A methodology to select and rank covariates in high-dimensional data under dependence

Aurélie Muller-Gueudin [1] & Anne Gégout-Petit [1]

[1] *IECL, INRIA, Université Lorraine, Nancy, UMR 7502, France;*
aurelie.gueudin@univ-lorraine.fr, anne.gegout-petit@univ-lorraine.fr.

Résumé. Nous proposons une méthode de sélection et de tri de covariables associées à une variable d'intérêt, dans le cadre de données dépendantes et de grande dimension, mais avec peu d'observations. Une première étape consiste à décorréler les covariables: après avoir effectué un clustering des covariables, nous décorrélons les covariables de chaque cluster via l'analyse en facteurs latents. La seconde étape sélectionne et trie les covariables en utilisant une agrégation de méthodes et tests statistiques. Après quelques simulations, nous appliquons notre méthode sur des données transcriptomiques ($p = 6810$ covariables) de $n = 37$ patients atteints d'un cancer du poumon, et ayant reçu un traitement. Notre méthode permet de sélectionner les covariables liées à la réussite ou non du traitement. Nous obtenons différents profils de patients suivant leur temps de survie.

Mots-clés. Sélection de variables correlées, Grande dimension, Tests multiples, Agrégation de méthodes, Classement, Profils génétiques.

Abstract. We propose a methodology to select and rank covariates associated to a variable of interest in a context of high-dimensional data under dependence but few observations. The methodology imbricates successively clustering of covariates, decorrelation of covariates using Factor Latent Analysis, selection using aggregation of adapted methods and finally ranking. After a simulation study, we apply our method on transcriptomic data of $n = 37$ patients with lung cancer, who have received chemotherapy. Our method selects the covariates that are the most linked with the outcome treatment among a set of 6810 transcriptomic covariates. We obtain different transcriptomic profiles of patients according to their survival time.

Keywords. Correlated covariates selection, High dimension, Multiple testing procedures, Aggregated methods, Ranking, Genetic profiles.

1 Introduction

We consider the problem to detect association between a variable of interest Y and p correlated covariates in a high dimensional dataset $\mathbf{X}$. Moreover, we are in a context of small sample size ($n \ll p$). Many statistical methods exist to select covariates in high-dimensional contexts: the control of false discoveries in multiple testing procedures is

very highly studied and many methods of regression are available (lasso (see Tibshirani *et al* (1996)), random forests (see Genuer *et al* (2010), ...).

Moreover, one way to deal with dependence is to model it by latent factors: Friguet *et al* (2009) propose a way to correct the data according to a regression link with the variable of interest Y in such a way that the corrected covariates are independent conditionally to Y. This method of correction is called FAMT correction (for Factor Analysis for Multiple Testing).

However, the framework of FAMT is to consider the data $\mathbf{X}$ as one block of correlated covariates and has to be adapted if $\mathbf{X}$ is structured in several independent clusters of correlated covariates. The FAMT does not give good results if the decomposition in independent clusters is not taken into account. We propose to identify the clusters of correlated covariates before performing FAMT correction on each of the clusters.

2 Methodology

2.1 Framework and model

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

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

On one hand, we model the dependence in each cluster as in Friguet *et al* (2009): inside each cluster $\mathbf{X}^{(k)}$, the common information between the p_k covariates is modeled by regression on a small set of q_k latent factors $\mathbf{Z}^{(k)}$:

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

where $\mathbf{Z}^{(k)}$ is a random centered q_k-vector such that $\mathbb{E}(\mathbf{Z}^{(k)}\mathbf{Z}^{(k)\prime}) = I_{q_k}$, $b_i^{(k)}$ is a q_k-vector, and $\varepsilon^{(\mathbf{k})} = (\varepsilon_1^{(k)}, \ldots, \varepsilon_{p_k}^{(k)})$ is a random centered p_k-vector with independent components, and independent of $\mathbf{Z}^{(k)}$. The covariance matrix of the data $\mathbf{X}^{(k)}$ conditionally to Y, is given by:

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

where $\Psi^{(k)}$ is a diagonal $p_k \times p_k$ matrix (the covariance matrix of $\varepsilon^{(k)}$) and $B^{(k)} = \left(b_i^{(k)}\right)_{i=1,\ldots,p_k}$ is a $p_k \times q_k$ matrix of factor loadings. Therefore, $B^{(k)}(B^{(k)})'$ appears as the shared variance in the common factor structure, and Friguet *et al* (2009) define the common variance by

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

On the other hand, we suppose that the informations specific at each cluster (that is vectors $(\mathbf{Z}^{(k)}, \varepsilon^{(k)})_{1 \leq k \leq K}$) are independent.

2.2 Main prodecure

2.2.1 Step 1: pretreatment of data

The aim is to perform a decorrelation of the covariates $\mathbf{X}$, to obtain corrected covariates $\mathbf{X}^\star$ that are suitable for testing and/or regression. The whole vector $\mathbf{X}$ satisfies assumption of Equation (1), and Friguet *et al* (2009) apply the decorrelation procedure (FAMT) on the whole set $\mathbf{X}$. But we propose to first detect the different clusters $(\mathbf{X}^{(k)})_{1\leq k\leq K}$ and then to apply the FAMT procedure on each cluster. Indeed, some simulation studies of Bastien *et al* (2018) have shown that the decorrelation was degraded by the dimension of the vector of covariates, whereas it was better after the detection of the independent clusters. By this way, the covariates selection procedure can be highly improved by clustering of covariates before applying factor analysis to correct the correlation within each cluster. In practice, we use `ClustOfVar` algorithm of Chavent *et al* (2012) to cluster covariates into homogeneous clusters. 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. Note that this correction of the data $\mathbf{X}$ is done conditionally on the variable of interest Y. At the end of this pretreatment procedure, we obtain corrected data, noted $\mathbf{X}^*$ in the sequel.

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

We perform L methods, then each covariate $X_j^\star$ obtains a score $S_j \in \{0, 1, \ldots, L\}$ 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. If $S_i = L$, then the corresponding variable has been selected by each of the L methods, whereas if $S_i = 0$, the corresponding variable has been selected by none of them.

In the examples proposed in the simulation studiy and in real data, Y is binary 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 Genuer *et al* (2010)). In the sequel, we call our procedure `ARMADA` for AggRegated Methods for covAriates selection under Dependence.

3 Simulation study

We consider a classification problem with $p = 1600$ covariates and sample size $n = 60$, where $Y = 1$ for $\frac{n}{2}$ subjects, and $Y = 0$ for $\frac{n}{2}$ subjects. This simulation design is inspired from Friguet *et al* (2009). One design in a regression case is given in Bastien *et al* (2018).

The covariates $\mathbf{X} = (\mathbf{X}^{(k)})_{k=1,\ldots,4}$ are clustered into four clusters of 400 covariates, which are independent conditionally to Y. For each cluster k, we generate $\tilde{\mathbf{X}}^{(k)}$ that is a

gaussian centered 400-vector. The correlation $\Sigma^{(k)}$ of $\tilde{\mathbf{X}}^{(k)}$ is designed by Equation (2), and we take high common variances (defined in Equation (3)) $\text{ComVar}^{(k)}$ equal to 0.8 in each cluster. The numbers of latent factors in each cluster are $(q^{(1)}, \ldots, q^{(4)}) = (4, 6, 8, 10)$.

Y is linked with 160 influential covariates in $\mathbf{X}$, the others being noise covariates. Y is the most strongly linked with the 10 first covariates of each cluster, and the strength of the link is decreasing in the successive groups of 10 influential covariates.

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

3.1 Interest of our data pretreatment

We compare the results of a Wilcoxon test after three different data pretreatments:

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

Procedure 2: the covariates $\mathbf{X}$ are decorrelated with the factor analysis procedure `FAMT` of Friguet *et al* (2009), taking Y into account.

Procedure 3: the 4 clusters are estimated with `ClustOfVar`; then the covariates are decorrelated in each cluster, taking Y into account, with `FAMT`. It is our data pretreatment.

We perform Wilcoxon tests on each of the p pretreated covariates of the dataset (to compare groups $Y = 0$ and $Y = 1$), this gives a three sets of p p-values. For each procedure, the selected covariates are those with p-values lower than 0.05. We apply these procedures on $N = 100$ runs of $(\mathbf{X}, Y)$. TP is the number of influential covariates that are correctly detected and FP the number of non-influential detected covariates. As shown in Figure 1, the Procedure 1 is the poorest. Our Procedure reduces the mean and the variability of the FP. The power of our Procedure is comparable with Procedure 2. This results show the interest of our proposed pretreatment before performing selection.

3.2 Results of the whole method (pretreatment and selection)

Figure 2 shows 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 "1.5", "1", "0.75", "0.5"; and the group of noise covariates is noted by "-"). The scores give a clear ranking of the covariates, according to the strength of their link with Y. The distribution of the

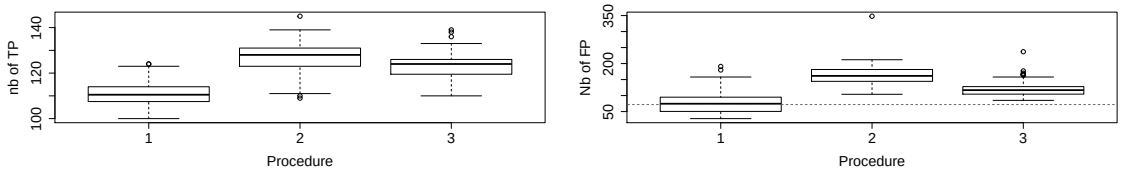

Figure 1: Number of TP (left), FP (right) according to the different pretreatment procedures. Dotted lines: expected number of FP. Boxplots are calculated on $N = 100$ runs.

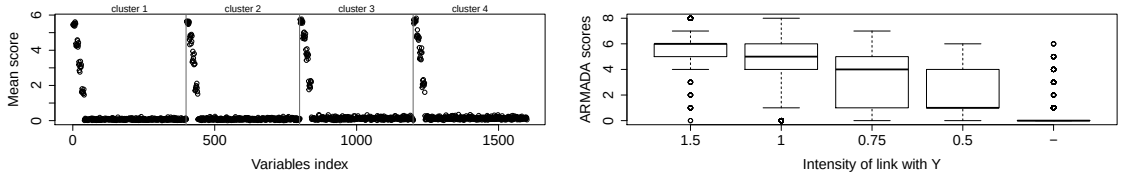

Figure 2: Left: mean of the `ARMADA` scores obtained by all the covariates. Right: boxplot of the scores of the covariates, ranked by levels of link with Y. Means and boxplots are calculated on $N = 100$ runs.

individual scores inside each group is given by the boxplots. The 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. Around 95% of the noise covariates obtained a `ARMADA`-score that was exactly 0.

We compare `ARMADA` with 2 other selection methods: the Wilcoxon test (the selected covariates are those with raw-pvalues lower than 0.05), and the `FAMT` procedure of Friguet *et al* (2009): the selected covariates are those with adjusted p-values lower than 0.05. The ROC curves given in Figure 3 shows 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). The ROC curves have been obtained by the mean of the $N = 100$ ROC curves obtained in the $N = 100$ runs of

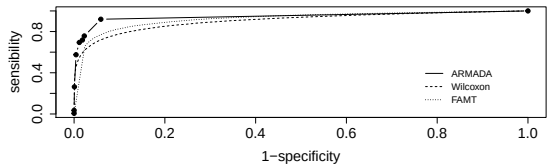

Figure 3: ROC curves for the three selection methods.

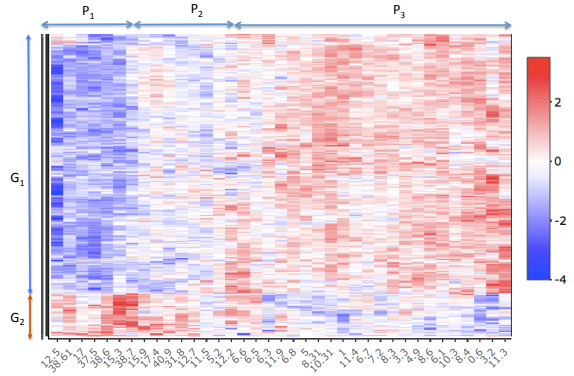

Figure 4: Heatmap of the covariates which have `ARMADA` scores greater or equal to 5. Each column corresponds to one patient. The x-axis represents the patients (marked with their survival time) and the y-axis represents the covariates.

$(\mathbf{X}, Y)$.

4 Real data

We have $p = 6810$ transcriptomic covariates of $n = 37$ patients with lung cancer, who have received chemotherapy. 24 (resp. 13) patients died before (resp. after) 12 months. The question is to find the genes which can explain a survival time greater or lower than 12 months. The select genes by our Procedure are shown in Figure 4. We obtain different transcriptomic profiles of patients according to their survival time.

Bibliographie

Tibshirani, R. (1996), Regression shrinkage and selection via the lasso, *Journal of the Royal Statistical Society. Series B.* pp. 267-288.
Genuer, R., Poggi, J.M. and Tuleau-Malot, C. (2010), Variable selection using random forests, *Pattern Recognition Letters*, 31 (14), pp. 2225-2236.
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).
Bastien, B., Chakir, H., Gégout-Petit, A., Muller-Gueudin, A. and Shi, Y. (2018) A statistical methodology to select covariates in high-dimensional data under dependence. Application to the classification of genetic profiles associated with outcome of a non-small-cell lung cancer treatment, *https://hal.archives-ouvertes.fr/hal-01939694*.
Chavent, M., Kuentz, V., Liquet, B. and Saracco, J. (2012), ClustOfVar: an R package for the clustering of variables, *Journal of Statistical Software*, 50,.