Correlation and variable importance in random forests

Article *in* Statistics and Computing · May 2017

DOI: 10.1007/s11222-016-9646-1 · Source: arXiv

CITATIONS
246

READS
9,818

3 authors, including:

Baptiste Gregorutti
Safety Line
7 PUBLICATIONS **315** CITATIONS
SEE PROFILE

Philippe Saint-Pierre
Paul Sabatier University - Toulouse III
52 PUBLICATIONS **1,039** CITATIONS
SEE PROFILE

Correlation and variable importance in random forests

Baptiste Gregorutti[12*], Bertrand Michel[2], Philippe Saint-Pierre[2]

[1] *Safety Line*
15 rue Jean-Baptiste Berlier, 75013 Paris, France

[2] *Laboratoire de Statistique Théorique et Appliquée*
Université Pierre et Marie Curie
4 place Jussieu, 75252 Paris Cedex 05, France

Abstract

This paper is about variable selection with the random forests algorithm in presence of correlated predictors. In high-dimensional regression or classification frameworks, variable selection is a difficult task, that becomes even more challenging in the presence of highly correlated predictors. Firstly we provide a theoretical study of the permutation importance measure for an additive regression model. This allows us to describe how the correlation between predictors impacts the permutation importance. Our results motivate the use of the Recursive Feature Elimination (RFE) algorithm for variable selection in this context. This algorithm recursively eliminates the variables using permutation importance measure as a ranking criterion. Next various simulation experiments illustrate the efficiency of the RFE algorithm for selecting a small number of variables together with a good prediction error. Finally, this selection algorithm is tested on a real life dataset from aviation safety where the flight data recorders are analysed for the prediction of a dangerous event.

Keywords: Random Forests, Supervised Learning, Variable Importance, Variable Selection.

1 Introduction

This paper is motivated by a real life problem in the context of aviation safety. The recent recommendations of the International Civil Aviation Organisation forces airlines to evaluate the risk of incidents and more generally to ensure the operational safety. Indeed, flight data recorders provide a large amount of raw data. The available data contains many flight parameters which are highly correlated. A challenging problem is to provide an efficient prediction of dangerous events. In addition, an interpretable model based on a small number of flight parameters is crucial for pilot training and developing new flight procedures. Variable selection techniques is a natural answer to this issue.

In such large scale learning problems, in particular when the number of variables is much larger than the number of observations, not all of the variables are relevant for predicting the outcome of interest. Some irrelevant variables may have a negative effect on the model accuracy. Variable selection techniques, also called feature selection or subset selection, involve eliminating irrelevant variables and provide two main advantages. First, a model with a small number of

*Corresponding author: baptiste.gregorutti@safety-line.fr

variables is more interpretable. Secondly, the model accuracy might be improved and then avoid the risk of overfitting.

Many studies about variable selection have been conducted during the last decade. In Guyon and Elisseeff [13], the authors review three existing approaches: filter, embedded and wrapper. A filter algorithm, also known as variable ranking, orders the variables in a preprocessing step and the selection is done independently of the choice of the learning technique. Two classical ranking criteria are the Pearson correlation coefficient and the mutual information criterion as mentioned in the recent survey of Lazar et al. [22]. The main drawback of this approach is that the choice of the selected variables is not induced by the performance of the learning method. The embedded approach selects the variables during the learning process. The two main examples are the Lasso (Tibshirani [33]) for regression problems (the selection process is done through the ℓ_1 regularization of the least square criterion) and decision trees (the selection is induced by the automatic selection of the splitting variables) such as CART algorithm (Breiman et al. [8]). A wrapper algorithm uses the learning process to identify an optimal set of variables among all possible subsets (see Kohavi and John [20], Blum and Langley [5]). The measure of optimality is usually defined by the error rate estimate. As it is impossible to evaluate all variable subsets when the dimension of the data is too large, the wrapper approach consists of using greedy strategies such as forward or backward algorithms. A heuristic is required to select the variables to be introduced or eliminated. This algorithm has been adapted for various contexts in the literature (see for instance Guyon et al. [14], Rakotomamonjy [29], Svetnik et al. [32], Díaz-Uriarte and Alvarez de Andrés [10], Louw and Steel [23], Genuer et al. [11]).

A classical issue of variable selection methods is their instability: a small perturbation of the training sample may completely change the set of selected variables. This instability is a consequence of the data complexity in high dimensional settings (see Kalousis et al. [19], Křížek et al. [21]). In particular, the instability of variable selection methods increases when the predictors are highly correlated. For instance, Bühlmann et al. [9] have shown that the lasso tends to discard most of the correlated variables even if they are discriminants and randomly selects one representative among a set of correlated predictors. In the context of random forests, the impact of correlated predictors on variable selection methods has been highlighted by several simulation studies, see for instance Toloşi and Lengauer [34]. For real life applications it is of first importance to select a subset of variables which is the most stable as possible. One popular solution to answer the instability issue of variable selection methods consists in using bootstrap samples: a selection is done on several bootstrap subsets of the training data and a stable solution is obtained by aggregation of these selections. Such generic approach aims to improve both the stability and the accuracy of the method. This procedure is known as "ensemble feature selection" in the machine learning community. Several classification and regression techniques based on this approach have been developed (Bi et al. [4] in the context of Support vector regression, Meinshausen and Bühlmann [24] with the stability selection). Haury et al. [16] provide a comparison of ensemble feature selections combined with several classification methods.

The random forests algorithm, introduced by Breiman [7], is a modification of bagging that aggregates a large collection of tree-based estimators. This strategy has better estimation performances than a single random tree: each tree estimator has low bias but high variance whereas the aggregation achieve a bias-variance trade-off. The random forests are very attractive for both classification and regression problems. Indeed, these methods have good predictive performances in practice, they work well for high dimensional problems and they can be used with multi-class output, categorical predictors and imbalanced problems. Moreover, the random forests provide some measures of the importance of the variables with respect to the prediction of the outcome variable.

Several studies have used the importance measures in variable selection algorithms (Svetnik

et al. [32], Díaz-Uriarte and Alvarez de Andrés [10], Genuer et al. [11]). The effect of the correlations on these measures has been studied in the last few years by Archer and Kimes [2], Strobl et al. [31], Nicodemus and Malley [27], Nicodemus et al. [28], Nicodemus [26], Auret and Aldrich [3] and Toloşi and Lengauer [34]. However, there is no consensus on the interpretation of the importance measures when the predictors are correlated and more precisely there is no consensus on what is the effect of this correlation on the importance measures (see e.g. Grömping [12], Neville [25]). One reason for this is that, as far as we know, no theoretical description of this effect has been proposed in the literature. This situation is particularly unsatisfactory as the importance measures are intensively used in practice for selecting the variables.

The contributions of this paper are two-fold. First, we give some theoretical descriptions of the effect of the correlations on the ranking of the variables produced by the permutation importance measure introduced in Breiman [7]. More precisely, we consider a particular additive regression model for which it is possible to express the permutation importance in function of the correlations between predictors. The results of this section are validated by a simulation study.

The second contribution of this paper is of algorithmic nature. We take advantage of the previous results to compare wrapper variable selection algorithms for random forests in the context of correlated predictors. Note that most of the variable selection procedures using the random forests are wrapper algorithms. It can be used also as a filter algorithm as in Hapfelmeier and Ulm [15]. Two main wrapper algorithms are considered in the literature. These two rely on backward elimination strategies based on the ranking produced by the permutation importance measure. The first algorithm computes the permutation importance measures in the full model which produces a ranking of the variables. This ranking is kept unchanged by the algorithm (Svetnik et al. [32], Díaz-Uriarte and Alvarez de Andrés [10], Genuer et al. [11]). The second algorithm was first proposed by Guyon et al. [14] in the context of support vector machines (SVM) and is referred to as Recursive Feature Elimination (RFE). This algorithm requires to update the ranking criterion at each step of a backward strategy: at each step the criterion is evaluated and the variable which minimizes this measure is eliminated. In the random forests setting, although less popular than the first one, this strategy has been implemented for instance in Jiang et al. [18]. As far as we know, only one study by Svetnik et al. [32] compared the two approaches and concluded that the non recursive algorithm provides better results. However, their findings are based on a real life dataset without taking into account the effect of correlated predictors and are not confirmed by simulation studies. Moreover, this position goes against the results we find in Section 3.

A simulation study has been performed to compare the performances of the recursive and the non recursive strategies. Several designs of correlated data encounters in the literature have been simulated for this purpose. As expected, the simulations indicate that the recursive algorithm provides better results.

The paper is organized as follows. We first introduce the statistical background of the random forests algorithm and the permutation importance measure. Section 3 provides some theoretical properties of this criterion in the special case of an additive regression model. Section 4 describes the RFE algorithm used for variable selection in a random forests analysis. Next, the effect of the correlations on the permutation importance and the good performances of the RFE algorithm in the case of correlated variables are emphasized in a simulation study. This algorithm is finally carried out for studying the risk of aircraft incidents using flight data.

2 Random forests and variable importance measures

Let us consider a variable of interest Y and a vector of random variables $\mathbf{X} = (X_1, \ldots, X_p)$. In the regression setting a rule $\hat{f}$ for predicting Y is a measurable function taking its values in $\mathbb{R}$. The prediction error of $\hat{f}$ is then defined by $R(\hat{f}) = \mathbb{E}\left[(\hat{f}(\mathbf{X}) - Y)^2\right]$ and our goal is to estimate the conditional expectation $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$.

Let $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ be a learning set of n i.i.d. replications of $(\mathbf{X}, Y)$ where $\mathbf{X}_i = (X_{i1}, \ldots, X_{ip})$. Since the true prediction error of $\hat{f}$ is unknown in practice, we consider an estimator based on the observation of a validation sample $\bar{\mathcal{D}}$:

$$\hat{R}(\hat{f}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}} (Y_i - \hat{f}(\mathbf{X}_i))^2.$$

Classification and regression trees, particularly CART algorithm due to Breiman et al. [8], are competitive techniques for estimating f. Nevertheless, these algorithms are known to be unstable insofar as a small perturbation of the training sample may change radically the predictions. For this reason, Breiman [7] introduced the random forests as a substantial improvement of the decision trees. It consists in aggregating a collection of such random trees, in the same way as the bagging method also proposed by Breiman [6]: the trees are built over n_{tree} bootstrap samples $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^{n_{tree}}$ of the training data $\mathcal{D}_n$. Instead of CART algorithm, a small number of variables is randomly chosen to determine the splitting rule at each node. Each tree is then fully grown or until each node is pure. The trees are not pruned. The resulting learning rule is the aggregation of all of the tree-based estimators denoted by $\hat{f}_1, \ldots, \hat{f}_{n_{tree}}$. The aggregation is based on the average of the predictions.

In parallel the random forests algorithm allows us to evaluate the relevance of a predictor thanks to variable importance measures. The original random forests algorithm computes three measures, the permutation importance, the z-score and the Gini importance. We focus here on the permutation importance due to Breiman [7]. Broadly speaking, a variable X_j can be considered as important for predicting Y if by breaking the link between X_j and Y the prediction error increases. To break the link between X_j and Y, Breiman proposes to randomly permute the observations of the X_j's. It should be noted that the random permutations also breaks the link between X_j and the other covariates. The empirical permutation importance measure can be formalized as follows: define a collection of out-of-bag samples $\{\bar{\mathcal{D}}_n^t = \mathcal{D}_n \setminus \mathcal{D}_n^t,\ t = 1, \ldots, n_{tree}\}$ which contains the observations not selected in the bootstrap subsets. Let $\{\bar{\mathcal{D}}_n^{tj},\ t = 1, \ldots, n_{tree}\}$ denotes the permuted out-of-bag samples by random permutations of the values of the j-th variable in each out-of-bag subsets. The empirical permutation importance of the variable X_j is defined by

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^t)\right]. \tag{2.1}$$

This quantity is the empirical counterpart of the permutation importance measure $I(X_j)$, as formalized recently in Zhu et al. [36]. Let $\mathbf{X}_{(j)} = (X_1, \ldots, X_j', \ldots, X_p)$ be the random vector such that X_j' is an independent replication of X_j which is also independent of Y and of all of the others predictors, the permutation importance measure is given by

$$I(X_j) = \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right].$$

The permutation of the values of X_j in the definition of $\hat{I}(X_j)$ mimics the independence and the identical copy of the distribution of X_j in the definition of $I(X_j)$. Zhu et al. [36] show that

$\hat{I}(X_j)$ converges to $I(X_j)$ at an exponential rate for particular tree-based algorithms, their result is given for the z-score in the regression but the proof given also provides the convergence for $\hat{I}(X_j)$.

The permutation importance measure can be used to rank or select the predictors. Among others criteria, the permutation importance measure has shown good performances for leading variable selection algorithms. Nevertheless variable selection is a difficult issue especially when the predictors are highly correlated. In the next section we investigate deeper the properties of the permutation importance measure in order to understand better how this quantity depends on the correlation between the predictors.

3 Permutation importance measure of correlated variables

Previous results about the impact of correlation on the importance measures are mostly based on experimental considerations. We give a non exhaustive review of these contributions and we compare them with our theoretical results. Archer and Kimes [2] observe that the Gini measure is less able to detect the most relevant variables when the correlation increases and they mention that the same is true for the permutation importance. The experiments of Auret and Aldrich [3] confirm these observations. Genuer et al. [11] study the sensitivity of the empirical permutation importance measure to many parameters, in particular they study the sensitivity to the number of correlated variables. Recently, Tolo̧si and Lengauer [34] identify what they call the "correlation bias". Note that it does not correspond to a statistical bias. More precisely, these authors observe two key effects of the correlation on the permutation importance measure: first, the importance values of the most discriminant correlated variables are not necessarily higher than a less discriminant one, and secondly the permutation importance measure depends on the size of the correlated groups.

Since previous studies are mainly based on numerical experiments, there is obviously a need to provide theoretical validations of these observations. We propose below a first theoretical analysis of this issue, in a particular statistical framework. In the rest of the section, we assume that the random vector $(\mathbf{X}, Y)$ satisfies the following additive regression model:

$$Y = \sum_{j=1}^{p} f_j(X_j) + \varepsilon, \tag{3.1}$$

where ε is such that $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$, $\mathbb{E}[\varepsilon^2|\mathbf{X}]$ is finite and the f_j's are measurable functions. In other words, we require that the regression function can be decomposed into $f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j)$. In the sequel, $\mathbb{V}$ and $\mathbb{C}$ denote variance and covariance.

Proposition 1. *1. Under model* (3.1)*, for any* $j \in \{1, \ldots, p\}$*, the permutation importance measure satisfies*

$$I(X_j) = 2\mathbb{V}[f_j(X_j)].$$

2. Assume moreover that for some $j \in \{1, \ldots, p\}$ *the variable* $f_j(X_j)$ *is centered. Then,*

$$I(X_j) = 2\mathbb{C}[Y, f_j(X_j)] - 2\sum_{k \neq j} \mathbb{C}[f_j(X_j), f_k(X_k)].$$

Proof. see Appendix A.1 □

In this framework, the permutation importance corresponds to the variance of $f_j(X_j)$, up to a factor 2. The second result of Proposition 1 is the key point to study the influence of the

correlation on the permutation measure. This result strongly depends on the additive structure of the regression function f and it seems difficult to give such a simple expression of the permutation importance without assuming this additive form for the regression function.

If $(\mathbf{X}, Y)$ is assumed to be a normal vector it is possible to specify the permutation importance measure. Note that in this context the conditional distribution of Y over $\mathbf{X}$ is also normal and the conditional mean f is a linear function: $f(\mathbf{x}) = \sum_{j=1}^{p} \alpha_j x_j$ with $\alpha = (\alpha_1, \ldots, \alpha_p)^t$ a sequence of deterministic coefficients (see for instance Rao [30], p. 522).

Proposition 2. *Consider a Gaussian random vector* $(\mathbf{X}, Y) \sim \mathcal{N}_{p+1}\left(0, \begin{pmatrix} C & \boldsymbol{\tau} \\ \boldsymbol{\tau}^t & \sigma_y^2 \end{pmatrix}\right)$, *where* $\boldsymbol{\tau} = (\tau_1, \ldots, \tau_p)^t$ *with* $\tau_j = \mathbb{C}(X_j, Y)$, $\sigma_y^2 > 0$ *and* $C = [\mathbb{C}(X_j, X_k)]$ *is the non degenerated variance-covariance matrix of* $\mathbf{X}$. *Then, for any* $j \in \{1, \ldots, p\}$,

$$I(X_j) = 2\alpha_j^2 \mathbb{V}(X_j) = 2\alpha_j \mathbb{C}(X_j, Y) - 2\alpha_j \sum_{k \neq j} \alpha_k \mathbb{C}(X_j, X_k),$$

where $\alpha_j = [C^{-1}\boldsymbol{\tau}]_j$.

Proof. see Appendix A.2 □

Note that if we consider a linear function $f : \mathbb{R}^p \mapsto \mathbb{R}$, a random vector $\mathbf{X}$ of $\mathbb{R}^p$ and a random variable ε such that $(\mathbf{X}, \varepsilon)$ is a multivariate normal vector, and if we define the outcome by $Y = f(\mathbf{X}) + \varepsilon$, then the vector $(\mathbf{X}, Y)$ is clearly a multivariate normal vector. Thus, the assumption on the joint distribution of $(\mathbf{X}, Y)$ is in fact mild in the regression framework. Note also that Proposition 2 corresponds to the criterion used in Guyon et al. [14] as a ranking criterion in the SVM-RFE algorithm with $\mathbb{V}(X_j) = 1$.

We now discuss the effect of the correlation between predictors on the importance measure by considering Gaussian regression models with various configurations of correlation between predictors:

Case 1: Two correlated variables. Consider the simple context where $(X_1, X_2, Y) \sim \mathcal{N}_3\left(0, \begin{pmatrix} C & \boldsymbol{\tau} \\ \boldsymbol{\tau}^t & 1 \end{pmatrix}\right)$ with

$$C = \begin{pmatrix} 1 & c \\ c & 1 \end{pmatrix},$$

and $\boldsymbol{\tau}^t = (\tau_0, \tau_0)$ with $\tau_0 \in (-1; 1)$. Since the two predictors have the same correlation τ_0 with the outcome, and according to Proposition 2, we have for $j \in \{1, 2\}$:

$$\alpha_j = \frac{\tau_0}{1 + c}. \tag{3.2}$$

Consequently, the permutation importance for both X_1 and X_2 is

$$I(X_j) = 2\left(\frac{\tau_0}{1 + c}\right)^2, \; j \in \{1, 2\}. \tag{3.3}$$

For positive correlations c, the importance of the two variables X_1 and X_2 decreases when c increases. This result is quite intuitive: when one of the two correlated variables is permuted, the error does not increase that much because of the presence of the other variable, which carries a similar information. The value of the prediction error after permutation is then close to the value of the prediction error without permutation and the importance is small.

Case 2: Two correlated and one independent variables. We add to the previous case an additional variable X_3 which is assumed to be independent of X_1 and X_2, X_1 and X_2 being unchanged. It corresponds to

$$C = \begin{pmatrix} 1 & c & 0 \\ c & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

and $\boldsymbol{\tau}^t = (\tau_0, \tau_0, \tau_3)$. It can be easily checked that

$$C^{-1}(\tau_0, \tau_0, \tau_3)^t = \left(\frac{\tau_0}{1+c}, \frac{\tau_0}{1+c}, \tau_3 \right)^t.$$

Thus, (3.2) and (3.3) still hold and $\alpha_3 = \tau_3$. As a consequence, $I(X_3) = 2\tau_3^2$ can be larger than $I(X_1)$ and $I(X_2)$ if the correlation c is sufficiency large even if $\tau_0 > \tau_3$. This phenomenon corresponds to the observation made by Tolo̧si and Lengauer [34].

Case 3: p correlated variables. We now consider p correlated variables where

$$C = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & 1 \end{pmatrix},$$

and $\boldsymbol{\tau}^t = (\tau_0, \ldots, \tau_0)$. In this context the α_j's are equal. The results of Case 1 can be generalized to that situation as shown by the following result:

Proposition 3. *Assume that the correlation matrix C can be decomposed into $C = (1-c)I_p + c\mathbb{1}\mathbb{1}^t$, where I_p is the identity matrix and $\mathbb{1} = (1, \ldots, 1)^t$. Let $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_0)^t \in \mathbb{R}^p$. Then for all $j \in \{1, \ldots, p\}$:*

$$[C^{-1}\boldsymbol{\tau}]_j = \frac{\tau_0}{1-c+pc},$$

and consequently

$$I(X_j) = 2\left(\frac{\tau_0}{1-c+pc} \right)^2. \tag{3.4}$$

Proof. see Appendix A.3 □

The proposition shows that the higher the number of correlated variables is, the faster the permutation importance of the variables decreases to zero (see Fig. 1). It confirms the second key observation of Tolo̧si and Lengauer [34].

Case 4: One group of correlated variables and one group of independent variables. Let us now assume that two independent blocks of predictors are observed. The first block corresponds to the p correlated variables $X_1, \ldots, X_p$ of Case 3 whereas the second block is composed of q independent variables $X_{p+1}, \ldots, X_{p+q}$. This case is thus a generalization of Case 2 where

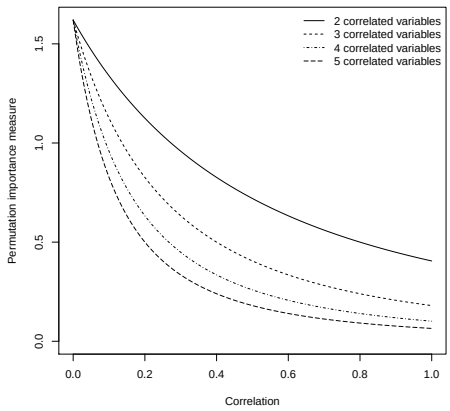


Figure 1: Case 3 - Permutation importance measure (3.4) versus the predictor correlation for $p \in \{2, 3, 4, 5\}$ and for $\tau_0 = 0.9$.

$$C = \begin{pmatrix} 1 & \cdots & c & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ c & \cdots & 1 & 0 & \cdots & 0 \\ 0 & \cdots & 0 & 1 & \cdots & 0 \\ \vdots & & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & 0 & \cdots & 1 \end{pmatrix},$$

and $\boldsymbol{\tau}^t = (\tau_0, \ldots, \tau_0, \tau_{p+1}, \ldots, \tau_{p+q})$. It can be checked that the importance of the correlated variables is still given by (3.4) and that $I(X_j) = 2\tau_j^2$ for $j \in \{p+1, \ldots, p+q\}$. Again, the independent variables may show higher importance values even if they are less informative than the correlated ones.

Case 5: Anti-correlation. All the previous cases consider positive correlations between the predictors. We now look at the effect of anti-correlation on the permutation measure. Let us consider two predictors X_1 and X_2 such that $X_2 = -\rho X_1 + \varepsilon$ where X_1 and ε are independent and $\rho \in (0, 1]$. The correlation between X_1 and X_2 equals $-\rho$, assuming that the variances of X_1 and X_2 are equal to 1. The permutation importance increases when ρ grows to 1 according to Equation (3.2). This surprising phenomenon can be explained intuitively: if ρ is close to -1, we need both X_1 and X_2 in the model to explain Y because they vary in two opposite directions. Consequently, the random permutations of one of the two variables induces a high prediction error. Finally, the permutation importance of these two variables is high for ρ close to -1.

Permutation importance measure for classification

We close this section by giving few elements regarding permutation importance measures in the classification framework. In this context, Y takes its values in $\{0, 1\}$. The error of a rule f for predicting Y is $R(f) = \mathbb{P}[f(\mathbf{X}) \neq Y]$. The function minimizing R is the Bayes classifier defined by $f(\mathbf{x}) = \mathbb{1}_{\eta(\mathbf{x})>0.5}$, where $\eta(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = \mathbb{P}[Y = 1|\mathbf{X} = \mathbf{x}]$ is the regression function. Given a classification rule $\hat{f}$, we consider its empirical error based on the learning set $\mathcal{D}_n$ and a

validation sample $\bar{\mathcal{D}}$:

$$\hat{R}(\hat{f}, \bar{\mathcal{D}}) = \frac{1}{|\bar{\mathcal{D}}|} \sum_{i:(\mathbf{X}_i, Y_i) \in \bar{\mathcal{D}}} \mathbb{1}_{\hat{f}(\mathbf{X}_i) \neq Y_i}.$$

Consequently, the permutation importance measure is

$$I(X_j) = \mathbb{P}\left[Y \neq f(\mathbf{X}_{(j)})\right] - \mathbb{P}\left[Y \neq f(\mathbf{X})\right],$$

and its empirical counterpart is

$$\hat{I}(X_j) = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} \left[\hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{tj}) - \hat{R}(\hat{f}_t, \bar{\mathcal{D}}_n^{t})\right],$$

as in Equation (2.1). We can equivalently rewrite the importance $I(X_j)$ as

$$\begin{aligned} I(X_j) &= \mathbb{E}[(Y - f(\mathbf{X}_{(j)}))^2] - \mathbb{E}[(Y - f(\mathbf{X}))^2] \\ &= \mathbb{E}[(Y - \eta(\mathbf{X}_{(j)}))^2] - \mathbb{E}[(Y - \eta(\mathbf{X}))^2] \end{aligned} \tag{3.5}$$

Of course the regression function does not satisfy the additive model (3.1) but we can consider alternatively the additive logistic regression model:

$$\text{Logit}(\eta(\mathbf{x})) = \sum_{j=1}^{p} f_j(X_j).$$

However, the permutation importance measure (3.5) cannot be easily related to the variance terms $\mathbb{V}[f_j(X_j)]$. In fact, this is possible by defining a permutation importance measure $\tilde{I}$ on the odd ratios $\frac{\eta(\mathbf{x})}{1-\eta(\mathbf{x})}$ rather than on the regression function as follows:

$$\tilde{I}(X_j) = \mathbb{E}\left[\left(\log \frac{\eta(\mathbf{X})}{1-\eta(\mathbf{X})} - \log \frac{\eta(\mathbf{X}_{(j)})}{1-\eta(\mathbf{X}_{(j)})}\right)^2\right].$$

Indeed, straightforward calculations show that

$$\tilde{I}(X_j) = 2\mathbb{V}[f_j(X_j)].$$

Roughly, the permutation of X_j has an impact in I only if the permutations change the predicted class (for instance when the odd ratios are close to 1). In contrast the perturbation of the odd ratio due to a permutation of X_j in $\tilde{I}$ is taken into account in $\tilde{I}$ whatever the value of the odd ratio. Nevertheless, the calculations we propose for the regression framework can be hardly adapted in this context, essentially because $\tilde{I}$ cannot be easily expressed in function of the correlations between variables. Moreover, as explained before, $\tilde{I}$ is less relevant than I for the classification purpose.

The results of this section show that the permutation importance is strongly sensitive to the correlation between the predictors. Our results also suggest that, for backward elimination strategies, the permutation importance measure should be recomputed each time a variable is eliminated. We study this question in the next section.

4 Wrapper algorithms for variable selection based on importance measures

In this section we study wrapper variable selection algorithms with random forests in the context of highly correlated predictors. In the applications we have in mind, the number of predictors is large and it is then impossible to evaluate the error of the all subsets of variables. Such an exhaustive exploration is indeed ruled out by the computational cost. One solution to this issue, which has been investigated in previous studies, is to first rank the variables according to some criterion and then to explore the subsets of variables according to this ranking. Several papers follow this strategy, they differ to each other first on the way the error is computed and second on the way the permutation importance of variables is updating during the algorithm.

Choosing the error estimator is out of the scope of this paper although various methods are proposed in the literature on this issue. For instance, Díaz-Uriarte and Alvarez de Andrés [10] and Genuer et al. [11] use the out-of-bag (OOB) error estimate whereas Jiang et al. [18] use both the OOB and a validation set to estimate the error. Finally, in order to avoid the selection bias described in Ambroise and McLachlan [1], Svetnik et al. [32] use an external 5-fold cross-validation procedure: they produce several variable selections on the 5 subsets of the training data and compute the averaged CV errors. In the sequel, the algorithms are performed by computing two kinds of errors : (i) OOB error which is widely used but often too optimistic as discussed in Breiman [7], (ii) validation set error which is more suitable but can not be considered in all practical situations.

We focus on the way the permutation importance measure is used in the algorithms. The first approach consists in computing the permutation importance only at the initialization of the algorithm and then to follow a backward strategy according to this "static" ranking. The method can be summarized as follows:

1. Rank the variables using the permutation importance measure
2. Train a random forest
3. Eliminate the less relevant variable(s)
4. Repeat steps 2 and 3 until no further variables remain

This strategy is called Non Recursive Feature Elimination (NRFE). Svetnik et al. [32], Díaz-Uriarte and Alvarez de Andrés [10] have developed such backward algorithms. More elaborated algorithms based on NRFE has been proposed in the literature as in Genuer et al. [11]. Since we are interested here in the effect of updating the measure importance, we only consider here the original version of NRFE.

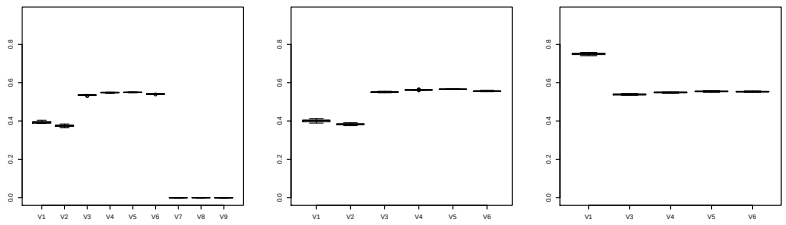
The second approach called Recursive Feature Elimination (RFE) is inspired by Guyon et al. [14] for SVM. It requires an updating of the permutation importance measures at each step of the algorithm. This strategy has been implemented in Jiang et al. [18]. The RFE algorithm implemented in this paper can be summarized as follows:

1. Train a random forest
2. Compute the permutation importance measure
3. Eliminate the less relevant variable(s)
4. Repeat steps 1 to 3 until no further variables remain

The two approaches are compared in Svetnik et al. [32]. The authors find that NRFE has better performance than RFE algorithm for their real life application. But as far as we know, no simulation studies have been carried out in the literature to confirm their observations. Moreover, this position goes against the theoretical considerations detailed above.

The results of the previous section show that the permutation importance measure of a given variable strongly depends on its correlation with the other ones and thus on the set of variables not yet eliminated in the backward process. As a consequence, RFE algorithm might be more reliable than NRFE since the ranking by the permutation importance measure is likely to change at each step. In the end, RFE algorithm can select smaller size models than NRFE since the most informative variables are well ranked in the last steps of the backward procedure even if they are correlated. In addition, by recomputing the permutation importance measure, we make sure that the ranking of the variables is consistent with their use in the current forest.

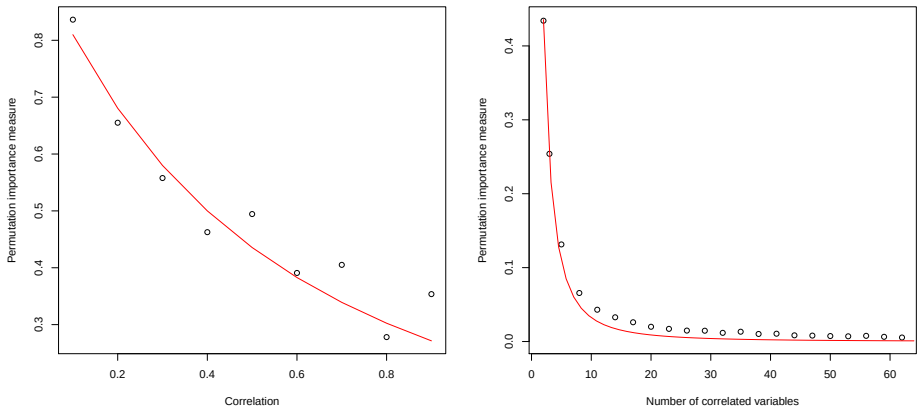
Let us consider a simple example to illustrate these ideas: we observe two correlated variables highly correlated with the outcome, four independent variables less correlated to the outcome and six irrelevant variables. More precisely, the variables are generated under the assumptions of Proposition 2: the correlation between the two relevant variables and the outcome is set to 0.7, the correlation between these variables is 0.9. The correlation between the independent variables and the outcome is 0.6. In addition, the variance of the outcome is set to 2 in order to have a positive-definite covariance matrix in the normal multivariate distribution. Figure 2 represents the boxplots of the permutation importance measures at several steps of the RFE algorithm. At the beginning of the algorithm, the permutation importance measure of the two first variables is lower than the independent ones (V3 to V6) even if they are more correlated to the outcome. Regarding the prediction performances of the selection procedure, one would like to select firstly one of the most informative variables (V1 or V2 in our example). At the last steps of the algorithm (Fig. 2c), one of the most relevant variables is eliminated and there is no more correlations between the remaining variables. The permutation importance of variable V1 becomes larger than the other variables. Consequently, RFE algorithm firstly selects this variable whereas it is selected in fifth position when using NRFE according to the ranking shown in Figure 2a.



(a) Step 1 (b) Step 4 (c) Step 5

Figure 2: RFE algorithm step by step with six relevant variables (two correlated and four independent) and three irrelevant variables.

In real life applications we usually need to find small size models with good performances in prediction. Thus, it is of first importance to efficiently reduce the effect of the correlations at the end of the backward procedure. By recomputing the variable importances at each step of the algorithm, the RFE algorithm manages to find small models which are efficient in term of prediction.



(a) (b)

Figure 3: Permutation importance measure versus the correlation (left) and the number of correlated predictors (right). The curves come from the expression of the permutation importance given in Proposition 3.

Remark. *Note that another importance measure has been proposed in Strobl et al. [31] for variable ranking with random forests in the context of correlated predictors. This importance measure, called conditional importance measure, consists in permuting variables conditionally to correlated ones. This method shows good performances for a small number of predictors. However it is computationally demanding and consequently it can be hardly implemented for problems with several hundreds of predictors.*

5 Numerical experiments

In this section, we verify with several experiments that the results proved in Section 3 for the permutation importance measure are also valid for its empirical version (2.1). RFE and NRFE approaches are compared for both classification and regression problems.

In the experiments, the number of trees in a forest is set to $n_{tree} = 1000$ and the number of variables randomly chosen for each split of the trees is set to the default value $m_{try} = \sqrt{p}$. For the NRFE algorithm, the permutation importance measure is averaged over 20 iterations as a preliminary ranking of the variables.

5.1 Correlation effect on the empirical permutation importance measure

This experiment is carried out under the assumptions of Proposition 2 and more precisely it corresponds to the regression problem presented in Case 1 and Case 3. Using the notations introduced in Section 3, the variance-covariance matrix C of the p covariates $X_1, \ldots, X_p$ has the form

$$C = \begin{pmatrix} 1 & c & \cdots & c \\ c & 1 & \cdots & c \\ \vdots & \vdots & \ddots & \vdots \\ c & c & \cdots & 1 \end{pmatrix},$$

where $c = \mathbb{C}(X_j, X_k)$, for $j \neq k$. The correlation between the X_j's and Y is denoted by τ_0. Two situations are considered. First, we take $p = 2$ (Case 1). The permutation importance measure is given by Proposition 3:

$$I(X_1) = I(X_2) = 2\left(\frac{\tau_0}{1+c}\right)^2. \tag{5.1}$$

Figure 3a represents the permutation importance measure of X_1 and its empirical counterpart versus the correlation c. The correlation τ_0 is set to 0.7 and c is varying between 0 and 1. We observe that the empirical permutation importance measure averaged over 100 simulations shares the same behaviour with the permutation importance measure (solid line in Fig. 3a) under predictor correlations.

Secondly, we consider p correlated predictors (Case 3) with $\tau_0 = 0.7$ and $c = 0.5$. The permutation importance is given by Proposition 3:

$$I(X_j) = 2\left(\frac{\tau_0}{1-c+pc}\right)^2. \tag{5.2}$$

In Figure 3b, the permutation importance measure and its empirical version are drawn versus the number of correlated predictors chosen among a grid between 2 and 62. We observe again that the empirical permutation importance measure fits with the permutation importance measure (solid line in Fig. 3b).

5.2 Variable selection for classification and regression problems

The RFE and NRFE algorithms are compared on several classification and regression experiments.

Experiment 1. This classification problem is inspired by Genuer et al. [11]. The procedure generates two groups of three relevant variables respectively highly, moderately and weakly discriminant and a group of irrelevant variables. The relevant variables are drawn conditionally to a realisation of the outcome Y. More precisely, the first three relevant variables are generated from the distribution $\mathcal{N}_1(Yj, 1)$ with probability 0.7 and from $\mathcal{N}_1(0, 1)$ with probability 0.3, $j \in \{1, 2, 3\}$. The variables 4 to 6 are simulated from the distribution $\mathcal{N}_1(0, 1)$ with probability 0.7 and from $\mathcal{N}_1(Y(j-3), 1)$ with probability 0.3, $j \in \{4, 5, 6\}$. The irrelevant variables are generated independently from the Gaussian distribution $\mathcal{N}_1(0, 20)$. Thus, conditionally to Y, the X_j's are drawn according to the following Gaussian mixtures densities:

$$p_{X_j}(x) = 0.7\varphi(x; Yj, 1) + 0.3\varphi(x; 0, 1),\ j \in \{1, 2, 3\},$$

$$p_{X_j}(x) = 0.7\varphi(x; 0, 1) + 0.3\varphi(x; Y(j-3), 1),\ j \in \{4, 5, 6\},$$

and

$$p_{X_j}(x) = \varphi(x; 0, 20),\ j \in \{7, \ldots, p\},$$

where $\varphi(\cdot; \mu, \sigma)$ is the normal density function with mean μ and standard error σ. We generate $n = 100$ samples and $p = 200$ variables.

Experiment 2. This classification problem is inspired by Toloşi and Lengauer [34]. Three groups of correlated variables are generated with a decreasing discriminative power: the variables are highly relevant in first group, they are weakly relevant in the second and irrelevant in the last group. The three groups respectively contain p_1, p_2 and p_3 variables. The simulation procedure is

the following: let $\mathbf{U} = (U_1, \ldots, U_n)^t$ be a vector of i.i.d. variables drawn according to the mixture density $\frac{1}{2}\varphi(\cdot\,; 0, 0.2) + \frac{1}{2}\varphi(\cdot\,; 1, 0.3)$. For $j \in \{1, \ldots, p_1\}$, let $\mathbf{U}^j$ a random vector defined by adding to a Gaussian noise $\mathcal{N}_1(0, 0.5)$ to 20 % of the elements of $\mathbf{U}$, the perturbated coordinates being chosen at random. Independently, some vectors $\mathbf{V}, \mathbf{V}^1, \ldots, \mathbf{V}^{p_2}$ and $\mathbf{R}$, $\mathbf{R}^1, \ldots, \mathbf{R}^{p_3}$ are drawn in the same way. Finally, the outcomes Y_i's are defined for $i \in \{1, \ldots, n\}$ by

$$Y_i = \begin{cases} 1 & \text{if } 5U_i + 4V_i - (\overline{5\mathbf{U} + 4\mathbf{V}}) + \varepsilon_i > 0 \\ 0 & \text{otherwise} \end{cases}, \tag{5.3}$$

where $\overline{5\mathbf{U} + 4\mathbf{V}}$ denotes the mean of the vector $5\mathbf{U} + 4\mathbf{V}$ and the ε_i's are i.i.d. random variables drawn according to $\mathcal{N}_1(0, 0.1)$. The problem considered here consists in predicting Y using as predictors all the $\mathbf{U}^j$, the $\mathbf{V}^j$ and the $\mathbf{R}^j$, but not $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{R}$. Since only $\mathbf{U}$ and $\mathbf{V}$ are involved in the definition of Y, it follows that only the $\mathbf{U}^j$ and the $\mathbf{V}^j$ are relevant. Moreover, the $\mathbf{U}^j$ are more relevant than the $\mathbf{V}^j$. This model is motivated in Toloşi and Lengauer [34] by a genomic application. It can be seen as slight perturbation of a linear discriminant protocol. We set the number of observations to $n = 100$, the number of variables to $p = 250$ with $p_1 = p_2 = 100$ and $p_3 = 50$.

Experiment 3. This experiment is an original classification problem based on Gaussian mixture distributions with a large number of predictors. We generate n_b groups $\mathcal{B}_1, \ldots, \mathcal{B}_{n_b}$ of correlated variables and an additional group $\mathcal{B}_{ind}$ of independent variables. Within a group $\mathcal{B}_\ell$, each variables are simulated from the mixture density

$$p_{X_j}(x) = \frac{1}{2}\varphi(x; 0, 1) + \frac{1}{2}\varphi(x; \mu_\ell, 1),\ j \in \mathcal{B}_\ell,\ \ell \in \{1, \ldots, n_b\},$$

and the variables in $\mathcal{B}_{ind}$ are simulated from

$$p_{X_j}(x) = \frac{1}{2}\varphi(x; 0, 1) + \frac{1}{2}\varphi(x; \mu_j, 1),\ j \in \mathcal{B}_{ind}.$$

The parameters μ_ℓ and μ_j give the discriminative power of the variables. We choose these mean parameters decreasing linearly from 1 to 0.5. In this way, the groups $\mathcal{B}_1, \ldots, \mathcal{B}_{n_b}$ have a decrease discriminative power which are higher than the independent group $\mathcal{B}_{ind}$. The variables in a group $\mathcal{B}_\ell$ of size q_ℓ are simulated from the multivariate Gaussian distribution $\mathcal{N}_{q_\ell}(0, C)$ with $C = (1 - c)Id + c\mathbb{1}\mathbb{1}^t$ where c is the correlation between two different variables of $\mathcal{B}_\ell$. For the experiment, we simulate $n = 250$ samples and $p = 500$ variables: 4 blocks of 15 variables highly correlated with $c = 0.9$, one block $\mathcal{B}_{ind}$ of 10 independent variables and 430 irrelevant variables.

Experiment 4. This classification problem is inspired by Archer and Kimes [2]. We simulate $n = 100$ independent vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ as follows: the mean vectors μ_i's are chosen uniformly distributed in $[6; 12]^p$ and the $\mathbf{X}_i$'s are drawn from the Gaussian distribution $\mathcal{N}_p(\mu_i, C)$. Each observation is composed of $L = 20$ independent blocks of $K = 40$ covariates ($p = 800$). The corresponding covariance matrix of the observed complete vector has the block-diagonal form

$$C = \begin{pmatrix} C_1 & 0 & \cdots & 0 \\ 0 & C_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & C_L \end{pmatrix}.$$

The covariance matrix C_ℓ of the ℓ-th group is

$$C_\ell = \begin{pmatrix} 1 & \rho_\ell & \cdots & \rho_\ell \\ \rho_\ell & 1 & \cdots & \rho_\ell \\ \vdots & \vdots & \ddots & \vdots \\ \rho_\ell & \rho_\ell & \cdots & 1 \end{pmatrix},$$

where the correlations are set to $\rho_\ell = 0.05\ell - 0.05$. The correlation of each block is taken from 0 to 0.95 by increments of 0.05. We then compute the probability that the observation $\mathbf{X}_i$ is from class 1 by

$$\pi_i = \frac{e^{\mathbf{X}_i^t \beta}}{1 + e^{\mathbf{X}_i^t \beta}},$$

for $i \in \{1, \ldots, n\}$. The regression coefficients β are sparse, that is $\beta_j = \beta_0$ if $j = (\ell - 1)K + 1$ for some $l \in \{1, \ldots, L\}$ and 0 otherwise. In other words, the posterior probability π_i is generated using only informations from the first variables of each group. Finally, for $i \in \{1, \ldots, n\}$ the response Y_i is generated from

$$Y_i = \begin{cases} 1 & \text{if } \pi_i < U_i, \text{ where } U_i \text{ is an uniform distribution on } [0, 1] \\ 0 & \text{otherwise.} \end{cases}$$

Experiment 5. We now propose a complex linear regression design which satisfies the assumptions of Proposition 2. We simulate $n = 100$ i.i.d. copy of the random vector $(\mathbf{X}, Y)$ from the multivariate Gaussian distribution $\mathcal{N}_{p+1}\left(0, \begin{pmatrix} C & \boldsymbol{\tau} \\ \boldsymbol{\tau}^t & 1 \end{pmatrix}\right)$ where the vector $\boldsymbol{\tau} = (\tau_0, \ldots, \tau_0, 0, \ldots, 0)^t$ contains the covariances between each predictor X_j and Y, $\tau_j = \tau_0$ for the relevant variables and $\tau_j = 0$ for the irrelevant ones. We consider L groups of correlated variables and some additional irrelevant and independent variables. The matrix C has a block-diagonal form

$$C = \begin{pmatrix} C_1 & 0 & \cdots & 0 & 0 \\ 0 & C_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & C_L & 0 \\ 0 & 0 & \cdots & 0 & Id \end{pmatrix},$$

where

$$C_\ell = \begin{pmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{pmatrix}.$$

This design differs from the experiment 5 since the correlations are all equal in each group. Note that the blocks can be simulated with different size to each other in order to highlight the effect of the number of correlated variables. We take 4 groups of 5 variables, 2 groups of 15 variables and 50 irrelevant variables. The correlation ρ is set to 0.9 and τ_0 is equal to 0.3 for the relevant variables and is equal to 0 for the irrelevant ones.

Variable selection

We compare the performances of the two wrapper algorithms RFE and NRFE for the five experiments described above. The model error corresponds to the misclassification rate for the

		Minimum error model		Parsimonious model		
		RFE	NRFE	p^*	RFE	NRFE
OOB	Exp. 1	**0.0139**	0.0175	5	**0.0188**	0.0208
	Exp. 2	**0.0385**	0.0628	8	**0.0445**	0.1487
	Exp. 3	**0.1220**	0.1383	4	**0.1664**	0.2398
	Exp. 4	**0.1726**	0.2080	5	**0.2219**	0.2820
	Exp. 5	**0.4860**	0.4880	8	**0.5324**	0.6591
Test	Exp. 1	**0.0167**	0.0215	5	**0.0242**	0.0267
	Exp. 2	**0.0794**	0.0843	8	**0.0950**	0.1788
	Exp. 3	**0.1462**	0.1508	8	**0.1822**	0.2488
	Exp. 4	**0.3821**	0.3910	6	**0.4472**	0.4667
	Exp. 5	**0.4995**	0.5051	8	**0.5706**	0.6954

Table 1: Averaged error estimates over 100 runs. The column p^* gives the number of variables in the parsimonious models we choose.

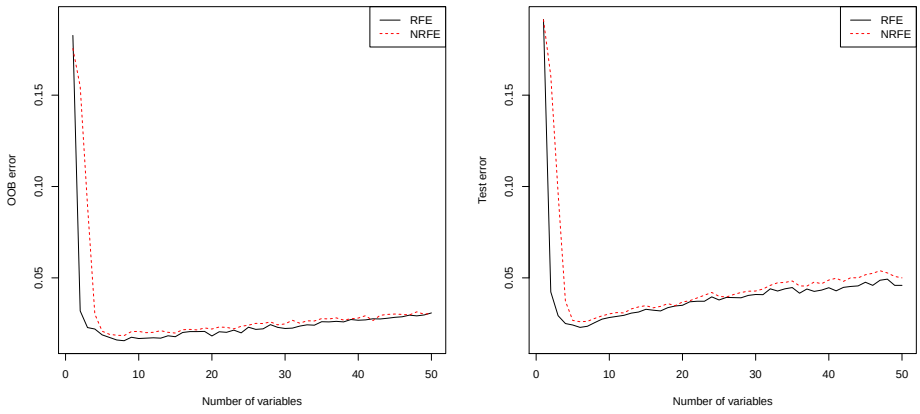
classification and to the mean square error in regression case. As mentioned in Section 4, the error is estimated in two different ways : the out-of-bag error embedded in the random forests and the error obtained using a test set simulated independently. For each algorithm, the errors are computed in function of the number of variables in the model. The procedure is repeated 100 times to eliminate the estimation variability. We also provide for each experiment the boxplots of the initial permutation importance measures, that is the permutation importance measures used by NRFE for ranking the variables.

Table 1 summarizes the performances of the two variable selection approaches. The errors are given for a model which minimizes the error among the selection path induced by the backward search (Minimum error model). We also consider the errors of a parsimonious model chosen with a few number of predictors. Clearly, RFE have better prediction performances for all experiments. As expected, the differences between the RFE and the NRFE errors are higher in the parsimonious models than in the minimum error models. Indeed, RFE is particularly useful to reduce the effect of correlations in the last steps of the variable selection procedure and tends to select firstly the variables which are the most able to predict the outcome.

As regards Experiment 1, RFE and NRFE achieve similar performances (see Fig. 4). This can be explained by the fact that the design used for this experiment does not involve correlations between the predictors since all of the variables are simulated to be independent. For this reason, RFE algorithm does not bring anything more than the initial ranking obtained by the permutation importance measure computed with all the variables as shown in Figure 5.

Concerning Experiments 2 and 3, Figures 6 and 8 show a meaningful difference between RFE and NRFE regarding the number of involved variables. NRFE gives a minimum error close to RFE but it needs 60 variables whereas RFE procedure only needs less than 20 variables to reach the minimum error rate. Indeed, the boxplots of the initial permutation importance measures highlight the effects of the correlations (see Fig. 7 and Fig. 9).

Experiment 4 is a tricky problem in a high dimensional setting: only 100 observations of 800 variables are available. Without surprise, as illustrated in Figure 11, it is a difficult task to find the relevant variables from the empirical permutation importances. In particular, we cannot clearly discriminate the first variables of each group as they do not show higher importances than the other variables in the same block. In this complex simulation design related to gene expressions [2], RFE and NRFE give similar performances according to the test error whereas RFE seems more efficient according to OOB error (see Fig. 10).



(a) (b)

Figure 4: Exp. 1 - Out-of-bag error estimate (left) and test set estimate (right) versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

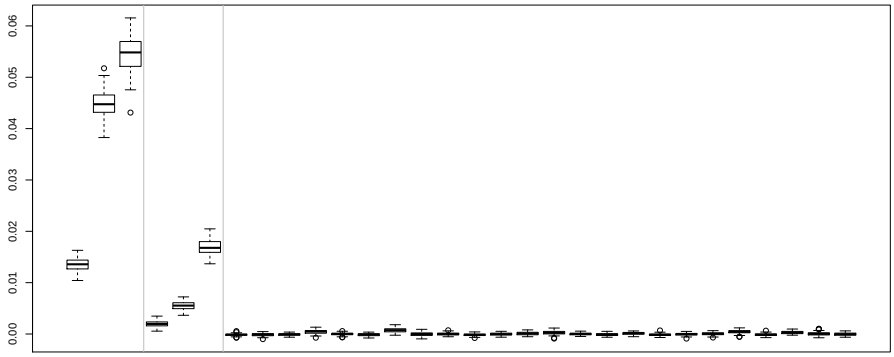
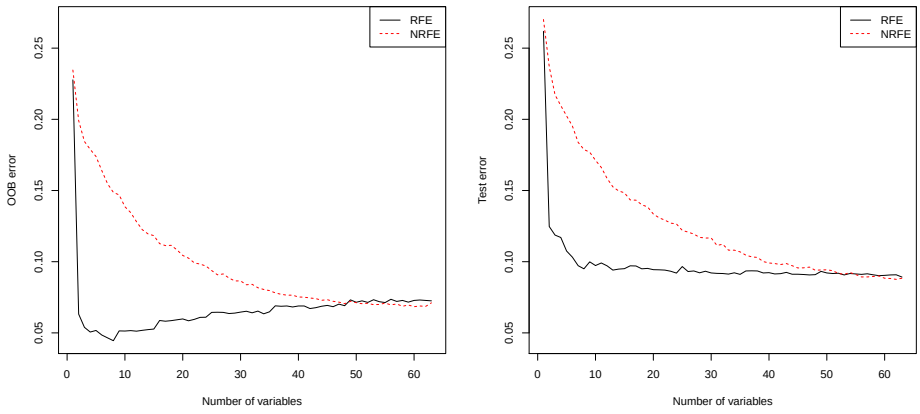


Figure 5: Exp. 1 - Boxplots of the initial permutation importance measures. Only the 6 relevant variables and 24 irrelevant variables are displayed.



(a) (b)

Figure 6: Exp. 2 - Out-of-bag error estimate (left) and test set estimate (right) versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

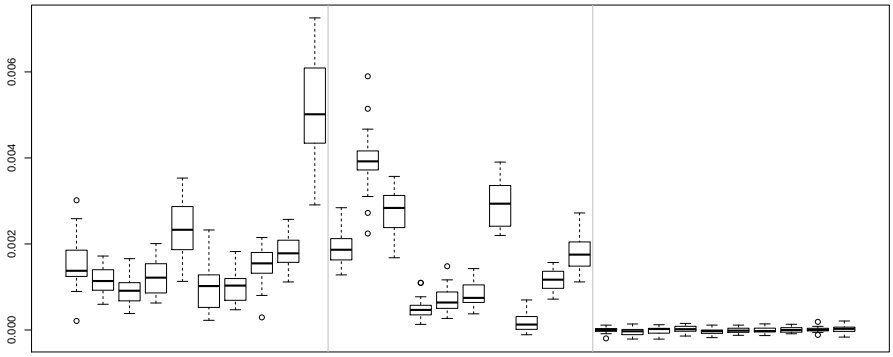
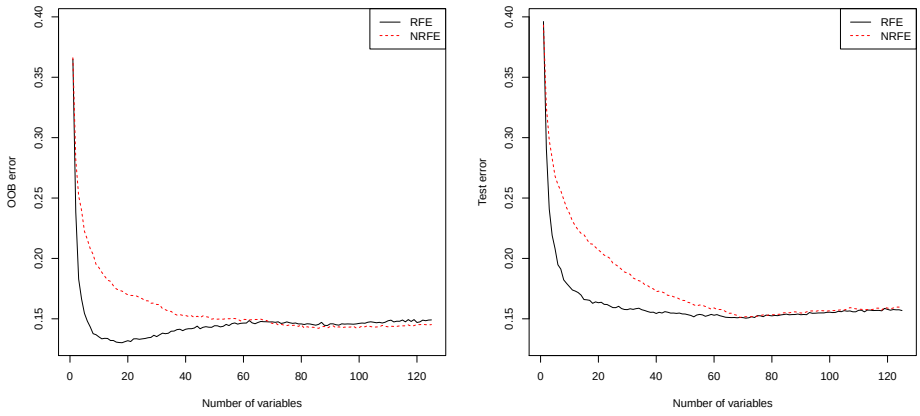


Figure 7: Exp. 2 - Boxplots of the initial permutation importance measures. Only 10 variables of each group are displayed.



(a) (b)

Figure 8: Exp. 3 - Out-of-bag error estimate (left) and test set estimate (right) versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

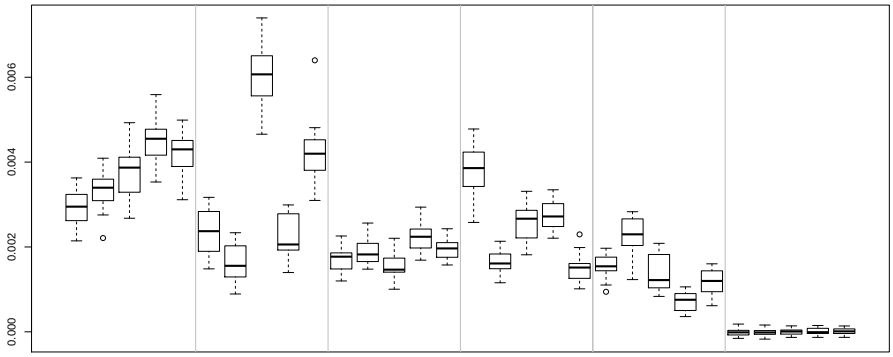
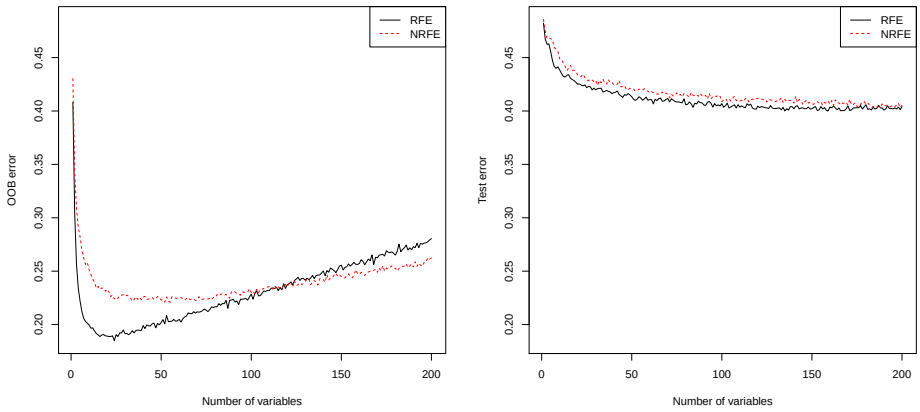


Figure 9: Exp. 3 - Boxplots of the initial permutation importance measures. Only 5 variables of each group are displayed.



(a) (b)

Figure 10: Exp. 4 - Out-of-bag error estimate (left) and test set estimate (right) versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

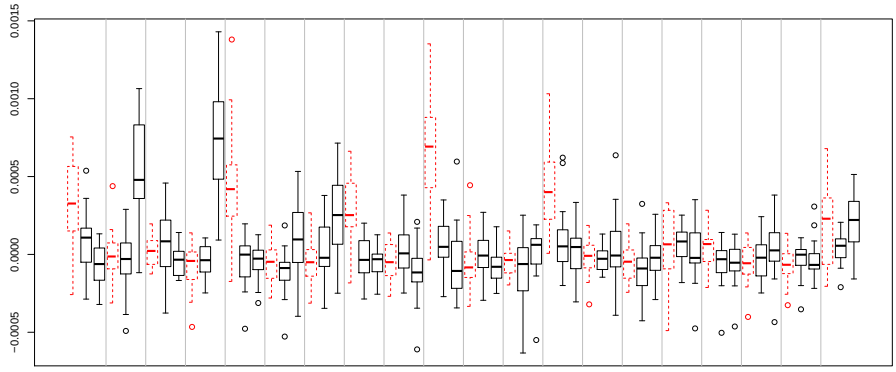
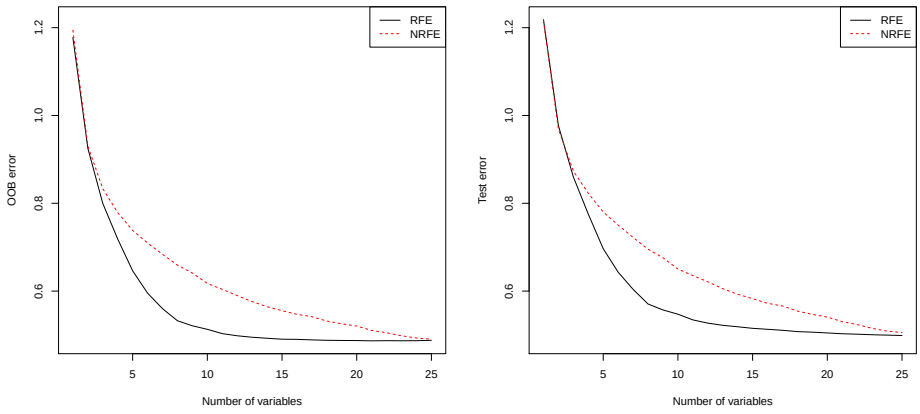


Figure 11: Exp. 4 - Boxplots of the initial permutation importance measures. Only the first variables in each group (dashed lines) and two additional variables of the same group (solid lines) are displayed.



(a) (b)

Figure 12: Exp. 5 - Out-of-bag MSE error (left) and test set (right) estimate versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

Regarding the regression problem studied in Experiment 5, RFE also outperforms NRFE according to mean squared error estimated from OOB sample and test sample (see Table 1). Figures 12 and 13 illustrates the fact that RFE reaches an error level close to the minimum error much faster than NRFE. Moreover, the relevant variables are well detected by the empirical permutation importance measure even if we cannot clearly identify the different groups of correlated variables (see Fig. 13).

These simulations confirm the well-known fact that OOB error tends to be too optimistic. We have also checked that the correlation increases the instability of the empirical permutation importance, as explained in Tolo̧si and Lengauer [34] and Genuer et al. [11].

6 Application to flight data analysis

In this section the two approaches are applied to a real life problem coming from aviation safety. Airlines collect many informations during their flights using flight data recorders. Since several years, airlines have to use these data for flight safety purposes. A large number of flight parameters (up to 1000) are recorded each second as for instance the aircraft speed and accelerations, the heading, the position, several warnings. A flight provides a set of time series corresponding to these variables.

An important goal in this context is to perform variable selection in order to obtain a simple model showing good performances in term of prediction. This interpretable model for predicting a particular airline risk could be used for pilot training or for developments of new flight procedures.

We focus here on the risk of long landing. We propose to describe the risk with two levels: a flight is labelled by 1 if the landing distance exceeds 60 % of the runway length and by 0 otherwise. This naive approach does not catch the whole complexity of the phenomenon but it is better understood by safety experts.

Following the recommendations of aviation experts, 22 numerical variables are preselected and the last ten minutes before touchdown are extracted from the recordings. A sample of

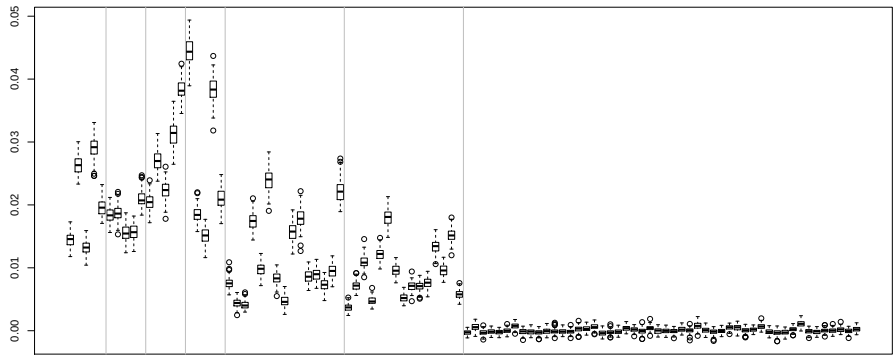


Figure 13: Exp. 5 - Boxplots of the initial permutation importance measures. Four groups of 5 correlated variables and two groups of 15 correlated variables are simulated.

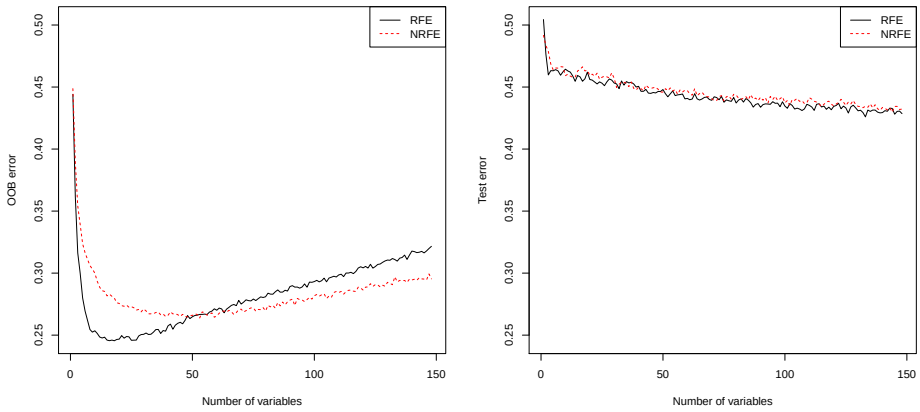
254 flights from the same airport and the same company is considered. The covariates obtained are highly correlated, due for instance to physical laws describing some relations between them. Some of these relations are linear but many of them are not. This problem is thus more complicated than those studied before in the paper.

The random forests algorithm can be adapted for longitudinal observations. Our approach is based on a projection of each time series in a Daubechies wavelet basis. All the variables are projected on a common wavelet basis. We consider the 8 first wavelets coefficients for each variable in order to speed up the algorithm. The obtained 176 wavelet coefficients are the input variables of the random forests algorithm.

For both RFE and NRFE, the OOB error and the classification error estimated using a test set are given in Figure 14. The test set is randomly chosen. It contains one third of the observations. The two algorithms provide similar results when the error is estimated using a test set (see Fig. 14b). As regards the OOB error, the results reinforce the simulation study. Indeed, the OOB errors for RFE algorithm are always lower for small size model (see Fig. 14a).

We now analyse the outputs of the RFE algorithm carried on with the test set approach. For improving the stability of the variable selection, we aggregate the procedure over 100 iterations using sub-sampling in the same spirit as the stability selection method from Meinshausen and Bühlmann [24]. By doing this, we compute the proportion of times each wavelet coefficient is selected. This procedure does not reveal the most relevant flight parameters for predicting long landings. One solution is to average for each flight parameter the selection frequencies of the eight coefficients. The resulted score provide the proportion of times at least one of the eight wavelet coefficients has been selected by the RFE algorithm. This indicator helps us to better understand the relationship between the flight parameters and the risk of long landing (see Fig. 15).

This procedure reveals four relevant flight parameters: the static air temperature (SAT), the altitude (ALT_STDC), the gross weight (GW_KG) and the wind speed (CASC_GSC). Figure 16 displays two profiles corresponding to a normal (solid lines) and an abnormal flight (dashed lines) for the variables ALT_STDC, CASC_GSC and SAT. The solid lines are consistent to normal operations. Indeed, aircraft have to level off at 4000 feet for stabilising before the final approach (see Fig. 16a). It is also necessary to have head wind for landing (positive values of the wind speed) in order to help the aircraft to reduce the airspeed. The abnormal profile



(a) (b)

Figure 14: Flight data - Out-of-bag error estimate (left) and test set estimate (right) versus the number of variables for RFE and NRFE algorithms. The curves are averaged over 100 runs of variable selections.

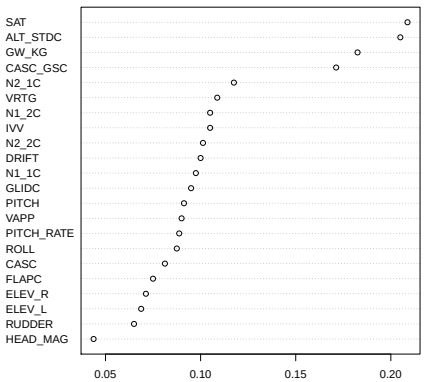
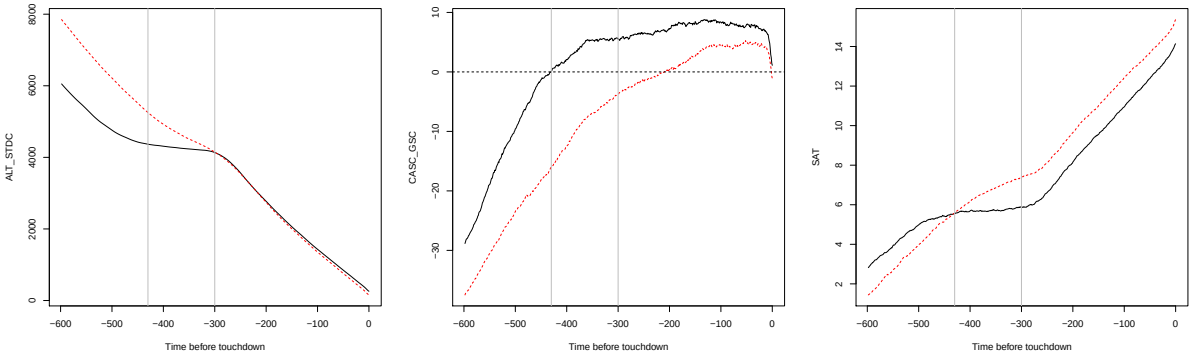


Figure 15: Flight data - Selection percentage of the flight parameters



(a) Altitude (b) Wind speed (c) Static air temperature

Figure 16: Flight data - Normal (solid lines) and abnormal (dashed lines) profiles for the risk of long landing. The most relevant time sequence is between the two vertical lines.

of ALT_STDC displayed in Figure 16a does not level off at 4000 feet. The abnormal profile of CASC_GSC displayed in Figure 16b shows a tail wind which has a negative effect on the airspeed reduction. In addition, the profiles of static air temperature in Figure 16c reveal two different atmospheric conditions which are related to the wind speed. Finally, the analysis of the gross weight shows that the average of the weights are 165 tonnes for the abnormal flights and 160 tonnes for the regular flights. This has also a negative effect on the deceleration efficiency on the runway.

7 Conclusion

In this paper, we studied the problem of variable selection using the permutation importance measure from the random forests. Several simulation studies in the literature have shown an effect of the correlations between predictors on this criterion.

We first provided some theoretical insights on the effect of the correlations on the permutation importance measure. Considering an additive regression model, we obtained a very simple expression of this criterion which depends on the correlation between the covariates and on the number of correlated variables. Extending our results to a more general context is a challenging problem, this question should be investigated deeply for improving our knowledge of this widely used criterion. Moreover, the impact of the correlations on other importance measures (see van der Laan [35], Ishwaran [17]) is a general question of great interest.

In a second step we focused on variable selection algorithm based on random forests analysis. A recursive and a non recursive approaches have been evaluated through an extensive simulation study on several classification and regression designs. As expected, the RFE algorithm provides better performances than the non recursive one in presence of correlated predictors: the prediction errors is always smaller with recursive strategy when small size models are selected. As a matter of fact RFE reduces the effect of the correlation on the importance measure.

In the RFE algorithm, updating the ranking is especially crucial at the last steps of the algorithm, when most of the irrelevant variables have been eliminated. In future works, the algorithm could be adapted by combining a non recursive strategy at the first steps and a recursive strategy at the end of the algorithm.

A Proofs

A.1 Proof of Proposition 1

The random variable X_j' and the vector $\mathbf{X}_{(j)}$ are defined as in Section 2:

$$\begin{aligned} I(X_j) &= \mathbb{E}[(Y - f(\mathbf{X}) + f(\mathbf{X}) - f(\mathbf{X}_{(j)}))^2] - \mathbb{E}[(Y - f(\mathbf{X}))^2] \\ &= \mathbb{E}[(f(\mathbf{X}) - f(\mathbf{X}_{(j)}))^2] + 2\mathbb{E}\left[\varepsilon(f(\mathbf{X}) - f(\mathbf{X}_{(j)}))\right] \\ &= \mathbb{E}[(f(\mathbf{X}) - f(\mathbf{X}_{(j)}))^2], \end{aligned}$$

since $\mathbb{E}[\varepsilon f(\mathbf{X})] = \mathbb{E}[f(\mathbf{X})\mathbb{E}[\varepsilon|\mathbf{X}]] = 0$ and $\mathbb{E}[\varepsilon f(\mathbf{X}_{(j)}] = \mathbb{E}(\varepsilon)\mathbb{E}[f(\mathbf{X}_{(j)}] = 0$. Since the model is additive, we have:

$$\begin{aligned} I(X_j) &= \mathbb{E}[(f_j(X_j) - f_j(X_j'))^2] \\ &= 2\mathbb{V}[f_j(X_j)], \end{aligned}$$

as X_j and X_j' are independent and identically distributed. For the second statement of the proposition, using the fact that $f_j(X_j)$ is centered we have:

$$\begin{aligned} \mathbb{C}[Y, f_j(X_j)] &= \mathbb{E}\left[f_j(X_j)\mathbb{E}[Y|\mathbf{X}]\right] = \mathbb{E}[f_j(X_j)\sum_{k=1}^{p} f_k(X_k)] \\ &= \mathbb{V}[f_j(X_j)] + \sum_{k\neq j} \mathbb{E}\left[f_j(X_j)f_k(X_k)\right] \\ &= \frac{I(X_j)}{2} + \sum_{k\neq j} \mathbb{C}\left[f_j(X_j), f_k(X_k)\right]. \end{aligned}$$

A.2 Proof of Proposition 2

This proposition is an application of Proposition 1 for a particular distribution. We only show that $\alpha = C^{-1}\boldsymbol{\tau}$ in that case.

Since $(\mathbf{X}, Y)$ is a normal multivariate vector, the conditional distribution of Y over $\mathbf{X}$ is also normal and the conditional mean $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ is a linear function: $f(\mathbf{x}) = \sum_{j=1}^{p} \alpha_j x_j$ (see for instance Rao [30], p. 522). Then, for any $j \in \{1, \ldots, p\}$,

$$\begin{aligned} \tau_j &= \mathbb{E}[X_j Y] \\ &= \mathbb{E}[\ X_j\mathbb{E}[Y|\mathbf{X}]\ ] \\ &= \alpha_1\mathbb{E}[X_1X_j] + \cdots + \alpha_j\mathbb{E}[X_j^2] + \cdots + \alpha_p\mathbb{E}[X_pX_j] \\ &= \alpha_1 c_{1j} + \cdots + \alpha_j c_{jj} + \cdots + \alpha_p c_{pj}. \end{aligned}$$

The vector α is thus solution of the equation $\boldsymbol{\tau} = C\alpha$ and the expected result is proven since the covariance matrix C is invertible.

A.3 Proof of Proposition 3

The correlation matrix C is assumed to have the form $C = (1-c)I_p + c\mathbb{1}\mathbb{1}^t$. We show that the invert of C can be decomposed in the same way. Let $M = aI_p + b\mathbb{1}\mathbb{1}^t$ where a and b are real numbers to be chosen later. Then

$$\begin{aligned} CM &= \big((1-c)I_p + c\mathbb{1}\mathbb{1}^t\big)\big(aI_p + b\mathbb{1}\mathbb{1}^t\big) \\ &= a(1-c)I_p + b(1-c)\mathbb{1}\mathbb{1}^t + ac\mathbb{1}\mathbb{1}^t + bc\mathbb{1}\mathbb{1}^t\mathbb{1}\mathbb{1}^t \\ &= a(1-c)I_p + (b(1-c) + ac + pbc)\mathbb{1}\mathbb{1}^t, \end{aligned}$$

since $\mathbb{1}^t\mathbb{1} = p$. Thus, $CM = I_d$ if and only if

$$\begin{cases} a(1-c) = 1 \\ b(1-c) + ac + pbc = 0, \end{cases}$$

which is equivalent to

$$\begin{cases} a = \dfrac{1}{(1-c)} \\ b = \dfrac{-c}{(1-c)(1-c+pc)}. \end{cases}$$

Consequently, $M^{-1}_{jk} = C^{-1}_{jk} = b$ if $j \neq k$ and $M^{-1}_{jk} = C^{-1}_{jj} = a + b$. Finally we find that for any $j \in \{1 \dots p\}$:

$$\begin{aligned} [C^{-1}\boldsymbol{\tau}]_j &= \tau_0(a+b) + \tau_0 b(p-1) \\ &= \tau_0(a+pb) \\ &= \tau_0\left(\frac{1}{(1-c)} - \frac{pc}{(1-c)(1-c+pc)}\right) \\ &= \frac{\tau_0}{1-c+pc}. \end{aligned}$$

The second point derives from Proposition 2.

Acknowledgements

The authors would like to thank Gérard Biau for helpful discussions and the Safety Line company for making the flight data available for this study.

References

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

[2] K. J. Archer and R. V. Kimes. Empirical characterization of random forest variable importance measures. *Computational Statistics and Data Analysis*, 52:2249–2260, 2008.

[3] L. Auret and C. Aldrich. Empirical comparison of tree ensemble variable importance measures. *Chemometrics and Intelligent Laboratory Systems*, 105:157–170, 2011.

[4] J. Bi, K. P. Bennett, M. Embrechts, C. M. Brenemanand, and M. Song. Dimensionality reduction via sparse support vector machines. *Journal of Machine Learning Research*, 3: 1229–1243, 2003.

[5] A. L. Blum and P. Langley. Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271, 1997.

[6] L. Breiman. Bagging predictors. *Machine Learning*, 24:123–140, 1996.

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

[8] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Advanced Books and Software, 1984.

[9] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang. Correlated variables in regression: Clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1858, 2013.

[10] R. Díaz-Uriarte and S. Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7:3, 2006.

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

[12] U. Grömping. Variable importance assessment in regression: linear regression versus random forest. *The American Statistician*, 63:308–319, 2009.

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

[14] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422, 2002.

[15] A. Hapfelmeier and K. Ulm. A new variable selection approach using random forests. *Computational Statistics and Data Analysis*, 60:50–69, 2013.

[16] A.-C. Haury, P. Gestraud, and J.-P. Vert. The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS ONE*, 6:1–12, 2011.

[17] H. Ishwaran. Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, 1:519–537, 2007.

[18] H. Jiang, Y. Deng, H.-S. Chen, L. Tao, Q. Sha, J. Chen, C.-J. Tsai, and S. Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5:81, 2004.

[19] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and Information Systems*, 12:95–116, 2007.

[20] R. Kohavi and G. H. John. Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324, 1997.

[21] P. Křížek, J. Kittler, and V. Hlaváč. Improving stability of feature selection methods. In *Computer Analysis of Images and Patterns*. Springer Berlin Heidelberg, 2007.

[22] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe. A survey on filter techniques for feature selection in gene expression microarray analysis. *IEEE/ACM Transansaction on Computation Biology and Bioinformatics*, 9:1106–1119, 2012.

[23] N. Louw and S.J. Steel. Variable selection in kernel fisher discriminant analysis by means of recursive feature elimination. *Computational Statistics and Data Analysis*, 51:2043–2055, 2006.

[24] N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society, Series B*, 72:417–473, 2010.

[25] P. G. Neville. Controversy of variable importance in random forests. *Journal of Unified Statistical Techniques*, 1:15–20, 2013.

[26] K. K. Nicodemus. Letter to the editor: On the stability and ranking of predictors from random forest variable importance measures. *Briefings in Bioinformatics*, 12:369–373, 2011.

[27] K. K. Nicodemus and J. D. Malley. Predictor correlation impacts machine learning algorithms: implications for genomic studies. *Bioinformatics*, 25:1884–1890, 2009.

[28] K. K. Nicodemus, J. D. Malley, C. Strobl, and A. Ziegler. The behaviour of random forest permutation-based variable importance measures under predictor correlation. *BMC Bioinformatics*, 11:110, 2010.

[29] A. Rakotomamonjy. Variable selection using svm based criteria. *Journal of Machine Learning Research*, 3:1357–1370, 2003.

[30] C.R. Rao. *Linear statistical inference and its applications*. Wiley series in probability and mathematical statistics: Probability and mathematical statistics. Wiley, 1973.

[31] C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307, 2008.

[32] V. Svetnik, A. Liaw, C. Tong, and T. Wang. Application of breiman's random forest to modeling structure-activity relationships of pharmaceutical molecules. In *Multiple Classifier Systems*, 2004.

[33] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1996.

[34] L. Tolo̧si and T. Lengauer. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics*, 27:1986–1994, 2011.

[35] M. J. van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2:1–33, 2006.

[36] R. Zhu, D. Zeng, and M. R. Kosorok. Reinforcement learning trees. Submitted to the Annals of Statistics, 2012.