A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles in oncology

Bérangère Bastien, Taha Boukhobza, Hélène Dumond, Anne Gégout-Petit, Aurélie Muller-Gueudin, Charlène Thiébaut

A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles in oncology.

B. Bastien[c], T. Boukhobza[b], H. Dumond[b], A. Gégout-Petit[a], A. Muller-Gueudin[a], and C. Thiébaut[b]

[a]Université de Lorraine, CNRS, Inria, IECL, F-54000 Nancy, France; [b]Université de Lorraine, CNRS, CRAN, F-54000 Nancy, France; [c]Transgene S.A., Boulevard Gonthier d'Andernach, Parc d'innovation, CS80166, F-67405 Illkirch-Graffenstaden Cedex, France

ARTICLE HISTORY
Compiled September 12, 2019

ABSTRACT
We propose a new methodology for selecting and ranking covariates associated with a variable of interest in a context of high-dimensional data under dependence but few observations. The methodology successively intertwines the clustering of covariates, decorrelation of covariates using Factor Latent Analysis, selection using aggregation of adapted methods and finally ranking. A simulation study shows the interest of the decorrelation inside the different clusters of covariates. We first apply our method to transcriptomic data of 37 patients with advanced non-small-cell lung cancer who have received chemotherapy, to select the transcriptomic covariates that explain the survival outcome of the treatment. Secondly, we apply our method to 79 breast tumor samples to define patient profiles for a new metastatic biomarker and associated gene network in order to personalize the treatments.

KEYWORDS
Aggregated methods; Correlated covariates selection; Genetic profiles; High dimension; Multiple testing procedures; Personalized medicine; Ranking; Variable selection.

1. Introduction

The purpose of personalized medicine is to select appropriate and optimal therapies based on the context of a patient's genetic content or other molecular or cellular analysis. One of the main challenges faced by biologist and mathematician consortium for the construction of explanatory models of multivariable biological processes, is the relatively low amount of experimental data available compared to the huge number of variables. The point is of great concern when the biological question deals with transcriptomic data in order to build gene networks and decipher the role of a rare isoform, for which no specific probe is currently available. In that context, the purpose of this paper is to propose a method to select the covariates that are linked with the outcome of a given therapy or a biological marker, among a set of more than tens of thousands covariates. For instance, the relevant dataset we will study in this paper are the following:

- 51336 transcriptomic data of 37 patients with advanced non-small-cell lung cancer who have received chemotherapy. The survival time being known, the objective is to select the transcriptomic covariates that explain the survival outcome of the treatment, then to define the profiles of the patients who survive the treatment.
- 54676 probes in 79 breast tumor samples. A new metastatic biomarker being known, the objective is to define patient profiles for this metastatic biomarker, and associated gene network in order to personalize the treatments.

The variable of interest being known (treatment outcome or biological marker), the question is

A. Muller-Gueudin. Email: Aurelie.gueudin@univ-lorraine.fr

to find its link with the transcriptomic profile 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 [2, 3, 11, 18, 19]. 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. In the two data studies of this paper, 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 functions. For this purpose, it is interesting to rank the genes according to the strength of their link with the variable of interest. Although we present biological studies, our goal 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 the context of transcriptomic data, 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 [6] to control the Family Wise Error Rate (FWER) has been emulated and we can find a review about these methods in [10]. Alternative methods focused on the control of the False Discovery Rate (FDR) ([4, 5]), or of the local FDR [12] or the q-value [26–28]. For a review (in french) of the methods, see [3]. Regarding regression in the framework of high dimensional data ($n \ll p$), many methods are available. For exemple, the PLS approach of [30] is a kind of principal component regression. The Lasso regression proposed by [32] 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 problems like logistic-regression in the case of binary data [21], or network inference [22]. Another versatile tool to select covariates in different non parametric contexts is given by the random forests, with the concept of importance of covariates [16].

Another important characteristic of the data that has to be taken into account is the structure of covariance of the covariates. Most of the multiple testing corrections make the assumption of the independence between the tests. However it is well-known that omics data are correlated by blocks. In the context of multiple testing, 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 [14, 15]. In [16] it is also shown that despite the robustness of the random forests, importance of covariates calculated by random forests is perturbed by adding other correlated covariates. One of the ways to deal with dependence 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 [15] and [14]. 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 (leading to the independence of the tests). After this correction, they propose a multiple testing procedure based on the method of [4] and [5]. 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 clusters with strong within correlation. 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 [9] 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 clusters of covariates by using the clustering of covariates proposed by [9], and (step 1.2) applying a "decorrelation" between the covariates inside each cluster using the factor analysis proposed by [7, 14, 15]. Their method performs a decorrelation of the covariates and calculates the corrected covariates suitable for statistical 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 [29].

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 the covariates scores 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 the case where the variable of interest is binary. Section 4 is dedicated to two real data analysis: the purpose of the first analysis is to select covariates that are linked with the outcome of a lung cancer treatment, whereas the second analysis selects covariates linked with a breast cancer biomarker. In both analysis, the selected covariates are used to define genetic profiles of patients. Section 5 gives some conclusions and perspectives. The Appendix gives two simulation studies in the case where the variable of interest is respectively a binary and a continuous quantitative variable (Sections A and B of Appendix). Technical details on the two real data applications are also given (Sections C and D of the Appendix).

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 the covariates are decomposed into K independent clusters:

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

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

On one hand, we model the dependence in the K clusters of covariates as in the framework of [15]: inside each cluster $\mathbf{X}^{(k)}$, the common information between the p_k covariates is modeled by regression on q_k latent factors:

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

where $\delta_i^{(k)}(Y)$ is a function of Y, $\mathbf{Z}^{(k)}$ is a random centered q_k-vector such that $\mathbb{E}(\mathbf{Z}^{(k)} \mathbf{Z}^{(k)\prime}) = \mathbf{I}_{q_k}$,

$\mathbf{b}_i^{(k)}$ is a q_k-vector, and $\varepsilon^{(k)} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{p_k}^{(k)})$ 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), we have:

$$\begin{aligned} \mathbf{\Sigma}^{(k)} := \mathbb{V}(\mathbf{X}^{(k)}|Y) &= \mathbf{B}^{(k)}(\mathbf{B}^{(k)})' + \mathbf{\Psi}^{(k)} & (2) \\ \mathbb{V}(\mathbf{X}^{(k)}|Y, \mathbf{Z}^{(k)}) &= \mathbf{\Psi}^{(k)} & (3) \\ \mathbb{C}ov(\varepsilon_i^{(k)}, Z_j^{(k)}) &= 0, \quad \forall i, j, k & (4) \end{aligned}$$

where $\mathbf{\Psi}^{(k)}$ is a diagonal $p_k \times p_k$ matrix (the covariance matrix of $\varepsilon^{(k)}$) and $\mathbf{B}^{(k)}$ is a $p_k \times q_k$ matrix of factor loadings $\mathbf{b}_i^{(k)}$ (cf Equation (1), the $\mathbf{b}_i^{(k)}$ being the ith row of $\mathbf{B}^{(k)}$). In the decomposition given in Equation (2), the diagonal element $\Psi_i^{(k)}$ is the specific variance of the response $X_i^{(k)}$ while $\mathbf{B}^{(k)}(\mathbf{B}^{(k)})'$ appears as the shared variance in the common factor structure. [15] define the common variance by

$$\text{ComVar}^{(k)} = \frac{\text{trace}(\mathbf{B}^{(k)}(\mathbf{B}^{(k)})')}{\text{trace}(\mathbf{\Sigma}^{(k)})}. \qquad (5)$$

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

2.2. Main prodecure

The procedure is decomposed in two steps: a pretreatment of the covariates (step 1) and 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 (the p-value adjustment methods proposed by [6], [4, 5], or the q-value proposed by [26–28], or the local FDR presented in [2], [3]) are all built on the assumption that tests are independent. As a results, they are no longer promising if the independence is not verified. A very detailed discussion can be found in the Friguet's thesis [13].

In estimating together the latent factors $\mathbf{Z}^{(k)}$ and the coefficients of regressions ($\mathbf{B}^{(k)}$, $\mathbf{\Psi}^{(k)}$) by an E.M. algorithm in model (1), the FAMT procedure of [15] 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 $X_i^{(k)\star} = X_i^{(k)} - \mathbf{b}_i^{(k)}\mathbf{Z}^{(k)} = \delta_i^{(k)}(Y) + \varepsilon_i^{(k)}$, $i = 1, \ldots, p_k$, 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 [15].

Of course, the whole vector $\mathbf{X}$ satisfies assumption of Equation (1), and [15] apply 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 $(\mathbf{X}^{(k)})_{1 \le k \le K}$ and then to apply the decorrelation method on each cluster. Indeed, [25] has shown with some simulation studies 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.

2.2.1.1. ***Step 1.1: clustering of covariates.*** We apply a clustering of covariates in the purpose to find clusters of correlated variables as we assumed in Section 2.1. We propose to use the algorithm of [9] 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.

2.2.1.2. ***Step 1.2: Factor analysis to correct dependency 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 $\mathbf{X}_Y^*$ in the sequel. Note that $\mathbf{X}_Y^*$ depends on Y. To simplify the notations, $\mathbf{X}_Y^*$ will be noted $\mathbf{X}^*$.

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 $\mathbf{X}^\star$. We perform L methods, then each covariate $X_i^\star$ obtains a score $S_i \in \{0, 1, \ldots, L\}$ that is the number of selections among the L methods. By this way, the covariates can be ranked 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 [16]). 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.

2.3. R Package `armada`

In the sequel, we call our procedure ARMADA for AggRegated Methods for covAriates selection under Dependence. Our procedure has been implemented in an R package, called `armada`, available on the CRAN [23]. The package proposes also a graphical representation of the selected covariates, through heatmaps, as presented in Figures 5 and 8.

3. Simulations

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

3.1. Simulation design

We propose a simulation study with $p = 1600$ covariates and sample size $n = 60$. We first describe how to create dependence in the covariates $\mathbf{X}$, then we present a simulation design in a classification study. Two other designs in classification and regression cases are given in Sections A and B of the Appendix.

The covariates $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4}$ are clustered into four independent clusters, 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 $\mathbf{\Sigma}^{(k)}$. The correlation between the covariates of $\tilde{\mathbf{X}}^{(k)}$ 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 Equation (2) can be found in [13]. We simulate data with common variances $\mathrm{ComVar}^{(k)}$ equal to 0.8 in each cluster (recall that the common variance is defined in Equation (5)). Moreover, the numbers of latent factors in each cluster are $(q^{(1)}, \dots, q^{(4)}) = (4, 6, 8, 10)$.

Now, we create the dependence between $\mathbf{X}$ and Y in perturbing some component of $\tilde{\mathbf{X}}$. We 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). Y is linked with 160 influential covariates in $\mathbf{X}$, whose links with the response variable Y have different intensities. The other covariates are noise. More precisely, in each cluster $k = 1, \dots, 4$, and for $i = 1, \dots, 400$,

$$X_i^{(k)} = \tilde{X}_i^{(k)} + \delta_i \mathbf{1}_{Y=0}$$

with $\delta = (\delta_i)_{i=1,\dots,400} = (\underbrace{1.5, \dots, 1.5}_{i=1,\dots,10}, \underbrace{1, \dots, 1}_{i=11,\dots,20}, \underbrace{0.75, \dots, 0.75}_{i=21,\dots,30}, \underbrace{0.5, \dots, 0.5}_{i=31,\dots,40}, \underbrace{0, \dots, 0}_{i=41,\dots,400})$. In other words, Y is linked with the $m_1 = 40$ first covariates of each cluster, and the $m_0 = 360$ remaining covariates of each cluster are independent of Y. Then, the 10 first covariates of each cluster are the most strongly linked with the response variable Y and the strength of the link is decreasing in the successive groups of 10 influential covariates.

We can remark that this design respects the covariance matrix given in Figure 1 and the model given by Equation (1).

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, taking Y into account, with the factor analysis procedure of [7, 15], implemented in the R package `FAMT`. This gives a new dataset $\mathbf{X}_Y^{\dagger}$.

Procedure 3: the 4 clusters are estimated with the procedure of [9], implemented in the R package `ClustOfVar`; then the covariates are decorrelated in each cluster, taking Y into account, with the factor analysis procedure of [7, 15], implemented in the R package `FAMT`. This gives a new dataset $\mathbf{X}^*{}_Y$ 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, $\mathbf{X}_Y^{\dagger}$ for Procedure 2, $\mathbf{X}^*{}_Y$ 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 here) and no False Positive. However, according to the detection threshold chosen for the p-value, the expected number of FP is $72 = 5\% \times (1600 - 160)$. The results are shown in Figure 2.

On the Figure 2, 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. Our Procedure reduces the mean and the variability of the distributions of the false positive rates, in comparison to the Procedure 2 (i.e. the FAMT procedure). The power of our Procedure is comparable with Procedure 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 Figure 3 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” (see Section 3.1); the group of noise covariates is noted by ”-”).

We can see on the Figure 3 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 performant: 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. The ARMADA 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. Note that around 95% of the noise covariates obtained an ARMADA score that was exactly 0.

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 [7]: the selected covariates are those with adjusted p-values lower than 0.05.

To compare the three selection methods, the Table 1 gives the rates of selection for each group of influential covariates, and for the group of noise covariates. The rates of selection have been computed on $N = 100$ runs of $(\mathbf{X}, Y)$. We can see that our method respects 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. 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.

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

	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)

Finally, we can conclude with the ROC curves given in Figure A3 that our method out-

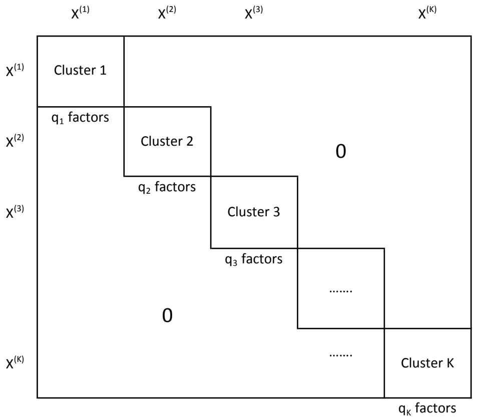


Figure 1. Covariance structure of covariates given Y

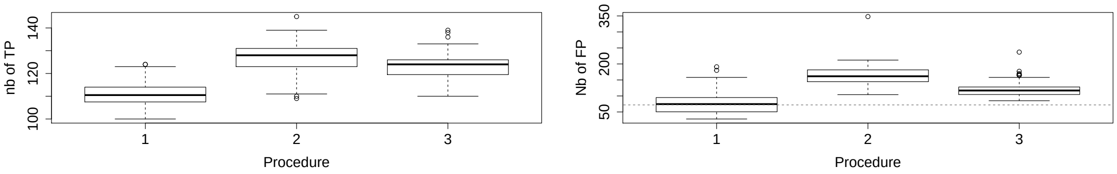


Figure 2. Number of true positive tests (left), false positive tests (right) according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 100$ runs.

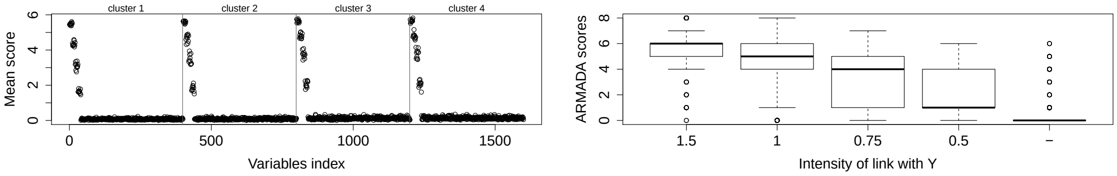


Figure 3. Left: mean of the ARMADA scores obtained by all the covariates. Right: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs.

performs 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). The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of $(\mathbf{X}, Y)$.

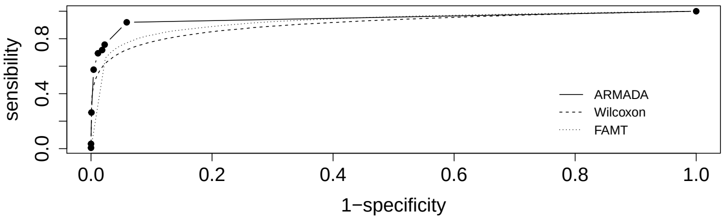


Figure 4. ROC curves for the three selection methods.

4. Application to real data

In this section we apply our method on two real datasets, both in oncology. The first one concerns the selection of transcriptomic covariates linked with the outcome of a chemotherapy for lung cancer. The second one concerns the selection of covariates linked with a quantitative biomarker ERα36 in breast cancer.

4.1. Outcome of chemotherapy for advanced non-small-cell lung cancer

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 51336 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. In the pretreatment step, we found that the $p = 6810$ covariates are decomposed in 3 independent group of covariates.

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: $Y = 1$ 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 response variable being binary, we first applied the selection method presented in Section 2.2.2.

Moreover, as the survival time was known for all the 37 patients without any censoring, 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 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 [16]).

The joint results of the classification and regression studies are given in Table 2. In the classification study, we can see that 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. 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 function of the 10 genes with a score at 7 has to be studied to understand its link with the ”success” of the treatment. The table 2 is a little disappointing, because regression and classification do not select the same covariates. Whatever, among the covariates with a C-score (score in the classification case) equal to 7, there is only one with a R-score (score in the regression case) lower than 4 (equal to 0!). But these two analyses are not looking for the same kind of link with the covariates. 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 5, we show the heatmap of the selected covariates which have a classification score **and** a regression score greater than five. For the visualisation of the results, we then build an heatmap 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` (Figure 5).

Table 2. 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 instance, 41 covariates have a R-score equal to 1, and a C-score equal to 0.

	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

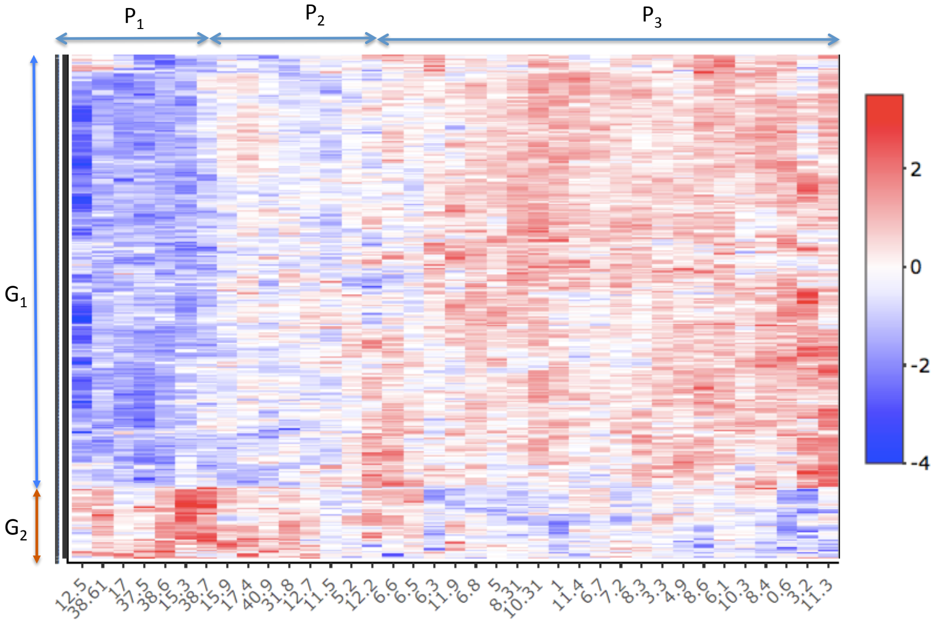


Figure 5. Heatmap of the 342 covariates which have ARMADA scores greater or equal to 5 in **both** classification **and** regression studies. Each column corresponds to one patient. The x-axis represents the patients (marked with their survival time) and the y-axis the covariates. Color gives the normalised expression of the covariates.

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 5) 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 and P_2 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 discriminates groups P_1 and P_3. Indeed, 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.

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 (reported in the Section C of the Appendix) show that our method is robust: the distributions of the bootstrapped scores faithfully reproduce the original scores.

4.2. Biological network involving ERα36 in breast cancer

ERα36 is a variant of the oestrogen receptor α encoded by the ESR1 locus and expressed only in humans [33]. ERα36 expression and activity have been mainly studied in vitro and in vivo in the context of breast cancer. However, due to the lack of comprehensive transcriptomic data that include ERα36, only sparse information is available on factors that could act up- and downstream ERα36 in biological networks. Our challenge from a statistical point of view was to explain the ERα36 expression variation obtained in a small number of breast tumors from a large number of potential explanatory variables that correspond to the 54676 transcriptomic probes. For this, we analysed the biological network involving ERα36 through the use of 4 sets of Affymetrix transcriptomic data obtained from breast tumors of different molecular subtypes: the triple negative (noted TN), ERa66+, PR+ and PR- datasets (details are provided in Section D of the Appendix).

The analysis was performed in three steps (Figure 6).

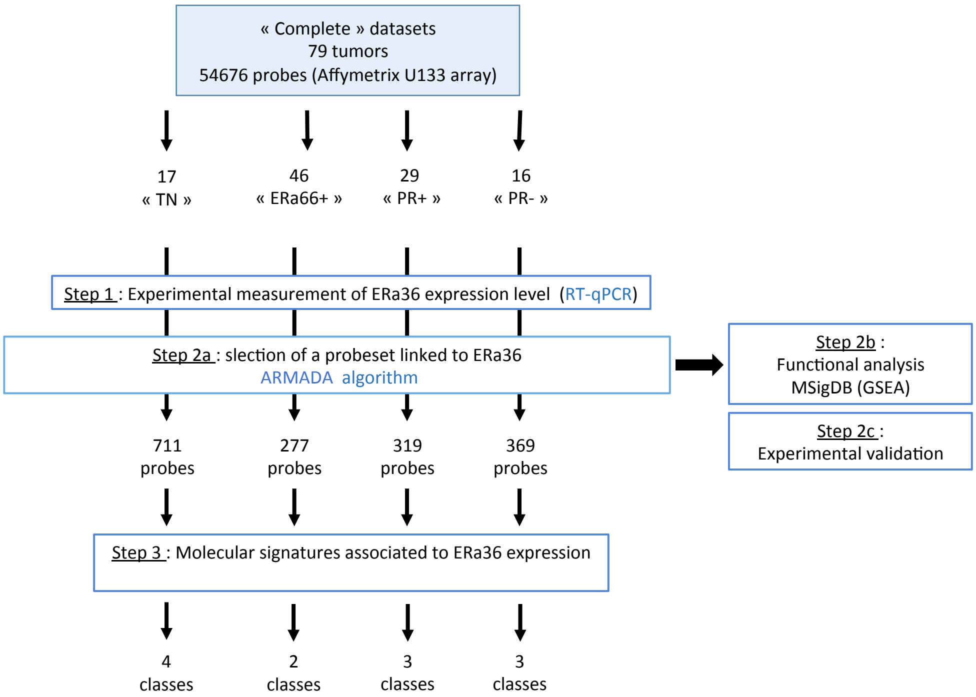


Figure 6. Workflow describing the 3 steps of four breast tumor transcriptomic datasets.

Step 1, clinical data completion. These 4 transcriptomic datasets were completed by the measurement of ERα36 expression level in each tumor. Biological details are given in Section

D of the Appendix.
Step 2a, statistical analysis. To explain the ERα36 expression variation obtained in a small number of tumors from a large number of potential explanatory variables, we used the R package `armada`, in its regression version (as in Section 4.1 where Y is quantitative), that allowed to select, among the 54676 initial probes, a few hundreds of genes whose expression is supposed to be correlated to that of ERα36 (ARMADA score ≥ 1). We obtained four lists of respectively 711, 277, 319 and 369 probes correlated to the expression of ERα36 in the TN, ERa66+, PR+ and PR- groups.
Step 2b, functional analysis. From these four lists of transcriptomic probes, we carried out a functional analysis using the MSigDB database (GSEA). In particular, we looked for transcription factors and microRNAs involved in the regulation of the majority of genes from the different lists (TN, ERa66+, PR+ and PR-). The results indicated that four transcription factors: NFAT, FOXO4, SP1 and LEF-1, were common regulators and could therefore be mediators of the ERα36 effect in all breast tumor subtypes. Interestingly, a study by [1] has shown that these transcription factors FOXO4, SP1 and LEF-1 are transcriptional hallmarks characteristic of cancer cells and associated to the Wnt signaling pathway (involved in metastasis and maintenance of cancer stem cells). Regarding the analysis of microRNAs, the results indicated that the majority of the genes whose expression correlated to ERα36 one in the TN set were regulated by the microRNAs: hsa-miR-106B, hsa-miR21 and hsa-miR-29A, listed as oncogenic microRNAs involved in metastatic processes, survival and self-sufficiency in growth factors of mammary tumors [20]. These results recalled those of a previous study of [8], which showed that a high ERα36 expression in mammary tumors is associated with an increased metastatic potential and an estrogen-independent tumor growth.
Step 2c, experimental validation. Subsequently, we provided an experimental confirmation of the biological reliability of the results: the correlation between the expression of ERα36 and that of ZEB1, FZD7, ZIC1 and TCF7LD genes, identified by `armada` as correlated to ERα36 in all tumor sets, was verified in vitro by RT-qPCR in two breast cancer cell lines (MCF-7 (ERa66+, PR+, PR-) and MDA-MB-231 (TN)). The results of Figure 7 confirmed the correlation (positive or negative) between the expression of ERα36 and that of the genes identified by `armada` in the both cell lines.

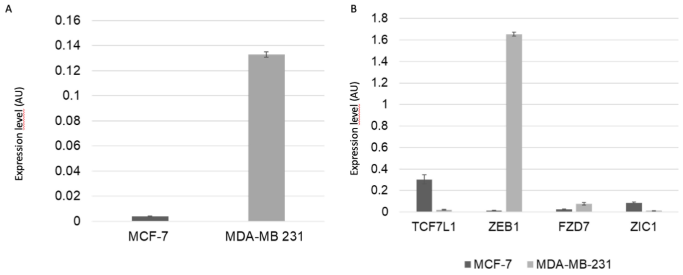


Figure 7. Experimental validation of biological data inferred from `armada`. (A) ERα36 expression level as measured by RT-qPCR in MCF-7 and MDA-MB-231 breast cancer cells. (B) TCF7L1, ZEB1, FZD7 and ZIC1 expression level as measured by RT-qPCR in MCF-7 and MDA-MB-231 breast cancer cells.

Step 3, tumor classification according to ERα36 expression level. The final goal of our study was to identify the molecular signatures accounting for the ERα36 expression level in the four different sets of tumors. These signatures were identified thanks to the R package `heatmaply` after co-clustering of the ERα36 expression (on the x-axis) and of the covariates (on the y-axis) with the function `hclust`. For each of the four tumor datasets, a heatmap was built which accounted for the expression level of the genes correlated to ERα36. Thanks to the associated dendogram, different classes of tumors were defined and characterized by both the level of ERα36 expression and an associated molecular signature. The Figure 8 illustrates the results for the study on the dataset ERa66+: two classes of tumors were identified, called ERα36^{++} and ERα36^{-}. Taken together, the `armada` package helped to cluster patients which breast tumor highly express ERα36 and associated genes. These patients could be treated by

Wnt signaling inhibitors or specific microARN modulators and therefore benefit such promising new personalized medicine.

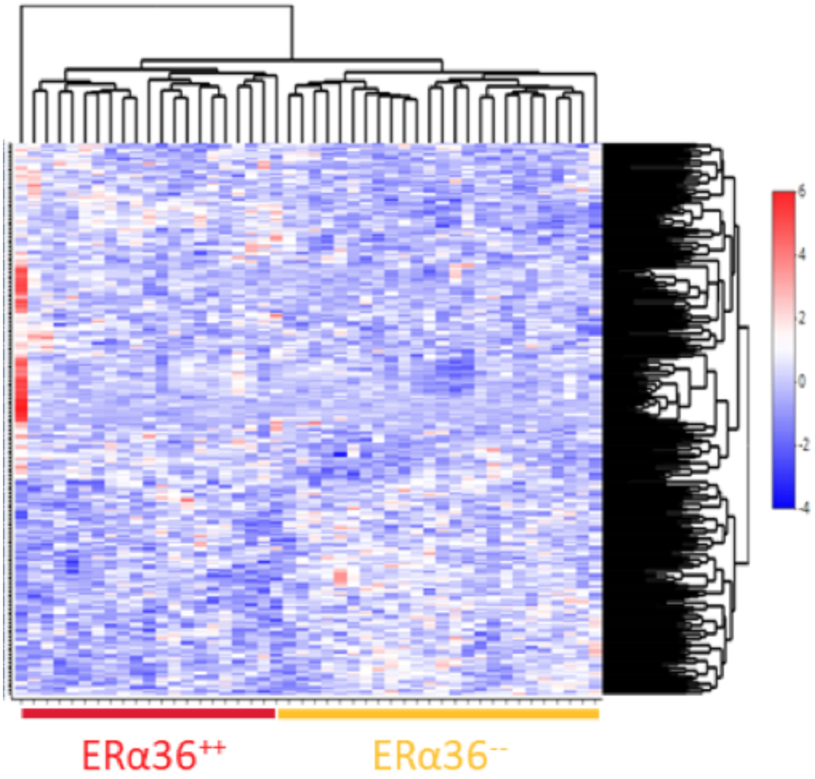


Figure 8. Heatmap of the 711 covariates which have ARMADA scores greater or equal to 1 in the study of the dataset ERa66+. Each column corresponds to one ERa66+ tumor. The x-axis represents the tumors (marked with their ERα36 expression level) and the y-axis represents the selected probes. Color gives the normalised expression of the 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 or a biological marker). The method is of particular interest in the high dimensional case and when the covariates are correlated. The algorithms corresponding to this method are available through the R package `armada`. After this selection obtained with our method, it is then easy to visualise the selected genes (or probes) for all the patients, and to classify the genetic profiles of patients with respect to their treatment outcome or biological marker. In the study 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. In the study of the mammary tumors, the covariates selection allows the biologist to study the functional role of selected probes and also to classify tumors and associated transcriptomic signatures. This kind of results is very promising for the identification of new therapeutic targets and the development of more efficient and personalized anti-cancer treatment.

Appendix

The Appendix gives additional simulations on another classification model, and on a regression model where Y is a continuous quantitative variable (Sections A and B of Appendix), some further analysis of the lung cancer data presented in Section 4.1 (Section C of Appendix), and technical informations on the biological material used in Section 4.2 (Section D of Appendix).

Appendix A. Classification design

A.1. Simulation design

As in Section 3, we propose a simulation study with $p = 1600$ covariates and sample size $n = 60$. The response variable Y is binary: $Y = 1$ for $\frac{n}{2}$ subjects, and $Y = 0$ for $\frac{n}{2}$ subjects. The covariates $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4}$ are clustered into four independent clusters, 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 $\mathbf{\Sigma}^{(k)}$. The correlation between the covariates of $\tilde{\mathbf{X}}^{(k)}$ inside the cluster k is designed by a factor analysis model, as in the Section 3. More precisions on the factor analysis model can be found in [13]. Now, we create the dependence between $\mathbf{X}$ and Y in perturbing some component of $\tilde{\mathbf{X}}$. This simulation design is inspired from the toys-data of [16]. The outcome Y is linked with 240 influential covariates in $\mathbf{X}$, the others being noise covariates. The links between the influential covariates and the response variable Y have different intensities. More precisely, the 10 first covariates of each cluster are the most strongly linked with the response variable Y and the strength of the link is decreasing in the successive groups of 10 influential covariates.

More precisely, let us define the simulation model by giving the conditional distribution of X_i given the value y of Y: in each cluster $k = 1, \dots, 4$, and for $i = 1, \dots, 400$,

$$X_i^{(k)} = \tilde{X}_i^{(k)} + \delta_i^{(k)}(y)$$

where $\delta_i(y)$ is a random variable.

- The relevant covariates are the $m_1 = 60$ first covariates of each cluster. The distribution of the $\delta_i^{(k)}(y)$ leading to the links between the relevant covariates and Y is given in Table A1.
- The $m_0 = 340$ remaining covariates of each cluster are independent of Y: $\delta_i^{(k)}(y) = 0$ whatever y for $i = 61, \dots, 400$.

Table A1. Links between the relevant covariates and Y in the classification design. The notation $\delta_i^{(k)} \sim 0.7\mathcal{N}(3y, 1) + 0.3\mathcal{N}(0, 1)$ means that, with probability 0.7, $\delta_i^{(k)} \sim \mathcal{N}(3y, 1)$, and with probability 0.3, $\delta_i^{(k)} \sim \mathcal{N}(0, 1)$.

i	model for δ_i^k
for $i = 1, \dots, 10$	$\delta_i^{(k)} \sim 0.7\mathcal{N}(3y, 1) + 0.3\mathcal{N}(0, 1)$
for $i = 11, \dots, 20$	$\delta_i^{(k)} \sim 0.7\mathcal{N}(2y, 1) + 0.3\mathcal{N}(0, 1)$
for $i = 21, \dots, 30$	$\delta_i^{(k)} \sim 0.7\mathcal{N}(y, 1) + 0.3\mathcal{N}(0, 1)$
for $i = 31, \dots, 40$	$\delta_i^{(k)} \sim 0.3\mathcal{N}(3y, 1) + 0.7\mathcal{N}(0, 1)$
for $i = 41, \dots, 50$	$\delta_i^{(k)} \sim 0.3\mathcal{N}(2y, 1) + 0.7\mathcal{N}(0, 1)$
for $i = 51, \dots, 60$	$\delta_i^{(k)} \sim 0.3\mathcal{N}(y, 1) + 0.7\mathcal{N}(0, 1)$

We can remark that this design respects the covariance matrix given in Figure 1. This design differs a little bit from the model of Equation (1), because $\delta_i^{(k)}(Y)$ is a random function of 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.

A.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, taking Y into account, with the factor analysis procedure of [7, 15], implemented in the R package `FAMT`. This gives a new dataset $\mathbf{X}_Y^{\dagger}$.

Procedure 3: the 4 clusters are estimated with the procedure of [9], implemented in the R package `ClustOfVar`; then the covariates are decorrelated in each cluster, taking Y into account, with the factor analysis procedure of [7, 15], implemented in the R package `FAMT`. This gives a new dataset $\mathbf{X}^{*}{}_Y$ 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, $\mathbf{X}_Y^{\dagger}$ for Procedure 2, $\mathbf{X}^{*}{}_Y$ 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 240 in this study) and no False Positive. However, according to the detection threshold chosen for the p-value, the expected number of FP is $68 = 5\% \times (1600 - 240)$. The results are shown in Figure A1.

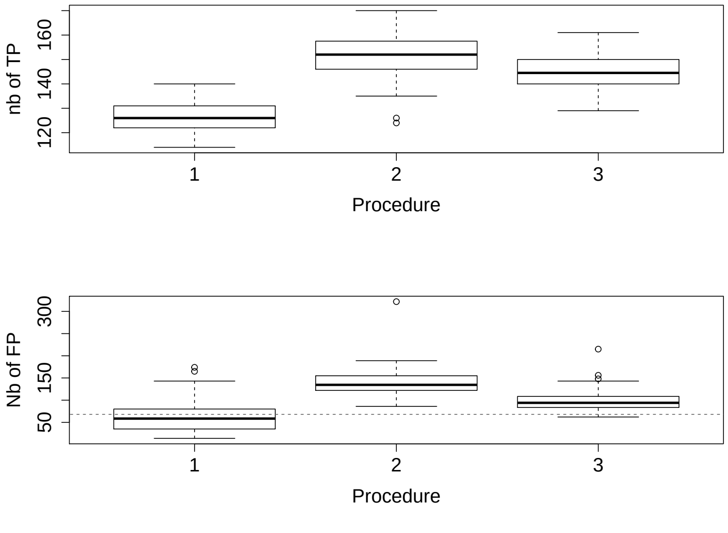


Figure A1. Number of true positive tests (top), false positive tests (bottom) in the classification design according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 100$ runs.

If we analyse the results given by Figure A1, 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 reduces the mean and the variability of the distributions of the false positive rates, in comparison to the Procedure 2 (i.e. the FAMT procedure). The power of our Procedure is comparable with Procedure 2. This results show the interest of our proposed pretreatment before performing selection.

A.3. Results of the whole method (pretreatment and selection)

In order to describe the performances of our method, we show in Figure A2 the mean ARMADA scores obtained on the $N = 100$ runs of $(\mathbf{X}, Y)$. 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 "(0.7,3)", "(0.7,2)", "(0.7,1)", etc.; the group of noise covariates is noted by "-").

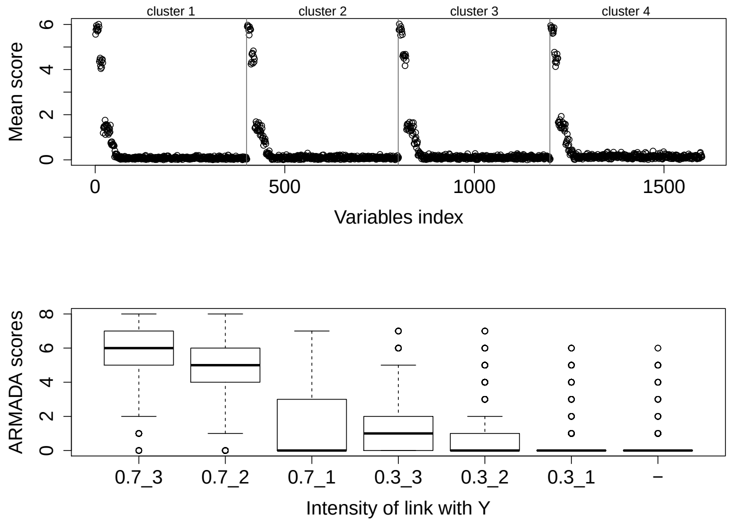


Figure A2. Top: mean of the ARMADA scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs. Simulation in the classification design.

We can see on the Figure A2 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 method is not so performant as in the design presented in Section 3, probably because we are note exactly in the model of the study, given by 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. We can precise that around 95% of the noise covariates obtained an ARMADA score that was exactly 0.

A.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 [7]: the selected covariates are those with adjusted p-values lower than 0.05.

To compare the three selection methods, the Table A2 gives the rates of selection for each group of influential covariates, and for the group of noise covariates. The rates of selection have been computed on $N = 100$ runs of $(\mathbf{X}, Y)$. 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 %. Our method is competitive with the FAMT procedure for the detection of influential covariates, but again FAMT procedure has more false positives than ours.

Table A2. Results of the $N = 100$ runs in the classification design: rates of selection of the different groups of influential and noise covariates by the ARMADA method, 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)

Finally, we can conclude with the ROC curves given in Figure A3 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 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 $(\text{1-specificity, sensibility})_{\text{ARMADA score}=0}$ and $(\text{1-specificity, sensibility})_{\text{ARMADA score}=1}$. The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of $(\mathbf{X}, Y)$.

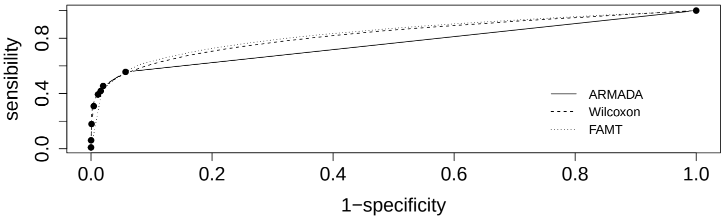


Figure A3. ROC curves for the three selection methods in the classification design.

Appendix B. Regression design

In this section, we give results of simulations to study the behavior of our algorithm to select covariates linked with a continuous variable of interest (like survival time here). We simulate $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}^{(k)})_{k=1,\ldots,4}$ as in Section 3, and Y as a standard gaussian variable. Now, we create the dependence with outcome Y in perturbing some component of $\tilde{\mathbf{X}}$: in all cluster $k = 1, \ldots, 4$, and for all $i = 1, \ldots, 400$:

$$X_i^{(k)} = \tilde{X}_i^{(k)} + \delta_i Y \tag{B1}$$

where $\delta = (\delta_j)_{j=1,\ldots,400} = (1, 0.8, 0.6, 0.4, 0.2, 0, 0, \ldots, 0)$. Only the first 5 covariates of each cluster are linked with Y.

We show the interest of our pretreatment, comparing the three procedures detailed in Section 3. As Y is a gaussian variable, we use the Pearson correlation test (instead of the Wilcoxon test used in Section 3). We produce $N = 100$ runs of $(\mathbf{X}, Y)$ and count the number of false and true positive, and the ARMADA scores (shown in Figures B1 and B2).

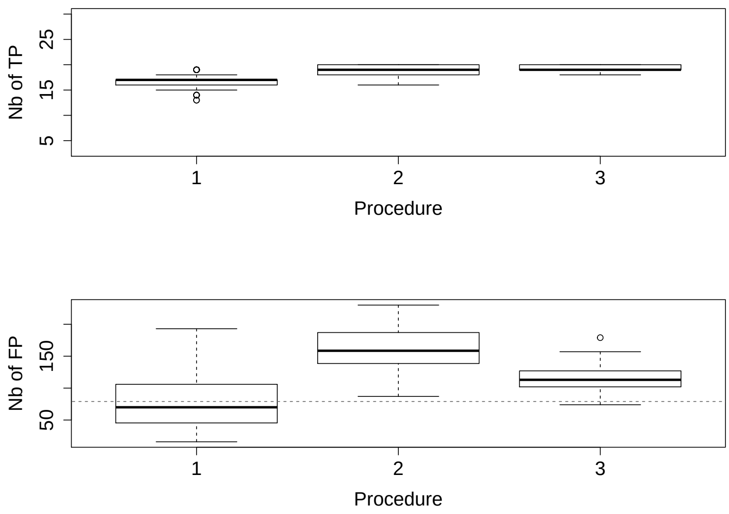


Figure B1. Number of: true positive tests (top), false positive tests (bottom) in the regression design. Boxplots are calculated on $N = 100$ runs.

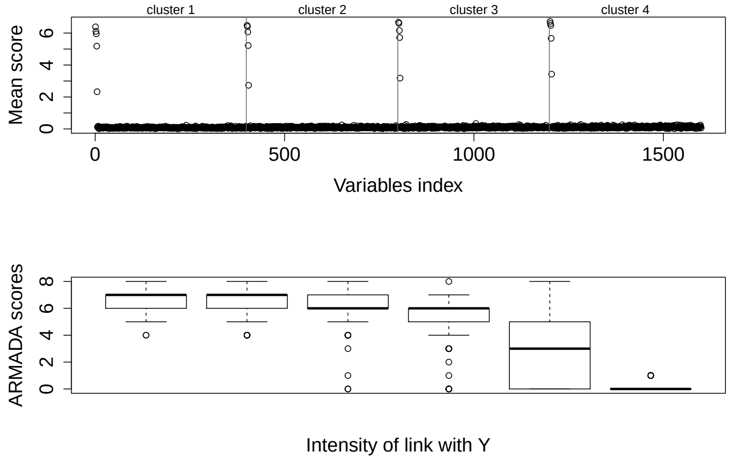


Figure B2. Top: mean of the ARMADA scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs. Simulation in the regression design.

Similarly to the classification studies presented in Section 3, our Procedure reduces the mean and the variability of the distributions of the false positive rates, in comparison to the Procedure 2 (i.e. the FAMTprocedure), and the power of our Procedure is comparable with Procedure 2.

The Figure B2 shows the ARMADA scores obtained on these $N = 100$ runs of $(\mathbf{X}, Y)$. Again, similarly to the Section 3, the scores give a ranking of the covariates, according to the intensity of their link with respect to the response variable Y. The true covariates are clearly separated of the noise covariates. We can also precise that 96% of the noise covariates obtained a score that was 0.

As in Section 3, the Table B1 and the ROC curve in Figure B3 allow us to compare our method with the Pearson test and the FAMT procedure. Our method seems to be a good compromise to have quite good detection rates for the true covariates, but small detection rates for the noise covariates. Even though true covariates are not always enough detected, compared to the FAMT procedure, detection rate of noisy covariates is lower than FAMT. The Pearson test has the lowest levels of detection rates, and the true covariates with a small link with Y are not well detected. On the whole, our method seems to be appropriate for sparse models particularly when the goal is to avoid false positive detections.

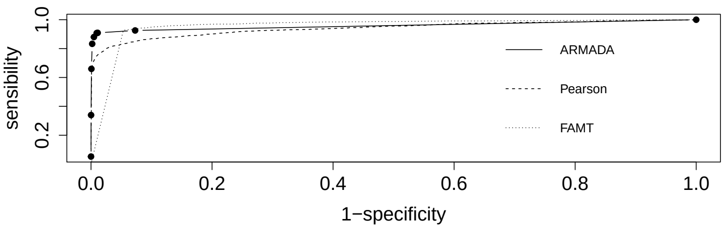


Figure B3. ROC curves for the three selection methods, in the case of regression design. The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of $(\mathbf{X}, Y)$.

Table B1. Results of the $N = 100$ runs in the regression design: rates of selection of the different groups of influential and noise covariates by the ARMADA method, the Pearson correlation test and the FAMT procedure. The corresponding standard deviations are given in brackets.

	ARMADA	Pearson	FAMT
1	1 (0)	1 (0)	1 (0)
0.8	1 (0)	1 (0)	1 (0)
0.6	0.99 (0.08)	0.99 (0.08)	1 (0)
0.4	0.97 (0.18)	0.82 (0.38)	0.98 (0.13)
0.2	0.67 (0.47)	0.33 (0.47)	0.76 (0.43)
-	0.07 (0.26)	0.05 (0.22)	0.10 (0.30)

Appendix C. Lung cancer real dataset: bootstrap analysis

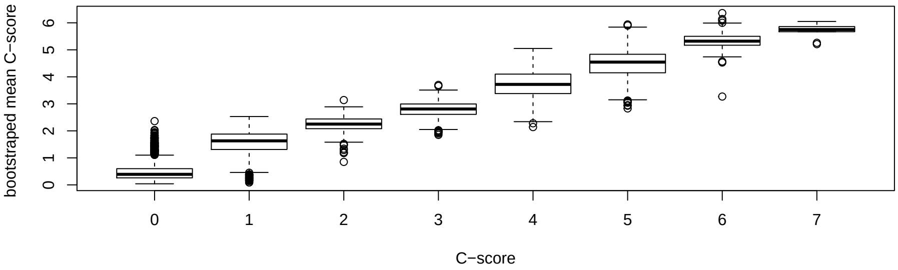
As the number of patients $n = 37$ is small compared to the number of covariates even after filtering ($p = 6810$), we have checked our results with a bootstrap study. We have calculated the C-scores and R-scores of each covariates on $B = 100$ bootstrap samples and the mean of the B results. We give the distribution of the bootstrapped means according to the original scores for the original dataset (Figure C1).

We can see that the distributions of the bootstrapped means of the scores have a quite small dispersion and faithfully reproduce the original scores. The same conclusion holds for the bootstrapped median scores (shown in Tables C1 and C2).

Moreover, we can emphasis that our method is robust to detect the most important covariates (for instance, the 10 covariates that have a C-score equal to 7, or the 6 covariates that have an R-score greater than 7): their corresponding bootstraped means of scores are also high, and their corresponding bootstraped median scores are greater than 5.

Appendix D. Biological material for the study of ERα36 in breast cancer

We analysed the biological network involving ERα36 through the use of 4 sets of Affymetrix transcriptomic data obtained from breast tumors of different molecular subtypes: the triple negative (noted TN), ER66+, PR+ and PR- datasets:



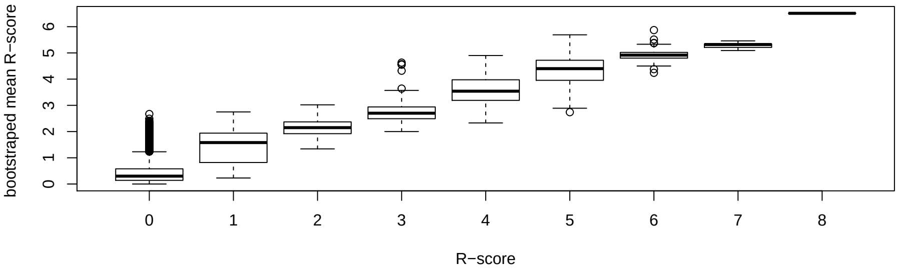


Figure C1. Distribution of the bootstraped mean of C-scores (resp. R-scores), i.e. means of C-(or R-)scores obtained on $B = 100$ bootstrap samples), according to the corresponding C-scores (resp. R-scores) in the original dataset for all the $p = 6810$ covariates.

Table C1. Distribution of the boostraped median C-scores of the $p = 6810$ covariates, obtained on $B = 100$ boostrap samples, versus the corresponding C-scores.

	ARMADA C-score							
Bootstraped median C-score	0	1	2	3	4	5	6	7
0	2698	53	0	0	0	0	0	0
0.5	11	1	0	0	0	0	0	0
1	108	315	29	1	0	0	0	0
1.5	0	19	5	0	0	0	0	0
2	9	162	308	76	5	0	0	0
2.5	0	2	28	14	0	0	0	0
3	1	1	87	321	55	2	0	0
3.5	0	0	1	29	12	1	0	0
4	0	0	2	155	922	218	1	0
4.5	0	0	0	0	19	6	0	0
5	0	0	0	0	157	644	221	2
5.5	0	0	0	0	0	0	6	0
6	0	0	0	0	0	17	78	8

Table C2. Distribution of the boostraped median R-scores of the $p = 6810$ covariates, obtained on $B = 100$ boostrap samples, versus the corresponding R-scores.

	ARMADA C-score								
Bootstraped median R-score	0	1	2	3	4	5	6	7	8
0	3773	29	11	0	0	0	0	0	0
0.5	20	2	5	0	0	0	0	0	0
1	67	22	17	0	0	0	0	0	0
1.5	8	1	9	0	0	0	0	0	0
2	109	32	243	40	1	0	0	0	0
2.5	4	0	22	8	2	0	0	0	0
3	7	3	147	295	80	2	0	0	0
3.5	0	0	0	14	13	2	0	0	0
4	0	0	2	149	788	210	0	0	0
4.5	0	0	0	0	10	14	0	0	0
5	0	0	0	3	90	462	85	5	0
5.5	0	0	0	0	0	1	0	0	0
6	0	0	0	0	0	1	1	0	1

- the TN dataset corresponding to Affymetrix transcriptomic comprehensive data from 17 patients derived xenografts (PDX) breast tumors was extracted from the Xentech$^{\text{TM}}$ database with the permission of Olivier Déas and Stefano Cairo (MTA CXT-295 Xentech SAS/University of Lorraine ; [24]).
- the 3 other datasets (46 tumors ER66+, 29 tumors PR+, 16 tumors PR-) were part of those from the Carte d'Identité des Tumeurs Program (CIT) from the Ligue Nationale Contre le Cancer described in [17]. Transcriptomic raw data were kindly provided by Aurélien De Reynies and Jacqueline Métral. One microgram of cDNAs from each tumor sample gathered at the Oncogenetics laboratory, INSERM U735, Institut Curie-Hôpital-Centre René Huguenin, St Cloud, France was also kindly provided by Ivan Bieche to measure ERα36 expression.

The measurement of ERα36 expression in each tumor (Step 1: clinical data completion) has been done as described in [31]. Total RNA extraction of PDX samples and qPCR analyses were performed. The following primers were used for qRT-PCR : GAPDH forward (Fw) 5'-TGC-ACC-ACC-AAC-TGC-TTA-GC -3', GAPDH reverse (Rev) 5'-GGC-ATG-GAC-TGT-GGT-CAT-GAG -3', ERα36 forward (Fw) 5'- ATG-AAT-CTG-CAG-GGA-GAG-GA-3', ERα36 reverse (Rev) 5'- GGC-TTT-AGA-CAC-GAG-GAA-ACC-3'. Assays were performed at least in triplicate, and the mean values were used to calculate expression levels, using the $\Delta\Delta$C(t) method referring to GAPDH housekeeping gene expression.

Notes on contributors

The first case study of Section 4.1 comes from Transgene team thanks to B. Bastien. T. Boukhobza, H. Dumond and C. Thiebaut conducted the biological study of Section 4.2 and performed the functional analysis that follows the covariates selection with `armada`. S. Cairo and O. Déas ; XenTech, Genopole, 91000 Evry (France) provided the TN dataset; A. De Reynies, I. Bieche and J. Métral from the Carte d'Identité des Tumeurs program provided the access to transcriptomic raw data and biological samples (ERa66+, PR+, PR- datasets). The statistical methodology has been developed by A. Gégout-Petit and A. Muller-Gueudin, and thanks to Y. Shi and H. Chakir during their Master internships.

References

[1] Antanavičiūtė I. , Mikalayeva V., Ceslevičienė I., Milašiūtė G., Skeberdis V.A. and Bordel S. (2017). Transcriptional hallmarks of cancer cell lines reveal an emerging role of branched chain amino acid catabolism. *Scientific reports* **7**(1), 7820.

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

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

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

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

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

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

[8] Chamard-Jovenin C., Jung A.C., Chesnel A., Abecassis J., Flament S., Ledrappier S., Macabre C., Boukhobza T. and Dumond H. (2015). From ERα66 to ERα36: a generic method for validating a prognosis marker of breast tumor progression. *BMC systems biology* **9**(1), 28.

[9] Chavent M., Kuentz V., Liquet B. and Saracco J. (2012). Clustofvar: an R package for the clustering of variables. *Journal of Statistical Software* **50**, 91–116.

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

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

[12] Efron B. (2005). *Local false discovery rates*. Division of Biostatistics, Stanford University.

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

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

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

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

[17] Guedj M., Marisa L., De Reynies A., Orsetti B., Schiappa R., Bibeau F., Macgrogan G., Lerebours F., Finetti P., Longy M. *and others*. (2012). A refined molecular taxonomy of breast cancer. *Oncogene* **31**(9), 1196.

[18] Günther O.P., Shin H., Ng R.T., McMaster W.R., McManus B.M., Keown P.A., Tebbutt S.J. and Lê Cao K.A. (2014). 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), 682–695.

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

[20] Le Quesne J. and Caldas C. (2010). Micro-rnas and breast cancer. *Molecular oncology* **4**(3), 230–241.

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

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

[23] Muller-Gueudin A. and Gégout-Petit A. (2019). *armada: A Statistical Methodology to Select Covariates in High-Dimensional Data under Dependence*. R package version 0.1.0.

[24] Reyal F., Guyader C., Decraene C., Lucchesi C., Auger N., Assayag F., De Plater L., Gentien D., Poupon M.F., Cottu P. *and others*. (2012). Molecular profiling of patient-derived breast cancer xenografts. *Breast cancer research* **14**(1), R11.

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

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

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

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

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

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

[31] Thiebaut C., Chamard-Jovenin C., Chesnel A., Morel M., Djermoune E.H., Boukhobza T. and Dumond H. (2017). Mammary epithelial cell phenotype disruption in vitro and in vivo through ERα36 overexpression. *PloS one* **12**(3), e0173931.

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

[33] Wang Z.Y., Zhang X.T., Shen P., Loggie B.W., Chang Y.C. and Deuel T.F. (2005). Identification, cloning, and expression of human estrogen receptor-α36, a novel variant of human

estrogen receptor-α66. *Biochemical and biophysical research communications* **336**(4), 1023–1027.