Data Analysis 48 (1) (2005) 159–205.

[15] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological) (1996) 267–288.

[16] L. Meier, S. Van De Geer, P. Bühlmann, The group lasso for logistic regression, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 70 (1) (2008) 53–71.

[17] N. Meinshausen, P. Bühlmann, High-dimensional graphs and variable selection with the lasso, The Annals of Statistics (2006) 1436–1462.

[18] R. Genuer, J.-M. Poggi, C. Tuleau-Malot, Variable selection using random forests, Pattern Recognition Letters 31 (14) (2010) 2225–2236.

[19] C. Friguet, M. Kloareg, D. Causeur, A factor model approach to multiple testing under dependence, Journal of the American Statistical Association 104 (488) (2009) 1406–1415.

[20] C. Friguet, D. Causeur, Estimation of the proportion of true null hypotheses in high-dimensional data under dependence, Computational Statistics & Data Analysis 55 (9) (2011) 2665–2676.

[21] M. Chavent, V. Kuentz, B. Liquet, L. Saracco, Clustofvar: an r package for the clustering of variables, Journal of Statistical Software 50 (2012) 91–116.

[22] D. Causeur, C. Friguet, M. Houee-Bigot, M. Kloareg, Factor analysis for multiple testing (famt): an r package for large-scale significance testing under dependence, Journal of Statistical Software. 40 (14) (2011) 19.

[23] Y. Su, T. Murali, V. Pavlovic, M. Schaffer, S. Kasif, Rankgene: identification of diagnostic genes based on expression data, Bioinformatics 19 (12) (2003) 1578–1579.

[24] C. Friguet, Impact of dependence in large-scale multiple testing, Ph.D. thesis, Universite de Bretagne-Sud (2012).

[25] Y. Shi, Microarray data analysis : feature selection, clustering and prediction, Master Internship report (2016) 1–40.