Grouped variable importance with random forests and application to multivariate functional data analysis

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*
Sorbonne Universités, UPMC Univ Paris 06
F-75005, Paris, France

Abstract

In this paper, we study the selection of grouped variables using the random forests algorithm. We first propose a new importance measure adapted for groups of variables. Theoretical insights of this criterion are given for additive regression models. The second contribution of this paper is an original method for selecting functional variables based on the grouped variable importance measure. Using a wavelet basis, we propose to regroup all of the wavelet coefficients for a given functional variable and use a wrapper selection algorithm with these groups. Various other groupings which take advantage of the frequency and time localisation of the wavelet basis are proposed. An extensive simulation study is performed to illustrate the use of the grouped importance measure in this context. The method is applied to a real life problem coming from aviation safety.

Keywords: random forests, functional data analysis, group permutation importance measure, group variable selection

1 Introduction

In high dimensional setting, the identification of the most relevant variables has been the subject of much research during the last two decades (Guyon and Elisseeff; 2003). For linear regression, the Lasso method (Tibshirani; 1996) is widely used. Many variable selection procedures have also been proposed for non linear methods. In the context of random forests (Breiman; 2001), it has been shown that the permutation importance measure introduced by Breiman, is an efficient tool for selecting variables (Díaz-Uriarte and Alvarez de Andrés; 2006; Genuer et al.; 2010; Gregorutti et al.; 2014).

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

In many situations, as in medical studies and genetics, groups of variables can be clearly identified and it is of interest to select groups of variables rather than to select variables individually (He and Yu; 2010). Indeed, the interpretation of the model may be improved as well as the prediction accuracy by grouping the variables according to an a priori knowledge on the data. In the end, grouping the variables can be seen as a solution to stabilize variable selection methods. In linear settings, and more particularly for the linear regression, the group Lasso has been developed to deal with groups of variables, see for instance (Yuan and Lin; 2006). Group variable selection have also been proposed for kernel methods (Zhang et al.; 2008) and neural networks (Chakraborty and Pal; 2008). As far as we know, this problem has not been studied for the random forests algorithm introduced by Breiman (2001). In this paper, we adapt the permutation importance measure for groups of variables in order to select groups of variables in the context of random forests.

The first contribution of this paper is a theoretical analysis of the grouped variable importance measure. Generally speaking, the grouped variable importance does not reduce to the sum of the individual importances and it can be hardly related to these last. However, in more particular models such as additive regression models, we derive exact decompositions of the grouped variable importance measure.

The second contribution of this work is an original method for selecting functional variables based on the grouped variable importance measure. Functional Data Analysis (FDA) is a field in statistics that analyzes data indexed by time (Ramsay and Silverman; 2005; Ferraty and Vieu; 2006; Ferraty; 2011). One standard approach in FDA consists in projecting the functional variables on a finite dimensional space spanned by a functional basis. Classical bases in this context are Splines, wavelets, Karhunen-Loève expansion for instance. Most of the papers about regression and classification methods for functional data consider only one functional predictor. Some references are Cardot et al. (1999, 2003); Rossi et al. (2006); Cai and Hall (2006) for linear regression methods, Amato et al. (2006); Araki et al. (2009) for logistic regression methods, Biau et al. (2005); Fromont and Tuleau (2006) for k-NN algorithms and Rossi and Villa (2006, 2008) for SVM classification. The multivariate FDA problem, where p functional variables are observed, has been less studied. Recently, Matsui and Konishi (2011); Fan and James (2013) have proposed answers to the linear regression problem with Lasso-like penalties. The logistic regression has been studied by Matsui (2014). Classification based on multivariate functional variables has also been considered using CART algorithm (Poggi and Tuleau; 2006) and SVM (Yang et al.; 2005; Yoon and Shahabi; 2006).

We propose a new approach of multivariate FDA using random forest and the grouped variable importance measure. Indeed, various groups of basis coefficients can be proposed for a given functional decomposition. For instance, one can choose to regroup all the coefficients of a given functional variable. In this case, the selection of a group of coefficients corresponds to the selection of a functional variable. Various other groupings are proposed for a wavelet decomposition. For a given family of groups, we adapt the recursive feature elimination algorithm (Guyon et al.; 2002) which is particularly efficient when the predictors are strongly correlated (Gregorutti et al.; 2014). In the context of random forests, this backward-like selection algorithm is guided by the grouped variable importance. Note that by regrouping the coefficients, the computational cost of the algorithm is drastically reduced compared to a backward strategy that would eliminate only one coefficient at each step.

An extensive simulation study illustrates the applications of the grouped importance measure for FDA. The method is finally applied to a real life problem coming from aviation safety. The aim of this study is to explain and predict landing distances. We select the most relevant flight parameters regarding the risk of long landings, which is a major issue for airlines. In order to speed up the algorithm, the dimension of the flight data is reduced in a preprocessing step. In Appendix B, we propose a modified version of the well-known shrinkage method Donoho and Johnstone (1994) that simultaneously shrinks to zero the coefficients of the observed curves of a functional variable.

The group permutation importance measure is introduced in Section 2. Section 3 deals with multivariate FDA using random forests and the grouped variable importance measure. The application to flight data analysis is developed in Section 4.

2 Grouped variable importance measure

Let $u^\top$ denote the transpose of the vector $u \in \mathbb{R}^p$. Let Y be a random variable in $\mathbb{R}$ and $\mathbf{X} = (X_1, \ldots, X_p)$ a random vector in $\mathbb{R}^p$. We denote by $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{X} = \mathbf{x}]$ the regression function. Let $\mathrm{Var}(\mathbf{X})$ and $\mathrm{Cov}(\mathbf{X})$ denote the variance and the variance-covariance matrix of $\mathbf{X}$.

2.1 Importance measure of a group of variables

The permutation importance introduced by Breiman (2001) measures the accuracy of each variable X_j for predicting Y. It is based on the elementary property that the quadratic risk $\mathbb{E}\left[(Y - f(\mathbf{X}))^2\right]$ is the minimum error for predicting Y knowing $\mathbf{X}$. The formal definition of the variable importance measure of X_j is:

$$\mathcal{I}(X_j) := \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(j)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right], \tag{2.1}$$

where $\mathbf{X}_{(j)} = (X_1, \ldots, X_j', \ldots, X_p)^\top$ is a random vector such that X_j' is an independent replication of X_j which is also independent of Y and of all of the other predictors. Such criterion evaluates the increase of the prediction error after breaking the link between the variable X_j and the outcome Y, see Zhu et al. (2012) for instance.

In this paper, we extent the permutation importance for a group of variables. Let $J = (j_1, \ldots j_k)$ be a k-tuple of increasing indexes in $\{1, \ldots, p\}$, with $k \leq p$. We define the permutation importance of the sub-vector $\mathbf{X}_J = (X_{j_1}, X_{j_2}, \ldots, X_{j_k})^\top$ of predictors by

$$\mathcal{I}(\mathbf{X}_J) := \mathbb{E}\left[\left(Y - f(\mathbf{X}_{(J)})\right)^2\right] - \mathbb{E}\left[(Y - f(\mathbf{X}))^2\right],$$

where $\mathbf{X}_{(J)} = (X_1, \ldots, X_{j_1}', X_{j_1+1}, \ldots, X_{j_2}', X_{j_2+1}, \ldots, X_{j_\ell}', X_{j_\ell+1}, \ldots, X_p)^\top$ is a random vector such that $\mathbf{X}_J' = (X_{j_1}', X_{j_2}', \ldots, X_{j_k}')^\top$ is an independent replication of $\mathbf{X}_J$, which is also independent of Y and of all of the other predictors. We call this quantity the grouped variable importance since it only depends on which variables appear in $\mathbf{X}_J$. By abusing the notation and ignoring the ranking, we may also refer to $\mathbf{X}_J$ as a group of variables.

2.2 Decomposition of the grouped variable importance

Let $\mathbf{X}_J$ be a subgroup of variables from the random vector $\mathbf{X}$. Let $\mathbf{X}_{\bar{J}}$ denote the group of variables that does not appear in $\mathbf{X}_J$. Assume that we observe Y and $\mathbf{X}$ in the following additive regression model:

$$\begin{aligned} Y &= f(\mathbf{X}) + \varepsilon \\ &= f_J(\mathbf{X}_J) + f_{\bar{J}}(\mathbf{X}_{\bar{J}}) + \varepsilon, \end{aligned} \tag{2.2}$$

where the f_J and $f_{\bar{J}}$ are two measurable functions, and ε is a random variable such that $\mathbb{E}[\varepsilon|\mathbf{X}] = 0$ and $\mathbb{E}[\varepsilon^2|\mathbf{X}]$ is finite. The results of Gregorutti et al. (2014) about the permutation importance of individual variables can be extended to the case of a group of variables.

Proposition 1. *Under model (2.2), the importance of the group J satisfies*

$$\mathcal{I}(\mathbf{X}_J) = 2\,\mathrm{Var}\left[f_J(\mathbf{X}_J)\right].$$

The proof is left to the reader, it follows the lines of Proposition 1 in Gregorutti et al. (2014). Next Proposition gives the grouped variable importance for more specific models. It can be easily deduced from Proposition 1.

Corollary 1. *Assume that we observe Y and $\mathbf{X}$ in the model (2.2) with*

$$f_J(\mathbf{x}_J) = \sum_{j \in J} f_j(x_j), \tag{2.3}$$

where the f_j's are measurable functions and $\mathbf{x}_J = (x_j)_{j\in J}$.

1. *If the random variables $(X_j)_{j\in J}$ are independent, then*

$$\mathcal{I}(\mathbf{X}_J) = 2\sum_{j\in J} \mathrm{Var}\left(f_j(X_j)\right) = \sum_{j\in J} \mathcal{I}\left(X_j\right).$$

2. *If for any $j \in J$, f_j is a linear function such that $f_j(x_j) = \alpha_j x_j$, then*

$$\mathcal{I}(\mathbf{X}_J) = 2\alpha_J^\top \mathrm{Cov}(\mathbf{X}_J)\alpha_J, \tag{2.4}$$

   where $\alpha_J = (\alpha_j)_{j\in J}$.

If f is additive and if the variables of the group are independent, the grouped variable importance is nothing more than the sum of the individual importances. As shown by the second point of Corollary 1, this property is lost as soon as the variables in the group are correlated. Section A in appendix allows us to compare the grouped variable importance with the individual importances in various models. To sum up, these experiments suggest that the grouped variable importance cannot be compared with the sum of the individual importances in general settings. This is not surprising since the grouped variable importance is a more accurate measure of the importance of a group of variables than a simple sum of the individual importances.

Corollary 1 also tells us that the importance may increase with the number of variables in the group. This remark motivates the introduction of the renormalised version of the grouped variable importance:

$$\mathcal{I}_{\text{nor}}(\mathbf{X}_J) := \frac{1}{|J|}\mathcal{I}(\mathbf{X}_J).$$

In Section 3, we propose a variable selection algorithm based on the grouped variable importance. This algorithm used the normalised version to take into account the size of the groups in the selection process. More generally, we will rather prefer the normalised version when comparing groups of variables of different sizes.

2.3 Grouped variable importances and Random Forests

Classification and regression trees are competitive techniques for estimating f. The most popular method in this field is the CART algorithm due to Breiman et al. (1984). Though efficient, tree methods are also known to be unstable insofar as a small perturbation of the training sample may change radically the predictions. For answering this issue, Breiman (2001) introduced the random forests as a substantial improvement of the decision trees. The permutation importance measure was also introduced in this seminal paper. We now recall how individual permutation importances can be estimated with random forests before giving the natural extension to the estimation of grouped variable importances.

Assume that we observe n i.i.d. replications $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$ of $(\mathbf{X}, Y)$. Random forests algorithm consists in aggregating a collection of random trees, in the same way as the bagging method also proposed by Breiman (1996): the trees are built over M bootstrap samples $\mathcal{D}_n^1, \ldots, \mathcal{D}_n^M$ of the training data $\mathcal{D}_n$. Instead of CART algorithm, a subset of variables is randomly chosen for 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 $\hat{f}_1, \ldots, \hat{f}_M$. In the regression setting, the aggregation is based on the average of the predictions.

For any $m \in \{1, \ldots, M\}$, let $\bar{\mathcal{D}}_n^m := \mathcal{D}_n \setminus \mathcal{D}_n^m$ be the corresponding out-of-bag sample. The risk of $\hat{f}_m$ is estimated on the out-of-bag sample as follows:

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

Let $\bar{\mathcal{D}}_n^{mj}$ be the permuted version of $\bar{\mathcal{D}}_n^m$ obtained by randomly permuting the variable X_j in each out-of-bag sample $\bar{\mathcal{D}}_n^m$. The estimation of the permutation importance measure of the variable X_j is finally obtained by

$$\hat{\mathcal{I}}(X_j) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mj}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^m) \right]. \tag{2.5}$$

This random permutation mimics the replacement of X_j by X_j' in (2.1) and breaks the link between X_j and Y and the other predictors.

We now extend the method for estimating the permutation importance of a group of variables $\mathbf{X}_J$. For any $m \in \{1, \dots, M\}$, let $\bar{\mathcal{D}}_n^{mJ}$ be the permuted version of $\bar{\mathcal{D}}_n^m$ obtained by randomly permuting the group $\mathbf{X}_J$ in each out-of-bag sample $\bar{\mathcal{D}}_n^{mj}$. Note that the same random permutation is used for each variable X_j of the group. By this way the (empirical) joint law of $\mathbf{X}_J$ is left unchanged by the permutation whereas the link between $\mathbf{X}_J$ and Y and the other predictors is broken. The importance of $\mathbf{X}_J$ can be estimated by

$$\hat{\mathcal{I}}(\mathbf{X}_J) = \frac{1}{M} \sum_{m=1}^{M} \left[\hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{mJ}) - \hat{R}(\hat{f}_m, \bar{\mathcal{D}}_n^{m}) \right]. \tag{2.6}$$

Let $\hat{\mathcal{I}}_{\text{nor}}(\mathbf{X}_J)$ be the normalized version of this estimation.

In the next section, we use the grouped variable importance as a criterion for selecting features in the context of multivariate functional data.

3 Multivariate functional data analysis using the grouped variable importance

In this section, we consider an application of the grouped variable selection for multivariate functional regression with scalar response Y. Each covariate $X^1, \dots, X^p$ takes its values in the Hilbert space $L^2([0,1])$ equipped with the inner product

$$\langle f, g \rangle_{L^2} = \int f(t) g(t) dt,$$

for $f, g \in L^2([0,1])$. One common approach of functional data analysis is to project the variables on a finite dimensional subspace of $L^2([0,1])$ and to use the basis coefficients in a learning algorithm.

3.1 Functional representation using wavelets

The wavelet transform is widely used in signal processing and for non parametric function estimation (see for instance Antoniadis et al.; 2001). Unlike Fourier basis or Splines, the wavelets are localised both in frequency and time.

For $j \geq 0$ and $k = 0, \dots, 2^j - 1$, define a sequence of functions ϕ_{jk} (resp. ψ_{jk}), obtained by translations and dilatations of a compactly supported function ϕ (resp. ψ), called scaling function (resp. wavelet function). For any $j_0 \geq 0$, the collection

$$\mathcal{B} = \{\phi_{j_0 k}, k = 0, \dots, 2^{j_0} - 1\} \cup \{\psi_{jk}, j \geq j_0, k = 0, \dots, 2^j - 1\}$$

forms an orthonormal basis of $L^2([0,1])$, see for instance Percival and Walden (2000). Then a function $s \in L^2([0,1])$ can be decomposed into

$$s(t) = \sum_{k=0}^{2^{j_0}-1} \langle s, \phi_{j_0 k} \rangle_{L^2} \phi_{j_0 k}(t) + \sum_{j \geq j_0} \sum_{k=0}^{2^j - 1} \langle s, \psi_{jk} \rangle_{L^2} \psi_{jk}(t). \tag{3.1}$$

The first term in Equation (3.1) is the smooth approximation of s at level j_0 while the second term is the detail part of the wavelet representation.

We assume that each covariate X is observed on a fine sampling grid $t_1, \dots, t_N$ with $t_\ell = \frac{\ell}{N}$. A wavelet decomposition of X can be given, in a similar form as in (3.1). For $j_0 = 0$, we have

$$X(t_\ell) = \zeta \phi_{00}(t_\ell) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \xi_{jk} \psi_{jk}(t_\ell), \tag{3.2}$$

where $J := \log_2(N)$ is the maximal number of wavelet levels and ζ and ξ_{jk} are respectively the scale and the wavelet coefficients of the discretized curve X at the position k for the resolution level j. These empirical coefficients can be efficiently computed using the discrete wavelet transform algorithm described in Percival and Walden (2000, Chap. 4).

For a given wavelet basis, we introduce the *wavelet support at time* t as the set of all the indices of wavelet functions that are not null at t:

$$\mathcal{S}(t) = \{(j,k) : \psi_{jk}(t) \neq 0\}.$$

Figure 1 displays the matrix giving the correspondence between a time location and the associated wavelet functions, for a Daubechies wavelet basis with two vanishing moments. In a similar way but for an interval $\mathcal{T}$, we define the *wavelet support of the time interval* $\mathcal{T}$ by

$$\begin{aligned} \mathcal{S}(\mathcal{T}) &= \{(j,k) : \psi_{jk}(t) \neq 0, \forall t \in \mathcal{T}\} \\ &= \bigcap_{t \in \mathcal{T}} \mathcal{S}(t). \end{aligned}$$

This set corresponds to the wavelet functions localised on the interval $\mathcal{T}$.

3.2 Grouped variable importance for functional variables

In this section, we show how the grouped variable importance can be fruitfully used for comparing the importances of wavelet coefficients in the context of functional predictors. Remember that p functional covariates $X^1, \dots, X^p$ are observed together with a scalar response Y. For the sake of simplicity, the covariates are decomposed on the same wavelet basis $\mathcal{B}$ but the methodology presented above could be also adapted with a particular basis for each covariable. For any $u \in \{1, \dots, p\}$, let $\mathbf{W}^u = (\zeta^u, \xi^u_{jk})_{jk}$ be the random vector composed of the wavelet coefficients of the functional variable X^u.

Groups of wavelet coefficients

The wavelet coefficients are characterised by their frequency, their time location and the functional variables they describe. Consequently, they can be grouped in many ways. We give below a non exhaustive list of groups for which we are interested in computing the importance:

- **A group related to a variable.** The vector $\mathbf{W}^u$ defines the group $G(u)$.

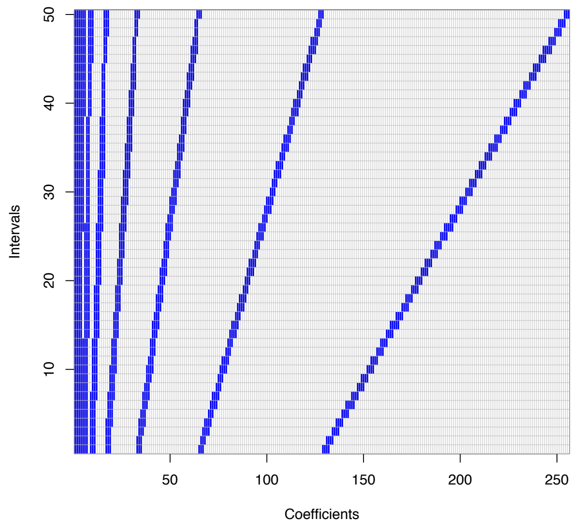


Figure 1: Correspondence between the time domain and the wavelet functions for a Daubechies wavelet basis with two vanishing moments. For a time t, the colored points correspond to the wavelet functions which are not null at time t.

- **A group related to a frequency level of a variable.** For a fixed variable X^u, the group is composed of the wavelet coefficients of frequency level j:

$$G(j,u) := \{\xi^u_{j,1}, \dots, \xi^u_{j,2^j-1}\}.$$

- **A group related to a frequency level.** The group is composed of all the wavelet coefficients of frequency level j for all the variables:

$$G(j) := \bigcup_{u=1,\dots,p} G(j,u).$$

- **A group related to a given time.** Define the group of "active" wavelet coefficients associated to a given time t by

$$G(t) := \bigcup_{u=1,\dots,p} \{\zeta^u\} \cup \bigcup_{u=1,\dots,p,\,(j,k)\in\mathcal{S}(t)} \{\xi^u_{jk}\}.$$

  Depending of the size of the support of ϕ and ψ, the group $G(t)$ may be very large. For instance with a Daubechies wavelet basis with two vanishing moments, on Figure 1 the group $G(t)$ is composed of the colored points of the row corresponding to time t.

- **A group related to a time interval.** Let $[a, b]$ be a time interval. The group of "active" wavelet coefficients associated to $[a, b]$ is

$$G([a, b]) := \bigcup_{t \in [a,b]} G(t).$$

Many other groupings could be proposed. For instance, one could regroup two correlated variables, or consider a group composed of the wavelet coefficients taken in a interval of frequencies, a group related to a given time and a fixed variable, etc...

By computing the importances of such groups, one directly obtain a rough detection of the most important groups of coefficients for predicting Y. When grouping by frequency levels or by time locations, all the groups do not have equal sizes. As explained in Section 2, it is preferable to use the normalised version of the grouped variable importance in order to compensate the effect of group size in the grouped variable importance measure.

Grouped variable selection

We now propose a more elaborated method for selecting groups of coefficients. The selection procedure is based on the Recursive Feature Elimination (RFE) algorithm proposed by Guyon et al. (2002) in the context of support vector machines. In this paper, we propose a random forests version of the RFE algorithm which is guided by the grouped variable importance. The procedure can be summarised in Algorithm 1. This backward grouped elimination approach produces a collection of nested subsets of groups. The selected groups are obtained by minimising the validation error computed in step 2.

Algorithm 1 Grouped Variable Selection

```
1: Train a random forest model
2: Compute the error using a validation sample
3: Compute the grouped variable importance measure
4: Eliminate the less important group of variables
5: Repeat 1–4 until no further groups remain
```

This algorithm is motivated by the results from our previous work (Gregorutti et al.; 2014) about variable selection using the permutation importance measure from the random forests. Strong correlations between predictors have a strong impact on the permutation importance measure. It was also shown in this previous paper that, when the predictors are srongly corrolated, the RFE algorithm provides better performances than the "non-recursive" strategy (NRFE) that computes the grouped variables importance just once and does not recompute the importance at each step of the algorithm. In the present paper, we continue this study by adapting the RFE algorithm for the grouped variable importance measure. We give below two applications of this algorithm.

- **Selection of functional variables.** Each vector $\mathbf{W}^u$ defines a group $G(u)$ and the goal is to perform a grouped variable selection over the groups $G(1), \ldots, G(p)$. The selection allows us to identify the most relevant functional variables.

- **Selection of the wavelet levels.** This problem is about the selection of the wavelet levels for a given functional variable. For a fixed u, we make a selection over the groups $G(j,u)$ to identify the frequency levels which yield predictive information.

Remark. *Algorithm 1 for grouped variable selection is appropriate for groups defining a partition over the wavelet coefficients. This is not the case for groups related to time locations. The algorithm can be hardly adapted with these groups because most of the wavelet coefficient belong to several groups and the elimination of a whole group might be a non efficient strategy. For instance the coefficient ζ^u, which approximates the smooth part of the curves and which is usually a good predictor, is common to all times t.*

3.3 Numerical experiments

In this section, we present various numerical experiments for illustrating the interest of the grouped variable importance for analysing functional data. We first describe the simulation designs.

Presentation of the general simulation design

The experiments presented below consider one or several functional covariates for predicting an outcome variable Y. Except for the second simulation of Experiment 1 which is presented in details in the next section, the functional covariates are defined in function of Y.

First, a n-sample of the outcome variable Y is simulated from a given distribution specified for each experiment. The realisation of a functional covariate X (denoted by X^u when there are several functional covariates) is a n-sample of independent discrete time random processes $X_i = (X_i(t_\ell))_{\ell=1,\dots,N}$, for all $i \in \{1,\dots,n\}$, according to a model of the form

$$X_i(t_\ell) = s(t_\ell, Z_i) + \sigma\varepsilon_{i,\ell},\ \ell = 1,\dots,N, \tag{3.3}$$

where the $\varepsilon_{i,\ell}$'s are i.i.d standard Gaussian random variables and $t_\ell = \frac{\ell}{N}$. The random variable Z_i is correlated to Y_i, it will be specified for each experiment. It is equal to the outcome variable Y_i for most of the experiments. The functional covariates are actually simulated in the wavelet domain from the following model: for $i = 1,\dots,n$, $j = 0,\dots,J-1$ and $k = 0,\dots,2^j-1$,

$$\zeta_i = \omega_0 + h_\zeta(Z_i) + \sigma\eta_{i\zeta}. \tag{3.4}$$

and

$$\xi_{ijk} = \left\{ \begin{array}{ll} \omega_{jk} + h_{jk}(Z_i) + \sigma\eta_{ijk} & \text{if } j \le j^\star,\ k \in \{0,\dots,2^j-1\}, \\ 0 & \text{if } j^\star < j \le J-1, \end{array} \right. \tag{3.5}$$

where $j^\star$ is the highest wavelet level of the signal. The random variables η_{ijk} and $\eta_{i\zeta}$ are i.i.d. standard Gaussian variables. The "signal" part of Equation (3.5) is the sum of a random coefficient ω_{jk} whose realisation is the same for all i, and a link function h_{jk}. The coefficients ω_0 and ω_{jk} in (3.4) and (3.5) are simulated as follows:

$$\left\{ \begin{array}{lll} \omega_{jk} & \sim & \mathcal{N}_1(0,\tau_j^2), \quad \text{if } j \le j^\star,\ k \in \{0,\dots,2^j-1\}, \\ \omega_0 & \sim & \mathcal{N}_1(3,1), \end{array} \right.$$

where $\tau_j = e^{-(j-1)}$. Note that the standard deviation τ_j decreases with j and thus less noise is added to the first wavelet levels. The link function h_{jk} describes the link between the wavelet coefficients ξ_{ijk} and the variable Z (or with the outcome variable Y). Two different link functions are considered in the experiments:

- a linear link $h_{jk}(z) = \theta_{jk} z$,
- a logistic link $h_{jk}(z) = \dfrac{\theta_{jk}}{1 + e^{-z}}$.

where the coefficients θ_{jk} parametrise the strength of the relation between Z and the wavelet coefficients. The n discrete processes $X_1, \ldots, X_i, \ldots X_n$ are simulated according to Equations(3.4) and (3.5) before applying the inverse wavelet transform.

We choose a Daubechies wavelet filter with four vanishing moments to simulate the observations. We use the same basis for the projection of the functional observations.

Experiment 1: detection of important time intervals

In this first experiment, we illustrate the use of the grouped variable importance for the detection of the most relevant time intervals. We simply estimate the importance of time intervals without applying Algorithm 1 (see Remark 3.2). We only consider one functional covariate X since it will be sufficient to illustrate the method. Let $\mathcal{T}^\star = [t_{50}, t_{55}]$, we propose two simulation designs for which the outcome Y is correlated to the signal X on the interval $\mathcal{T}^\star$.

- **Simulation 1.** For this first simulation, we follow the general simulation design presented before by considering linear link functions h for all the wavelet coefficients belonging to the wavelet support $\mathcal{S}(\mathcal{T}^\star)$. The outcome variable Y is simulated from a Gaussian distribution $\mathcal{N}_1(0, 3)$. We simulate the wavelet coefficients as in (3.5) and the scaling coefficients as in (3.4) with $Z = Y$. We take linear link functions. We take $n = 1000$, $\sigma = 0.01$, $N = 2^8$ and thus $J = 8$. We take $j^\star = 7$ which means that even the wavelet coefficients of highest level $j = 7$ are not Dirac distributions at zero. The wavelet coefficients and the scaling coefficient are generated as follows: for any $j \in \{0, \ldots, J-1\}$ and any $k \in \{0, \ldots, 2^j - 1\}$,

$$\xi_{ijk} = \begin{cases} \omega_{jk} + Y_i + \sigma\eta_{ijk}, & \text{if } (j,k) \in \mathcal{S}(\mathcal{T}^\star) \\ \omega_{jk} + \sigma\eta_{ijk}, & \text{otherwise,} \end{cases} \tag{3.6}$$

  and

$$\zeta_i = \omega_0 + Y_i + \sigma\eta_{i\zeta}, \tag{3.7}$$

  for $i = 1, \ldots, n$.

- **Simulation 2.** Contrary to the previous simulation, we first simulate the functional variable X and then we simulate the outcome variable Y in function of X. The functional variable is simulated in the wavelet domain according to Equations (3.4) and (3.5) with $h_{jk} = h_\zeta = 0$ for all j, k. We also take $n = 1000$, $\sigma = 0.01$, $N = 2^8$ and $j^\star = 7$. By applying the wavelet inverse transform for any i, we obtain an n-sample of discrete time random processes $X_i = (X_i(t_\ell))_{\ell=1,\ldots,N}$. Figure 2 displays a set of 10 of these processes.

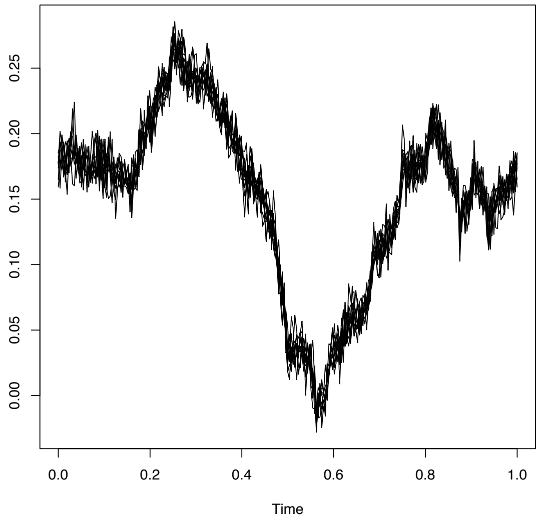


Figure 2: Experiment 1 – Example of 10 processes drawn from the protocol used for Simulation 2.

The outcome variable Y is finally obtained by the relation

$$Y_i = \frac{1000}{|\mathcal{T}^\star|} \sum_{t_\ell \in \mathcal{T}^\star} |X_i(t_\ell) - X_i(t_{\ell-1})|.$$

Thus, Y_i is a measure of the oscillations of the curve X_i over the interval $\mathcal{T}^\star$.

The aim is to detect $\mathcal{T}^\star$ using the grouped variable importance. In both cases, the grouped variable importance $\mathcal{I}(G(t))$ is evaluated at 50 equally spaced time points. Figure 3 displays the importance of the time points, averaged over 100 iterations. The first and third quartiles are also represented for highlighting the estimation variability. In the two cases, the importance estimation makes it possible to detect $\mathcal{T}^\star$.

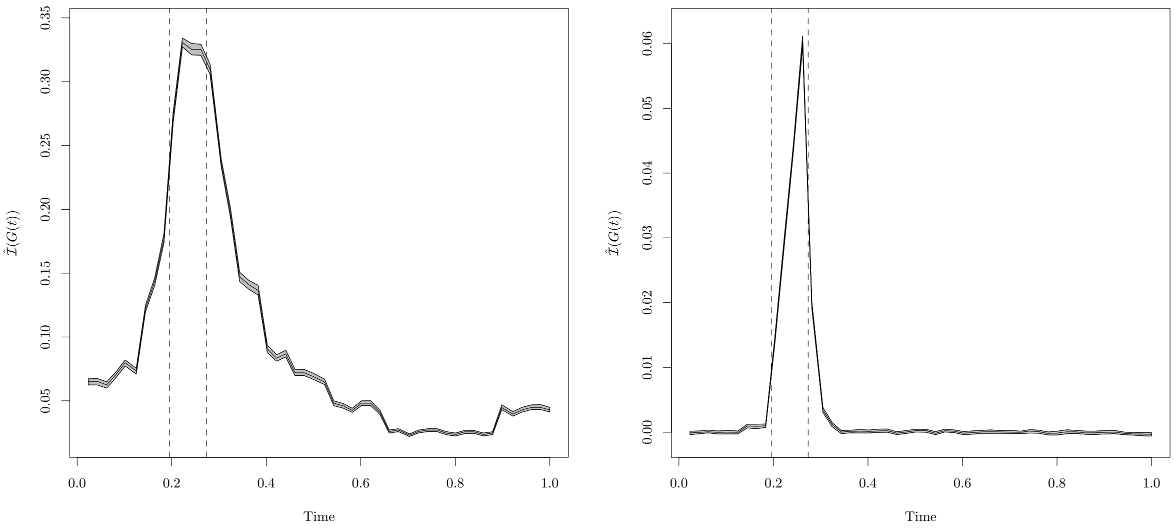
Note that the detection problem is tricky in the second case because the link between Y and the wavelet coefficients is complex in this simulation. Consequently the estimated importances are low and the important intervals are difficult to detect.

Experiment 2: Selection of wavelet levels

This simulation is about the selection of wavelet levels for one functional variable. We follow the general simulation design presented before. The outcome variable Y is simulated from a Gaussian distribution $\mathcal{N}_1(0, 3)$. We simulate the wavelet coefficients as in (3.5) and the scaling coefficients as in (3.4) with $Z = Y$. We make two simulations.

- In the first case we use linear link functions:

$$h_\zeta(y) = 0.1y$$



(a) Case 1

(b) Case 2

Figure 3: Experiment 1 – Averaged time importances, first and third quartiles over 100 iterations. The time interval $\mathcal{T}^\star$ is located between the two vertical lines.

and

$$h_{jk}(y) = \begin{cases} \theta_j y & \text{if } j \leq 3,\ k \in \{0, \dots, 2^j - 1\}, \\ 0 & \text{otherwise,} \end{cases}$$

where the θ_j's decrease linearly from 0.1 to 0.01.

- For the second simulation, we use logistic link functions:

$$h_\zeta(y) = \frac{0.1}{1 + e^{-y}}$$

and

$$h_{jk}(y) = \begin{cases} \dfrac{\theta_{jk}}{1 + e^{-y}} & \text{if } j \leq 3,\ k \in \{0, \dots, 2^j - 1\}, \\ 0 & \text{otherwise,} \end{cases}$$

where the θ_j's decrease linearly from 0.1 to 0.01.

We take $n = 1000$, $\sigma = 0.05$, $N = 2^8$ (thus $J = 8$) and $j^\star = 7$ in the two cases.

The aim is to identify the most relevant wavelet levels for the prediction of Y, using the grouped importance. We regroup the wavelet coefficients by wavelet levels: for $j \in \{0, \dots, J-1\}$,

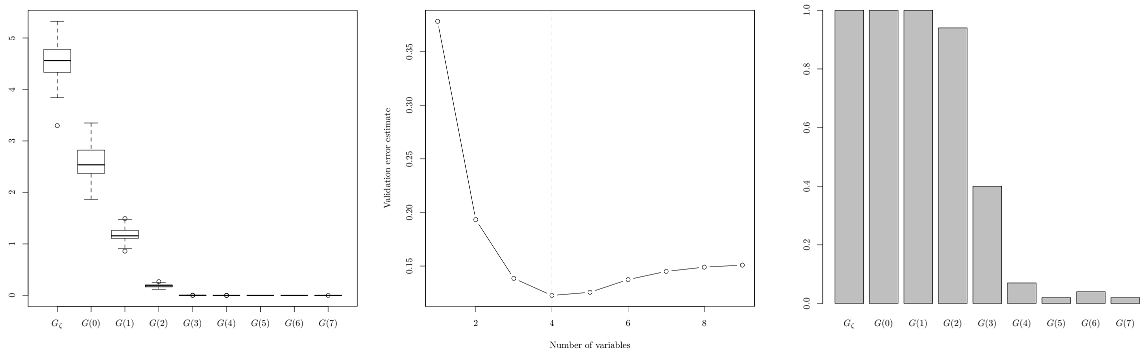
$$G(j) = \left\{\xi_{jk}, k \in \{1, \dots, 2^j - 1\}\right\}$$

and

$$G_\zeta = \{\zeta\}.$$

We apply Algorithm 1 with these groups. As the group sizes are of different, the normalised grouped importance criterion given in Section 2.3 is used.

The experiences are both repeated 100 times. Figure 4 and 5 respectively give the results for the linear link and the logistic link. We start with the experience with linear link. The boxplots of the grouped permutation importances at the first step of the algorithm over the 100 experiences are given on Figure 4a. The fifth group $G(3)$ being not strongly correlated with Y, its importance is close to zero. It is selected 40 times out of the 100 simulations (Fig. 4c) whereas G_ζ, $G(0)$, $G(1)$ and $G(2)$ are almost always selected. The other groups are not correlated with Y and are almost never selected. For each experience, the mean squared error (MSE) is computed in function of the number of variables in the model. Figure 4b shows the average of the errors over the 100 simulations. On average, the model selected by minimising the MSE includes four groups but the model with five groups also has an error close to the minimum.



(a) (b) (c)

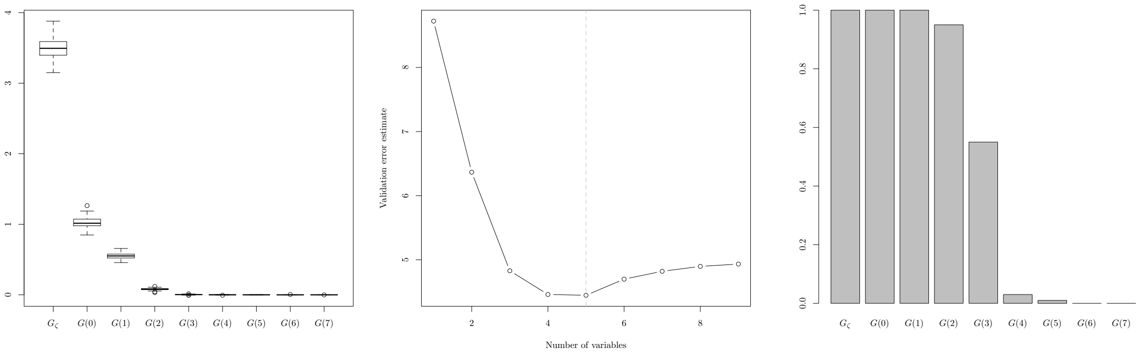
Figure 4: Experiment 2, linear links – Selection of the wavelet levels. From the left to the right: (a) Boxplots of the grouped variable importances, (b) MSE error versus the number of groups and (c) Selection frequencies.

The experience with the logistic link gives similar results. However the fifth group $G(3)$ is more frequently selected (Fig. 5c). The minimisation of the MSE leads to select five groups as shown in Figure 5b. Note that this approach based on the random forests and grouped variable importance performs well even with a non linear link.

In both experiences, the grouped variable importances obtained at the first step of the algorithm are ranked in the same order as the θ_j's. Indeed the impact of the correlation between predictors is not too strong in the two cases. In this context, the backward Algorithm 1 does not provide additional information compared to the “non-recursive” strategy (see the disussion following Algorithm 1).

Experiment 3: Selection of functional variables in presence of strong correlation

This simulation illustrates the interest of Algorithm 1 for selecting functional variables in presence of correlation.



(a) (b) (c)

Figure 5: Experiment 2, logistic links – Selection of the wavelet levels. From the left to the right: (a) Boxplots of the grouped variable importances, (b) MSE error versus the number of groups and (c) Selection frequencies.

First, we simulate $n = 1000$ i.i.d. realisations of $p = 20$ functional variables $X^1, \ldots, X^p$ according to the general simulation design detailed before. For all $i \in \{1, \ldots, n\}$, let $Z_i^1, \ldots, Z_i^p$ be some latent variables drawn from a standard Gaussian distribution. The outcome variable Y is defined as:

$$Y_i = 3.5\, Z_i^1 + 3\, Z_i^2 + 2.5\, Z_i^3 + 2.5\, Z_i^4.$$

Then for $u \in \{1, \ldots, p\}$, the wavelet coefficients are simulated according to (3.5) and (3.4) with a linear link:

$$\xi_{ijk}^u = \omega_{jk}^u + Z_i^u + \sigma \eta_{ijk}^u \qquad \text{if } j \leq j^\star,\ k \in \{0, \ldots, 2^j - 1\}$$

and

$$\zeta_i^u = \omega_0^u + Z_i^u + \sigma \eta_{i\zeta}^u,$$

with $\sigma = 0.1$, $N = 2^9$ and thus $J = 9$. We take $j^\star = 3$ in order to make the functional variables smooth enough. Among the 20 variables X^u, the first four variables have a decreasing predictive power whereas the others are independent of Y. Next, we add $q = 10$ i.i.d. variables $X^{1,1}, \ldots, X^{1,q}$ which are strongly correlated with X^1: for any $v \in \{1, \ldots, q\}$, any $i \in \{1, \ldots, n\}$,

$$\zeta_i^{1,v} = \zeta_i^1 + \tilde{\sigma} \eta_{i\zeta}^{1,v}$$

and

$$\xi_{ijk}^{1,v} = \begin{cases} \xi_{ijk}^1 + \tilde{\sigma} \eta_{ijk}^{1,v} & \text{if } j \leq j^\star,\ k \in \{0, \ldots, 2^j - 1\} \\ 0 & \text{otherwise,} \end{cases}$$

where $\tilde{\sigma} = 0.05$. The $\eta_{ijk}^{1,v}$'s and the $\eta_{i\zeta}^{1,v}$'s are i.i.d. standard Gaussian random variables. The discrete processes $X_i^{1,v}$ are obtained using the inverse wavelet transform. In the same way, we add $q = 10$ i.i.d. variables $X^{2,1}, \ldots, X^{2,q}$ which are strongly correlated with X^2.

To sum

up, the vector of predictors is composed of 40 variables:

$$X^1, X^{1,1}, \dots, X^{1,q}, X^2, X^{2,1}, \dots, X^{2,q}, X^3, X^4, \dots, X^p.$$

The aim is to identify the most relevant functional variables for the prediction of Y. For each functional variable, we regroup all the wavelet coefficients and we apply Algorithm 1. The experience is repeated 100 times.

The boxplots of the group permutation importances at the first step of the algorithm, over the 100 experiences and for each functional variable, are given on Figure 6a. We see that the importances of the variables X^1 and X^2 and their noisy replications are much lower than the importances of the variables X^3 and X^4. This is due to the strong correlations between the two first variables and their noisy replications and it confirms the results obtained in Gregorutti et al. (2014) for the individual importance measure. Indeed, it is shown in this last paper that the importance measure decreases when the correlation or the number of correlated variables increase. Note that the importances of X^1 and X^2 are slightly lower than the ones of their noisy replications. This can be explained by the fact that the correlation between X^1 and their noisy replications is higher than, for instance, the correlation of $X^{1,1}$ with $X^1, X^{1,2}, \dots, X^{1,q}$.

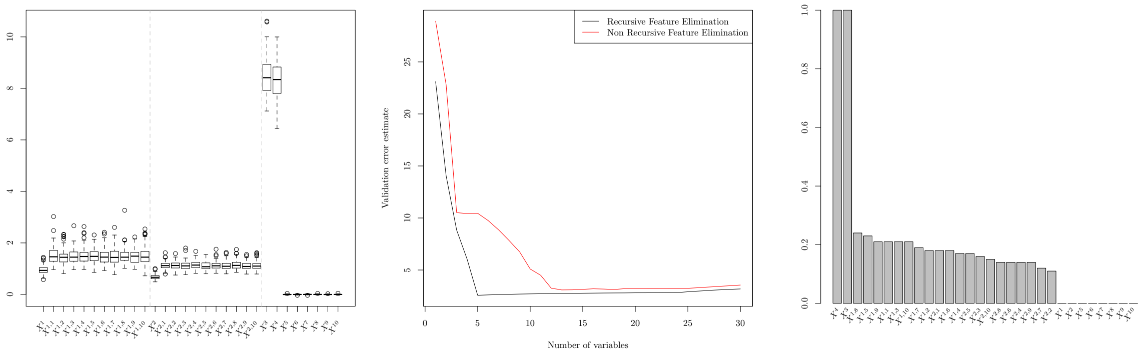
Figure 6b is a comparison of the performances of Algorithm 1 and the "non-recursive" strategy (NRFE). Algorithm 1 clearly shows better prediction performances. In particular, Algorithm 1 reaches a minimum error faster than the NRFE: only five variables for Algorithm 1 whereas NRFE needs about twelve variables. This observation is consistent with the conclusion of Gregorutti et al. (2014): the RFE procedure is more efficient than the NRFE when the predictors are highly correlated.

Additional informations are displayed in Figure 6c. The selection frequencies using Algorithm 1 show that the variables X^3 and X^4 are always selected. Indeed, these two variables have predictive power and they are not correlated to the other predictors. Note that the variables X^1 and X^2 are less selected then their replications, even if they are more correlated with Y than their replications are. This also comes from the fact that the correlation between X^1 and their replications is higher than the correlation of $X^{1,1}$ with $X^1, X^{1,2}, \dots, X^{1,q}$. We observe that X^1 and X^2 are eliminated in the first steps of the backward procedure, but this has no consequences on the prediction performances of Algorithm 1.

These results motivate the use of Algorithm 1 in practice. It reduces the effect of the correlation between predictors on the selection process and it provides better prediction performances.

4 A case study: variable selection for aviation safety

In this section, we study a real 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. One flight provides a multivariate time series corresponding to these family of functional variables.



(a) (b) (c)

Figure 6: Experiment 3 – Selection of functional variables. From the left to the right: (a) Boxplots of the grouped variable importances, (b) MSE error versus the number of groups and (c) Selection frequencies using Algorithm 1.

We focus here on the risk of long landing. A sequence of $N = 512$ seconds before touchdown is observed for predicting the landing distance. The evaluation of the risk the long landings is crucial for safety managers to avoid runway excursions and more generally to keep a high level of safety. One answer to this problem is to select the flight parameters that better explain the risk of long landings. By this way, we attempt to find a sparse model showing good predictive performances. Down the road, the analysis of the flight data could be used for pilot training or for developments of new flight procedures during approach.

Following the aviation experts, 23 variables are preselected and a sample of 1868 fights from the same airport and the same company is considered. The functional variables are projected on a Daubechies wavelet basis with four vanishing moments using the discrete wavelet transform algorithm as in Section 3. The choice of the wavelet basis is conducted by the nature of the flight data. Indeed, the data contains informations on the time scale as well as the frequency scale and it is important to well retrieve it.

Preliminary dimension reduction

The design matrix formed by the wavelet coefficients for all of the flight parameters has dimension $23 \times 512 = 11\ 776$. Selecting the variables directly from the whole coefficients is prohibitive, we first need to reduce significantly the dimension. The naive method that shrinks the n curves independently according to Donoho and Johnstone (1994) and then brings the non-zero coefficients together in a second step would lead to consider a large block of coefficients with many zero values. This first solution is not relevant in our context. We propose an alternative method which consists in shrinking the wavelet coefficients of the n curves simultaneously. More precisely, this method is adapted from Donoho and Johnstone (1994) for the particular context of n independent (but non necessary identically distributed) discrete random processes. The shrinkage is done on the norm of the n-dimensional vector containing the wavelet coefficients. The complete method is described in Appendix B.

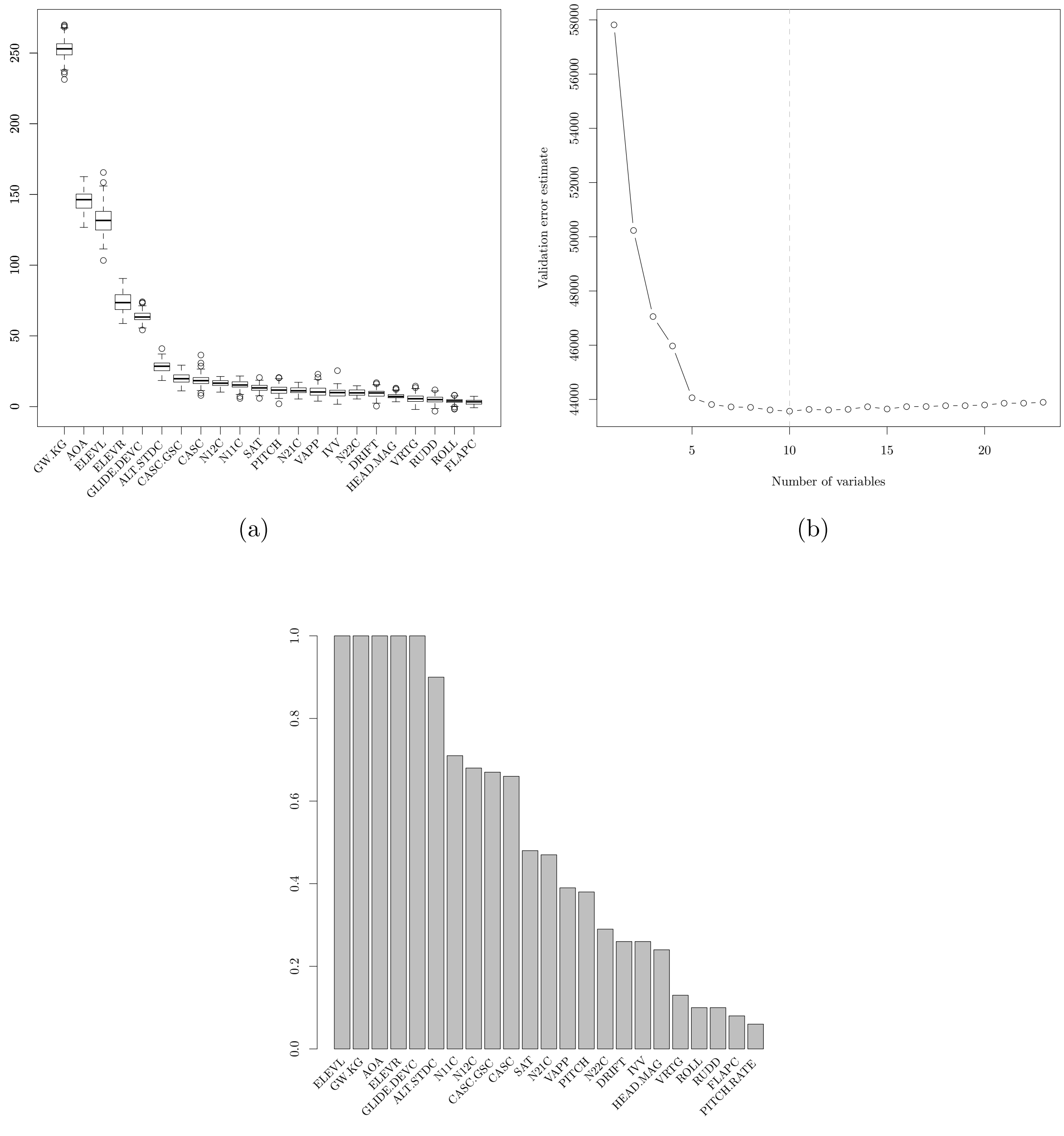
Selection of flight parameters

We obtain a selection of the functional parameters by grouping together the wavelet coefficients of each flight parameter and applying Algorithm 1 with these groups. At each iteration, we randomly split the dataset into a training set containing 90 % of the data and a validation set containing the remaining 10 %. In the backward algorithm the grouped variable importance is computed on the training set and the validation set is only used to compute the MSE errors. The selection procedure is repeated 100 times to reduce the variability of the selection. The final model is chosen by minimising the averaged prediction error. Figure 7a represents the boxplots of the grouped variable importance values computed on the 100 runs of the selection algorithm. According to this ranking, five variables are found significantly relevant. Looking at the averaged MSE estimate on Figure 7b, we see that the averaged number of selected variables is ten but taking only five variables is sufficient to get a risk close to the minimum.

Figure 7c gives additional informations by displaying the proportion of times each flight parameter is selected. Firstly, it confirms the previous remarks: five variables are always selected by the algorithm and the ten first are selected more than 60 times over the 100 runs. Secondly, it shows that the flight parameters related to the aircraft trajectory during the approach are among the most relevant variables for predicting the long landing. Indeed, the elevators (ELEVL, ELEVR) are used by the pilots to control the pitch of the aircraft. It has an effect on the angle of attack (AOA) and consequently on the landing. The variable GLIDE.DEVC is the glide slope deviation, that is the deviation between the aircraft trajectory and a glide path of approximatively three degrees above horizontal. It indicates how the aircraft approaches the airport. Another significant variables related to the airspeed reduction are the gross weight (GW.KG) which has an effect on the deceleration efficiency, the airspeed (CASC) and the engine rating (N11C, N12C).

It should be noted that the ranking due to the selection frequency is close to the direct ranking given by the importance measures when all the variables are taken in the model. This suggests that for this data the correlation between the predictors are not strong enough to influence their permutation importance measures. Moreover, if we regroup several variables as for example the flight parameters N11C and N12C, N21C and N22C and ELEVR and ELEVL into three new variables N1, N2 and ELEV, Figure 8 shows that the ranking is kept unchanged.

A similar study (Gregorutti et al.; 2014) used the classification version of the Random Forests Algorithm and (individual) permutation importance measures for variable selection. The database in Gregorutti et al. (2014) corresponds to a set of flights from a different company and a different aircraft type. Nevertheless we found in this previous study that the gross weight, the altitude (ALT.STDC) and the wind speed (CASC.GSC) were the most relevant for predicting the landing classes. More precisely, a too high altitude combined to a tail wind was decisive during the final approach and can lead to a long landing. The two analysis are thus consistent. The variable selection procedure used for the previous study is the “non grouped” version of Algorithm 1. It consists in directly selecting the wavelet coefficients (without grouping them) and then aggregating the selected coefficients in a second step to obtain a selection of the flight parameters.

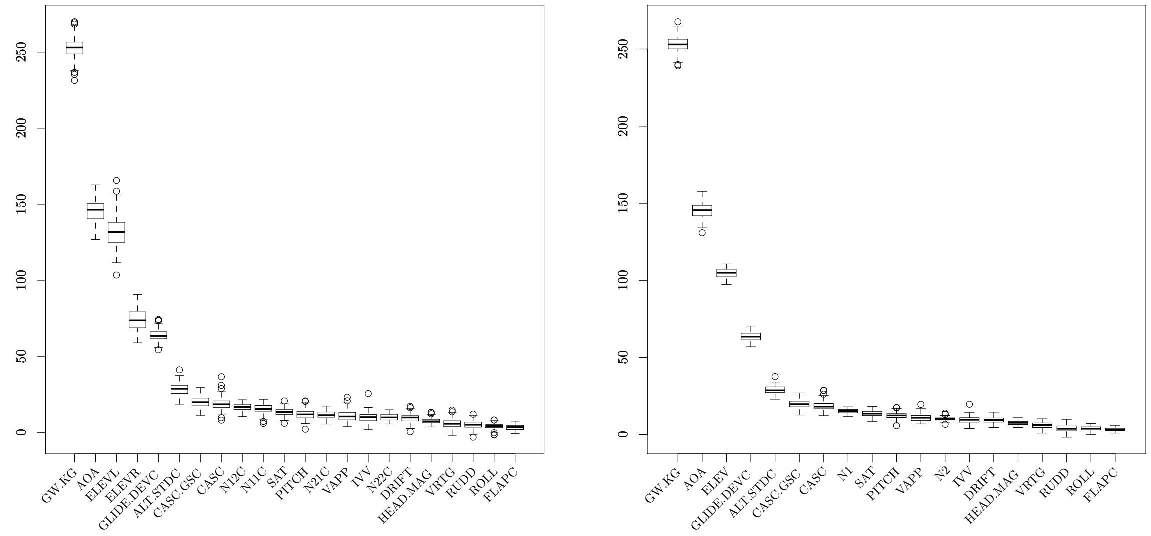


(a)

(b)

(c)

Figure 7: Application to long landing – From the left to the right: (a) Boxplots of the grouped variable importance (b) MSE error versus the number of groups and (c) selection frequencies.



(a)

(b)

Figure 8: Application to long landing – Grouped variable importance measure before and after grouping the correlated flight parameters N11C and N12C, N21C and N22C, ELEVR and ELEVL into N1, N2 and ELEV.

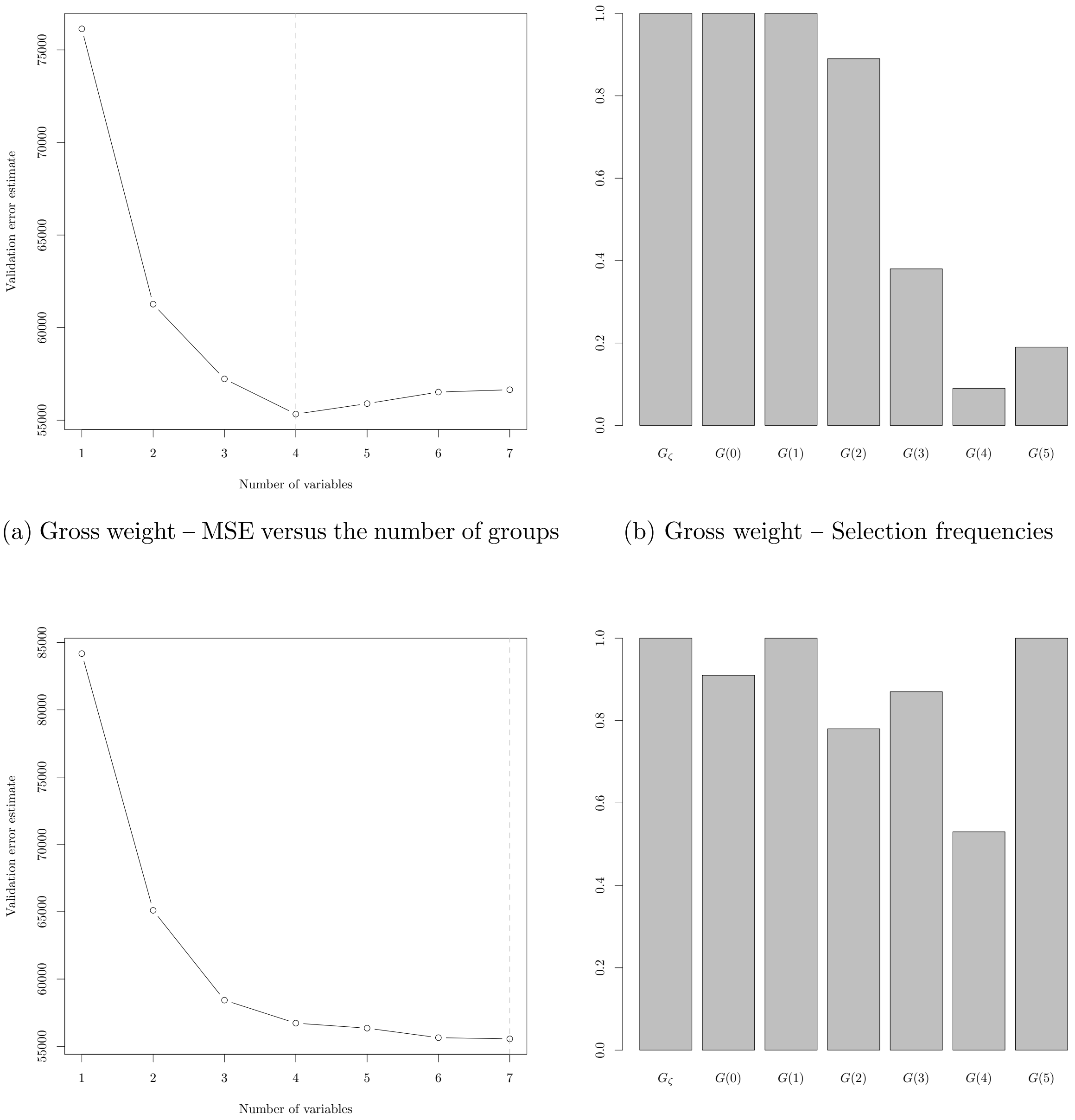
Such an approach is computationally demanding. Moreover, aggregating the selection in the second step is not obvious. For these reasons, Algorithm 1 is much efficient and satisfactory for analysing this kind of multivariate functional data.

Selection of wavelet levels

We now determine for a flight parameter which wavelet levels are the most able to predict the risk of long landing. The selection of the wavelet levels is done independently for the gross weight (GW.KG) and the angle of attack (AOA), which are among the most selected flight parameters (Fig. 9). Figures 9a and 9c show the averaged number of selected levels for GW.KG is less than for AOA. Indeed, the selection frequencies in Figure 9b indicate that for GW.KG, the first approximation levels are selected at each run (groups $\zeta, G(0)$ and $G(1)$) whereas the last levels are selected less than 40 times over 100. The situation is quite different for AOA: all the levels are selected more than 50 times over the 100 runs. The predictive power of this functional variable is shared by both the high levels of approximation and in the details of the wavelet decomposition (Fig. 9d).

Detection of important time intervals

We now compute the importance of time intervals for the altitude (ALT.STDC) and for the angle of attack (AOA). Figure 10 displays the averaged grouped importance $G(t)$ evaluated on 50 equally spaced times points (renormalised in $[0, 1]$). The time $t = 1$ stands for the touchdown of the aircraft and $t = 0$ corresponds to 512 seconds before touchdown.



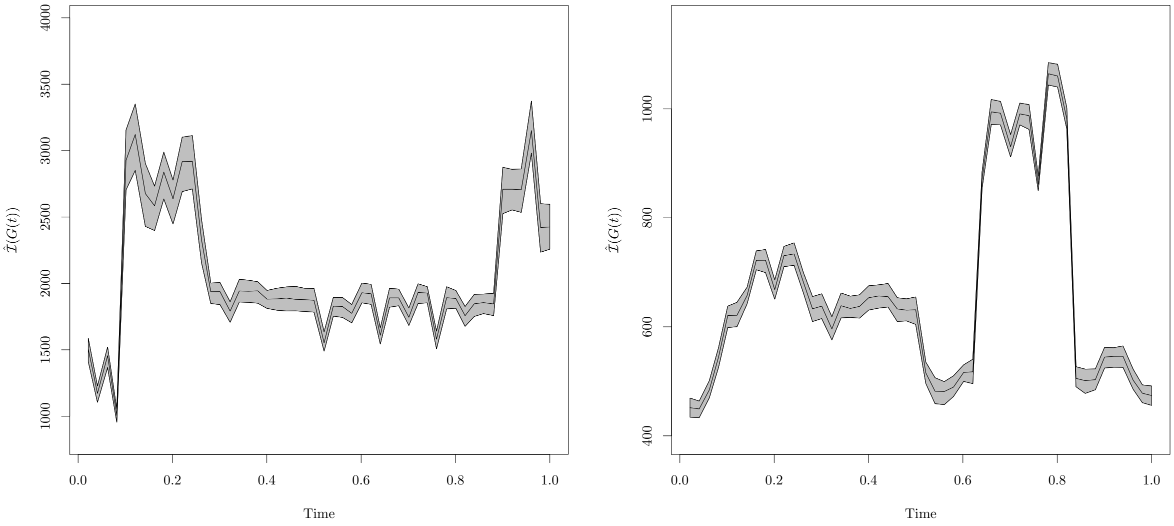
(a) Gross weight – MSE versus the number of groups

(b) Gross weight – Selection frequencies

(c) Angle of attack – MSE versus the number of groups

(d) Angle of attack – Selection frequencies

Figure 9: Application to long landing – Selection of the wavelet levels for the altitude and the angle of attack.



(a) Altitude

(b) Angle of attack

Figure 10: Application to long landing – Averaged time importance, first and third quartiles for 100 iterations.

Two intervals are detected with high predictive power for the altitude. These results are consistent with the view of aviation safety experts. Indeed, during the interval $[0.1, 0.25]$, the aircraft has to level off for stabilising before the final approach. A too high altitude at this moment can induce a long landing. During the interval $[0.9, 1]$, few seconds before touchdown, a too high altitude can also induce a long landing (Gregorutti et al.; 2014).

The interval detected for the angle of attack is $[0.6, 0.8]$. This make sense because the pilots have to reduce the airspeed few seconds before touchdown.

A Additional experiments about the Grouped Variable Importance

In this section, we investigate the properties of the permutation importance measure of groups of variables with numerical experiments, in addition to the theoretical results given before. In particular, we compare this quantity with a sum of individual importances in various models. We also study how this quantity behaves in “sparse situations” where only a small number of variables in the group are relevant for the prediction the outcome.

The general framework of the experiments is the following. For a fixed $p \geq 1$, let $\mathbf{X}^\top := (\mathbf{W}^\top, \mathbf{Z}^\top)$ where $\mathbf{W}$ and $\mathbf{Z}$ are two random vectors both of length p. Some of the components of $\mathbf{W}$ are correlated with Y whereas those in $\mathbf{Z}$ are all independent of Y. Let C_w be the variance-covariance matrix of $\mathbf{W}$. By incorporating the group $\mathbf{Z}$ in the model, we present a realistic framework where not all the X_j's have a link with Y. For each experiment, we simulate $n = 1000$ samples of Y and $\mathbf{X}$ and we compute the importance $\mathcal{I}(\mathbf{W})$ of the group $\mathbf{W}$, the normalised grouped variable importance $\mathcal{I}(\mathbf{W})$ and the sum of the individual importances of the variables in $\mathbf{W}$. We repeat each experiment 500 times. The boxplots of the importances over the 500 repetitions are drawn on Figures 11 to 14 with values p between 1 and 16.

Let 0_p and I_p denote the null vector and the identity matrix of $\mathbb{R}^p$. Let $\mathbb{1}_p$ the vector of $\mathbb{R}^p$ with all coordinates equal to one and let $0_{p,q}$ denote the null matrix of dimension $p \times q$.

Experiment 1: linear link function.

We simulate $\mathbf{X}$ and Y from a multivariate Gaussian distribution. More precisely, we simulate samples from to the joint distribution

$$\begin{pmatrix} \mathbf{X} \\ Y \end{pmatrix} = \begin{pmatrix} \mathbf{W} \\ \mathbf{Z} \\ Y \end{pmatrix} \sim \mathcal{N}_{2p+1} \left(0_{2p+1} \, , \begin{pmatrix} C_w & 0_{p,p} & \tau \\ 0_{p,p} & I_p & 0_p \\ \tau^\top & 0_p^\top & 1 \end{pmatrix} \right)$$

where $\boldsymbol{\tau}$ is the vector of the covariances between $\mathbf{W}$ and Y. In this context, the conditional distribution of Y over $\mathbf{X}$ is normal and the conditional mean f is a linear function: $f(\mathbf{x}) = \sum_{j=1}^{p+q} \alpha_j x_j$ with $\alpha = (\alpha_1, \dots, \alpha_p, 0, \dots 0)^\top$ a sequence of deterministic coefficients (see for instance Rao (1973), p. 522 and Section 3 in Gregorutti et al. (2014)).

- **Experiment 1a: independent predictors.** We take $\boldsymbol{\tau} = 0.9 \, \mathbb{1}_p$ and $C_w = I_p$. All the variables of $\mathbf{W}$ are independent and correlated with Y.

- **Experiment 1b: correlated predictors.** We take $\boldsymbol{\tau} = 0.9 \, \mathbb{1}_p$ and $C_w = (1 - 0.9)I_p + 0.9\mathbb{1}_p\mathbb{1}_p^\top$. The variables of $\mathbf{W}$ are correlated. They are also correlated with Y.

- **Experiment 1c: independent predictors, sparse case**. We take $\boldsymbol{\tau} = (0.9, 0, \dots, 0)^\top$ and $C_w = I_p$. Only the first variable in the group $\mathbf{W}$ is correlated with Y.

- **Experiment 1d: correlated predictors, sparse case**. We take $\boldsymbol{\tau} = (0.9, 0, \dots, 0)^\top$ and $C_w = (1 - 0.9)I_p + 0.9\mathbb{1}_p\mathbb{1}_p^\top$. The variables of $\mathbf{W}$ are correlated. Only the first variable in the group $\mathbf{W}$ is correlated with Y.

Experiment 2: additive link function.

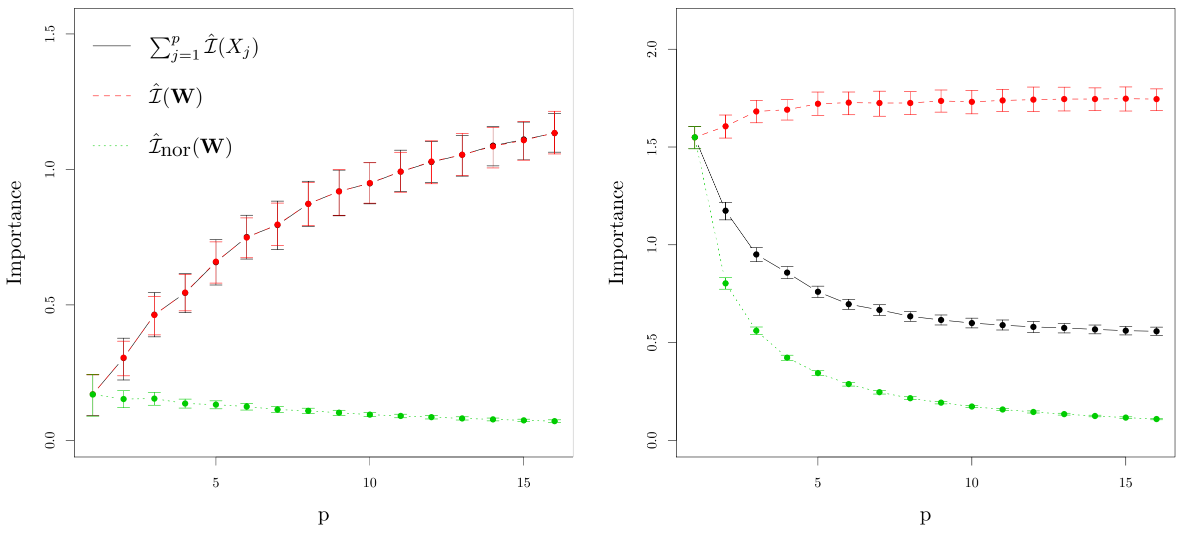
We simulate $\mathbf{X}$ from a multivariate Gaussian distribution:

$$\mathbf{W} = \begin{pmatrix} \mathbf{W} \\ \mathbf{Z} \end{pmatrix} \sim \mathcal{N}_{2p} \left(0_{2p} \, , \begin{pmatrix} C_w & 0_{p,p} \\ 0_{p,p} & I_p \end{pmatrix} \right) .$$

and the conditional distribution of Y is

$$(Y|\mathbf{X}) \sim \mathcal{N} \left(\sum_{j=1}^{p} f_j(X_j), 1 \right) ,$$

where $f_j(x) = \sin(2x) + j$ for $j < p/2$ and $f_j(x) = \cos(2x) + j$ for $j \geq p/2$.



(a) Experiments 1a

(b) Experiments 1b

Figure 11: Boxplots of the importance measures for Experiments 1a and 1b. The number of variables in $\mathbf{W}$ varies from 1 to 16. For Experiment 1a, the sum of the individual importances and $\hat{\mathcal{I}}(\mathbf{W})$ overlap.

- **Experiment 2a: independent predictors.** We take $C_w = I_p$. All the variables of $\mathbf{W}$ are independent and correlated with Y.
- **Experiment 2b: correlated predictors.** We take $C_w = (1 - 0.9)I_p + 0.9\mathbb{1}_p\mathbb{1}_p^\top$. The variables of $\mathbf{W}$ are correlated and also correlated with Y.

Experiment 3: link function with interactions.

We simulate $\mathbf{X}$ from a multivariate Gaussian distribution:

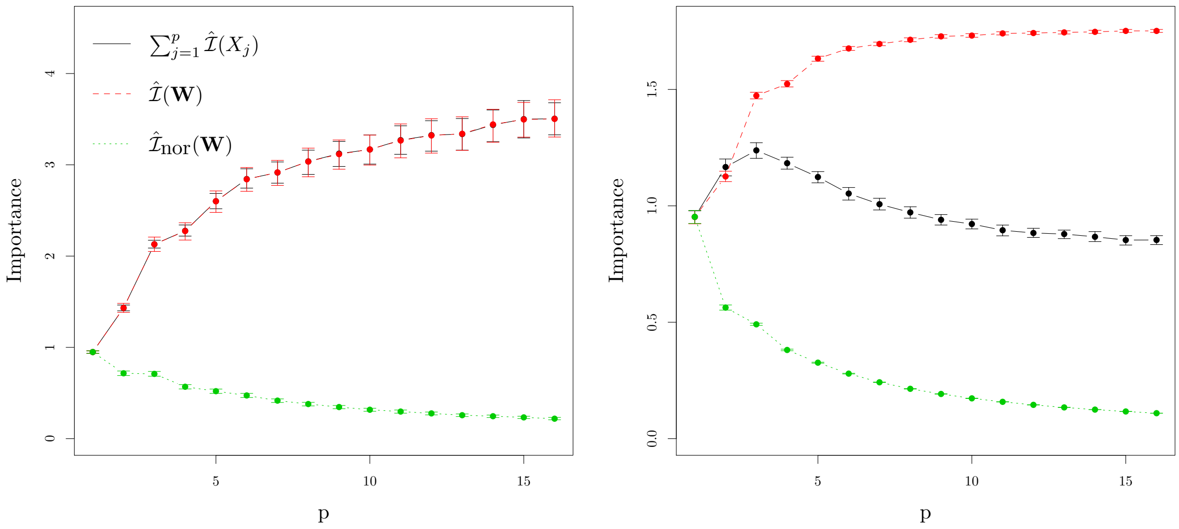
$$\mathbf{W} = \begin{pmatrix} \mathbf{W} \\ \mathbf{Z} \end{pmatrix} \sim \mathcal{N}_{2p}\left(0_{2p}\, I_{2p}\right).$$

and the conditional distribution of Y is

$$(Y|\mathbf{X}) \sim \mathcal{N}\left(\sum_{j=1}^{p} X_j + X_p X_1 + \sum_{j=1}^{p-1} X_j X_{j+1}\,,\, 1\right).$$

Results

Experiments 1a-b and 2a-b illustrate the results of Corollary 1 (see Figure 11 and 12). Indeed the regression function of both cases satisfies the additive property (2.3) for these experiments. In Experiments 1a and 2a, the variables of the group $\mathbf{W}$ are independent



(a) Experiments 2a

(b) Experiments 2b

Figure 12: Boxplots of the importance measures for Experiments 2a and 2b. The number of variables in $\mathbf{W}$ varies from 1 to 16. For Experiment 2a, the sum of the individual importances and $\hat{\mathcal{I}}(\mathbf{W})$ overlap.

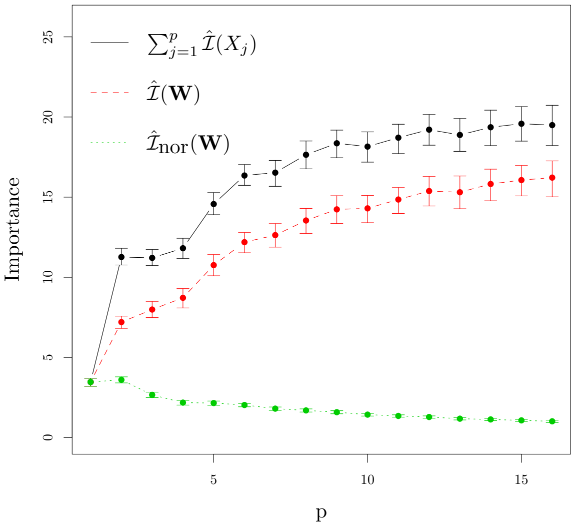
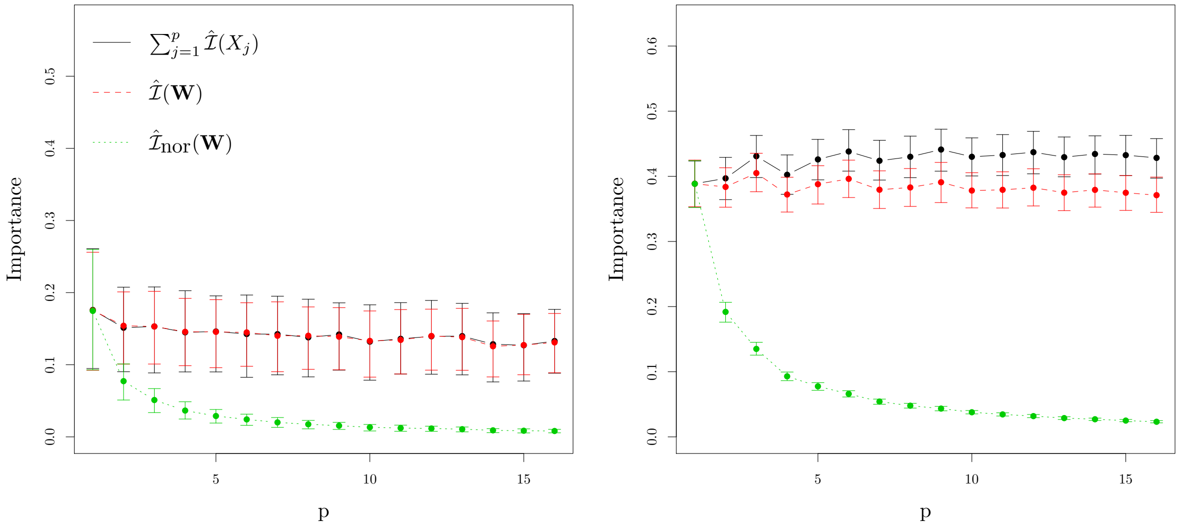


Figure 13: Boxplots of the importance measures for Experiment 3. The number of variables in $\mathbf{W}$ varies from 1 to 16.



(a) Experiment 1c

(b) Experiment 1d

Figure 14: Boxplots of the importance measures for Experiments 1c and 1d. The number of variables in $\mathbf{W}$ varies from 1 to 16. For Experiment 2a, the sum of the individual importances and $\hat{\mathcal{I}}(\mathbf{W})$ overlap.

and the grouped variable importance is nothing more than the sum of the individual importances in this case. In Experiments 1b and 2b, the variables of the group $\mathbf{W}$ are positively correlated. In these situations, the grouped variable importance is larger than the sum of the individual importances, which agrees with Equation 2.4. Note that the grouped variable importance increases with p, which is natural because the amount of information for predicting Y increases with the group size in these models. On the other hand, it was shown in Gregorutti et al. (2014) that individual importances decrease with correlation between the predictors. Indeed we observe that the sum of the individual importances decreases with p in the correlated cases.

The regression function of Experiments 3 does not satisfy the additive form (2.3). Although the variables in the group are independent, the grouped variable importance is not equal to the sum of the individual importances (Figure 13). In a general setting, it appears that these two quantities differ.

We now comment the results of the sparse Experiments 1c and 1d (Figure 14). It is clear that $\mathcal{I}(\mathbf{W}) = \mathcal{I}(X_1) = \sum_{j=1\ldots p} \mathcal{I}(X_j)$ for these two experiments (see Proposition 1 for instance). Regarding Experiment 1c, the boxplots of the estimated values $\hat{\mathcal{I}}(\mathbf{W})$ and $\sum_{j=1\ldots p} \hat{\mathcal{I}}(X_j)$ agree with this equality. On the other hand, in Experiment 1d, we observe that $\sum_{j=1\ldots p} \hat{\mathcal{I}}(X_j)$ is significantly higher than $\hat{\mathcal{I}}(\mathbf{W})$. Indeed, it has been noticed by Nicodemus et al. (2010) that the individual importances of predictors that are not associated with the outcome tend to be overestimated by the random forests, when there is correlation between predictors. In contrast, the estimator $\hat{\mathcal{I}}(\mathbf{W})$ seems to correctly estimate $\mathcal{I}(\mathbf{W})$ even for large p. Indeed, for both experiments 1c and 1d, the importance

$\hat{\mathcal{I}}(\mathbf{W})$ is unchanged when the size of the group p varies from 2 to 16. Note that for variable selection, we may prefer to consider the normalized importance $\hat{\mathcal{I}}_{\text{nor}}$ to select in priority small group of variables.

B Curve dimension reduction with wavelets

The analysis of flights data in Section 4 required, for computational reasons, to preliminary reduce the dimension of the wavelet decomposition of the flight parameters. We need to adapt the famous wavelet shrinkage method to the context of independent random processes. Using the notations of Section 3.1, we first recall the hard-thresholding estimator introduced by Donoho and Johnstone (1994) in the case of one random signal. This approach is then extended to deal with n independent random signals.

B.1 Signal denoising via wavelet shrinkage

The problem of signal denoising can be summarised as follows. Suppose that we observe N noisy samples $X(t_1), \dots, X(t_N)$ of a deterministic function s (the signal):

$$X(t_\ell) = s(t_\ell) + \sigma\varepsilon_\ell, \quad \ell = 1, \dots, N \tag{B.1}$$

where the ε_ℓ's are independent standard Gaussian random variables. We assume that s belongs to $L^2([0,1])$. The goal is to recover the underlying function s from the noisy data $\{X(t_\ell), \ell \in \{1, \dots, N\}\}$ with small error. Using the discrete wavelet transform, this model can be rewritten in the wavelet domain as

$$\xi_{jk} = \omega_{jk} + \sigma\eta_{jk}, \quad \forall j \in \{0, \dots, J-1\}, \forall k \in \{0, \dots, 2^j - 1\},$$

and the scaling domain as

$$\zeta = \omega_0 + \sigma\eta_0,$$

where ξ_{jk} and ζ are the empirical wavelet and scaling coefficients of $X(t_\ell)$ as in Equation (3.2). The random variables η_{jk} and η_0 are i.i.d. random variables from the distribution $\mathcal{N}_1(0,1)$.

A natural approach for estimating ω_{jk} is to shrink the coefficients ξ_{jk} to zero. An estimator of s in this context has the form

$$\hat{s}(t_\ell) = \hat{\omega}_0\phi(t_\ell) + \sum_{j=0}^{J-1}\sum_{k=0}^{2^j-1} \hat{\omega}_{jk}\psi_{jk}(t_\ell) \tag{B.2}$$

with $\hat{\omega}_0 = \zeta$ and

$$\hat{\omega}_{jk} = \begin{cases} \xi_{jk} & \text{if } |\xi_{jk}| > \delta_N \\ 0 & \text{otherwise.} \end{cases}$$

This method refers as the *hard-thresholding estimator* in the literature. Donoho and Johnstone (1994) propose the universal threshold $\delta_N = \sigma\sqrt{2\log(N)}$. In addition, the standard deviation σ can be estimated by the median absolute deviation (MAD) estimate of the wavelet coefficients at the finest levels, i.e.

$$\hat{\sigma} = \frac{\mathrm{Med}(|\xi_{jk} - \mathrm{Med}(\xi_{jk})| : j = J-1, k = 0, \ldots, 2^{J-1} - 1)}{0.6745},$$

where the normalization factor 0.6745 comes from the normality assumption in (B.1). This estimator is known to be a robust and consistent estimator of σ. The underlying idea is that the variability of the wavelet coefficients is essentially concentrated at the finest level.

B.2 Consistent wavelet thresholding for independent random signals

This section presents a natural extension of the hard-thresholding method when n independent random processes $X_1, \ldots, X_n$ are observed. The aim is to reduce the dimension of the problem by shrinking the wavelet coefficients of the n signals. One simple solution consists in applying independently the hard-thresholding rule to each wavelet decomposition. By doing so, the n shrunk decompositions have no reason to be identical: the wavelet coefficients set to zero will be different for the n signals. In some situations (as for Section 4 in this paper), it is preferable to shrink the same coefficients of the n observed signals. We adapt the hard-thresholding method to answer this problem.

Identically distributed case

We start by assuming that the observations come from the same distribution: for any $i \in \{1, \ldots, n\}$,

$$X_i(t_\ell) = s(t_\ell) + \sigma \varepsilon_{i,\ell}, \quad \ell = 1, \ldots, N, \tag{B.3}$$

where the $\varepsilon_{i,\ell}$'s are independent standard Gaussian random variables. The wavelet coefficients of s can be easily deduced from the mean signal $\bar{X} := \frac{1}{n}\sum_{i=1}^{n} X_i$ which satisfies

$$\bar{X}(t_\ell) = s(t_\ell) + \frac{\sigma}{\sqrt{n}} \varepsilon_\ell, \quad \ell = 1, \ldots, N,$$

where the ε_ℓ's are independent standard Gaussian random variables. By applying the hard-thresholding rule to this signal, we obtain the following estimation of the wavelet parameters of s: $\hat{\omega}_0 = \zeta$ and

$$\hat{\omega}_{jk} = \begin{cases} \bar{\xi}_{jk} & \text{if } |\bar{\xi}_{jk}| > \bar{\delta}_N \\ 0 & \text{otherwise,} \end{cases}$$

where $\bar{\xi}_{jk}$ is the wavelet coefficient (j,k) of $\bar{X}$. Here the threshold is $\bar{\delta}_N = \frac{\sigma}{\sqrt{n}}\sqrt{2\log(N)}$.

Non identically distributed case

In many real life situations, assuming that the n signals are identically distributed is not a realistic assumption. For the study presented in Section 4 for instance, the flight parameters have no reason to follow the same distribution in safe and unsafe conditions. We propose a generalisation of the model (B.3) by introducing a latent random variable Z

taking it value in a set $\mathcal{Z}$. Roughly speaking, the variable Z represents all the phenomena that have an effect on the mean signal. Conditionally to $Z_i = z_i$, the distribution of the process X_i is now defined, for any $i \in \{1, \dots, n\}$, by

$$X_i(t_\ell) = s(t_\ell, z_i) + \sigma\varepsilon_{i,\ell}\,, \quad \ell = 1, \dots, N, \tag{B.4}$$

where the $\varepsilon_{i,\ell}$'s are independent standard Gaussian random variables. This regression model allows us to consider various situations of interest arising in Functional Data Analysis. In supervised settings where a variable Y has to be predicted using X, one reasonable modeling is taking $Z = Y$. We now propose an hard-thresholding method which simultaneously shrinks the wavelet decomposition of the n signals.

Let $\|\cdot\|_n$ denote the ℓ_2-norm in $\mathbb{R}^n$: $\|\boldsymbol{u}\|_n := \sqrt{\sum_{i=1}^n u_i^2}$ for any $\boldsymbol{u} \in \mathbb{R}^n$. Let $\boldsymbol{\xi}_{jk}$ be the vector $(\xi_{1jk}, \dots, \xi_{ijk}, \dots, \xi_{njk})^\top$ where ξ_{ijk} is the coefficient of level (j,k) in the wavelet decomposition of the signal X_i.

For any $z \in \mathcal{Z}$, let $\omega_{jk}(z)$ be the wavelet coefficient of level (j,k) of $s(\cdot, z)$, and $\boldsymbol{\omega}_{jk} := (\omega_{jk}(Z_1), \dots, \omega_{jk}(Z_n))^\top$. We define the common wavelet support of s by

$$L := \{(j,k) \,|\, \omega_{jk}(Z) = 0 \; a.s\}\,.$$

If $(j,k) \in L$, then $\boldsymbol{\omega}_{jk} = (0, \dots, 0)^\top$ almost surely and $\|\boldsymbol{\xi}_{jk}\|_n^2$ has a centered chi-square distribution with n degrees of freedom. Otherwise, $\boldsymbol{\omega}_{jk}$ can be not null and in this case $\|\boldsymbol{\xi}_{jk}\|_n^2$ has the distribution of a sum of n independent uncentered chi-square distributions. We thus propose a thresholding rule for the statistic $\|\boldsymbol{\xi}_{jk}\|_n$. For any $j \in \{0, \dots, J-1\}$ and any $k \in \{0, \dots, 2^j - 1\}$, let

$$\hat{\boldsymbol{\omega}}_{jk} = \begin{cases} \boldsymbol{\xi}_{jk} & \text{if } \|\boldsymbol{\xi}_{jk}\|_n > \delta_{N,n} \\ (0, \dots, 0)^\top & \text{otherwise,} \end{cases} \tag{B.5}$$

where the threshold $\delta_{N,n}$ depends on N, n and σ.

Proving adaptive results in the spirit of Donoho et al. (1995) for this method is beyond the scope of the paper. However, an elementary consistent result can be proved. We would like $\hat{\boldsymbol{\omega}}_{jk}$ to be a zero vector with high probability when $(j,k) \in L$. For some $x \geq 0$, take $\delta_{N,n}^2 = \delta_{N,n}^2(x) = \sigma^2(2x + 2\sqrt{nx} + n)$, then

$$\begin{aligned} \mathbb{P}\left[\bigcup_{(j,k)\in L} \{\hat{\boldsymbol{\omega}}_{jk} \neq (0, \dots, 0)^\top\}\right] &\leq \sum_{(j,k)\in L} \mathbb{P}\left[\|\boldsymbol{\xi}_{jk}\|_n^2 \geq \delta_{N,n}^2(x)\right] \\ &= \sum_{(j,k)\in L} \mathbb{P}\left[\frac{\|\boldsymbol{\xi}_{jk}\|_n^2}{\sigma^2} - n \geq 2x + 2\sqrt{nx}\right] \\ &\leq |\bar{L}|e^{-x} \leq Ne^{-x} \end{aligned} \tag{B.6}$$

where we have used a deviation bound for central chi-square distributions from Laurent and Massart (2000, p. 1325). If the signal is exactly zero, it can be recovered with high probability by taking $x \gg \log(N)$. In particular, if we choose $x = 2\log(N)$, the threshold is $\delta_{N,n}^2 = \sigma^2(4\log(N) + 2\sqrt{2n\log(N)} + n)$ and the convergence rate in (B.6) is of order $O(\frac{1}{N})$. In practice, σ can be estimated by a MAD estimator computed on the coefficients of the highest level of all the n wavelet decompositions. Next, x and $\delta_{N,n}$ can be chosen such that (B.6) is lower than a given probability q. Letting $Ne^{-x} = q$, we obtain the threshold

$$\delta_{N,n} = \hat{\sigma}\left(2\log\left(\frac{N}{q}\right) + 2\sqrt{n\log\left(\frac{N}{q}\right)} + n\right)^{\frac{1}{2}}.$$

Assuming that $\omega_{j,k}(Z) = 0$ almost surely for some level (j,k) is a strong assumption that can be hardly met in practice. Hopefully, this method still works if the wavelet support of $s(\cdot, z)$ does not vary too much with z. In particular, it may be applied if there exists a common set S of indexes (j,k) such that, for any z, the projection of $s(\cdot, z)$ on $\mathrm{Vect}(\psi_{jk} \,|\, (j,k) \in S)$ is not too far from $s(\cdot, z)$ for the L^2 norm.

References

Amato, U., Antoniadis, A. and De Feis, I. (2006). Dimension reduction in functional regression with applications, *Computational Statistics and Data Analysis* **50**: 2422–2446.

Antoniadis, A., Bigot, J. and Sapatinas, T. (2001). Wavelet estimators in nonparametric regression: A comparative simulation study, *Journal of Statistical Software* pp. 1–83.

Araki, Y., Konishi, S., Kawano, S. and Matsui, H. (2009). Functional logistic discrimination via regularized basis expansions, *Communications in Statistics, Theory and Methods* **38**: 2944–2957.

Biau, G., Bunea, F. and Wegkamp, M. (2005). Functional classification in hilbert spaces, *IEEE Transactions on Information Theory* **51**: 2163–2172.

Breiman, L. (1996). Bagging predictors, *Machine Learning* **24**: 123–140.

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

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

Cai, T. and Hall, P. (2006). Prediction in functional linear regression, *The Annals of Statistics* **34**(5): 2159–2179.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model, *Statistics and Probability Letters* **45**: 11–22.

Cardot, H., Ferraty, F. and Sarda, P. (2003). Spline estimators for the functional linear model, *Statistica Sinica* **13**: 571–592.

Chakraborty, D. and Pal, N. R. (2008). Selecting useful groups of features in a connectionist framework, *IEEE Transactions on Neural Networks* **19**(3): 381–396.

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

Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage, *Biometrika* **81**: 425–455.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G. and Picard, D. (1995). Wavelet shrinkage: asymptopia, *Journal of the Royal Statistical Society, Series B* **57**: 301–369.

Fan, Y. and James, G. (2013). Functional additive regression. Preprint.

Ferraty, F. (ed.) (2011). *Recent Advances in Functional Data Analysis and Related Topics*, Springer-Verlag.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis: Theory and Practice (Springer Series in Statistics)*, Springer-Verlag New York, Inc.

Fromont, M. and Tuleau, C. (2006). Functional classification with margin conditions, *19th Annual Conference on Learning Theory.*

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

Gregorutti, B., Michel, B. and Saint Pierre, P. (2014). Correlation and variable importance in random forests. arXiv:1310.5726.

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

Guyon, I., Weston, J., Barnhill, S. and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines, *Machine Learning* **46**(1-3): 389–422.

He, Z. and Yu, W. (2010). Stable feature selection for biomarker discovery, *Computational biology and chemistry* **34**(4): 215–225.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection, *The Annals of Statistics* **28**: 1245–1501.

Matsui, H. (2014). Variable and boundary selection for functional data via multiclass logistic regression modeling, *Computational Statistics and Data Analysis* **78**(0): 176 – 185.

Matsui, H. and Konishi (2011). Variable selection for functional regression models via the regularization, *Computational Statistics and Data Analysis* **55**(12): 3304–3310.

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

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*, Cambridge University Press.

Poggi, J.-M. and Tuleau, C. (2006). Classification supervisée en grande dimension. application à l'agrément de conduite automobile, *Revue de Statistique Appliquée* **4**: 39–58.

Ramsay, J. O. and Silverman, B. W. (2005). *Functional Data Analysis*, Springer Series in Statistics, Springer.

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

Rossi, F., François, D., Wertz, V. and Verleysen, M. (2006). A functional approach to variable selection in spectrometric problems, *Proceedings of 16th International Conference on Artificial Neural Networks, ICANN 2006*, pp. 11–20.

Rossi, F. and Villa, N. (2006). Support vector machine for functional data classification, *Neurocomputing* **69**: 730–742.

Rossi, F. and Villa, N. (2008). Recent advances in the use of svm for functional data classification, *Proceedings of 1rst International Workshop on Functional and Operatorial Statistics, IWFOS*.

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

Yang, K., Yoon, H. and Shahabi, C. (2005). A supervised feature subset selection technique for multivariate time series, *Proceedings of the Workshop on Feature Selection for Data Mining: Interfacing Machine Learning with Statistics*.

Yoon, H. and Shahabi, C. (2006). Feature subset selection on multivariate time series with extremely large spatial features, *Proceedings of the Sixth IEEE International Conference on Data Mining - Workshops*, ICDMW '06, pp. 337–342.

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

Zhang, H. H., Liu, Y., Wu, Y., Zhu, J. et al. (2008). Variable selection for the multicategory svm via adaptive sup-norm regularization, *Electronic Journal of Statistics* **2**: 149–167.

Zhu, R., Zeng, D. and Kosorok, M. R. (2012). Reinforcement learning trees, *Technical report*, University of North Carolina.