Statistica Sinica **25** (2015), 1477-1502
doi:http://dx.doi.org/10.5705/ss.2013.310

$\mathbb{L}_2$-BOOSTING FOR SENSITIVITY ANALYSIS WITH DEPENDENT INPUTS

Magali Champion[1,3], Gaelle Chastaing[1,2], Sébastien Gadat[4] and Clémentine Prieur[2]

[1] *Université Paul Sabatier,* [2] *Université Grenoble Alpes CNRS,* [3] *INRA and* [4] *Université Toulouse I Capitole*

Abstract: This paper is dedicated to the study of an estimator of the generalized Hoeffding decomposition. We build an estimator using an empirical Gram-Schmidt approach and derive a consistency rate in a large dimensional setting. We then apply a greedy algorithm with these previous estimators to a sensitivity analysis. We also establish the consistency of this $\mathbb{L}_2$-boosting under sparsity assumptions of the signal to be analyzed. The paper concludes with numerical experiments that demonstrate the low computational cost of our method, as well as its efficiency on the standard benchmark of sensitivity analysis.

Key words and phrases: $\mathbb{L}_2$-boosting, convergence, dependent variables, generalized ANOVA decomposition, sensitivity analysis.

1. Introduction

In many scientific fields, it is desirable to extend a multivariate regression model as a specific sum of increasing dimension functions. Functional ANOVA decompositions and High Dimensional Representation Models (HDMR) (Hooker (2007); Li et al. (2010)) are well known expansions that make it possible to understand model behavior and to detect how inputs interact with each other.

When input variables are independent, Hoeffding establishes the uniqueness of the decomposition provided the summands are mutually orthogonal (Hoeffding (1948)). However, in practice, this assumption is sometimes difficult to justify, or even wrong (see Li and Rabitz (2012) for an application to correlated ionosonde data, or Jacques, Lavergne, and Devictor (2006), who studied an adjusted neutron spectrum inferred from a correlated dependent nuclear dataset).

When inputs are correlated, the orthogonality properties of the classical Sobol decomposition (Sobol (1993)) are no longer satisfied. As pointed out by several authors (Hooker (2007); Da Veiga, Wahl, and Gamboa (2009)), a global sensitivity analysis based on this decomposition may lead to erroneous conclusions. Following the work of Stone (1994), later applied to Machine Learning by Hooker (2007), and to sensitivity analysis by Chastaing, Gamboa, and Prieur

(2012), we consider a hierarchically orthogonal decomposition whose uniqueness has been proven under mild conditions on the dependence structure of the inputs (Chastaing, Gamboa, and Prieur (2012)). Thus, any model function can be uniquely decomposed as a sum of *hierarchically orthogonal* component functions. Two summands are considered to be *hierarchically orthogonal* whenever all of the variables included in one of them are also involved in the other. For a better understanding, this generalized ANOVA expansion is referred to as a Hierarchically Orthogonal Functional Decomposition (HOFD).

It is important to develop estimation procedures since the analytical formulation for HOFD is rarely available. We focus on a method proposed in Chastaing, Gamboa, and Prieur (2013) to estimate the HOFD components. It consists of constructing a hierarchically orthogonal basis from a suitable Hilbert orthonormal basis. The procedure recursively builds a multidimensional basis for each component that satisfies the identifiability constraints imposed on this summand. Each component is then well approximated on a truncated basis where the unknown coefficients are deduced by solving an ordinary least-squares regression. While in a high-dimensional paradigm, this procedure suffers from the curse of dimensionality, numerically, it is observed that only a few of the coefficients are not close to zero, so a small number of predictors can restore the major part of the information contained in the components. Thus, it is important to be able to select the most relevant representative functions and to then identify the HOFD with a limited computational budget.

We suggest here how to transform an ordinary least-squares regression into a penalized regression, as has been proposed in Chastaing, Gamboa, and Prieur (2013). In the present paper, we focus on the $\mathbb{L}_2$-boosting developed by Friedman (2001) to deal with the ℓ_0 penalization. $\mathbb{L}_2$-boosting is a greedy strategy that performs variable selection and shrinkage, is intuitive, and easy to implement. It is closely related to the LARS algorithm proposed by Efron et al. (2004), used for Lasso regression (Tibshirani (1996); Bühlmann and van de Geer (2011)); $\mathbb{L}_2$-boosting and LARS both select predictors using the maximal correlation with the current residuals.

The goal of this paper is to provide an overall consistent estimation of a signal spanned in a large-dimensional dictionary derived from a HOFD, thereby improving the results of Chastaing, Gamboa, and Prieur (2013). We first address the convergence rate of the empirical HOFD and then use this result to obtain a sparse estimator of the unknown signal. We need to manage sufficient theoretical conditions to ensure the consistency of our estimator. We discuss these conditions and provide some numerical examples in which such conditions are fulfilled.

The article is organized as follows. Notation is set in Section 2.1, and Section 2.2 provides the HOFD representation of the model function. In Section 2.3, we

review the procedure detailed in Chastaing, Gamboa, and Prieur (2013) that consists of constructing well-tailored hierarchically orthogonal bases to represent the components of the HOFD, then, highlight the curse of dimensionality that we are exposed to, and present the $\mathbb{L}_2$-boosting. Section 3 gives our main results on the proposed algorithms, and an interesting application of the general theory to global sensitivity analysis (SA). In Section 4, we apply $\mathbb{L}_2$-boosting to estimate the generalized sensitivity indices defined in Chastaing, Gamboa, and Prieur (2012, 2013), then, numerically compare the $\mathbb{L}_2$-boosting performance with a Lasso strategy and the Forward-Backward algorithm proposed by Zhang (2011). Section 5 concludes. The proofs of the main theorems are in the Supplementary Material.

2. Estimation of the Generalized Hoeffding Decomposition Components

2.1. Notation

We consider a measurable function f of a random real vector $\mathbf{X} = (X_1, \ldots, X_p)$ of $\mathbb{R}^p$, $p \geq 1$. The response variable Y is a real-valued random variable defined as:

$$Y = f(\mathbf{X}) + \varepsilon, \tag{2.1}$$

where ε is a centered random variable independent of $\mathbf{X}$ that models the variability of the response around its theoretical unknown value f. We denote the distribution law of $\mathbf{X}$ by $P_{\mathbf{X}}$, unknown in our setting, and we assume that $P_{\mathbf{X}}$ admits a density function $p_{\mathbf{X}}$ with respect to Lebesgue measure on $\mathbb{R}^p$. Here the components of $\mathbf{X}$ may be correlated.

We suppose that $f \in L^2_{\mathbb{R}}(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_{\mathbf{X}})$, where $\mathcal{B}(\mathbb{R}^p)$ denotes the Borel set of $\mathbb{R}^p$. The Hilbert space $L^2_{\mathbb{R}}(\mathbb{R}^p, \mathcal{B}(\mathbb{R}^p), P_{\mathbf{X}})$ is denoted by $L^2_{\mathbb{R}}$, for which we use the inner product $\langle \cdot, \cdot \rangle$, and the norm $\|\cdot\|$. We use $\mathbb{E}(\cdot)$ for expected value, $V(\cdot) = \mathbb{E}[(\cdot - \mathbb{E}(\cdot))^2]$ for variance, and $\mathrm{Cov}(\cdot, *) = \mathbb{E}[(\cdot - \mathbb{E}(\cdot))(* - \mathbb{E}(*))]$ for covariance.

For any $1 \leq i \leq p$, we denote the marginal distribution of X_i by $P_{\mathbf{X}_i}$ and extend our notation to $L^2_{\mathbb{R}}(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{\mathbf{X}_i}) := L^2_{\mathbb{R},i}$.

2.2. The generalized Hoeffding decomposition

Let $[1 : k] := \{1, 2, \ldots, k\}$, with $k \in \mathbb{N}^*$, and let S be the collection of all subsets of $[1 : p]$. Take $S^* := S \setminus \{\emptyset\}$. For $u \in S$, the subvector $\mathbf{X}_u$ of $\mathbf{X}$ is defined as $\mathbf{X}_u := (X_i)_{i \in u}$. Conventionally, for $u = \emptyset$, $\mathbf{X}_u = 1$. The marginal distribution (*resp.* density) of $\mathbf{X}_u$ is denoted by $P_{\mathbf{X}_u}$ (*resp.* $p_{\mathbf{X}_u}$).

A functional ANOVA decomposition consists in expanding f as a sum of increasing dimension functions:

$$\begin{aligned} f(\mathbf{X}) &= f_\emptyset + \sum_{i=1}^{p} f_i(X_i) + \sum_{1\le i<j\le p} f_{ij}(X_i, X_j) + \cdots + f_{1,\ldots,p}(\mathbf{X}) \\ &= \sum_{u\in S} f_u(\mathbf{X}_u), \end{aligned} \tag{2.2}$$

where $f_\emptyset$ is a constant term, f_i, $i \in [1:p]$ are the main effects, $f_{ij}, f_{ijk}, \cdots$, $i, j, k \in [1:p]$ are the interaction effects, and the last component $f_{1,\ldots,p}$ is the residual. Decomposition (2.2) is generally not unique. However, under mild assumptions on the joint density $p_{\mathbf{X}}$ (see Assumptions (C.1) and (C.2) in Chastaing, Gamboa, and Prieur (2012)), the decomposition is unique under some additional orthogonality assumptions.

Let $H_\emptyset = H_\emptyset^0$ be the set of constant functions and, for all $u \in S^*$, $H_u := L^2_{\mathbb{R}}(\mathbb{R}^u, \mathcal{B}(\mathbb{R}^u), P_{\mathbf{X}_u})$. For $u \in S \setminus \emptyset$, we take

$$H_u^0 = \left\{ h_u \in H_u,\ \langle h_u, h_v\rangle = 0, \forall\ v \subset u, \forall\ h_v \in H_v^0 \right\},$$

where $\subset$ denotes the strict inclusion. Under Assumptions (C.1) and (C.2) in Chastaing, Gamboa, and Prieur (2012), the decomposition (2.2) is unique if $f_u \in H_u^0$ for all $u \in S$. It is termed the Hierarchical Orthogonal Functional Decomposition (HOFD).

Remark 1. The components of the HOFD (2.2) are assumed to be hierarchically orthogonal, that is, $\langle f_u, f_v\rangle = 0\ \forall v \subset u$.

There is more information about the HOFD in Hooker (2007) and Chastaing, Gamboa, and Prieur (2012). Here, we are interested in estimating the summands in (2.2). As underlined in Huang (1998), estimating all components of (2.2) is an intractable problem in practice. To bypass this, we assume (without loss of generality) that f is centered, $f_\emptyset = 0$. Most models are governed by low-order interaction effects, as pointed out in Crestaux, Le Maître, and Martinez (2009) Blatman (2009), and Li et al. (2010), and we suppose that f is well approximated as

$$f(\mathbf{X}) \simeq \sum_{\substack{u\in S^* \\ |u|\le d}} f_u(\mathbf{X}_u), \quad d \ll p, \tag{2.3}$$

meaning that interactions of order $\ge d+1$ can be neglected. The choice of d, which is directly related to the notion of effective dimension in the superposition sense (see Definition 1 in Wang and Fang (2003)), is addressed in Zuniga, Kucherenko, and Shah (2013), but we assume it is fixed by the user. Even by

choosing a small d, the number of components in (2.3) can become prohibitive if the number of inputs p is high. We are interested in estimation procedures under sparse assumptions when the number of variables p is large.

2.3. Practical determination of the sparse HOFD

General description of the procedure

We propose a two-step estimation procedure to identify the components in (2.3): the first one is a simplified version of the Hierarchical Orthogonal Gram-Schmidt (HOGS) procedure developed in Chastaing, Gamboa, and Prieur (2013), and the second consists of a sparse estimation in the dictionary learned by the empirical HOGS.

We have chosen to use the so-called $\mathbb{L}_2$-boosting procedure instead of the widely used Lasso estimator. Our motivation is as follows.

From a technical point of view, the empirical HOGS produces a noisy estimation of the theoretical dictionary in which the true signal f is expanded. The arguments for Lasso estimation would have to be adjusted to this situation with errors in the variables. Moreover, as an M-estimator, such a modification is far from trivial (see Cavalier and Hengartner (2005) for an example of oracle inequalities derived from M estimators with noise in the variables). In contrast, the approximation obtained in the empirical HOGS can be easily handled with the boosting algorithm since we just have to quantify how the empirical inner products built with noisy variables are close to theoretical ones. Our proofs rely precisely on this strategy: we obtain a uniform bound on our statistical estimation of the HOGS dictionary, and then take advantage of the sequential description of the boosting with empirical inner products.

In order to obtain consistent estimation with the boosting procedure, we do not need to make any coherence assumption on the dictionary (such as the RIP assumption of Candes and Tao (2007) or the weakest RE(s, c_0) assumption of Bickel, Ritov, and Tsybakov (2009)). Such assumptions are generally necessary to assert some consistency results for the Dantzig and Lasso procedures, such as Sparse Oracle Inequalities (SOI). Nevertheless, it would be reasonable to impose these assumptions on the *theoretical* version of the HOGS although it seems difficult to deduce coherence results on the *empirical* HOGS from coherence results on the *theoretical* version of the HOGS. Our Theorem 2 below does not produce a SOI in expectation, our results are expressed in probability. We discuss the asymptotics of Theorem 2 after its statement, and underline the differences with the state-of-the-art results on the Lasso estimator.

To carry out the two-step procedure, we assume that we observe two independent and identically distributed samples, $(y^r, \mathbf{x}^r)_{r=1,\dots,n_1}$ and $(y^s, \mathbf{x}^s)_{s=1,\dots,n_2}$,

from the distribution of $(Y, \mathbf{X})$ (the initial sample can be split into such two samples). The empirical inner product $\langle \cdot, \cdot \rangle_n$ and the empirical norm $\|\cdot\|_n$ are

$$\langle h, g \rangle_n = \frac{1}{n} \sum_{s=1}^{n} h(\mathbf{x}^s) g(\mathbf{x}^s), \quad \|h\|_n = \langle h, h \rangle_n.$$

For $u = (u_1, \ldots, u_t) \in S$, we define the multi-index $\boldsymbol{l_u} = (l_{u_1}, \ldots, l_{u_t}) \in \mathbb{N}^t$, while $\text{Span}\{B\}$ is the linear span of the elements of B.

Step 1 and Step 2 of our sparse HOFD procedure will be described in detail below.

Remark 2. Our procedure can be extended to higher order approximations, but the description of the methodology for $d = 2$ provides better clarity.

Step 1: Hierarchically orthogonal Gram-Schmidt procedure

For each $i \in [1:p]$, let $\{1\,,\ \psi^i_{l_i}\,,\ l_i \in \mathbb{N}^*\}$ denote an orthonormal basis of $H_i := L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$. For $L \in \mathbb{N}^*$, for $i \neq j \in [1:p]$, we write

$$H^L_\emptyset = \text{Span}\{1\} \quad \text{and} \quad H^L_i = \text{Span}\left\{1, \psi^i_1, \ldots, \psi^i_L\right\},$$
$$H^L_{ij} = \text{Span}\left\{1, \psi^i_1, \ldots, \psi^i_L, \psi^j_1, \ldots, \psi^j_L, \psi^i_1 \otimes \psi^j_1, \ldots, \psi^i_L \otimes \psi^j_L\right\},$$

where $\otimes$ denotes the tensor product between two elements of the basis. The approximation of H^0_u is

$$H^{L,0}_u = \left\{h_u \in H^L_u,\ \langle h_u, h_v \rangle = 0, \forall\ v \subset u, \forall\ h_v \in H^{L,0}_v\right\}.$$

The recursive procedure aims to construct a basis for $H^{L,0}_i$, and a basis for $H^{L,0}_{ij}$ for any $i \neq j \in [1:p]$.

Initialization

For any $1 \leq i \leq p$, let $\phi^i_{l_i} := \psi^i_{l_i}$, $l_i \in [1:L]$. As a result of the orthogonality of $\{\psi^i_{l_i}\,,\ l_i \in \mathbb{N}\}$, we obtain $H^{L,0}_i := \text{Span}\left\{\phi^i_1, \cdots \phi^i_L\right\}$. For this step, we just need the orthogonality of the constant function with each of the $\psi^i_{l_i}\,,\ l_i \in \mathbb{N}^*$. However, orthogonality is needed for the proof of the consistency of the $\mathbb{L}_2$-boosting procedure.

Second order interactions

Let $u = \{i, j\}$, with $i \neq j \in [1:p]$. Since the dimension of H^L_{ij} is $L^2 + 2L + 1$, and the approximation space $H^{L,0}_{ij}$ is subject to $2L + 1$ constraints, its dimension is L^2. We want to construct a basis for $H^{L,0}_{ij}$ that satisfies the hierarchical orthogonal constraints and is of the form

$$\phi^{ij}_{\boldsymbol{l_{ij}}}(X_i, X_j) = \phi^i_{l_i}(X_i) \times \phi^j_{l_j}(X_j) + \sum_{k=1}^{L} \lambda^i_{k,\boldsymbol{l_{ij}}} \phi^i_k(X_i) + \sum_{k=1}^{L} \lambda^j_{k,\boldsymbol{l_{ij}}} \phi^j_k(X_j) + C_{\boldsymbol{l_{ij}}}, \quad (2.4)$$

with $\boldsymbol{l_{ij}} = (l_i, l_j) \in [1:L]^2$.

The constants $(C_{\boldsymbol{l_{ij}}}, (\lambda^i_{k,\boldsymbol{l_{ij}}})_{k=1}^L, (\lambda^j_{k,\boldsymbol{l_{ij}}})_{k=1}^L)$ are determined by resolving the constraints

$$\begin{aligned} \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^i_k \rangle &= 0, \quad \forall\ k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, \phi^j_k \rangle &= 0, \quad \forall\ k \in [1:L] \\ \langle \phi^{ij}_{\boldsymbol{l_{ij}}}, 1 \rangle &= 0. \end{aligned} \tag{2.5}$$

We first solve the linear system

$$A^{ij} \boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = D^{\boldsymbol{l_{ij}}}, \tag{2.6}$$

with $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}} = {}^t\left(\lambda^i_{1,\boldsymbol{l_{ij}}} \cdots \lambda^i_{L,\boldsymbol{l_{ij}}}\ \lambda^j_{1,\boldsymbol{l_{ij}}} \cdots \lambda^j_{L,\boldsymbol{l_{ij}}}\right)$, and

$$D^{\boldsymbol{l_{ij}}} = -\begin{pmatrix} \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_1 \rangle \\ \vdots \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^i_L \rangle \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_1 \rangle \\ \vdots \\ \langle \phi^i_{l_i} \times \phi^j_{l_j}, \phi^j_L \rangle \end{pmatrix}, A^{ij} = \begin{pmatrix} B^{ii} & B^{ij} \\ {}^tB^{ij} & B^{jj} \end{pmatrix}, \text{ with } B^{ij} = \begin{pmatrix} \langle \phi^i_1, \phi^j_1 \rangle \cdots \langle \phi^i_1, \phi^j_L \rangle \\ \vdots \qquad\qquad \vdots \\ \langle \phi^i_L, \phi^j_1 \rangle \cdots \langle \phi^i_L, \phi^j_L \rangle \end{pmatrix}.$$

As shown in Chastaing, Gamboa, and Prieur (2013), $A^{\boldsymbol{l_{ij}}}$ is a definite positive Gramian matrix and (2.6) provides a unique solution in $\boldsymbol{\lambda}^{\boldsymbol{l_{ij}}}$. Then

$$C_{\boldsymbol{l_{ij}}} = -\mathbb{E}\left[\phi^i_{l_i} \otimes \phi^j_{l_j}(X_i, X_j) + \sum_{k=1}^L \lambda^i_{k,\boldsymbol{l_{ij}}} \phi^i_k(X_i) + \sum_{k=1}^L \lambda^j_{k,\boldsymbol{l_{ij}}} \phi^j_k(X_j)\right]. \tag{2.7}$$

Higher interactions

This construction can be extended to any $|u| \geq 3$, see Chastaing, Gamboa, and Prieur (2013). Just note that the dimension of the approximation space $H_u^{L,0}$ is $L_u = L^{|u|}$, where $|u|$ denotes the cardinality of u.

Empirical procedure

Algorithm 1 proposes an empirical version of the HOGS procedure. It consists in substituting the inner product $\langle \cdot, \cdot \rangle$ with its empirical version $\langle \cdot, \cdot \rangle_{n_1}$ obtained with the first dataset $(y^r, \mathbf{x}^r)_{r=1,\ldots,n_1}$.

Algorithm 1: Empirical HOFD (EHOFD)

Input: Orthonormal system $(\psi_{l_i}^i)_{l_i=0}^L$ of H_i, $i \in [1:p]$, i.i.d. observations $\mathcal{O}_1 := (y^r, \mathbf{x}^r)_{r=1,\dots,n_1}$ of (2.1), threshold $|u_{max}|$

Initialization: for any $i \in [1:p]$ and $l_i \in [1:L]$, let $\hat{\phi}_{l_i,n_1}^i = \psi_{l_i}^i$.

- For any u such that $2 \leq |u| \leq |u_{max}|$, write the matrix $\left(\hat{A}_{n_1}^u\right)$ as well as $\left(\hat{D}_{n_1}^{\boldsymbol{l_u}}\right)$ obtained using the former expressions with $\langle \cdot, \cdot \rangle_{n_1}$.
- Solve (2.6) with the empirical inner product $\langle \cdot, \cdot \rangle_{n_1}$. Compute $\left(\hat{\boldsymbol{\lambda}}_{n_1}^{\boldsymbol{l_{ij}}}\right)$ and $\hat{C}_{\boldsymbol{l_{ij}}}^{n_1}$ with (2.7).
- The empirical version of the basis given by (2.4) is

$$\forall u \in [2:|u_{max}|] \quad \hat{H}_u^{L,0,n_1} = \mathrm{Span}\left\{\hat{\phi}_{1,n_1}^u, \dots, \hat{\phi}_{L^{|u|},n_1}^u\right\}.$$

Step 2: Greedy selection of sparse HOFD

Each component f_u of the HOFD is a projection onto H_u^0. Since for $u \in S^*$, the space $\hat{H}_u^{L,0,n_1}$ well approximates H_u^0, it is natural to approximate f as

$$f(\mathbf{x}) \simeq \bar{f}(\mathbf{x}) = \sum_{\substack{u \in S^* \\ |u| \leq d}} \bar{f}_u(\mathbf{x}_u), \text{ with } \bar{f}_u(\mathbf{x}_u) = \sum_{\boldsymbol{l_u}} \beta_{\boldsymbol{l_u}}^u \hat{\phi}_{\boldsymbol{l_u},n_1}^u(\mathbf{x}_u),$$

where $\boldsymbol{l_u}$ is the multi-index $\boldsymbol{l_u} = (l_i)_{i \in u} \in [1:L]^{|u|}$. Since there is no ambiguity, we omit the summation support of $\boldsymbol{l_u}$ in the sequel.

With the second sample $(y^s, \mathbf{x}^s)_{s=1,\dots,n_2}$, we attempt to recover the unknown coefficients $\left(\beta_{\boldsymbol{l_u}}^u\right)_{\boldsymbol{l_u}, |u| \leq d}$ in the regression $y^s = \bar{f}(\mathbf{x}^s) + \varepsilon^s, \quad s = 1, \dots, n_2$. However, the number of such coefficients is $\sum_{k=1}^d \binom{p}{k} L^k$ and, when p is large, the usual least-squares estimator is not adapted to estimate the coefficients $(\beta_{\boldsymbol{l_u}}^u)_{\boldsymbol{l_u},u}$. We then use the penalized regression

$$(\hat{\beta}_{\boldsymbol{l_u}}^u) \in \underset{\beta_{\boldsymbol{l_u}}^u \in \mathbb{R}}{\mathrm{Argmin}} \frac{1}{n_2} \sum_{s=1}^{n_2} \left[y^s - \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \beta_{\boldsymbol{l_u}}^u \hat{\phi}_{\boldsymbol{l_u},n_1}^u(\mathbf{x}_u^s) \right]^2 + \lambda J(\beta_1^1, \dots, \beta_{\boldsymbol{l_u}}^u, \dots), \quad (2.8)$$

where $J(\cdot)$ is the ℓ_0-penalty

$$J(\beta_1^1, \dots, \beta_{\boldsymbol{l_u}}^u, \dots) = \sum_{\substack{u \in S^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \mathbb{1}(\beta_{\boldsymbol{l_u}}^u \neq 0).$$

Such an optimization procedure is not tractable and, instead, we consider relaxed $\mathbb{L}_2$-boosting (Friedman (2001)) to solve the penalized problem. Mim-

icking the notation of Temlyakov (2000) and Champion et al. (2014), take the dictionary D of functions as

$$\mathcal{D} = \{\hat{\phi}^1_{1,n_1}, \cdots \hat{\phi}^1_{L,n_1}, \ldots, \hat{\phi}^u_{1,n_1}, \ldots, \hat{\phi}^u_{L_u,n_1}, \ldots\}.$$

The quantity $G_k(\bar{f})$ denotes the approximation of $\bar{f}$ at step k as a linear combination of elements of $\mathcal{D}$. The $\mathbb{L}_2$-boosting is described in Algorithm 2, with resulting estimate of $\bar{f}$ denoted by $\hat{f}$.

Algorithm 2: The $\mathbb{L}_2$-boosting

Input: Observations $\mathcal{O}_2 := (y^s, \mathbf{x}^s)_{s=1,\ldots,n_2}$, shrinkage parameter $\gamma \in]0,1]$, number of iterations $k_{up} \in \mathbb{N}^*$.

Initialization: $G_0(\bar{f}) = 0$.

for $k = 1$ *to* k_{up} **do**

1. Select $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \in \mathcal{D}$ such that:

$$\left|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2}\right| = \max_{\hat{\phi}^u_{\boldsymbol{l_u},n_1} \in \mathcal{D}} \left|\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^u_{\boldsymbol{l_u},n_1}\rangle_{n_2}\right|. \quad (2.9)$$

2. Compute the new approximation of $\bar{f}$ as:

$$G_k(\bar{f}) = G_{k-1}(\bar{f}) + \gamma\langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}\rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}. \quad (2.10)$$

end

Output: $\hat{f} = G_{k_{up}}(\bar{f})$.

For any step k, Algorithm 2 selects a function from $\mathcal{D}$ that provides sufficient information about the residual $Y - G_{k-1}(\bar{f})$. The shrinkage parameter γ is the standard step-length parameter of the boosting algorithm. It smoothly inserts the next predictor into the model, making a refinement of the greedy algorithm possible and statistically guaranteeing its convergence rate.

Remark 3. In a deterministic setting, the shrinkage parameter is not really useful and may be set to 1 (see Temlyakov (2000) for further details). It is particularly useful from a practical point of view to smooth the boosting iterations.

An algorithm for our new sparse HOFD procedure

Algorithm 3 provides a simplified description of our sparse HOFD procedure.

Algorithm 3: Greedy Hierarchically Orthogonal Functional Decomposition

Input: Orthonormal system $(\psi_{l_i}^i)_{l_i=0}^L$ of $L^2(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_{X_i})$, $i \in [1:p]$, i.i.d. observations $\mathcal{O} := (y^j, \mathbf{x}^j)_{j=1\dots n}$ of (2.1)

Initialization: Split $\mathcal{O}$ in a partition $\mathcal{O}_1 \cup \mathcal{O}_2$ of size (n_1, n_2).

- For any $u \in S$, use Step 1 with observations $\mathcal{O}_1$ to construct the approximation $\hat{H}_u^{L,0,n_1} := \text{Span}\left\{\hat{\phi}_{1,n_1}^u, \dots, \hat{\phi}_{L_u,n_1}^u\right\}$ of $H_u^{L,0}$ (see Algorithm 1).
- Use an $\mathbb{L}_2$-boosting algorithm on $\mathcal{O}_2$ with the random dictionary $\mathcal{D} = \{\hat{\phi}_{1,n_1}^1, \cdots \hat{\phi}_{L,n_1}^1, \dots, \hat{\phi}_{1,n_1}^u, \dots, \hat{\phi}_{L_u,n_1}^u, \dots\}$ to obtain the Sparse Hierarchically Orthogonal Decomposition (see Algorithm 2).

We turn to the asymptotic properties of the estimators.

3. Consistency of the Estimator

We restrict our study to the case of $d = 2$ and assume that f is well approximated by first and second order interaction components, see Remark 4. The observed signal Y can be represented as

$$Y = \sum_{\substack{u \in S^* \\ |u| \leq 2}} \sum_{\boldsymbol{l_u}} \beta_{\boldsymbol{l_u}}^{u,0} \phi_{\boldsymbol{l_u}}^u(\mathbf{X}_u) + \varepsilon, \quad \mathbb{E}(\varepsilon) = 0,\ \mathbb{E}(\varepsilon^2) = \sigma^2,$$

where $\boldsymbol{\beta}^0 = \left(\beta_{\boldsymbol{l_u}}^{u,0}\right)_{\boldsymbol{l_u},u}$ is the true parameter, and the functions $\left(\phi_{\boldsymbol{l_u}}^u\right)_{\boldsymbol{l_u}}$, $|u| \leq 2$ are constructed according to the HOFD described in Section 2.3. We assume that we have an n-sample of observations divided equally into samples, $\mathcal{O}_1$ and $\mathcal{O}_2$, used for the construction of $\left(\hat{\phi}_{\boldsymbol{l_u},n_1}^u\right)_{\boldsymbol{l_u},u}$ as in Algorithm 1, and for the estimate $\left(\beta_{\boldsymbol{l_u}}^u\right)_{\boldsymbol{l_u},u}$ as in Algorithm 2.

The goal of this section is to study the consistency of $\hat{f} = G_{k_n}(\bar{f})$ when the sample size n tends to infinity, and to determine an optimal number of steps k_n necessary to obtain a consistent estimator from Algorithm 2.

Remark 4. We take $d = 2$ to simplify the presentation, but it can be extended to arbitrary larger thresholds independent of the sample size n. This choice is legitimate if f is well approximated by low interaction components; this assumption is well suited for many practical situations (Rabitz et al. (1999); Sobol (2001)).

3.1. Assumptions

For all sequences $(a_n)_{n\geq 0}$, $(b_n)_{n\geq 0}$, we write $a_n = \underset{n\to+\infty}{\mathcal{O}}(b_n)$ when a_n/b_n is a bounded sequence for large enough n. For any random sequence $(X_n)_{n\geq 0}$, $X_n = \mathcal{O}_P(a_n)$ means that $|X_n/a_n|$ is bounded in probability.

Bounded Assumptions ($\mathbf{H_b}$)

These assumptions concern bounded support for the random variable X. They are collectively referred to as ($\mathbf{H_b}$) and correspond to

($\mathbf{H_b^1}$) $M := \sup_{\substack{i\in[1:p]\\ l_i\in[1:L]}} \left\|\phi_{l_i}^i(X_i)\right\|_\infty < +\infty$;

($\mathbf{H_b^2}$) The number of variables p_n satisfies

$$p_n = \underset{n\to+\infty}{\mathcal{O}}(\exp(Cn^{1-\xi})), \text{ where } 0<\xi\leq 1 \text{ and } C>0;$$

($\mathbf{H_b^{3,\vartheta}}$) The Gram matrices A^{ij} introduced in (2.6) satisfy

$$\exists C>0 \quad \forall (i,j)\in[1:p_n]^2 \qquad det(A^{ij})\geq Cn^{-\vartheta},$$

where *det* denotes determinant.

Regardless of the joint law of the random variables $(X_1,\ldots,X_p)$, it is always possible to build an orthonormal basis $(\phi_{l_i}^i)_{1\leq l_i\leq L}$ from a bounded (frequency truncated) Fourier basis and, therefore, ($\mathbf{H_b^1}$) is not restrictive in practice.

Assumption ($\mathbf{H_b^2}$) deals with the high-dimensional situation, in which, in our study, the number of variables p_n can grow exponentially fast with the number of observations n. The collection of subsets u, designated as S_n^*, also increases rapidly.

It is shown in Chastaing, Gamboa, and Prieur (2013) that each of A^{ij} is invertible, but, if $\vartheta=0$, ($\mathbf{H_b^{3,\vartheta}}$) has this invertibility *uniform* over all choices of (i,j). This hypothesis may be too strong for a large number of variables $p_n\to+\infty$ when $\vartheta=0$. When $\vartheta>0$, verification of ($\mathbf{H_b^{3,\vartheta}}$) requires the computation of an order of p_n^2 determinants of size $L^2\times L^2$. We have checked this assumption in our experiments, but, for very large values of n, this becomes impossible from a numerical point of view.

Noise Assumption ($\mathbf{H_{\varepsilon,q}}$):

$\mathbb{E}(|\varepsilon|^q)<\infty$, for one $q\in\mathbb{R}_+$.

Sparsity Assumption ($\mathbf{H_{s,\alpha}}$):

There exists $\alpha>0$ such that the parameter $\boldsymbol{\beta}^0$ satisfies

$$\|\beta^0\|_{\ell_1} := \sum_{\substack{u\in S_n^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \left|\beta_{\boldsymbol{l_u}}^{u,0}\right| = \underset{n\to+\infty}{\mathcal{O}}(n^\alpha).$$

3.2. Main results

Theorem 1. *Assume* $(\mathbf{H_b})$ *holds with* ξ *(resp.* ϑ*) given by* $(\mathbf{H_b^2})$ *(resp.* $(\mathbf{H_b^{3,\vartheta}})$*), and that a constant* Λ *exists such that* $\left\|\boldsymbol{\lambda}^{\boldsymbol{l}_{ij}}\right\|_2 \leq \Lambda$ *for any couple* (i,j)*. Then, if* $\vartheta < \xi/2$*, the sequence of estimators* $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}\right)_u$ *satisfies*

$$\sup_{\substack{u\in S_n^*,|u|\leq d\\ \boldsymbol{l_u}}} \left\|\hat{\phi}^u_{\boldsymbol{l_u},n_1} - \phi^u_{\boldsymbol{l_u}}\right\| = \zeta_{n,0} = \mathcal{O}_P(n^{\vartheta-\xi/2}).$$

The proof of this theorem can be found in the Supplementary Material.

Proposition 5.1 of Chastaing, Gamboa, and Prieur (2013) finds almost sure convergence of their estimator without any quantitative rate when the number of functions in the HOFD is held fixed as the number of observations goes to infinity n. In our high-dimensional paradigm, we allow S_n^* to grow with n and obtain an almost sure result associated with a convergence rate.

A second result concerns the $\mathbb{L}_2$-boosting that recovers the unknown $\tilde{f}$ up to a preprocessing estimation of $(\hat{\phi}^u_{\boldsymbol{l_u},n_1})_{\boldsymbol{l_u},u}$ on a first sample $\mathcal{O}_1$. Such a result is satisfied provided the sparsity assumption $(\mathbf{H_{s,\alpha}})$ holds. We take

$$Y = \tilde{f}(\mathbf{X}) + \varepsilon, \quad \tilde{f}(\mathbf{X}) = \sum_{\substack{u\in S_n^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \beta^{u,0}_{\boldsymbol{l_u}} \phi^u_{\boldsymbol{l_u}}(\mathbf{X}_u) \in H^L_u,$$

where $\boldsymbol{\beta}^0 = (\beta^{u,0}_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u}$ is the true parameter that expands $\tilde{f}$. Such a high-dimensional inference with noise in the variables appears to be novel.

Theorem 2 (Consistency of the $\mathbb{L}_2$-boosting)**.** *Consider an estimation* $\hat{f}$ *of* $\tilde{f}$ *from an i.i.d. n-sample broken down into* $\mathcal{O}_1 \cup \mathcal{O}_2$*. Assume that functions* $\left(\hat{\phi}^u_{\boldsymbol{l_u},n_1}\right)_{\boldsymbol{l_u},u}$ *are estimated from the first sample* $\mathcal{O}_1$ *under* $(\mathbf{H_b})$ *with* $\vartheta < \xi/2$*, and that a constant* Λ *exists such that* $\left\|\boldsymbol{\lambda}^{\boldsymbol{l}_{ij}}\right\|_2 \leq \Lambda$ *for any couple* (i,j)*.*
If $\hat{f}$ *is defined by* (2.10) *of Algorithm* 2 *on* $\mathcal{O}_2$ *as*

$$\hat{f}(\mathbf{X}) = G_{k_n}(\bar{f}), \quad \textit{with} \quad \bar{f} = \sum_{\substack{u\in S_n^*\\ |u|\leq d}} \sum_{\boldsymbol{l_u}} \beta^{u,0}_{\boldsymbol{l_u}} \hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{X}_u),$$

and if $(\mathbf{H_{s,\alpha}})$ *and* $(\mathbf{H_{\varepsilon,q}})$ *are satisfied with* $q > 4/\xi$ *and* $\alpha < \xi/4 - \vartheta/2$*, then a sequence* $k_n := C \log n$ *exists, where* $C < (\xi/2 - \vartheta - 2\alpha)/2 \cdot \log 3$*, such that*

$$\left\|\hat{f} - \tilde{f}\right\| \xrightarrow{\mathbb{P}} 0, \textit{when} \quad n \to +\infty.$$

In particular, for Gaussian noise, the constraint on q disappears and Theorem 2 can be applied as soon as $\xi < 1$.

Our result is a result in probability rather than in expectation. It is a frequently encountered fact that SOI in expectation is derived with additional assumptions on the coherence of the dictionary. With some coherence and boundedness assumptions, Bickel, Ritov, and Tsybakov (2009) deduced convergence rates of the Lasso estimator, in expectation, as soon as

$$\|\beta^0\|_{\ell_0} \frac{\log(p)}{n} \longrightarrow 0. \tag{3.1}$$

Rigollet and Tsybakov (2011) extended the study of the Lasso behavior with a result on the Lasso estimator on bounded variables without any coherence assumption, and showed consistency in probability when

$$\|\beta^0\|_{\ell_1} \sqrt{\frac{\log(p)}{n}} \longrightarrow 0. \tag{3.2}$$

Champion et al. (2014) also obtained consistency results in probability under the asymptotic setting given for (3.2) without a coherence assumption. Our results with a noisy dictionary require that

$$\left(\inf_{i,j} \det(A^{ij})\right)^{-1} \|\beta^0\|_{\ell_1}^2 \sqrt{\frac{\log p}{n}} \longrightarrow 0 \text{ as } n \longrightarrow +\infty, \tag{3.3}$$

which is stronger than (3.2).

When all linear systems defined through the Gram matrices A^{ij} are well conditioned, $\vartheta = 0$ and the condition becomes $\|\beta^0\|_{\ell_1}^2 \sqrt{\log p/n} \longrightarrow 0$; there is still a price to pay for the preliminary estimation of the elements of the HOGS. Theorem 2 can be applied only for sequences of coefficients such that $\|\beta^{u,0}_{\boldsymbol{l_u}}\|_{L_1} \lesssim n^{1/4}$. Note also that the degeneracy of the Gram determinants must be strictly larger than $n^{-1/2}$. For example, when $\vartheta = 1/4$, the norm $\|\beta^{u,0}_{\boldsymbol{l_u}}\|_{L_1}$ cannot be larger than $n^{1/8}$.

Sketch of Proof of Theorem 2. Mimicking the scheme of Bühlmann (2006) and Champion et al. (2014), the proof consists in defining the theoretical residual of Algorithm 2 at step k as

$$\begin{aligned} R_k(\bar{f}) &= \bar{f} - G_k(\bar{f}) \\ &= \bar{f} - G_{k-1}(\bar{f}) - \gamma \langle Y - G_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle_{n_2} \cdot \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}. \end{aligned} \tag{3.4}$$

Following the work of Champion et al. (2014), we introduce a *phantom* residual in order to reproduce the behavior of a deterministic boosting as in Temlyakov

(2000). This *phantom* algorithm is the theoretical $\mathbb{L}_2$-boosting, performed using randomly chosen elements of the dictionary by (2.9) and (2.10), but updated using the deterministic inner product. The *phantom* residuals $\tilde{R}_k(\bar{f})$, $k \geq 0$, are

$$\begin{cases} \tilde{R}_0(\bar{f}) = \bar{f}, \\ \tilde{R}_k(\bar{f}) = \tilde{R}_{k-1}(\bar{f}) - \gamma \langle \tilde{R}_{k-1}(\bar{f}), \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1} \rangle \hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}, \end{cases} \tag{3.5}$$

where $\hat{\phi}^{u_k}_{\boldsymbol{l_{u_k}},n_1}$ has been selected with (2.9). The aim is to decompose the quantity $\left\| \hat{f} - \tilde{f} \right\|$ to introduce the theoretical residuals and the *phantom* ones:

$$\left\| \hat{f} - \tilde{f} \right\| = \left\| G_{k_n}(\bar{f}) - \tilde{f} \right\| \leq \left\| \bar{f} - \tilde{f} \right\| + \left\| R_{k_n}(\bar{f}) - \tilde{R}_{k_n}(\bar{f}) \right\| + \left\| \tilde{R}_{k_n}(\bar{f}) \right\|. \tag{3.6}$$

We then have to show that each term on the right-hand side of (3.6) converges to zero in probability. For further details, see the Supplementary Material.

4. Numerical Applications

This section is devoted to the numerical efficiency of the two-step procedure of Section 2, and primarily focuses on the practical use of the HOFD through sensitivity analysis (SA). SA aims to identify the most contributive variables to the variability of a regression model (Saltelli, Chan, and Scott (2000); Cacuci, Ionescu-Bujor, and Navon (2005)). The most common quantification of this is the variance-based Sobol index (Sobol (1993)). This measure relies on the Hoeffding decomposition that provides an elegant and meaningful framework when inputs are known to be independent. However, it may be irrelevant when strong dependencies arise. The HOFD provides a general and rigorous multivariate regression extension that can be used to define sensitivity indices well-tailored to dependent inputs. As detailed in Chastaing, Gamboa, and Prieur (2013), the model variance can be expanded as

$$V(Y) = \sum_{u \in S_n^*} \left[V(f_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \mathrm{Cov}(f_u(\mathbf{X}_u), f_v(\mathbf{X}_v)) \right].$$

To measure the contribution of $\mathbf{X}_u$, for $|u| \geq 1$, in terms of model variability, it is then natural to define a sensitivity index S_u as

$$S_u = \frac{V(f_u(\mathbf{X}_u)) + \sum_{u \cap v \neq u,v} \mathrm{Cov}(f_u(\mathbf{X}_u), f_v(\mathbf{X}_v))}{V(Y)}. \tag{4.1}$$

We deduce the empirical estimation of (4.1) once we have applied the procedure described in Algorithm 3 to obtain $(\hat{f}_u, \hat{f}_v, u \cap v \neq u, v)$.

4.1. Description

We have done a short simulation study focused primarily on the performance of the greedy selection algorithm for the prediction of generalized sensitivity indices. Since the estimation of these indices consists in estimating the summands of the HOFD, we begin by constructing a hierarchically orthogonal system of functions to approximate the components. The invertibility of each linear system plays an important role in our theoretical study. For each situation, we therefore measured the degeneracy of the matrices involved through $d(A) = \inf_{i,j\in[1:p]} \det(A^{ij})$.

We used a variable selection method to select a sparse number of predictors. The goal was to numerically compare $\mathbb{L}_2$-boosting, the Forward-Backward greedy algorithm (referred to as FoBa below), and the Lasso estimator. We considered n-samples of i.i.d. observations $(y^s, \mathbf{x}^s)_{s=1,\dots,n}$ broken down into two samples of size $n_1 = n_2 = n/2$, first used to construct the system of functions according to Algorithm 1, second sample used to solve

$$(\hat{\beta}^u_{\boldsymbol{l_u}})_{\boldsymbol{l_u},u} \in \underset{\beta^u_{\boldsymbol{l_u}}\in\mathbb{R}}{\operatorname{Argmin}} \frac{1}{n_2}\sum_{s=1}^{n_2}\left[y^s - \sum_{\substack{u\in S\\ |u|\le d}}\sum_{\boldsymbol{l_u}} \beta^u_{\boldsymbol{l_u}}\hat{\phi}^u_{\boldsymbol{l_u},n_1}(\mathbf{x}^s_u)\right]^2 + \lambda J(\beta^1_1,\dots,\beta^u_{\boldsymbol{l_u}},\dots).$$

We briefly describe how we used the Lasso, the FoBa and the Boosting.

4.2. Feature selection algorithms

FoBa procedure

The FoBa algorithm, as well as the $\mathbb{L}_2$-boosting, uses a greedy exploration to minimize the previous criterion when $J(\cdot)$ is a ℓ_0 penalty,

$$J(\beta^1_1,\dots,\beta^u_{\boldsymbol{l_u}},\dots) = \sum_{\substack{u\in S^*_n\\ |u|\le d}}\sum_{\boldsymbol{l_u}} \mathbb{1}(\beta^u_{\boldsymbol{l_u}} \neq 0).$$

This algorithm is an iterative scheme that sequentially selects or deletes an element of $\mathcal{D}$ that has the least impact on the model residual. It is described in Zhang (2011), and used for HOFD in Chastaing, Gamboa, and Prieur (2013). The procedure depends on the shrinkage parameters, ϵ and δ: ϵ is the stopping criterion that predetermines if a large number of predictors is going to be introduced into the model; $\delta \in]0,1]$ offers a flexibility in the *backward* step since it allows the algorithm to smoothly eliminate a predictor at each step.

In our numerical experiments, we found a well-suited behavior of the FoBa procedure with $\epsilon = 10^{-2}$ and $\delta = 1/2$.

Calibration of the boosting

We fixed the shrinkage parameter to $\gamma = 0.7$ since it provides a suitable value for high dimensional regression, though we did not find any extreme differences when γ varies in $[0.5; 1[$. Since the optimal value for k_{up} is unknown in practice, we used a C_p Mallows-type criterion to fix the optimal number of iterations. This stopping criterion is much more important than the choice of the shrinkage parameter. It is induced by γ since it depends on the sequence of the boosting iterations.

As with the LARS algorithm, we followed the recommendations of Efron et al. (2004) to select the best solution. First, we took a large number of iterations, say K. For each step $k \in \{1, \ldots, K\}$, the boosting algorithm computes an estimation of the solution $\hat{\boldsymbol{\beta}}(k)$. On the basis of this, we computed

$$E_k^{\text{Boost}} = \frac{1}{n} \sum_{s=1}^{n_2} \left[y^s - \sum_{\hat{\phi}^u_{\boldsymbol{l_u}, n_1} \in \mathcal{D}} \hat{\beta}^u_{\boldsymbol{l_u}}(k) \hat{\phi}^u_{\boldsymbol{l_u}, n_1}(\mathbf{x}^s_u) \right]^2 - n_2 + 2k,$$

where the implied set of functions $\hat{\phi}^u_{\boldsymbol{l_u}, n_1}$ has been selected through the first k steps of the algorithm. We then chose the optimal number of selected functions

$$\hat{k}_{\text{up}} = \underset{k=1,\ldots,K}{\text{Argmin}}\, E_k^{\text{Boost}}.$$

Lasso algorithm

The ℓ_0 penalty is often replaced by the $\lambda \times \ell_1$ strategy that yields the Lasso estimator for a given penalization parameter $\lambda > 0$

$$J(\beta_1^1, \ldots, \beta^u_{\boldsymbol{l_u}}, \ldots) = \sum_{\substack{u \in S_n^* \\ |u| \leq d}} \sum_{\boldsymbol{l_u}} \left| \beta^u_{\boldsymbol{l_u}} \neq 0 \right|.$$

Several algorithms have been proposed to solve the Lasso regression. One of the most popular is the LARS method, but it is very expensive in large Lasso problems. To make a good numerical comparison with the greedy algorithms, we chose to perform a coordinate descent algorithm proposed by Fu (1998) and Friedman et al. (2007) because of its low computational cost compared to the LARS implementation. The tuning parameter λ was first selected by generalized cross-validation, and the Lasso Coordinate Descent (LCD) algorithm was performed with the R `lassoshooting` package.

4.3. Datasets

Each experiment was randomly reproduced 50 times to compute Monte-Carlo errors. Since each dataset has very few instances, the size L of the initial

orthonormal systems is limited. Here we arbitrarily chose $5 \leq L \leq 8$; the approximation performance did not suffer from sensitivity to L in these models.

First dataset: the Ishigami function

Well known in sensitivity analysis, the Ishigami model is

$$Y = \sin(X_1) + a\sin^2(X_2) + bX_3^4 \sin(X_1),$$

where we set $a = 7$ and $b = 0.1$, and where it is assumed that the inputs are independent. We considered the following cases. First, for all $i = 1, 2, 3$, inputs were uniformly distributed on $[-\pi, \pi]$, and we chose $n = 300$ observations, with the first eight Legendre basis functions ($L = 8$). For the second case, for all $i = 1, 2, 3$, inputs were uniformly distributed on $[-\pi, \pi]$, and we chose $n = 300$ observations, with the first eight Fourier basis functions. In each instance, the number of predictors was $m_n = pL + \binom{p}{2}L^2 = 408 \geq n$.

Second Dataset: the g-Sobol function

Referred to in Saltelli, Chan, and Scott (2000), the function is

$$Y = \prod_{i=1}^{p} \frac{|4X_i - 2| + a_i}{1 + a_i}, \quad a_i \geq 0,$$

where the inputs X_i are independent and uniformly distributed over $[0, 1]$. The analytical Sobol indices are given by:

$$S_u = \frac{1}{D}\prod_{i \in u} D_i, \quad D_i = \frac{1}{3(1 + a_i)^2}, \quad D = \prod_{i=1}^{p}(D_i + 1) - 1, \ \forall\ u \subseteq [1 : p].$$

Here, we took $p = 25$ and $a = (0, 0, 0, 1, 1, 2, 3, 4.5, 4.5, 4.5, 9, 9, 9, 9, 9, 99, \ldots, 99)$. For the construction of the hierarchical basis functions, we chose the first five Legendre polynomials ($L = 5$). We used $n = 2{,}000$ evaluations of the model and the number of predictors $m_n = pL + \binom{p}{2}L^2 = 7625$, exceeding the sample size n.

Third dataset: dependent inputs

As proposed by Mara and Tarantola (2012), we generated a sample set with X_1 and X_2 uniformly sampled in the set

$$\mathcal{S} := \left\{(x_1, x_2) \in [-1, 1]^2 \,|\, 2x_1^2 - 1 \leq x_2 \leq 2x_1^2\right\}.$$

X_3 was sampled uniformly in $[-1; 1]$; Y was taken as $Y = X_1 + X_2 + X_3$. The inputs X_1 and X_2 are not independent and we do not know the analytical Sobol indices. We chose $n = 100$ observations, with the first six Legendre basis functions ($L = 6$).

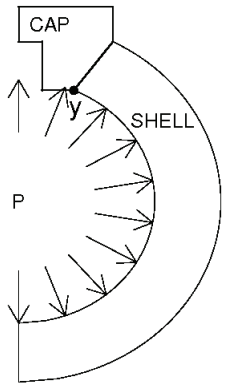


Figure 1. Tank distortion at point y.

4.4. The tank pressure model

This study concerns a shell closed by a cap and subject to internal pressure. Figure 1 illustrates a tank distortion. We are interested in the von Mises stress (von Mises (1913)) on the point y. The von Mises stress makes it possible to predict material yielding that occurs when the material yield strength is reached. The selected point y corresponds to the point for which the von Mises stress is maximal in the tank. One wants to prevent the tank from material damage induced by plastic deformations. To provide a large panel of tanks able to resist the internal pressure, a manufacturer wants to know the parameters that contribute the most to the von Mises criterion variability. In our model, the von Mises criterion depends on the shell internal radius (R_{int}), the shell thickness (T_{shell}), and the cap thickness (T_{cap}). It also depends on physical parameters concerning Young's modulus (E_{shell} and E_{cap}) and the yield strength ($\sigma_{y,shell}$ and $\sigma_{y,cap}$) of the shell and the cap. A last parameter is the internal pressure (P_{int}) applied to the shell. Some strong correlations exist between some of the inputs in the system as a result of the constraints of manufacturing processes, for example, between the shell radius and its thickness. The system is modeled by a 2D finite element ASTER code. Input distributions are provided in Table 1.

The geometrical parameters were taken as uniformly distributed because of the large choice left for tank construction. The correlation γ between them is induced by the constraints linked to manufacturing processes. The physical inputs were taken as normally distributed, the uncertainty due to the manufacturing process and the properties of the elementary constituent variabilities. The large variability of P_{int} in the model corresponds to the different internal pressure values that could be applied to the shell by the user.

Table 1. Inputs of the shell model.

Inputs	Distribution
R_{int}	$\mathcal{U}([1800; 2200])$, $\gamma(R_{int}, T_{shell}) = 0.85$
T_{shell}	$\mathcal{U}([360; 440])$, $\gamma(T_{shell}, T_{cap}) = 0.3$
T_{cap}	$\mathcal{U}([180; 220])$, $\gamma(T_{cap}, R_{int}) = 0.3$
E_{cap}	$\alpha N(\mu, \Sigma) + (1-\alpha)N(\mu, \Omega)$
$\sigma_{y,cap}$	$\alpha = 0.02$, $\mu = \begin{pmatrix} 210 \\ 500 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 350 & 0 \\ 0 & 29 \end{pmatrix}$, $\Omega = \begin{pmatrix} 175 & 81 \\ 81 & 417 \end{pmatrix}$
E_{shell}	$\alpha N(\mu, \Sigma) + (1-\alpha)N(\mu, \Omega)$
$\sigma_{y,shell}$	$\alpha = 0.02$, $\mu = \begin{pmatrix} 70 \\ 300 \end{pmatrix}$, $\Sigma = \begin{pmatrix} 117 & 0 \\ 0 & 500 \end{pmatrix}$, $\Omega = \begin{pmatrix} 58 & 37 \\ 37 & 250 \end{pmatrix}$
P_{int}	$N(80, 10)$

To measure the contribution of the correlated inputs to the output variability, we estimated the generalized sensitivity indices. We did $n = 1{,}000$ simulations, using the first Hermite basis functions, whose maximum degree is 5 for every parameter.

4.5. Results

We considered the estimation of the sensitivity indices, the ability to select the good representation of the different signals, and the computation time needed to obtain the sparse representation. "Greedy" refers to the Foba procedure and "LCD" refers to the Lasso coordinate descent method. Our method is referred to as "Boosting".

Sensitivity estimation

Figures 2 and 3 provide the dispersion of the sensitivity indices estimated by our three methods on the Ishigami function. We can see that the three methods behave well with the two basis functions. Handling the Fourier basis is, as expected, more suitable for the Ishigami function than the Legendre basis (see the sensitivity index S_3 in Figures 2 and 3). For the sake of clarity, Figure 4 only represents the first ten sensitivity indices. We can also draw similar conclusions with Figure 4, where the three methods lead to the same conclusion. The standard deviations of each method seem to be relatively equivalent. Figure 5 represents the estimated sensitivity indices when the inputs are correlated. The analytical results are obviously unknown, but we obtain similar results for the three methods.

As illustrated in Figure 6, the most contributive parameter to the von Mises criterion variability is the internal pressure P_{int}, which is not surprising. Concerning the geometric characteristics, the main parameters of the three methods are cap thickness, T_{cap} and shell thickness, T_{shell}, using their expensive code, although the shell internal radius does not seem to be that important.

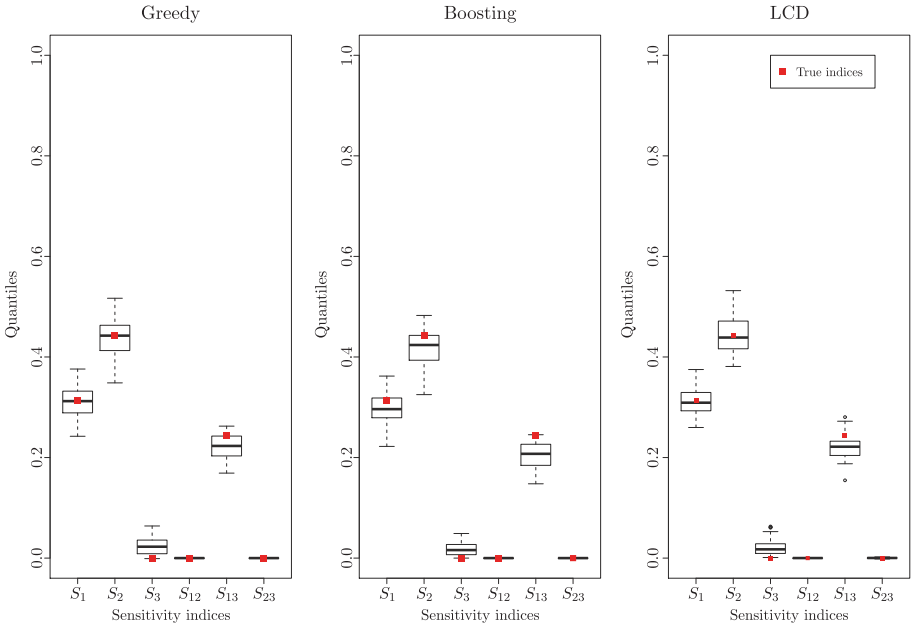


Figure 2. Representation of the first-order components on the first dataset (Ishigami function) described through the Legendre basis.

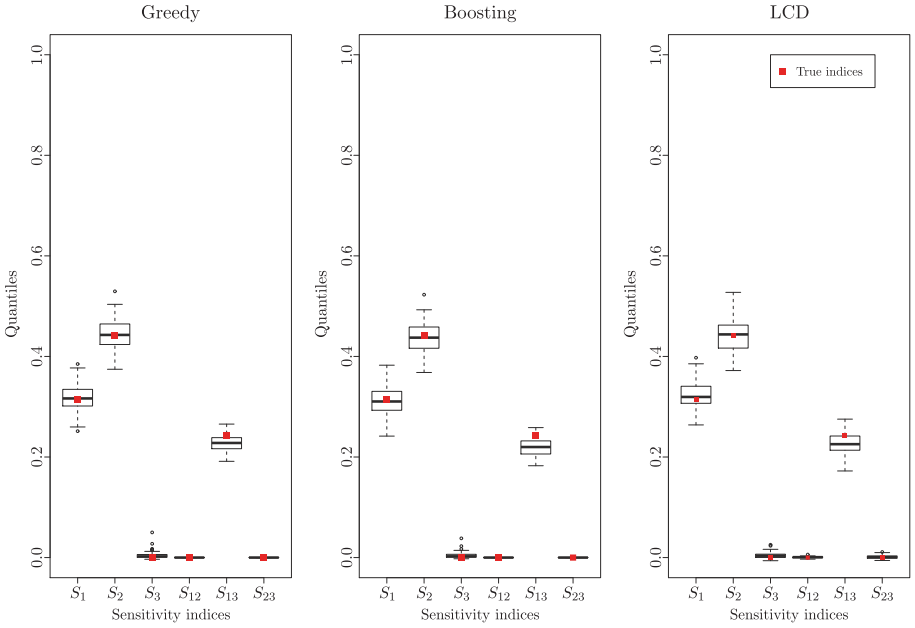


Figure 3. Representation of the first-order components on the first dataset (Ishigami function) described through the Fourier basis.

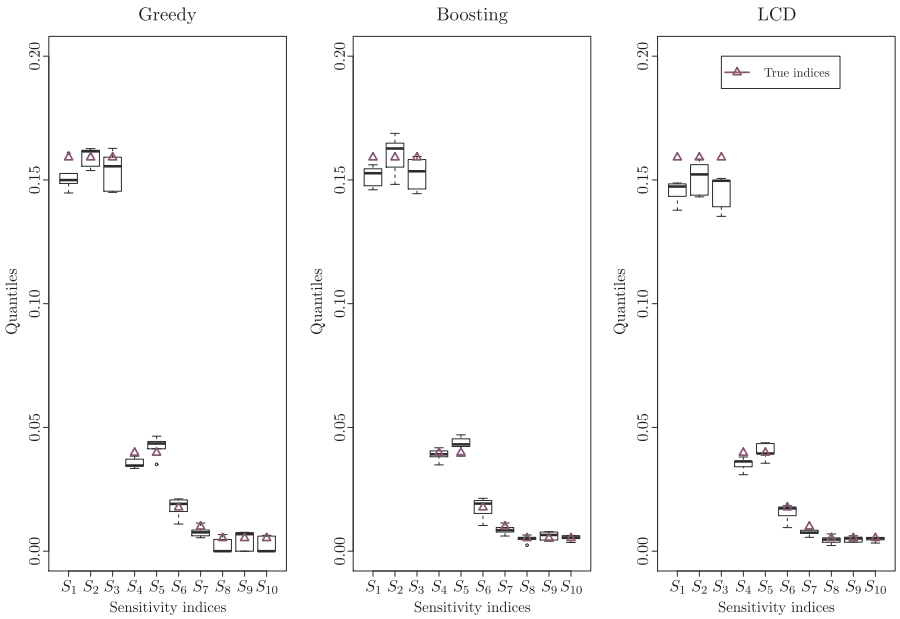


Figure 4. Representation of the first-order components on the second dataset (g-Sobol function).

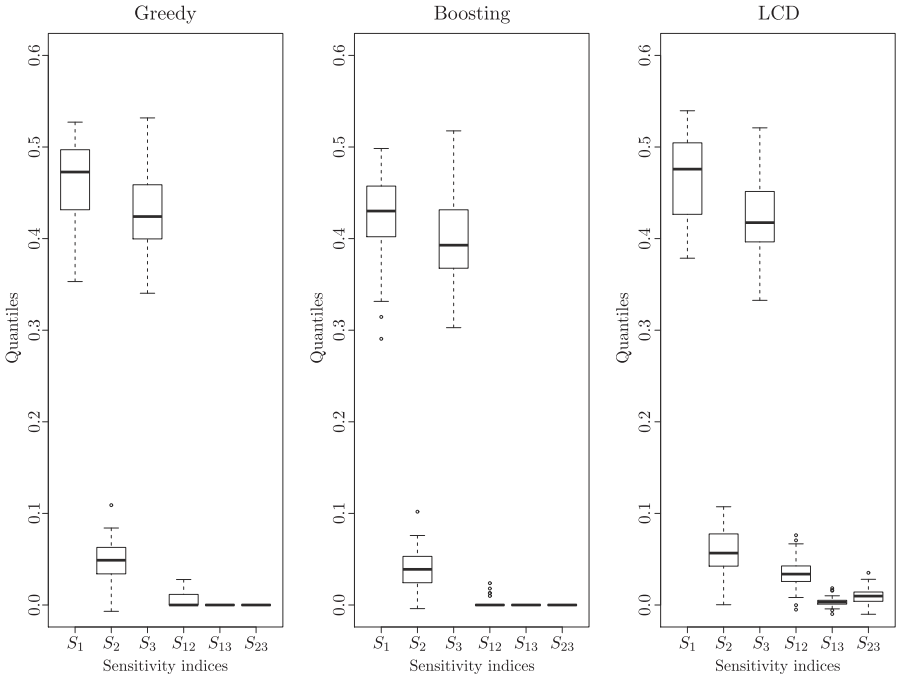
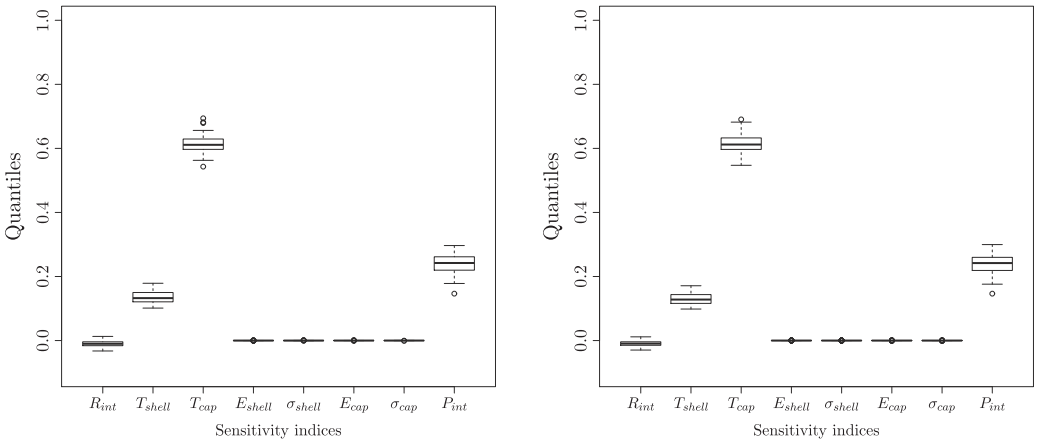
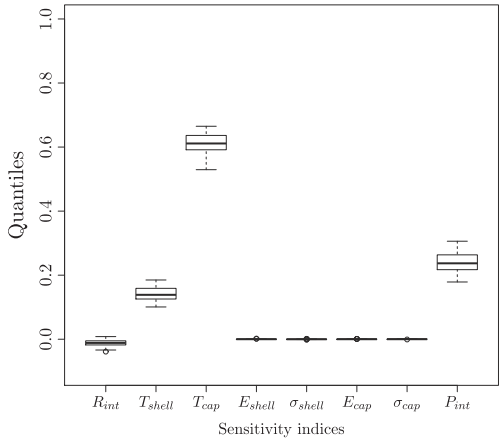


Figure 5. Representation of the first-order components on the third dataset (dependent inputs).



(a) Greedy (b) Boosting



(c) LCD

Figure 6. Dispersion of the first order sensitivity indices of the tank model parameters for the three methods.

Computation time and accuracy

The performances of the three methods are illustrated in Table 2 on the basis of their computational cost and the accuracy of the feature selection.

Regarding statistical accuracy, each estimator of high-dimensional regression possesses a comparable dispersion on all the datasets, and performs quite similarly on the first dataset. The Lasso estimator seems a little bit imprecise in the third dataset in comparison with the FoBa and Boosting methods. The LCD method is also outperformed on the third dataset (with dependent inputs): it selects a significantly larger number of sensitivity indices in comparison with the Boosting and FoBa methods (for example, the indices S_{13} and S_{23} are certainly equal to 0 as a result of the definition of Y). This may be due to the influence

Table 2. Features of the three algorithms.

Dataset	Procedure	$\left\|\hat{\boldsymbol{\beta}}\right\|_0$	Elapsed Time (in sec.)
Ishigami function Case 1	$\mathbb{L}_2$-boosting	19	0.0941
	FoBa	21	2.2917
	LCD	20	2.25
Ishigami function Case 2	$\mathbb{L}_2$-boosting	15	0.0884
	FoBa	12	1.0752
	LCD	13.9	0.41
g-Sobol function	$\mathbb{L}_2$-boosting	99	49.8
	FoBa	22.4	827.9
	LCD	91.8	5047.4
Dependent inputs	$\mathbb{L}_2$-boosting	4.14	0.028
	FoBa	4.76	0.1056
	LCD	24.1	0.061
Tank pressure model	$\mathbb{L}_2$-boosting	10	0.0266
	FoBa	22	0.3741
	LCD	23	0.15

of the dependence, among the inputs X_1 and X_2 in this dataset, on the Lasso estimator.

In Table 2, our proposed $\mathbb{L}_2$-boosting is the fastest method. This is particularly true on the 25-dimension g-Sobol function, where the fraction of additional time required by the LCD algorithm in comparison to the $\mathbb{L}_2$-boosting is about 100. Although we do not have access to the theoretical support recovery $\|\boldsymbol{\beta}\|_0$, we can observe that the results of the $\mathbb{L}_2$-boosting are equivalent to those of other algorithms in terms of its feature selection ability. Hence, for the same degree of accuracy, our method seems to be much faster.

We have computed the maximal "degeneracy" involved in the resolution of the linear systems, and quantified by Assumption $(\mathbf{H}_{\mathbf{b}}^{\mathbf{3},\vartheta})$, in column 2 of Table 3. In many cases, we obtain a significantly larger value than 0. The third column of Table 3 shows the admissible size of the parameter ϑ, and we can check that the number of variables p_n allowed by $(\mathbf{H}_{\mathbf{b}}^{\mathbf{2}})$ and the balance between ξ and ϑ (ξ should be greater than 2ϑ in our theoretical results) is not restrictive since $n^{1-2\vartheta}$ is always significantly greater than $log(m_n)$ in Table 3.

5. Conclusions and Perspectives

This paper provides a rigorous framework for the hierarchically orthogonal Gram-Schmidt procedure in a high-dimensional paradigm, with the use of the greedy $\mathbb{L}_2$-boosting. Overall, the procedure falls into the category of sparse estimation with a noisy dictionary, and we demonstrate its consistency up to some

Table 3. Degeneracy of the linear systems and admissible size of m_n ($n^{1-2\vartheta}$ should be greater than $\log(m_n)$.

Dataset	Degeneracy $d(A)$	$\vartheta \geq \frac{\log(1/d(A))}{log(n)}$	$n^{1-2\vartheta}$	$\log(m_n)$
Ishigami function Case 1	0.6388	$[0.0786, +\infty[$	122.3821	6.0113
Ishigami function Case 2	0.76	$[0.0481, +\infty[$	173.3094	6.0113
g-Sobol function	0.9745	$[0.0034, +\infty[$	1899	8.9392
Dependent inputs	0.628	$[0.101, +\infty[$	39.4457	4.8363

mild assumptions about the structure of the underlying basis. From a mathematical point of view, assumption ($\mathbf{H_b^1}$) presents a restrictive condition, and to relax it would open a wider class of basis functions for applications. We leave this development open for a future study, which could be based either on the development of a concentration inequality for unbounded random matrices or on a truncating argument. It also appears that our algorithm produces very satisfactory numerical results through our three datasets as a result of its very low computational cost. It can also be extended with some further numerical work to a larger truncation order of $d \geq 3$. Such an improvement may also be of interest from a theoretical point of view when dealing with a function that smoothly depends on the interaction order. In particular, a data-driven adaptive choice of d may be of practical interest in the future.

Acknowledgement

The authors are indebted to Fabrice Gamboa for his stimulating discussions and his numerous suggestions on the subject. The authors also thank the two anonymous referees and the associate editor for their helpful comments and suggestions on an earlier version of the manuscript.

References

Bickel, P.J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and Dantzig selector. *Ann. Statist.* **37**, 1705-1732.

Blatman, G. (2009). Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis. PhD thesis, Université BLAISE PASCAL - Clermont II.

Bühlmann, P. (2006). Boosting for high-dimensional linear models. *Ann. Statist.* **34**, 559-583.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, Berlin.

Cacuci, D. G., Ionescu-Bujor, M. and Navon, I. M. (2005). *Sensitivity and Uncertainty Analysis, Volume II: Applications to Large-Scale Systems*. Chapman and Hall/CRC.

Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**, 2313-2351.

Cavalier, L. and Hengartner, N. W. (2005). Adaptive estimation for inverse problems with noisy operators. *Inverse Problems* **21**, 1345-1361.

Champion, M., Cierco-Ayrolles, C., Gadat, S. and Vignes, M. (2014). Sparse regression and support recovery with $\mathbb{L}_2$-boosting algorithm. *J. Statist. Plann. Inference* **155**, 19-41.

Chastaing, G., Gamboa, F. and Prieur, C. (2012). Generalized Hoeffding-Sobol decomposition for dependent variables - Application to sensitivity analysis. *Electronic J. Statist.* **6**, 2420-2448.

Chastaing, G., Gamboa, F. and Prieur, C. (2013). Generalized sobol sensitivity indices for dependent variables: Numerical methods. Available at `http://arxiv.org/abs/1303.4372`.

Crestaux, T., Le Maïtre, O. and Martinez, J. (2009). Polynomial chaos expansion for sensitivity analysis. *Reliability Engineering and System Safety* **94**, 1161-1172.

Da Veiga, S., Wahl, F. and Gamboa, F. (2009). Local polynomial estimation for sensitivity analysis on models with correlated inputs. *Technometrics* **51**, 452-463.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004). Least angle regression. *Ann. Statist.* **32**, 407-451.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Ann. Statist.* **29**, 1189-1232.

Friedman, J., Hastie, T., H"ofling, H. and Tibshirani, R. (2007). Pathwise coordinate optimization. *Ann. Appl. Statist.* **1**, 302-332.

Fu, W. (1998). Penalized regressions: the bridge versus the lasso. *J. Comput. Graph. Statist.* **7**, 397-416.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19**, 293-325.

Hooker, G. (2007). Generalized functional anova diagnostics for high-dimensional functions of dependent variables. *J. Comput. Graph. Statist.* **16**, 709-732.

Huang, J. (1998). Projection estimation in multiple regression with application to functional anova models. *Ann. Statist.* **26**, 242-272.

Jacques, J., Lavergne, C. and Devictor, N. (2006). Sensitivity analysis in presence of model uncertainty and correlated inputs. *Reliability Engineering and System Safety* **91**, 1126-1134.

Li, G. and Rabitz, H. (2012). D-morph regression: application to modeling with unknown parameters more than observation data. *J. Math. Chemistry* **48**, 1010-1035.

Li, G., Rabitz, H., Yelvington, P., Oluwole, O., Bacon, F., Kolb, C. E. and Schoendorf, J. (2010). Global sensitivity analysis with independent and/or correlated inputs. *J. Phys. Chemistry A* **114**, 6022-6032.

Mara, T. and Tarantola, S. (2012). Variance-based sensitivity analysis of computer models with dependent inputs. *Reliability Engineering and System Safety* **107**, 115-121.

Rabitz, H., Ali, O., Shorter, J. and Shim, K. (1999). Efficient input-output model representations. *Computer Phys. Comm.* **117**, 11-20.

Rigollet, P. and Tsybakov, A. (2011). Exponential screening and optimal rates of sparse estimation. *Ann. Statist.* **39**, 731-771.

Saltelli, A., Chan, K. and Scott, E. M. (2000). *Sensitivity Analysis*. Wiley, West Sussex.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Model. Comput. Experiment* **1**, 407-414.

Sobol, I. M. (2001). Global sensitivity indices for nonlinear mathematical models and their monte carlo estimates. *Math. Comput. Simulations* **55**, 271-280.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.* **22**, 118-171.

Temlyakov, V. N. (2000). Weak greedy algorithms. *Adv. Comput. Math.* **12**, 213-227.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc.* **58**, 267-288.

von Mises, R. (1913). Mechanik der festen körper im plastisch deformablen zustand. *Göttin. Nachr. Math. Phys.* **1**, 582-592.

Wang, X. and Fang, K. (2003). The effective dimension and quasi-Monte Carlo integration. *J. Complexity* **19**, 101-124.

Zhang, T. (2011). Adaptive forward-backward algorithm for learning sparse representations. *IEEE Trans. Inform. Theory* **57**, 4689-4708.

Zuniga, M., Kucherenko, S. and Shah, N. (2013). Metamodelling with independent and dependent inputs. *Computer Phys. Comm.* **184**, 1570-1580.

Institut de Mathématiques de Toulouse, Université Paul Sabatier, 118, route de Narbonne F-31062 Toulouse Cedex 9, France.

E-mail: magali.champion@math.univ-toulouse.fr

EDF R&D, Batiment S, 6 quai Watier, 78400 Chatou, France.

E-mail: gal.chastaing@gmail.com

GREMAQ, Toulouse School of Economics, Université Toulouse I Capitole, 21 allées de Brienne, F-31000 Toulouse, France.

E-mail: sebastien.gadat@tse-fr.eu

Université Joseph Fourier, LJK/MOISE BP 53, 38041 Grenoble Cedex, France.

E-mail: clementine.prieur@imag.fr

(Received October 2013; accepted September 2014)