Combining clustering of variables and feature selection using random forests

Marie Chavent, Robin Genuer, Jerome Saracco

▶ **To cite this version:**

Marie Chavent, Robin Genuer, Jerome Saracco. Combining clustering of variables and feature selection using random forests. 2016. hal-01345840v3

HAL Id: hal-01345840
https://hal.science/hal-01345840v3

Preprint submitted on 6 Nov 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Combining clustering of variables and feature selection using random forests

Marie Chavent[1,2], Robin Genuer[3,4], Jérôme Saracco[1,2]

[1] Institut de Mathématiques de Bordeaux, UMR CNRS 5251, Université de Bordeaux, 351 cours de la libération, 33405 Talence Cedex, France.

[2] INRIA Bordeaux Sud-Ouest, CQFD team, France.

[3] ISPED, INSERM U-1219, Université de Bordeaux, 146 rue Léo Saignat, 33076 Bordeaux, France.

[4] INRIA Bordeaux Sud-Ouest, SISTM team, France.

e-mail : Marie.Chavent,Jerome.Saracco@math.u-bordeaux.fr, Robin.Genuer@u-bordeaux.fr

Abstract

Standard approaches to tackle high-dimensional supervised classification problem often include variable selection and dimension reduction procedures. The novel methodology proposed in this paper combines clustering of variables and feature selection. More precisely, hierarchical clustering of variables procedure allows to build groups of correlated variables in order to reduce the redundancy of information and summarizes each group by a synthetic numerical variable. Originality is that the groups of variables (and the number of groups) are unknown a priori. Moreover the clustering approach used can deal with both numerical and categorical variables (i.e. mixed dataset). Among all the possible partitions resulting from dendrogram cuts, the most relevant synthetic variables (i.e. groups of variables) are selected with a variable selection procedure using random forests. Numerical performances of the proposed approach are compared with direct applications of random forests and variable selection using random forests on the original p variables. Improvements obtained with the proposed methodology are illustrated on two simulated mixed datasets (cases $n > p$ and $n < p$, where n is the sample size) and on a real proteomic dataset. Via the selection of groups of variables (based on the synthetic variables), interpretability of the results becomes easier.

Keywords: *clustering of variables; random forests; supervised classification; variable selection*

1 Introduction

This paper addresses the problems of dimension reduction and variable selection in the context of supervised classification. In this framework, there are often two objectives : prediction (to be able to predict classes associated to new observations) and feature selection (to extract the most interesting variables). Typically in a medical context, the first goal could be to succeed in predicting if patients will or will not respond well to a treatment, given e.g. their gene expression profile, whereas the second aim could be to determine which part of the genome is responsible for the good or bad response to the treatment. These two objectives can be related, because it may be easier to perform prediction if useless variables have already been eliminated. Those goals are indeed especially relevant when dealing with e.g. genomics or proteomics data, where the number p of variables largely exceeds the number n of available observations.

A classical way of addressing such issues is to use a variable selection technique (see e.g. Guyon and Elisseeff, 2003; Tibshirani, 1996; Zou and Hastie, 2005). The hope is that the method will be able to select the most interesting variables, while preserving good prediction performances. Moreover, data often come with many highly correlated variables and succeeding in selecting all variables in a group of correlated variables can be very difficult. Even if obtaining all these variables does not always appear to be interesting for prediction purpose, it can be useful for interpretation purpose, depending on the application.

We stress that the framework considered in this work is the *supervised* classification framework, in which an output variable is observed. Hence, we do not address the problem of variable selection in a clustering (or unsupervised classification) context as e.g. in Raftery and Dean (2006); Tadesse et al. (2005); Maugis et al. (2009).

In this article, we propose and evaluate a novel methodology for dimension reduction and variable selection, which combines clustering of variables and feature selection using random forests. The clustering of variables method, introduced in (Chavent et al., 2012a) and denoted by **CoV** hereafter, allows to eliminate redundancy of explanatory variables. This clustering approach can deal with both numerical and categorical variables (i.e. mixed dataset). Note that this method clusters the *variables*, which differs from the (more classical) problem of *individuals* clustering (Kaufman and Rousseeuw, 1990; Gordon, 1999). The clustering of variables groups together highly correlated variables and provides for each group (cluster) a synthetic variable which is a numerical variable summarizing the variables within the cluster. The main advantage of this approach is to eliminate redundancy and to keep all the variables together in a cluster during the rest of the analysis. Moreover it reduces the dimension of the data by replacing the p original variables by K synthetic variables (where K denotes the selected number of clusters). In the proposed methodology the number K of synthetic variables is optimized according to prediction performance of a random forest

classifier trained with those synthetic variables. Let K^* denote the optimal number of clusters of variables. Note that the reduction of dimension provides K^* synthetic variables which only use the original variables within the selected clusters, unlike the principal components in principal component analysis (PCA) (Jolliffe, 2002). Hence, in **CoV**, an original variable takes action in the construction of a unique synthetic variable, which makes the interpretation easier. Once the dimension reduction is done, the most important synthetic variables are then selected using a procedure based on random forests (**RF**), introduced in Genuer et al. (2010). This variable selection procedure, denoted **VSURF** hereafter, is applied to the reduced dataset consisting of the n observations described with the K^* synthetic variables, and leads to provide $m \leq K^*$ relevant synthetic variables. Thus the prediction for new observations can be done with a predictor built on these m selected synthetic variables (i.e. a list of clusters of variables). Hence, the output of the proposed method is a classifier and a set of selected variables with the additional information of the group structure of those variables. This additional information can be of great interest depending of the application, and as far as we know, this type of methodology has not already been introduced in the literature.

Note that the proposed approach *does not* require definition of a priori groups of variables. Hence, it is different from (sparse) group lasso or (sparse) group partial least squares (PLS) approaches (Yuan and Lin, 2006; Simon et al., 2013; Liquet et al., 2015). In addition, our method also differs from sparse PLS techniques (Chun and Keleş, 2010; Lê Cao et al., 2011), which perform dimension reduction and variable selection without any group structure information.

In the proposed methodology combining **CoV** and **VSURF**, the synthetic variables are linear combinations of the variables within a group, and the **VSURF** procedure is purely non parametric. This avoids to make assumptions between the response variable and synthetic variables (which are unknown at the beginning of the algorithm). Moreover, since we heavily use the classifier (to choose the optimal number K^* of synthetic variables and to select the m most important of them), **RF** is chosen as an easy-to-use and well-performing algorithm (Fernández-Delgado et al., 2014).

Finally, let us stress that the main objective of this paper is to introduce a novel procedure which

- selects groups of informative variables,
- can deal with mixed data,
- provides good numerical performance,
- and eases interpretation of results.

The rest of the paper is organized as follows. Sections 2.1 and 2.2 give an overview of both the clustering of variables method and the feature selection procedure. The methodology combining **CoV** and **VSURF** is

then described in Section 2.3. In Section 3, the numerical performances of this methodology are compared, via a simulation study, with a straightforward applications of **VSURF** or **RF** (on the original p variables) and with application of **RF** on all the synthetic variables given by **CoV**. Two cases are addressed: $n > p$ and $n < p$. A real proteomic dataset is analyzed in Section 4. Finally, let us mention that the proposed methodology is available and implemented in the R package `CoVVSURF`[1] including the simulation procedure exhibited in Section 3.

2 Description of the methodology

First, the two underlying methods, **CoV** for unsupervised dimension reduction and **VSURF** for both variable selection and prediction, are presented respectively in Sections 2.1 and 2.2. Then, the proposed methodology combining **CoV** and **VSURF**, named **CoV/VSURF** hereafter, is described in Section 2.3.

Let us consider a p-dimensional explanatory variable $X = (X^1, \ldots, X^j, \ldots, X^p)'$ and a univariate response variable Y, which takes its values in $\{1, \ldots, L\}$. Let n be the number of observations of these variables. More precisely, let us consider a set of p_1 numerical variables measured on the n observations denoted by $\{\mathbf{x}^1, \ldots, \mathbf{x}^{p_1}\}$, and a set of p_2 categorical variables denoted by $\{\tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^{p_2}\}$ with $p_1 + p_2 = p$. Let $\mathbf{X} = [\mathbf{x}^1, \ldots, \mathbf{x}^{p_1}, \tilde{\mathbf{x}}^1, \ldots, \tilde{\mathbf{x}}^{p_2}]$ be the corresponding data matrix of dimension $n \times p$. The i-th row of $\mathbf{X}$ is denoted $\mathbf{x}_i$. Let $\mathbf{y}$ be the vector of the n observations of the response variable.

2.1 Clustering of variables

The objective of clustering of variables is to sort variables into homogeneous clusters, that is to construct clusters of variables which are strongly related to each other and thus provide similar information. The idea is to summarize all the variables belonging to a cluster by a synthetic numerical variable which is the most "linked" to all the variables within this cluster. In this section, we focus on a clustering method based on the principal component analysis method **PCAmix** (Kiers, 1991; Chavent et al., 2012b) defined for a mixture of categorical and numerical variables. More precisely, we present the ascendant hierarchical clustering algorithm implemented in the R (R Core Team, 2017) package `ClustOfVar` (Chavent et al., 2012a).

Note that like PCA, **CoV** is a dimension reduction method but, contrary to PCA, it can be helpful for variable selection. Indeed, each synthetic variable of **CoV** is a linear combination of a subset of variables (the variables within the corresponding cluster) whereas the principal components in PCA are linear combination of all the original variables. Hence, selecting synthetic variables of **CoV** means selecting of subsets of original variables, which is not the case when selecting principal components in PCA.

[1] `https://github.com/robingenuer/CoVVSURF`

Synthetic variable of a cluster C_k

This variable is defined as the numerical variable $\mathbf{f}^k \in \mathbb{R}^n$ which is the "most linked" to all the variables in C_k:

$$\mathbf{f}^k = \arg\max_{\mathbf{u}\in\mathbb{R}^n} \left\{ \sum_{\mathbf{x}^j \in C_k} r^2_{\mathbf{u},\mathbf{x}^j} + \sum_{\tilde{\mathbf{x}}^j \in C_k} \eta^2_{\mathbf{u}|\tilde{\mathbf{x}}^j} \right\}, \tag{1}$$

where $r^2_{\mathbf{u},\mathbf{x}^j} \in [0,1]$ is the squared Pearson correlation between the numerical variables $\mathbf{u}$ and $\mathbf{x}^j$, and $\eta^2_{\mathbf{u}|\tilde{\mathbf{x}}^j} \in [0,1]$ is the correlation ratio between $\mathbf{u}$ and the categorical variable $\tilde{\mathbf{x}}^j$ (which measures the part of the variance of $\mathbf{u}$ explained by the levels of $\tilde{\mathbf{x}}^j$). It has been shown that:

- $\mathbf{f}^k$ is the first principal component of the **PCAmix** method applied to the variables in C_k.
- $\mathbf{f}^k$ is a linear combination of the numerical variables and of the dummy (indicator) variables of the levels of the categorical variables in C_k. This linear combination can be used to predict the value (score) of a new observation on the synthetic variable of C_k.

Details on the **PCAmix** method and on the prediction of principal component scores can be found in the Appendix. Note that the **PCAmix** algorithm is implemented in the R package `PCAmixdata` (Chavent et al., 2017).

Homogeneity H of a cluster C_k

The following criterion H measures the adequacy between the variables within the cluster and its associated synthetic variable $\mathbf{f}^k$. It is defined as follows:

$$H(C_k) = \sum_{\mathbf{x}^j \in C_k} r^2_{\mathbf{x}^j,\mathbf{f}^k} + \sum_{\tilde{\mathbf{x}}^j \in C_k} \eta^2_{\mathbf{f}^k|\tilde{\mathbf{x}}^j} = \lambda^k_1, \tag{2}$$

where λ^k_1 denotes the first eigenvalue of **PCAmix** applied to the cluster C_k.

The first term (based on the squared Pearson correlation r^2) quantifies the link between the numerical variables in C_k and $\mathbf{f}^k$, independently of the sign of the relationship, whereas the second term (based on the correlation ratio η^2) measures the link between the categorical variables in C_k and $\mathbf{f}^k$. The homogeneity of a cluster is then maximized when all the numerical variables are perfectly correlated (or anti-correlated) to $\mathbf{f}^k$ and when all the correlation ratios of the categorical variables are equal to 1. In that case, all variables in cluster C_k bring the same information which is summarized by the corresponding synthetic variable $\mathbf{f}^k$.

Homogeneity $\mathcal{H}$ of a partition $P_K = \{C_1, \ldots, C_K\}$

The homogeneity criterion $\mathcal{H}$ is defined as the sum of the homogeneities of its clusters:

$$\mathcal{H}(P_K) = \sum_{k=1}^{K} H(C_k) = \lambda_1^1 + \ldots + \lambda_1^K. \tag{3}$$

A hierarchical clustering algorithm

The objective of this algorithm is to find a partition of the p available (numerical and/or categorical) variables. This partition must be such that the variables within a cluster are strongly related to each other in the sense of the homogeneity criterion introduced previously. More specifically, for a given number K of clusters, the aim is to find a partition P_K which maximizes the homogeneity function $\mathcal{H}$ defined in (3). To this end, a hierarchical clustering algorithm can be used and is described hereafter.

This algorithm builds a set of p nested partitions of variables as follows.

- Step $l = 0$: Initialization. Start with the partition in p clusters (i.e. one variable per cluster).
- Step $l = 1, \ldots, p-2$: Aggregation of two clusters. The objective is to aggregate two clusters of the partition in $p - l + 1$ clusters to get a new partition in $p - l$ clusters. To this end, we have to choose the two clusters A and B which provide the smallest dissimilarity $d(A, B)$ defined as:

$$d(A, B) = H(A) + H(B) - H(A \cup B) = \lambda_1^A + \lambda_1^B - \lambda_1^{A \cup B}. \tag{4}$$

  This dissimilarity (aggregation measure) quantifies the lost of homogeneity observed when the two clusters A and B merge. Based on this measure of aggregation, the new partition in $p - l$ clusters maximizes $\mathcal{H}$ among all the partitions in $p - l$ clusters obtained by aggregation of two clusters of the partition in $p - l + 1$ clusters.

- Step $l = p-1$: Stop. The final partition in one cluster (i.e. containing all the p variables) is obtained.

2.2 Variable selection using random forests

In this section, we focus on the **VSURF** procedure. This method is based on random forests (Breiman, 2001), which provide a non-parametric predictor, with very good prediction performance in lots of applied situations including high-dimensional data (see e.g. Verikas et al., 2011). Implemented in the R package `VSURF` (Genuer et al., 2015), the **VSURF** procedure is fully automatic (i.e. it does not need a pre-specified number of variables to select) and can be applied for both supervised classification and regression problems. In this paper, we focus on the supervised classification case.

Random Forests

A Random Forests (RF in the sequel) predictor is obtained by aggregating a collection of randomly perturbed decision trees. The randomness comes at two levels: the individuals level with a preliminary bootstrap sample draw, and the variables level with random variables sub-samples draws before splitting a node of a tree.

Let $\hat{h}_1, \ldots, \hat{h}_q$ be an ensemble of q binary tree predictors. The number of trees q is usually set to several hundreds. Each tree is a piece-wise constant function, which associates a class label for every input vector. A binary tree is made of internal nodes (which are split in two children nodes) and leaves (also called terminal nodes). The collection of the leaves forms a partition of the input space. The first node (which contains all data) is called the root of the tree, and a node is pure if it contains observations belonging to the same class. Finally each internal node t is split according to a splitting variable X^{j_t} and a splitting value v_t. RF methodology is detailed in the following algorithm, which is implemented in the `randomForest` R package (Liaw and Wiener, 2002).

INPUT: a learning sample of size n and an input vector X.

GOAL: predict the class label associated to X.

- For each $l = 1, \ldots, q$:
  - BOOTSTRAP SAMPLE. Draw a bootstrap sample of the learning set, by randomly choosing n observations among the n available, with replacement.
  - TREE CONSTRUCTION.
    - Initialize the tree by putting all observations of the bootstrap sample in the root node.
    - *While it exists an impure node among current terminal nodes*, for each impure terminal node, randomly choose a subset of `mtry` variables among the p variables without replacement, seek the best split of the node only among splits involving the selected variables, and split the node into two children nodes according to the best split (see details below).
  - TREE PREDICTION.

    Let X go down the tree and note the leaf it falls into. Return $\hat{h}_l(X)$, the class label of the observations of the learning set belonging to this leaf.

- AGGREGATION.

  Return the majority class among trees predictions: $\hat{f}(X) = \underset{c \in \{1,\ldots,L\}}{\operatorname{argmax}} \sum_{l=1}^{q} \mathbf{1}_{\hat{h}_l(X)=c}$.

For supervised classification, the Gini index is used for heterogeneity measure of a node, in terms of class labels. It is defined, for a node t, as $\text{Gini}(t) = \sum_{c \in \{1,\ldots,L\}} p_c(t)(1 - p_c(t))$, where $p_c(t)$ is the proportion of observations of class c in node t. Hence, the best split of a node t is the one minimizing $\text{Gini}(t)$ over all possible

splits. For a numerical variable, possible splits are of the form $\{X^j \leq v\}$ with $v \in [\min(\mathbf{x}^j), \max(\mathbf{x}^j)[$. This split means that observations with a j-th variable value not larger than v are sent to the left child node and the others to the right one. For a categorical variable, possible splits are of the form $\{X^j \in \mathcal{M}_j\}$ where $\mathcal{M}_j$ is a subset of the j-th variable levels (except empty set and total set).

Note that compared to Classification And Regression Trees (CART) introduced in Breiman et al. (1984), trees are fully developed and are not pruned.

OOB error and variable importance

During a RF run, an estimation of the prediction error and a measure of variable importance can be computed. Out-Of-Bag (OOB) error of a RF $\hat{f}$ is defined as $\text{OOBerror}(\hat{f}) = \frac{1}{n}\text{Card}\{i \in \{1, \ldots, n\} \mid y_i \neq \hat{y}_i\}$, where $\hat{y}_i$ is the majority class label among predictions of the trees $\hat{h}_l$ for which y_i is OOB, that is for which y_i was not chosen in the bootstrap sample used to build $\hat{h}_l$. Variable Importance (VI) also uses the OOB samples (all observations of the learning set not included in a bootstrap sample) to compute a measure of the link between Y and a variable X^j. The idea is that the more the mean error of a tree on its OOB sample increases when the link between X^j and Y is broken, the more important the variable is.

Variable Selection Using Random Forests

The **VSURF** procedure works with three steps.

- The first one begins by sorting variables based on random forests VI, and eliminates useless variables by an adaptive thresholding. The threshold is set to the estimation of the standard deviation (over multiple RF runs) of the VI of an unimportant variable.
- The second one starts with the previously kept variables and performs an ascendant variable introduction strategy, which builds embedded RF models. The model which attains the minimum OOB error rate is then selected, and the variables set on which it is based is called the interpretation set.
- The third step consists in eliminating the redundancy of interpretation variables and leads to a smaller variables set called the prediction set. It consists in a step-wise ascendant strategy, which at each step verifies that the next variable to introduce helps to decrease enough the OOB error rate.

More details on those three steps can be found in Genuer et al. (2015).

2.3 The CoV/VSURF procedure

We now describe the proposed methodology, which combines **CoV** and **VSURF**, in the following algorithm. First, the algorithm performs "groups of informative variables" selection in (a). It encompasses the ascendant

hierarchical clustering of variables, the optimal choice of the number of clusters of variables according to prediction performance, and the selection of the most important synthetic variables. Then, in (b), the algorithm computes a prediction of a new observation $\mathbf{x}$, by building a classifier based on the groups of variables selected in (a).

INPUT: a dataset $(\mathbf{X}, \mathbf{y})$ and a new observation $\mathbf{x}$ of X.

GOAL: select groups of informative variables and predict the class label of $\mathbf{x}$.

(a) GROUPS OF INFORMATIVE VARIABLES SELECTION:

1. Apply **CoV** on $\mathbf{X}$ to obtain a hierarchy (a tree) of variables.
2. For each $K = 2, \ldots, p$, cut **CoV** tree in K clusters, train a RF with the K synthetic variables $\mathbf{f}^1, \ldots, \mathbf{f}^K$ as predictors and $\mathbf{y}$ as output variable and compute its OOB error rate.
3. Choose the optimal number K^* of clusters, which leads to the minimum OOB error rate.

   Cut **CoV** tree in K^* clusters.
4. Perform **VSURF** with the K^* synthetic variables $\mathbf{f}^1, \ldots, \mathbf{f}^{K^*}$ as predictors and $\mathbf{y}$ as output variable. Denote by $m \leq K^*$ the number of selected informative synthetic variables (corresponding to the interpretation set of **VSURF**).

(b) PREDICTION OF A NEW OBSERVATION $\mathbf{x}$:

1. Train a random forest, $\hat{f}$, on the dataset consisting of the m selected synthetic variables and $\mathbf{y}$.
2. Compute the scores of $\mathbf{x}$ on the m selected synthetic variables and predict its class label using $\hat{f}$.

The main feature of this algorithm is that it outputs a list of m selected informative synthetic variables. Since each synthetic variable is built on a subset of original variables, the algorithm implicitly leads to select groups of original variables, the group structure being a priori unknown. Hence, in addition to perform variable selection, it takes advantage of the clustering of variables to give a list of selected variables with the additionnal group structure information.

Another interesting feature of the procedure is that even if the ascendant hierarchical clustering of variables algorithm is *unsupervised* (in the sense that it does not use response variable $\mathbf{y}$), the final variables partition is *supervised* since the number of clusters is optimized in terms of RF classifier prediction error. This choice is justified by the fact that our main goal is prediction, so we are primarily interested in groups of informative variables, rather than groups of all variables (e.g., a complex group structure only involving

non-informative variables does not have to be found by the method to give good prediction). We stress that each group of informative variables is summarized by its synthetic variable (the first principal component of the group). These synthetic variables are then used to build the predictive model. Typically, informative variables should be well represented by their associated synthetic variables. Indeed, assume that the information of a predictive variable X^j is not captured by the first principal component of its group C (for instance X^j is orthogonal—not correlated—to the component). The procedure would then select a higher number of clusters (K^*) in order to split C into two new informative groups C_1 and C_2, with e.g. C_1 containing variables highly correlated to the synthetic variable associated to C, and C_2 with a synthetic variable that retrieves the prediction power of X^j.

This procedure is illustrated on simulated samples and a real proteomic dataset, respectively in Section 3 and Section 4. Finally, let us mention that the method **CoV/VSURF** has been implemented in an R package, which is available online together with a vignette guide [2].

3 Simulation study

In this section, we evaluate numerical performance of the **Cov/VSURF** procedure in the framework of supervised classification. In the following simulation model, explanatory variables are structured in informative or non-informative groups. The underlying groups of variables can be numerical, categorical or mixed. Let us recall that **Cov/VSURF** does not take into account a priori the group structure information. We first describe the simulation model used to generate the data. We then consider an "$n > p$" (resp. an "$n < p$") simulated dataset to evaluate the proposed methodology in comparison with alternative approaches.

3.1 Simulation model

Let us consider a binary response variable Y and a p-dimensional explanatory variable X. The conditional probability $P(Y = 1|X = x) =: p(x)$ is modeled as a function of x using the well known logistic regression model:

$$\log\left(\frac{p(x)}{1-p(x)}\right) = x'\beta - 9, \tag{5}$$

where $\beta \in \mathbb{R}^p$. Note that the term "-9" in (5) is used to center the index $X'\beta$ (for the considered choice of parameters below) and therefore to obtain equibalanced response for Y.

Let us now specify how we construct the p-dimensional explanatory variable X (with $p = 120$) in order to get 10 groups of variables (components X^j of X) of several types (numerical, categorical or mixed) such that

[2] `https://github.com/robingenuer/CoVVSURF/blob/master/vignettes/intro_CoVVSURF.Rmd`

some groups of variables are informative or not. Table 1 provides a complete description of the 10 groups of these $p = 120$ variables. To sum up, we have a binary response variable and nine groups of correlated variables: one small and one large informative groups of numerical variables, one moderate non-informative group of numerical variables, and this structure is repeated for the three groups of categorical variables and the three groups of mixed variables. In addition, we have 30 non-informative and non-structured (independent) variables in the last group. Details on data generating process are given below.

Table 1: Overview of the 10 groups of variables in the simulation study.

Type of variables	Size of group	Informative group	Components of β	Names of variables
Numerical	Small (3)	Yes	$(1,\ 2,\ 3)$	`NumS`j with $j = 1, \ldots, 3$
Numerical	Moderate (12)	No	$(0, \ldots, 0)$	`NumM`j with $j = 1, \ldots, 12$
Numerical	Large (15)	Yes	$\frac{1}{5}(\underbrace{1, \ldots, 1}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{3, \ldots, 3}_{5})$	`NumL`j with $j = 1, \ldots, 15$
Categorical	Small (3)	Yes	$(1,\ 2,\ 3)$	`CategS`j with $j = 1, \ldots, 3$
Categorical	Moderate (12)	No	$(0, \ldots, 0)$	`CategM`j with $j = 1, \ldots, 12$
Categorical	Large (15)	Yes	$\frac{1}{5}(\underbrace{1, \ldots, 1}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{3, \ldots, 3}_{5})$	`CategL`j with $j = 1, \ldots, 15$
Mixed	Small (3)	Yes	$(1,\ 2,\ 3)$	`MixedS`j with $j = 1, \ldots, 3$
Mixed	Moderate (12)	No	$(0, \ldots, 0)$	`MixedM`j with $j = 1, \ldots, 12$
Mixed	Large (15)	Yes	$\frac{1}{5}(\underbrace{1, \ldots, 1}_{5}, \underbrace{2, \ldots, 2}_{5}, \underbrace{3, \ldots, 3}_{5})$	`MixedL`j with $j = 1, \ldots, 15$
Numerical	Big (30)	No	$(0, \ldots, 0)$	`Noise`j with $j = 1, \ldots, 30$

Let Z be the following multivariate Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$ where $p = 120$, $\mu = 0_p$ and the covariance matrix Σ is a block-diagonal matrix in order to get 10 groups of variables. These groups are independent from each other. Let us also introduce the parameter $\rho \in [-1, 1]$ to control the link between the variables within a group. The covariance matrix of Z is then defined as:

$$\Sigma = \text{diag}\left(\Sigma_{3,\rho}, \Sigma_{15,\rho}, \Sigma_{12,\rho}, \Sigma_{3,\rho}, \Sigma_{15,\rho}, \Sigma_{12,\rho}, \Sigma_{3,\rho}, \Sigma_{15,\rho}, \Sigma_{12,\rho}, \sigma^2 I_{30}\right), \tag{6}$$

where $\Sigma_{s,\rho}$ is an $s \times s$ matrix, whose coefficients are all equal to ρ except on the diagonal where they are equal to 1, $\sigma^2 > 0$ and I_{30} is the identity matrix of dimension 30. When ρ is close to 1, we obtain nine groups of highly correlated numerical variables and one group of 30 variables which are independent. In the following, we set $\rho = 0.9$.

We then generate n mutually independent random p-dimensional vectors $\mathbf{z}_1, \ldots, \mathbf{z}_n$ from the Gaussian distribution $\mathcal{N}_p(\mu, \Sigma)$. From the $\mathbf{z}_i$'s, we first construct the output variable values $y_i \in \{0, 1\}$ using the logistic regression model (5) and β coordinates given in Table 1. Since the $\mathbf{z}_i$'s original data are numerical, we have to binarize some numerical variables Z^j to obtain categorical ones by thresholding at their median

value: if the variable value is less than the median, the value of the corresponding binarized variable X^j is 0, and this value is 1 otherwise. In the 3 groups of mixed variables, one third of the variables are binarized (the last, last 4, last 5 variables in the small, medium and large group respectively).

3.2 An "$n > p$" simulated dataset

Application of CoV/VSURF to a learning dataset

From the previous model, a learning sample of $n = 600$ observations of the $p = 120$ explanatory variables and of the binary response variable Y is generated. The method **CoV** is applied to the 600×120 matrix of explanatory variables. The dendrogram of the hierarchy of the 120 variables (also called the **CoV** tree) is given in Figure 1. This dendrogram suggests a partition in 9 clusters. However, in our methodology, the number of clusters is not chosen according to the shape of the dendrogram but according to the prediction of the binary response variable Y. For each value of K between 2 and 120, we cut the **CoV** tree, build a random forest on the K synthetic variables of the corresponding K-clusters partition and compute the RF OOB error rate. This procedure is illustrated Figure 2. The (nearly) optimal value we get for this learning dataset is $K^* = 9$. This partition in $K^* = 9$ clusters retrieves almost the complete structure of the data. We recover 8 of the 9 groups of correlated variables, while all noise variables are pulled together in a large cluster with the last group of correlated variables (the small one). However, the synthetic variable of this cluster does not really take the noise variables into account, whose loadings (coefficients in the linear combination) are very low contrary to the loadings of the variables of the small group of correlated variables. Since this small group of numerical variables is informative, it means that this synthetic variable is also informative.

VSURF is then applied on the $K^* = 9$ synthetic variables of the previously chosen partition. **VSURF** selects 6 synthetic variables, corresponding to the 6 informative groups of variables. From the interpretation point of view, we succeed in selecting all informative variables, with in addition the clustering structure. For comparison, **VSURF** directly applied on the 120 original variables selects 39 variables among the 54 informative ones, with at least one per informative group (the large group of categorical variables being the least recovered).

Prediction performances on a test dataset

In addition to the learning dataset, a test sample of $n = 600$ observations is now generated. We focus on the prediction performances of the proposed methodology **CoV/VSURF** on this test sample and do comparisons with 3 other methodologies:

- **VSURF**: variable selection using random forests is applied on the original 120 variables,
- **RF**: random forests are applied on the original 120 variables,

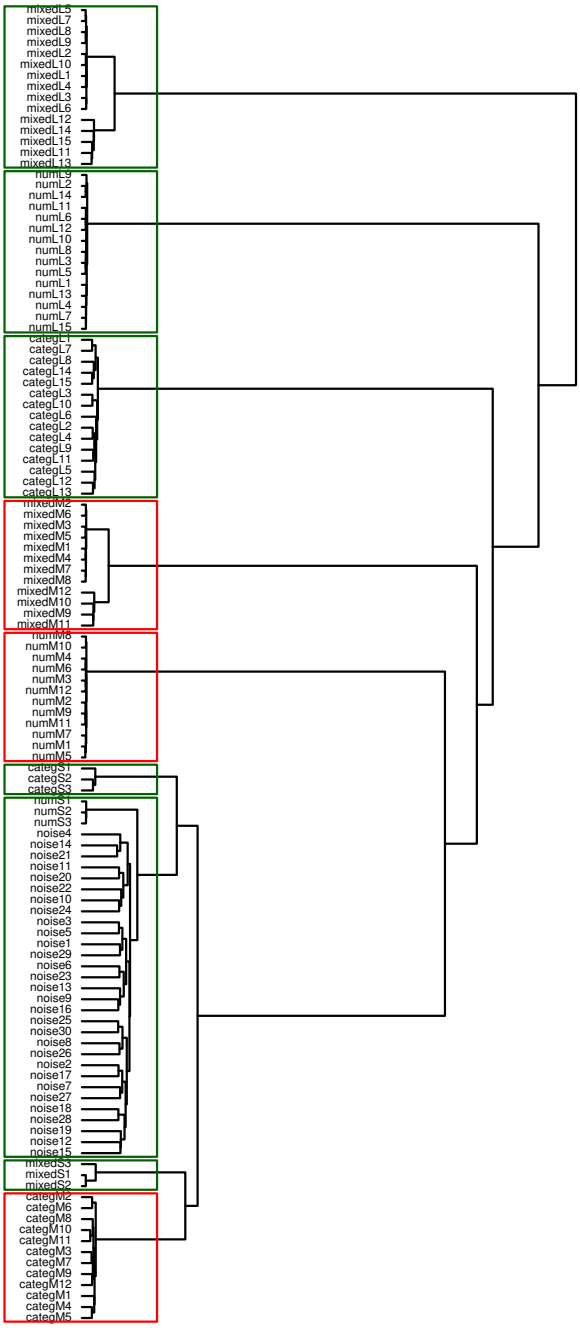


Figure 1: **CoV** tree of the $p = 120$ variables based on the learning sample of size $n = 600$. Heights correspond to dissimilarities between two aggregated clusters. Rectangles indicate the partition in $K^* = 9$ clusters (in green, the $m = 6$ clusters selected by **VSURF**, and in red the remaining unrelevant clusters).

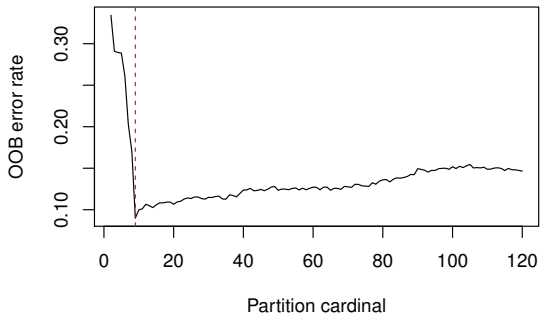
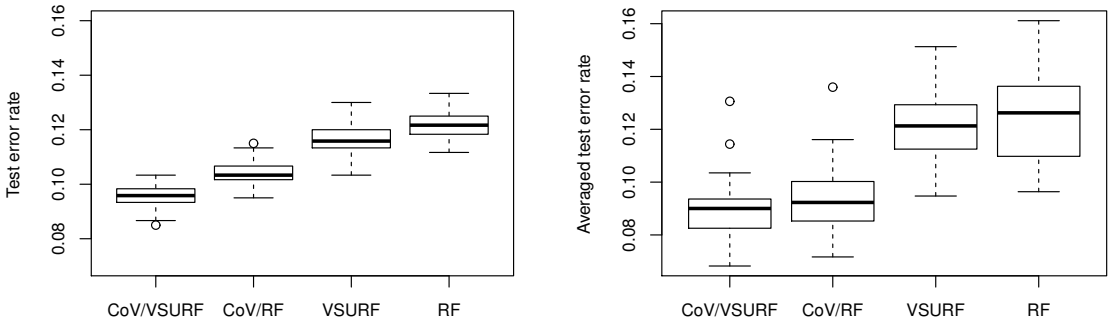


Figure 2: Random Forests OOB error rate according to the number of clusters of the partitions obtained by cutting the **CoV** tree for the simulated learning dataset of $n = 600$ observations. The dashed red vertical line corresponds to $K^* = 9$ clusters.

- **CoV/RF**: random forests are applied on the K^* synthetic variables obtained by **CoV**.

Note that the test sample of size $n = 600$ is fixed all along this simulation study. We obtain in Figure 3 the boxplots of test error rates.



(a) Single learning dataset (b) Replication of 50 learning datasets

Figure 3: Comparison of test error rates with learning datasets of size $n = 600$. **CoV/VSURF** and **CoV/RF** correspond to **VSURF** and random forests (RF) applied on the K^* synthetic variables given by **CoV**. **VSURF** and **RF** refer to methods applied on the original 120 variables. (a) Test error rates of 100 forests trained on one learning dataset. (b) Averaged (over 100 forests) test error rates computed on 50 learning datasets.

On the left-hand side of Figure 3, boxplots correspond to 100 runs of random forests. More precisely,

- For **CoV/VSURF** and **CoV/RF**, the clustering of variables is performed once and the optimal number K^* of clusters is chosen automatically using OOB error rate. For **Cov/RF**, 100 forests are trained on the K^* synthetic variables, while for **Cov/VSURF**, 100 forests are trained on the $m < K^*$ synthetic variables selected by `VSURF` (which is then performed also a single time).

- For **VSURF**, 100 forests are trained on the variables selected with one run of `VSURF` applied on the $p = 120$ original variables.

- For **RF**, 100 forests are trained on the original variables.

The results are better with **CoV/VSURF** and **CoV/RF** i.e. when **VSURF** and **RF** are applied after the dimension reduction using **CoV**. In addition, the approach **CoV/VSURF** is slightly better than **CoV/RF** in this setting. Also, the variability due to the random nature of the **RF** is relatively small here and very similar with the four approaches.

On the right-hand side of Figure 3, 50 learning datasets are generated in order to take into account the variability of the results due to the random nature of the simulation procedure. The four methodologies explained above for a single learning dataset are applied on each dataset. Boxplots correspond then to the 50 averaged (over 100 forests) test error rates. The results are better here with **CoV/VSURF** and **CoV/RF**. The improvement obtained by grouping variables together before the classifier run is confirmed here, despite the variability due to the random simulation procedure.

To sum up, using **CoV** is in a dimension reduction step leads — at least for this example — to a gain in prediction both with **RF** and **VSURF**. Moreover applying **VSURF** in **CoV/VSURF** procedure permits to select informative groups of linked variables without loss in prediction. In the next section, the exact same experiment is repeated but in a more challenging situation where only $n = 60$ ($< p = 120$) observations are generated.

3.3 An "$n < p$" simulated dataset

Now, we simulate a learning sample of $n = 60$ observations of the $p = 120$ explanatory variables and of the binary response variable Y. The **CoV** method is applied to the 60×120 matrix of explanatory variables. Note that we set the number of trees q to 2000 in all our experiments. The dendrogram of the hierarchy is represented in Figure 4. This tree suggests to retain the partition in $K = 9$ clusters which matches with the underlying partition in 9 groups of correlated variables (i.e. the informative and non informative structured groups of variables) while the 30 noise variables are pulled together either with the small group of numerical variables or with the small group of mixed variables.

Our methodology chooses to cut the dendrogram in $K^* = 10$ clusters (see Figure 5). The structure in 9 groups of correlated variables is conserved and the 10-th cluster only contains 8 noise variables. **VSURF** applied on the $K^* = 10$ associated synthetic variables selects $m = 4$ of them. The two groups of informative numerical variables and the two groups of informative mixed variables are retained.

Thus, contrary to the case with $n = 600$ observations (see Figure 1), where the group structure is perfectly identified and all informative variables selected, in this more difficult context where $n < p$, the **CoV/VSURF** methodology still retrieves the group structure in the dendrogram (see Figure 4) but misses

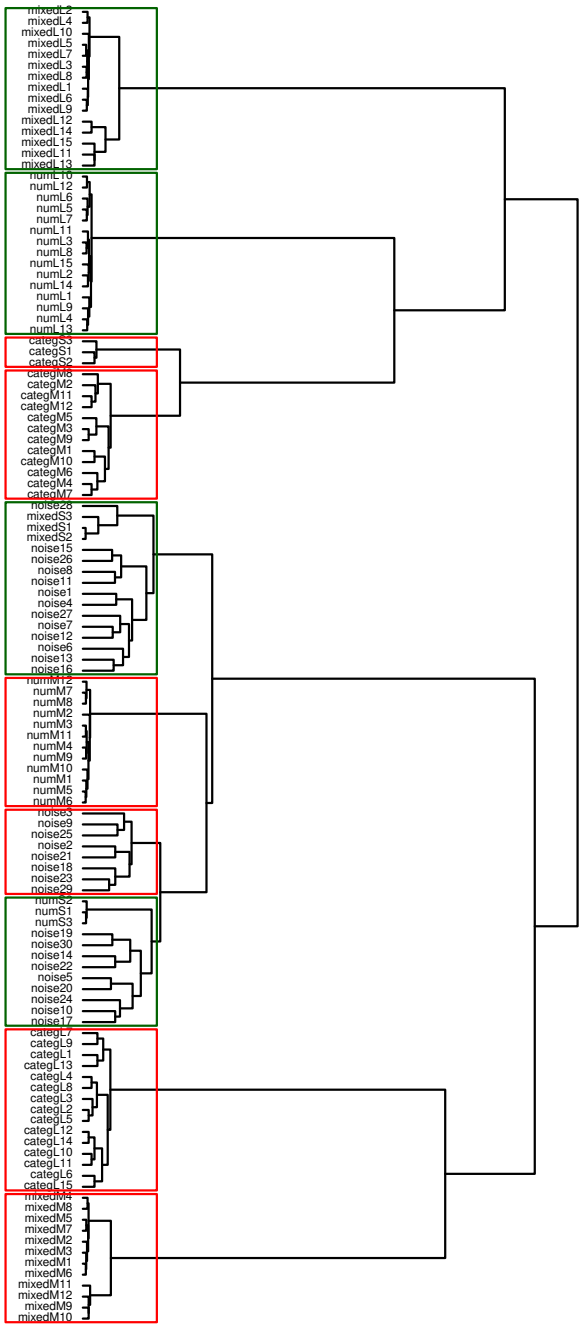


Figure 4: **CoV** tree of the $p = 120$ variables based on the learning sample of size $n = 60$. Heights correspond to dissimilarities between two aggregated clusters. Rectangles indicate the partition in $K^* = 10$ clusters (in green, the $m = 4$ clusters selected by **VSURF**, and in red the remaining unrelevant clusters).

the informative groups of categorical variables in the selection step. This can be explained by the fact that those variables are binary, hence less explanatory than the numerical ones. **VSURF** applied on the 120 original variables only selects 8 variables, with at least one per group selected by our approach, and still no informative categorical variable are retained.

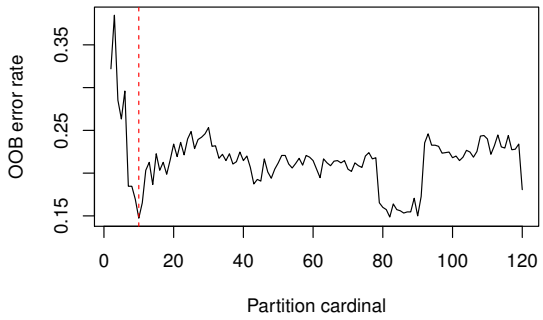
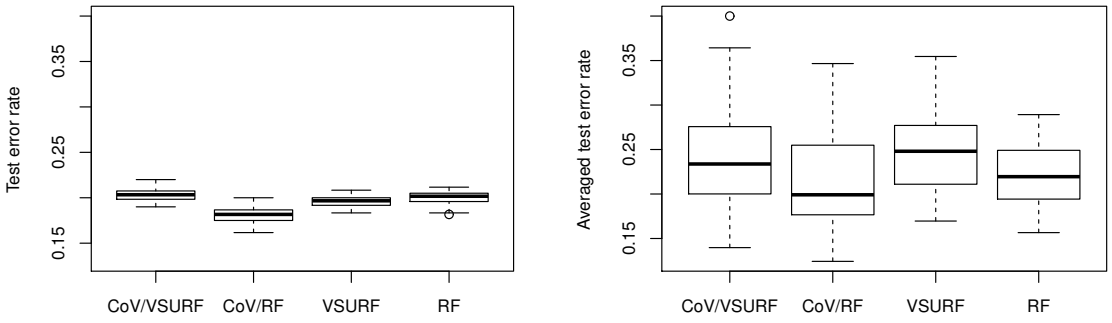


Figure 5: Random Forests OOB error rate according to **CoV** partition cardinal for the learning dataset of $n = 60$ observations. The dashed red vertical line corresponds to $K^* = 10$ clusters.

To study prediction performance of the proposed methodology **CoV/VSURF**, the same test sample as in the previous section is used, and Figure 6 is obtained in the same way than in the $n = 600$ case. Those graphs show that the variability due to random forests (left-hand side) remains relatively small compared to the variability due to random generation of the 50 learning dataset (right-hand side). In this more difficult situation, all methodologies are fairly comparable in prediction, with a slight advantage to **CoV/RF**. However, the methodology **CoV/VSURF** has the advantage to select groups of informative variables, without increasing prediction error rate.



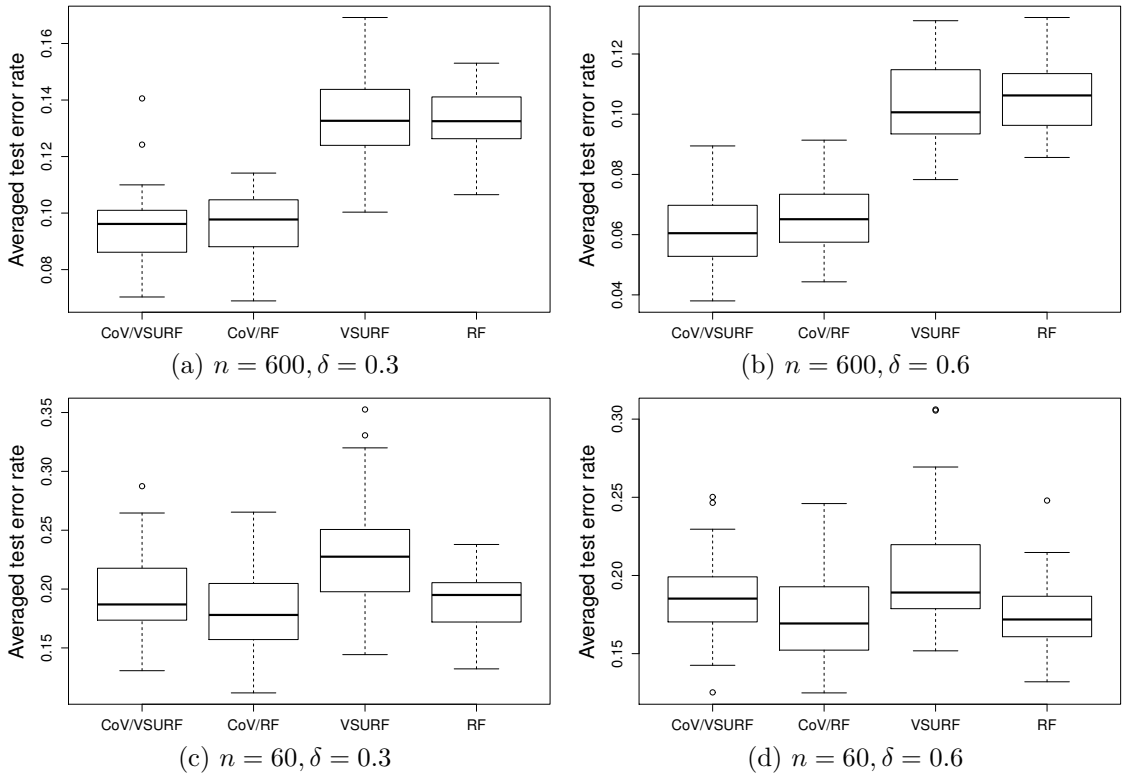
(a) Single learning dataset

(b) Replication of 50 learning datasets

Figure 6: Comparison of test error rates with learning datasets of size $n = 60$. **CoV/VSURF** and **CoV/RF** correspond to **VSURF** and random forests (RF) applied on the K^* synthetic variables given by **CoV**. **VSURF** and **RF** refer to methods applied on the original 120 variables. (a) Test error rates of 100 forests trained on one learning dataset. (b) Averaged (over 100 forests) test error rates computed on 50 learning datasets.

3.4 Sensitivity analysis

To conclude this simulation study, we conduct a sensitivity analysis with respect to the simulated group structure. Indeed, in the previous simulation model, the covariance matrix Σ (see Equation 6) induces a very clear group structure in the data, which could be somehow unrealistic for real applications. Thus, we relax this structure by adding a between-groups correlation coefficent $\delta \geq 0$. More precisely, the intra-groups correlaction structure is unchanged, but two variables from two different groups have now a correlation of δ, while noise variables remain independent with every other variables.



(a) $n = 600, \delta = 0.3$ (b) $n = 600, \delta = 0.6$

(c) $n = 60, \delta = 0.3$ (d) $n = 60, \delta = 0.6$

Figure 7: Comparison of test error rates with learning datasets of size $n = 600$ (top) and $n = 60$ (bottom) including a between-groups correlation of $\delta = 0.3$ (left) and $\delta = 0.6$ (right). Boxplots correspond to the averaged (over 100 forests) test error rates computed on 50 learning datasets.

Figure 7 illustrates the results in the two previsous cases with $n = 600$ and $n = 60$, where δ takes the values 0.3 and 0.6. Boxplots correspond to the 50 averaged (over 100 forests) test error rates associated to 50 randomly generated datasets as in Figures 3(b) and 6(b).

Those results show that the comparison between the four methods **CoV/VSURF**, **CoV/RF**, **VSURF** and **RF** is quite stable, with a clear advantage in terms of prediction for methods using clustering of variables in the case $n = 600$ and similar performances of the four methods in the case $n = 60$.

Furthermore, we investigate the results of the **CoV/VSURF** procedure applied to one dataset (results

not shown here) for several values of δ. It reveals that the method is not perturbed by values of δ up to 0.6 (the clustering of variables is accurate and the selection of synthetic variables is satisfactory), and that for larger values it naturally begins to give very small number of groups. For example in the extreme case of $\delta = 0.9$, the method groups together all variables (except noise variables) as expected.

4 Proteomic data application

In this section, the **CoV/VSURF** procedure is illustrated with an application on a real proteomic dataset. This dataset comes from a clinical study sponsored by University Hospital in Bordeaux. It involved $n = 44$ patients with a rectum cancer who undertook a treatment of chemotherapy and radiotherapy, before a surgery intervention. Some patients responded favorably to the treatment and hence had a smaller tumor at t_1, the time of surgery, compared to t_0, a time just before the beginning of the treatment, and some patients did not. The main goal of this study was to predict if the patient will be a good treatment responder or not, using proteomic information, measured at t_0. Indeed, it is useless to give the treatment to a bad responder, for which other alternatives should be tried. A secondary objective is to select the proteins which best discriminate the two kinds of treatment response.

The dataset contains $p = 4786$ numerical explanatory variables which measure protein abundances. Measurements are done on peptides (one protein is made of several peptides). The total number of peptides is $p = 4786$, whereas total number of proteins is 868. A priori knowledge of what peptide is part of which protein is not used in our approach, but it will be used for interpretation purpose. This helps us to interpretate the resulting group structure and permits to illustrate the interest of our method for other problems where no a priori group structure is known.

The **CoV/VSURF** methodology is then applied to this dataset. Since the number p of variables is relatively large, the number K^* of clusters of peptides is chosen among 2 and 2000. The resulting random forests OOB error rates are displayed in Figure 8 and the (nearly) optimal number of clusters is $K^* = 68$. The **VSURF** method, applied on those $K^* = 68$ synthetic variables, selects 4 of them. These 4 synthetic variables are sorted by decreasing order of their variable importance (VI). The corresponding 4 groups of peptides gather respectively 37, 61, 20, 25 peptides, that is a total of 143 peptides. Those 143 peptides come from 73 different proteins. The most homogeneous one is the 4th with 22 peptides (over 25) coming from the same protein, *PZP* (Pregnancy Zone Protein). The 3rd one contains 18 peptides (over 20) coming from two different proteins, *IGHD* (Ig Delta Chain) protein and *THRB* (Prothrombin) protein. The second cluster is the more heterogeneous with 61 peptides coming from 50 different proteins. The first one contains 13 peptides (over 37) coming from *A2GL* (Alpha 2 Glycoprotein), 6 from *CRP* (C-reactive protein).

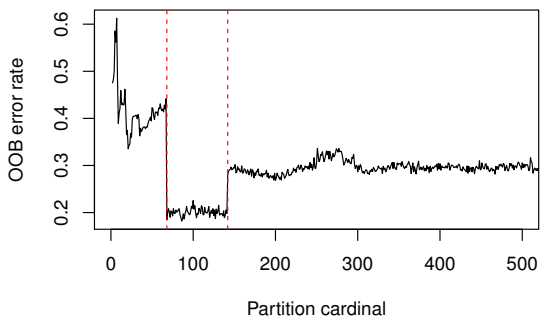


Figure 8: Random forests OOB error rate according to **CoV** partition cardinal (number of peptides clusters) for the proteomic data. The two dashed red vertical lines correspond respectively to partitions in $K^* = 68$ and $K = 142$ clusters. For visibility, the x-axis is truncated around $K = 500$ clusters (the error rate remains quite stable until $K = 2000$).

The cluster with highest VI, made of 37 peptides, is the one responsible of the large decrease of OOB error rates in Figure 8: it appears when the number of synthetic variables goes from 67 to 68 i.e. when the OOB error rate decreases from around 40% to 20%. In addition, when this particular cluster is split into two new clusters (when the partition cardinal K reaches 142), the OOB error increases significantly from 20% to 30%. This confirms the importance of this cluster of peptides to predict treatment response. Note that, among the peptides selected by **CoV/VSURF**, we find peptides that come from *APOA1* (Apolipoprotein A1), *HPT* (Haptoglobin), *TRFE* (Serotransferrin) and *PGRP2* (N-Acetylmuramoyl L-Alanine Amidase). Those proteins make sense to oncologists in this context, who are currently investigating in more depth the relation between those proteins and the response to the treatment. In addition, when **VSURF** is applied alone on the original data, only 35 peptides are selected. So, **VSURF** gives a sparse variable selection (less peptides are selected), but with no group structure.

To get an estimation of classification error rates for this real dataset, we perform an external leave-one-out cross-validation to fairly estimate prediction performance of our approach (see Ambroise and McLachlan, 2002). This means that, for each patient in this study, we perform the entire methodology **CoV/VSURF** on the dataset containing all other patients, before predicting him. As before, we compare the performance with the three approaches **CoV/RF**, **VSURF** and **RF**. Figure 9 provides the results. According to these boxplots, there is almost no difference between the performances of **CoV/VSURF**, **CoV/RF** and **VSURF** (recall that since $n = 44$, one more good classification of an individual gives approximately a 0.02 decrease in error rate). It is important to highlight that only the **CoV/VSURF** procedure selects groups of informative peptides without increasing the prediction error. To conclude this real data study, **CoV/VSURF** achieves the double objective of predicting if the patient is a good treatment responder and determining groups of peptides responsible for the good or the bad response to the treatment.

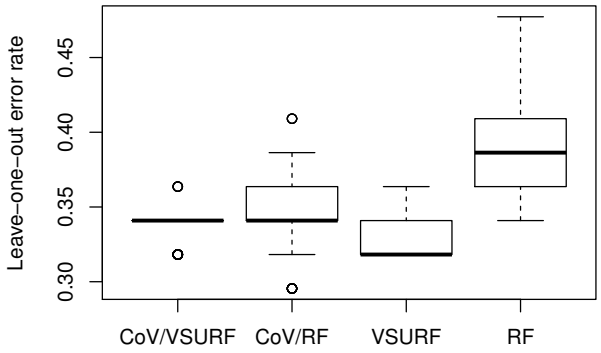


Figure 9: Comparison of leave-one out error rates of 100 forests for proteomic data. **CoV/VSURF** and **CoV/RF** correspond to **VSURF** and random forests (RF) applied on the $K^* = 68$ synthetic variables given by **CoV**. **VSURF** and **RF** refer to methods applied on the original $p = 4786$ peptides.

Acknowledgments.

The authors would like to thank members of the USMR unit of Bordeaux University Hospital (especially Hélène Savel and Paul Perez) for allowing access to the clinical study data and for insightful discussions. In addition, the authors also thank the editor and two anonymous reviewers who contribute to greatly improve the manuscript.

References

Ambroise, C. and McLachlan, G. J. (2002). Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences*, 99(10):6562–6566.

Breiman, L. (2001). Random forests. *Machine Learning*, 45(1):5–32.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Chavent, M., Kuentz-Simonet, V., Labenne, A., and Saracco, J. (2017). Multivariate analysis of mixed data: The R package PCAmixdata. *arXiv preprint arXiv:1411.4911*.

Chavent, M., Kuentz-Simonet, V., Liquet, B., and Saracco, J. (2012a). ClustOfVar: An R package for the clustering of variables. *Journal of Statistical Software*, 50(13):1–16.

Chavent, M., Kuentz-Simonet, V., and Saracco, J. (2012b). Orthogonal rotation in PCAMIX. *Advances in Data Analysis and Classification*, 6(2):131–146.

Chun, H. and Keleş, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(1):3–25.

Fernández-Delgado, M., Cernadas, E., Barro, S., and Amorim, D. (2014). Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, 15(1):3133–3181.

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

Genuer, R., Poggi, J.-M., and Tuleau-Malot, C. (2015). VSURF: An R package for variable selection using random forests. *The R Journal*, 7(2):19–33.

Gordon, A. D. (1999). *Classification*. Chapman & Hall/CRC.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of Machine Learning Research*, 3:1157–1182.

Jolliffe, I. (2002). *Principal Component Analysis*. Springer Verlag, New York.

Kaufman, L. and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Kiers, H. A. (1991). Simple structure in component analysis techniques for mixtures of qualitative and quantitative variables. *Psychometrika*, 56(2):197–212.

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.

Liaw, A. and Wiener, M. (2002). Classification and regression by randomForest. *R News*, 2(3):18–22.

Liquet, B., de Micheaux, P. L., Hejblum, B. P., and Thiébaut, R. (2015). Group and sparse group partial least square approaches applied in genomics context. *Bioinformatics*, 32(1):35–42.

Maugis, C., Celeux, G., and Martin-Magniette, M.-L. (2009). Variable selection for clustering with gaussian mixture models. *Biometrics*, 65(3):701–709.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Raftery, A. E. and Dean, N. (2006). Variable selection for model-based clustering. *Journal of the American Statistical Association*, 101(473):168–178.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Tadesse, M. G., Sha, N., and Vannucci, M. (2005). Bayesian variable selection in clustering high-dimensional data. *Journal of the American Statistical Association*, 100(470):602–617.

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

Verikas, A., Gelzinis, A., and Bacauskiene, M. (2011). Mining data with random forests: A survey and results of new tests. *Pattern Recognition*, 44(2):330–349.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320.