APPLICATION NOTE (SUPPLEMENTAL MATERIAL)

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

B. Bastien[c], T. Boukhobza[b], H. Dumont[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 August 10, 2020

1. Classification design

1.1. Simulation design

As in Section 3.1 of the Main Article, 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 3.1 of the Main article. More precisions on the factor analysis model can be found in [3]. 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 [5]. 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 1.
- The $m_0 = 340$ remaining covariates of each cluster are independent of Y: $\delta_i^{(k)}(y) = 0$ whatever y for $i = 61, \dots, 400$.

We can remark that this design respects the covariance matrix given in Figure 1 of the Main Article. This design differs a little bit from the model of Equation (1) of the Main Article,

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

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.

1.2. 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 [1, 4], implemented in the R package `FAMT`. This gives a new dataset $\mathbf{X}^{\dagger}_{Y}$.

Procedure 3: the 4 clusters are estimated with the procedure of [2], implemented in the R package `ClustOfVar`; then the covariates are decorrelated in each cluster, taking Y into account, with the factor analysis procedure of [1, 4], 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, $\mathbf{X}^{\dagger}_{Y}$ 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 $N = 500$ 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 1.

If we analyse the results given by Figure 1, 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.

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

In order to describe the performances of our method, we show in Figure 2 the mean ARMADA scores obtained on the $N = 500$ 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 "-").

We can see on the Figure 2 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 3.1 of the Main Article, probably because we are note exactly in the model of the study (Equation (1) of the Main Article) 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.

1.4. 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 [1]: the selected covariates are those with adjusted p-values lower than 0.05.

To compare the three selection methods, the Table 2 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 $N = 500$ 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.

Finally, we can conclude with the ROC curves given in Figure 3 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 $(1\text{-specificity, sensibility})_{\text{ARMADA score}=0}$ and $(1\text{-specificity, sensibility})_{\text{ARMADA score}=1}$. The ROC curves have been obtained by the mean of the $N = 500$ ROC curves obtained in the $N = 500$ runs of $(\mathbf{X}, Y)$.

2. 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 $\tilde{\mathbf{X}} = (\tilde{\mathbf{X}}^{(k)})_{k=1,\dots,4}$ as in Section 1, 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 $k = 1, \dots, 4$, and for all $i = 1, \dots, 400$:

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

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

As in Section 1, we show the interest of our pretreatment, comparing the three procedures detailed in Section 1. As Y is a gaussian variable, we use the Pearson correlation test (instead of the Wilcoxon test used in Section 1). We produce $N = 500$ runs of $(\mathbf{X}, Y)$ and count the number of false and true positive, and the ARMADA scores (shown in Figures 4 and 5).

Similarly to the classification studies presented in Section 1 of this Supplemental Material, and in Section 3 of the Main Article, 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), and the power of our Procedure is comparable with Procedure 2.

The Figure 5 shows the ARMADA scores obtained on these $N = 500$ runs of $(\mathbf{X}, Y)$. Again, similarly to the Section 1, 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 1, the Table 3 and the ROC curve in Figure 6 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.

3. Simulations when the factor modelling fails

We show here the performances of our method to select variable when the factor modelling fails inside the independent clusters. The principle of the simulated model is the following: for each independent block, we generate the latent factors $Z^{(k)}$, some of them have direct influence on the outcome Y and have also influence on a few number of covariates and no influence on the others. The other latent factors creates correlation between all the covariates of the cluster. In path analysis, we will say that there is an indirect correlation from some covariates on the outcome through the latent factors while in the model (1) of the Main Article, the correlation between the influential covariates and the outcome is direct. We describe the simulation design in the following section.

3.1. Simulation design

The idea behind the following model is inspired from biological models: some latent factors govern particular biological functions. We note them $\mathbf{Z} = (Z_1, \ldots, Z_Q)$. The variable Y is directly linked with some factors Z_i, $i \in \{1, \ldots, Q\}$. The covariates $\mathbf{X}$ are also related to $\mathbf{Z}$, via a factor model. More precisely, as in the previous simulation designs, the covariates $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\ldots,4}$ are clustered into four independent clusters, each of them containing $p_k = 400$ covariates. Inside each cluster, the correlation between the covariates of $\mathbf{X}^{(k)}$ is designed by a factor analysis model described in Equation (2) of the Main Article. The dimension of the latent factors $\mathbf{Z}^{(k)}$ in each cluster are $(q_1, \ldots, q_4) = (4, 6, 8, 10)$. Then, the whole set of latent factors $\mathbf{Z} = (\mathbf{Z}^{(k)})_{k=1,\ldots,4}$ is composed of $Q = \sum_k q_k = 28$ latent factors. 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) of the Main Article). Now, our simulation design is the following:

$$Y = \sum_{k=1}^{4} \sum_{i=1}^{2} Z_i^{(k)} \tag{2}$$

$$X_i^{(k)} = \mathbf{b}_i^{(k)} \mathbf{Z}^{(k)} + \epsilon_i^{(k)}, \qquad \text{for } i = 1, \ldots, p_k \tag{3}$$

where $\mathbf{b}_i^{(k)}$ is a q_k vector, corresponding to the ith row of a $p_k \times q_k$ matrix $\mathbf{B}^{(k)}$ simulated as in the previous simulation study, but with the following contraints:

$$\mathbf{B}^{(k)} = \left(\begin{array}{ccccc} 1.5 & 1.5 & B_{1,3} & \dots & B_{1,q_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1.5 & 1.5 & B_{5,3} & \dots & B_{5,q_k} \\ 1 & 1 & B_{6,3} & \dots & B_{6,q_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & B_{10,3} & \dots & B_{10,q_k} \\ 0.75 & 0.75 & B_{11,3} & \dots & B_{11,q_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.75 & 0.75 & B_{15,3} & \dots & B_{15,q_k} \\ 0.5 & 0.5 & B_{16,3} & \dots & B_{16,q_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0.5 & 0.5 & B_{20,3} & \dots & B_{20,q_k} \\ \hline 0 & 0 & B_{21,3} & \dots & B_{21,q_k} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & B_{p_k,3} & \dots & B_{p_k,q_k} \end{array}\right) \tag{4}$$

In other words, Y is linked with two latent factors of each cluster, noted $Z_1^{(k)}$ and $Z_2^{(k)}$. Whereas, in each cluster, $Z_1^{(k)}$ and $Z_2^{(k)}$ are highly linked with the covariates $X_i^{(k)}$ for $i = 1, \dots, 20$ (and the intensity of the link decreases with i), and are independant of the other covariates $X_i^{(k)}$ for $i = 21, \dots, p_k$. We then have an indirect link between Y and the covariates $\left(X_i^{(k)}\right)_{i=1,\dots,20;k=1,\dots,4}$, through the latent factors $\left(Z_1^{(k)}, Z_2^{(k)}\right)_{k=1,\dots,4}$. Clearly, this simulation design differs from the model (1) of the Main Article.

3.2. Results

Figures 7 show the interest of the pretreatment of the covariates, and again, our pretreatment is a good compromise between the true positive (TP) rate (that is better than that of Procedure 1) and the False Positive rate (that is better than that of Procedure 2). Figures 8 show the scores of the covariates: the median score is around four and six for the group of significants covariates and is zero for the covariates with no link with the outcome. We see that the variability of the scores is larger when the link with the outcome is lower. The ROC curve and Table 4 show that our method is not the best one but it stays specific with a score of two.

4. Simulations when the gaussian hypothesis fails

In this section we investigate another scenario when the modelling fails because of the gaussian hypothesis of the factors. We propose a Student with three degrees of freedom for each of them, it is known to have a heavy tail.

4.1. Simulation design

We simulate $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\dots,4}$ with four independent clusters, each of them containing $p_k = 100$ covariates. As in Section 2.1 of the main article, the dependence in each cluster of covariates is

modeled with a factor model:

$$X_i^{(k)} = \delta_i Y + \mathbf{b}_i^{(k)} \mathbf{Z}^{(k)} + \varepsilon_i^{(k)}, \qquad \text{for } i = 1, \ldots, p_k, \tag{5}$$

but now the factors $\mathbf{Z}^{(k)}$ are a q_k-Student centered vector with covariance matrix given by the identity matrix $S = I_{q_k}$, and with 3 degrees of freedom. The number of factors in the four clusters are respectively $q_1 = 4, q_2 = 6, q_3 = 8, q_4 = 10$. As in our preceding simulation in the regression case, Y is a standard gaussian variable. The dependence between $\mathbf{X}$ and the outcome Y is given by $\delta = (\delta_i)_{i=1,\ldots,400} = (1, 0.8, 0.6, 0.4, 0.2, 0, 0, \ldots, 0)$. In other words, only the first 5 covariates of each cluster are linked with Y.

4.2. Results

The results are similar as those obtained in a normal case (cf Figures 10, 11, 12 and Table 5). This first result is encouraging to apply our method, even on non-normal data.

5. Check of the covariance structure

In this section, we propose a way to check that the data support the covariance structure given in Figure 1 and by Equations (2-3) of the Main Article. One way to check the covariance structure is to map the heatmap of the data: before clustering, after clustering, and after decorrelation into clusters by FAMT. We give here a pedagogical exemple, with only $p = 880$ covariates, which correspond to peptide data, observed on $n = 62$ patients with non-small cell lung cancer, who received a treatment developed by the Transgene society. The variable of interest is Y: the outcome of the treatment. Among the 62 patients who received treatment, 27 died before 12 months ($Y = 1$), and 35 survived to 12 months ($Y = 0$). In that study, our method detects 4 clusters. We can see on the Figure 13 that the peptide data respect our proposed correlation structure.

In our article, the dimension of the datasets of the two proposed applications (Sections 4.1 and 4.2 of the Main Article) are of the order of 50,000 covariates, and then of the order of 6,000 after an initial filtering. It is impossible to represent a heatmap of the correlation matrix for such a large dimension. That is why, and only to illustrate (even it is partial) the correlation structure of our data, we sub-sampled in the covariates, in order to reduce the dimension: for the first application (Section 4.1), we took 800 covariances at random, among the 6,810 covariates of the Transgene dataset. On that smaller dataset, our method detects two clusters. The heatmaps of the correlation matrix are given in the Figure 14. As for the peptide dataset, we can check the correlation structure by clusters. For the second application (Section 4.2), we took the 277 covariates discussed in Figure 8 of the Main Article, to illustrate the correlation structure for the ERa66+ dataset. The heatmaps of the correlation matrix are given in the Figure 15. Again, we can check the correlation of the data, structured by clusters.

6. 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 16). 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 6 and 7).

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.

7. 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:

- the TN dataset corresponding to Affymetrix transcriptomic comprehensive data from 17 patients derived xenografts (PDX) breast tumors was extracted from the Xentech^{TM} database with the permission of Olivier Déas and Stefano Cairo (MTA CXT-295 Xentech SAS/University of Lorraine ; [7]).
- 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 [6]. 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 [8]. 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.

References

[1] 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.

[2] 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.

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

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

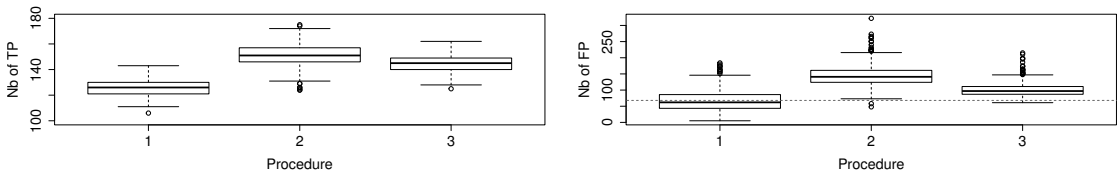

Figure 1. Number of true positive tests (top), false positive tests (bottom) in the classification design according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 500$ runs.

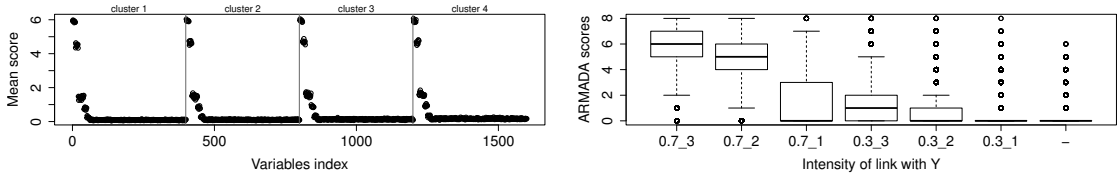

Figure 2. Top: mean of the ARMADA scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 500$ runs. Simulation in the classification design.

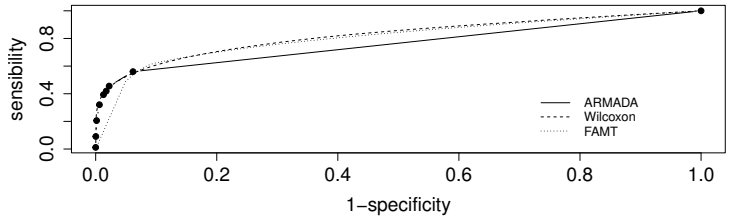

Figure 3. ROC curves for the three selection methods in the classification design. The ROC curves have been obtained by the mean of the $N = 500$ ROC curves obtained in the $N = 500$ runs of $(\mathbf{X}, Y)$.

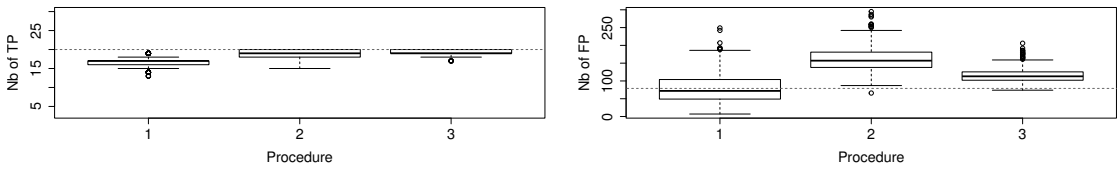

Figure 4. Number of: true positive tests (top), false positive tests (bottom) in the regression design. Boxplots are calculated on $N = 500$ runs.

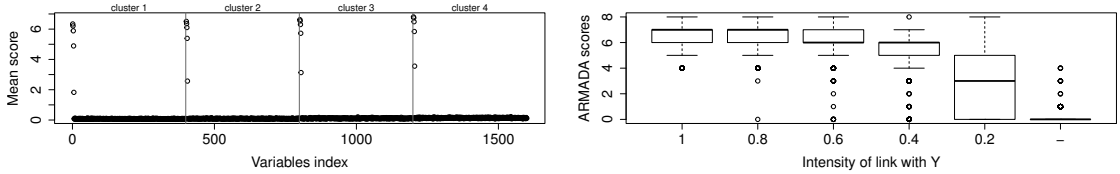

Figure 5. Top: mean of the ARMADA scores obtained by all the covariates. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 500$ runs. Simulation in the regression design.

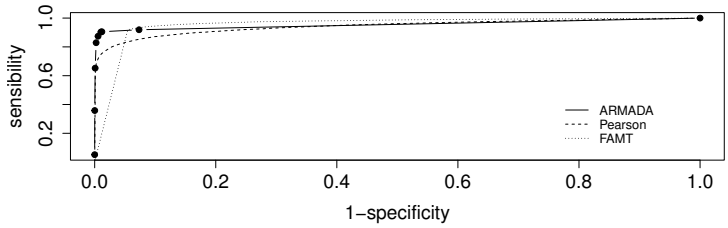

Figure 6. ROC curves for the three selection methods, in the case of regression design. The ROC curves have been obtained by the mean of the $N = 500$ ROC curves obtained in the $N = 500$ runs of $(\mathbf{X}, Y)$.

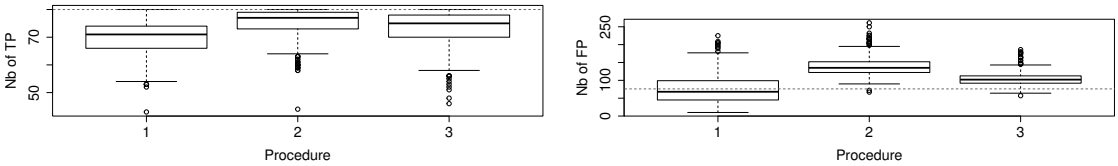

Figure 7. Number of: true positive tests (top), false positive tests (bottom) in the "out of factor modelling" design. Boxplots are calculated on $N = 500$ runs.

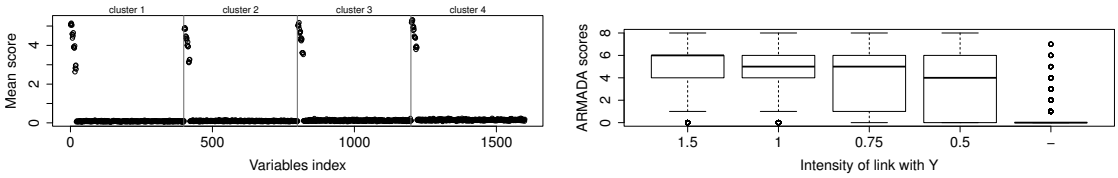

Figure 8. Top: mean of the ARMADA scores obtained by all the covariates in the "out of factor modelling" design. Bottom: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 500$ runs. Simulation in the regression design.

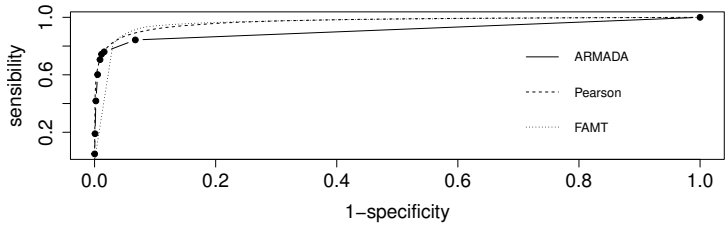

Figure 9. ROC curves for the three selection methods, in the "out of factor modelling" design. The ROC curves have been obtained by the mean of the $N = 500$ ROC curves obtained in the $N = 500$ runs of $(\mathbf{X}, Y)$.

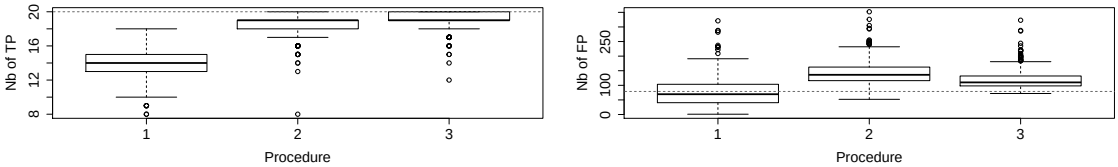

Figure 10. Number of true positive tests (left), false positive tests (right) according to the different pretreatment procedures (1: Nothing, 2: FAMT, 3: clustering followed by FAMT in each cluster). Dotted lines: expected number of FP. Boxplots are calculated on $N = 500$ runs. Regression design, with non-normal covariates (factors $\mathbf{Z}$ are centered Gamma vectors).

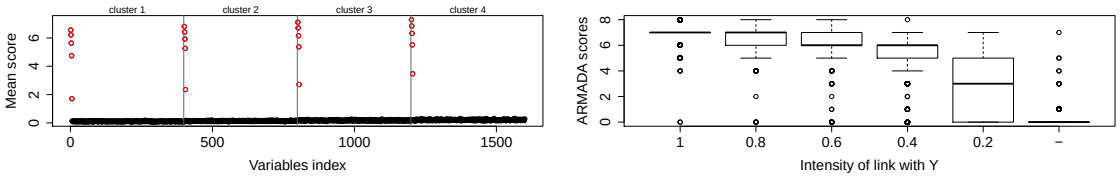

Figure 11. Left: mean of the ARMADA scores obtained by all the covariates. Right: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 500$ runs. Regression design, with non-normal covariates (factors $\mathbf{Z}$ are Student vectors).

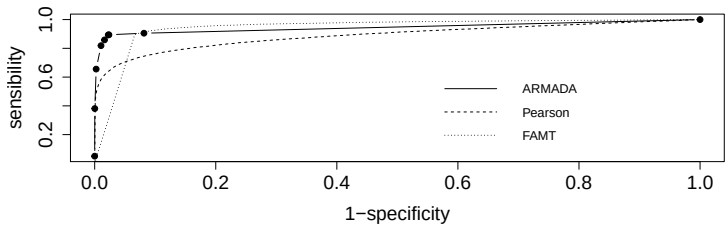

Figure 12. ROC curves for the three selection methods. The ROC curves have been obtained by the mean of the $N = 500$ ROC curves obtained in the $N = 500$ runs of $(\mathbf{X}, Y)$. Regression design, with non-normal covariates (factors $\mathbf{Z}$ are Student vectors).

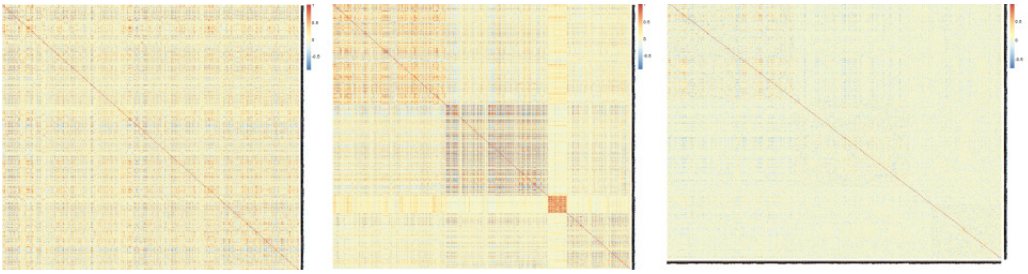

Figure 13. Heatmap of the correlation matrix of the 880 peptide covariates: before clustering, i.e. on the original dataset (left), after clustering and sorting the covariates according to their clusters (middle), and after decorrelation into the clusters (right).

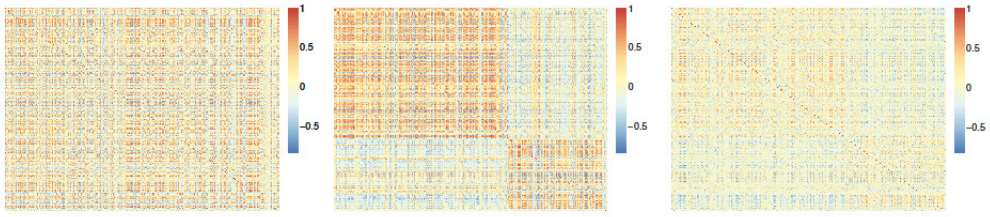

Figure 14. Heatmaps of the correlation matrix of 800 covariates of the Transgene dataset: before the application of our procedure (left), after clustering and sorting the covariates according to their clusters (middle), and after decorrelation into the clusters (right).

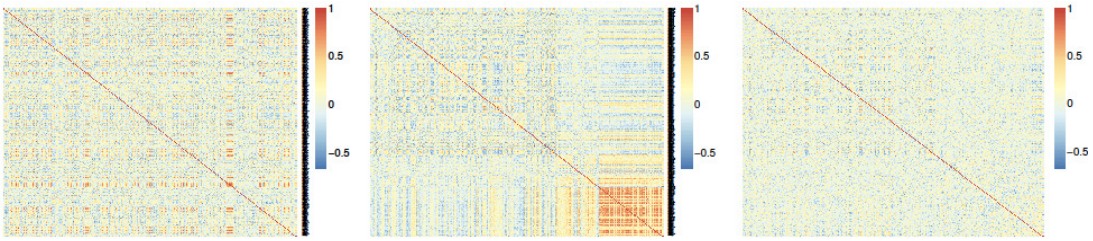

Figure 15. Heatmaps of the correlation matrix of 277 covariates of the ERa66+ dataset: before the application of our procedure (left), after clustering and sorting the covariates according to their clusters (middle), and after decorrelation into the clusters (right).

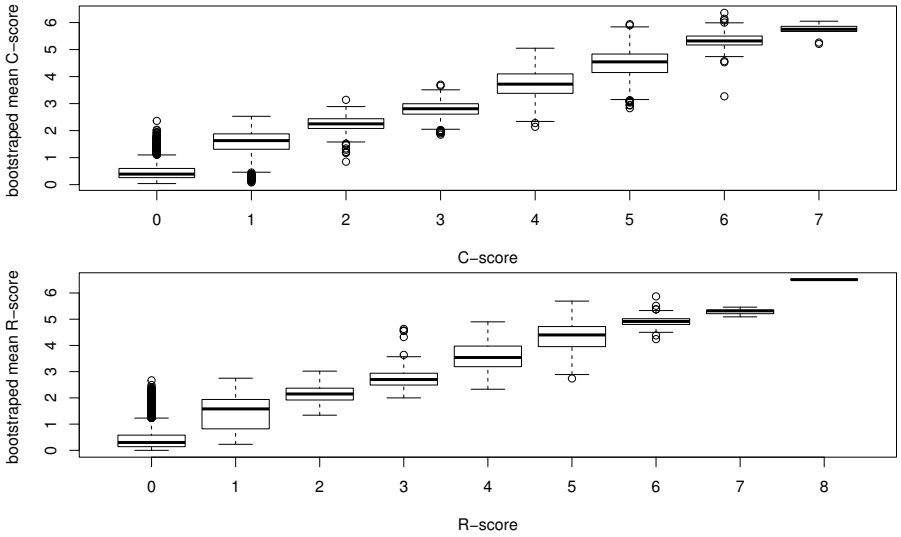

Figure 16. Distribution of the bootstraped mean of C-scores (resp. R-scores), i.e. means of C-(or R-)scores obtained on $B = 100$ bootstrap samples, according to the corresponding C-scores (resp. R-scores) in the original dataset for all the $p = 6810$ covariates.

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

Table 2. Results of the $N = 500$ 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 (0.08)	0.99 (0.04)
(0.7-2)	0.92 (0.27)	0.92 (0.27)	0.96 (0.17)
(0.7-1)	0.45 (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)	0.33 (0.47)	0.28 (0.45)	0.42 (0.49)
(0.3-1)	0.13 (0.32)	0.12 (0.32)	0.20 (0.40)
-	0.06 (0.24)	0.05 (0.22)	0.10 (0.31)

Table 3. Results of the $N = 500$ 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 (0.02)	1 (0)	1 (0)
0.6	0.99 (0.06)	0.99 (0.08)	1 (0.02)
0.4	0.97 (0.17)	0.84 (0.36)	0.98 (0.12)
0.2	0.63 (0.48)	0.32 (0.47)	0.74 (0.44)
-	0.07 (0.26)	0.05 (0.22)	0.10 (0.30)

Table 4. Results of the $N = 500$ runs in the "out of factor modelling" 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
1.5	0.92 (0.26)	0.96 (0.17)	0.97 (0.15)
1	0.88 (0.32)	0.93 (0.24)	0.96 (0.18)
0.75	0.83 (0.37)	0.87 (0.33)	0.94 (0.24)
0.5	0.73 (0.44)	0.69 (0.46)	0.86 (0.34)
-	0.06 (0.25)	0.05 (0.22)	0.09 (0.29)

Table 5. Results of the $N = 500$ runs in the non-normal 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 (0.03)	0.98 (0.13)	1 (0.02)
0.8	1 (0.04)	0.95 (0.22)	1 (0.05)
0.6	0.99 (0.10)	0.84 (0.36)	0.99 (0.06)
0.4	0.95 (0.21)	0.55 (0.50)	0.96 (0.19)
0.2	0.59 (0.49)	0.19 (0.39)	0.64 (0.48)
-	0.08 (0.27)	0.04 (0.22)	0.09 (0.28)

Table 6. 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 7. 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