Indices de Sobol généralisés aux variables dépendantes : tests de performance de l'algorithme HOGS couplé à plusieurs estimateurs paramétriques

Julien Sainte-Marie, Gautier Viaud, Paul-Henry Cournède

Journal de la Société Française de Statistique
Vol. 158 No. 1 (2017)

Indices de Sobol généralisés aux variables dépendantes : tests de performance de l'algorithme HOGS couplé à plusieurs estimateurs paramétriques

Title: Performances de l'algorithme HOGS

Julien Sainte-Marie[1], Gautier Viaud[1] et Paul-Henry Cournède[1]

Résumé : L'algorithme "Hierarchically Orthogonal Gram–Schmidt" (HOGS) (Chastaing *et al.*, 2015) estime des indices de Sobol généralisés aux modèles à entrées dépendantes, quantifiant explicitement la sensibilité du modèle due aux corrélations. HOGS construit un méta-modèle pour chaque variable d'intérêt par projection sur une base fonctionnelle bien choisie pour le calcul des indices. Les coefficients de projection sont obtenus par l'estimateur des moindres carrés (OLS) ou les régressions pénalisées lasso, ridge et Elastic Net (EN). Quatre cas d'étude sont proposés : trois modèles simples permettent d'appréhender le fonctionnement de HOGS, et le modèle LNAS (Log-Normal Allocation and Senescence) dédié à la dynamique complexe de la croissance des plantes. Plusieurs configurations de HOGS et la pertinence du méta-modèle sont étudiées grâce à un indice de consistance. L'interprétation des indices de Sobol est illustrée grâce à LNAS. En conclusion, HOGS-OLS est la méthode la plus performante lorsque les ressources informatiques ne sont pas limitantes. Dans le cas contraire, la question de l'estimation paramétrique avec hypothèse de sparsité met en évidence que : i) EN est plus robuste mais plus coûteux numériquement que le Lasso, ii) HOGS génère une base trop grande, créant de la sparsité artificielle. Un amendement de HOGS a été proposé pour réduire la dimension de la base.

Abstract: The algorithm "Hierarchically Orthogonal Gram-Schmidt" (HOGS) (Chastaing *et al.*, 2015) estimates generalized Sobol indices dedicated to models with dependent inputs, quantifying explicitly the model sensitivity due to correlations. HOGS constructs a meta-model for each variable of interest by projection on a functional basis suited to indices calculation. Regression coefficients are obtained with the ordinary least-square estimator (OLS) or penalized regression methods Lasso, Ridge and Elastic Net (EN). Four study cases are proposed: three toy models allowing to investigate HOGS functioning and numerical properties, and the LNAS (Log-Normal Allocation and Senescence) model dedicated to the complex dynamics of plant growth. Several HOGS configurations and meta-model accuracy are tested by means of a consistency index. An interpretation of Sobol indices is given for LNAS. It appears that HOGS-OLS is the most efficient method when simulation resources are not limited. Otherwise, considering the issue of parameter estimation with sparsity highlights that: i) EN is more robust but more costly than Lasso, ii) the basis constructed by HOGS is too large which creates artificial sparsity. A modification of HOGS has been proposed to reduce the dimension of the basis.

Mots-clés : analyse de sensibilité, HDMR, HOGS, régression pénalisée

Keywords: sensitivity analysis, HDMR, HOGS, penalized regression

Classification AMS 2000 : 49Q12

[1] Laboratoire de Mathématiques et Informatique pour la Complexité et les Systèmes, CentraleSupélec, Université Paris-Saclay
E-mail : juliensaintemarie@gmail.com, E-mail : gautier.viaud@centralesupelec.fr et
E-mail : paul-henry.cournede@centralesupelec.fr

 ISSN: 2102-6238

1. Introduction

L'analyse de sensibilité globale pour les modèles à entrées indépendantes compte de nombreuses approches bien comprises par la communauté des modélisateurs (Saltelli *et al.*, 2008). Notamment, les décompositions d'Hoeffding–Sobol et de la variance (Hoeffding, 1948) ainsi que les indices de sensibilité qui en découlent (Sobol, 1993) offrent une représentation très pertinente et interprétable de l'influence de l'incertitude de chacun des paramètres sur la variabilité des sorties d'intérêt de tels modèles. Cependant, l'hypothèse d'indépendance nécessaire à l'analyse de Sobol ne peut pas être satisfaite en règle générale ce qui pose un problème méthodologique pour l'étude de modèles à entrées dépendantes. En effet, les paramètres sont souvent estimés à partir de mesures de laboratoire ou de terrain et ne sont pas indépendants à priori (Li *et al.*, 2010). Cette tendance est d'autant plus marquée que les systèmes étudiés sont complexes et ouverts. Ce phénomène s'observe notamment pour les modèles issus de la biologie et en particulier dans le cas des modèles de croissance des plantes. Lorsqu'un même modèle de croissance est utilisé pour plusieurs génotypes d'une même plante, la dépendance entre les paramètres émergeant des différents génotypes est significative (Lecoeur *et al.*, 2011). Il est donc nécessaire pour la communauté des modélisateurs de comprendre et de maîtriser l'influence des corrélations sur les méthodes classiques d'analyse de sensibilité (Saltelli *et al.*, 2008).

L'analyse de sensibilité globale pour les modèles à entrées dépendantes connaît de nombreux développements depuis une dizaine d'années. Plusieurs généralisations des indices de Sobol ont été proposées dont, par exemple, les travaux de Kucherenko *et al.* (2012) qui introduisent une généralisation des indices de premier ordre et totaux de Sobol. Bien que cette approche soit parcimonieuse en temps de calcul, elle ne quantifie pas explicitement la part de sensibilité due aux corrélations. Nous nous intéresserons dans cet article à une seconde classe de méthodes capables de quantifier l'influence de la dépendance.

Li *et al.* (2010) ont introduit une généralisation des indices de Sobol appelée *structural and correlative sensitivity analysis* (SCSA). Dans cette approche, les indices de Sobol sont dissociés en deux indices explicitant la part de sensibilité structurelle du modèle et celle due aux corrélations. La formulation de ces indices repose sur une décomposition fonctionnelle théorique du modèle inspirée des résultats obtenus dans le cas indépendant (Hoeffding, 1948; Sobol, 1993). La méthode a été amendée par Li et Rabitz (2012) avec l'introduction de la *condition d'orthogonalité hiérarchique* (COH) (Stone, 1994; Hooker, 2007) donnant un critère d'unicité pour la décomposition. L'existence de cette décomposition d'Hoeffding–Sobol hiérarchique généralisée (HSHG), a été formellement démontrée par Chastaing *et al.* (2012) sous des conditions précises de régularité des distributions de probabilité des entrées dépendantes.

La méthode SCSA requiert l'utilisation d'un méta-modèle nommé *high dimensional model representation* (HDMR) (Li *et al.*, 2001) permettant d'approcher les termes de la décomposition HSHG, introduisant naturellement un biais ainsi qu'un coût numérique important. Dans le prolongement de Li *et al.* (2010), la question de l'HDMR à utiliser a été abordée dans plusieurs travaux. Une difficulté pratique de cette méthode de projection est l'augmentation exponentielle du nombre de termes à estimer en fonction du nombre d'entrées étudiées. Caniou (2012) propose l'utilisation d'une surface de réponse basée sur un développement en polynômes de chaos. Elle est déterminée dans le cas indépendant puis est utilisée par la suite pour étudier le cas dépendant. Contrairement aux travaux de Chastaing *et al.* (2012) que nous considérerons dans cet article,

 ISSN: 2102-6238

cette approche souffre de l'absence d'un résultat d'unicité dans le cas dépendant. Chastaing *et al.* (2015) développent l'algorithme *Hierachically Orthogonal Gram-Schmidt* (HOGS) déterminant une base fonctionnelle de projection satisfaisant la COH et permettant d'obtenir une projection des termes de la décomposition HSHG. Trois critères déterminent le dimensionnement de la base et le coût numérique de l'algorithme : (1) la taille des échantillons aléatoires des entrées/sorties du modèle, dépendant essentiellement du coût de calcul du modèle étudié, (2) la dimension des bases de projection pour les termes fonctionnels d'ordre 1 de la décomposition HSHG, la dimension des bases de projection pour les ordres supérieurs étant alors fixée par construction, (3) le nombre maximal d'interactions entre variables considéré par l'algorithme permettant de contrôler l'augmentation exponentielle de la dimension de la base de projection.

Les travaux cités précédemment font appel à différentes méthodes d'estimation des coefficients de l'HDMR. L'estimateur des moindres carrés (OLS) est la base de toutes de ces méthodes. Cependant, les contraintes en ressources informatiques, l'augmentation exponentielle de la taille de la base ou encore la complexité et la lenteur d'exécution du modèle étudié peuvent conduire à résoudre des problèmes d'estimation paramétrique sous-déterminés. Plusieurs solutions ont été proposées dans les travaux cités précédemment pour gérer l'estimation paramétrique sous hypothèse de sparsité. Chastaing (2013) considère plusieurs procédures de pénalisation. L'ajout d'un terme de pénalité - norme-l_0, norme-l_1 (*lasso*) - à l'estimateur des moindres carrés permet de sélectionner les termes les plus informatifs de l'HDMR. L'auteur précise que l'utilisation de la pénalisation l_0 apparaît hasardeuse à cause de son instabilité numérique. La régression par pénalisation l_1 (lasso) peut, quant à elle, être traitée par plusieurs algorithmes comme *forward stagewise regression*, *least-angle regression* (LARS) (Hastie *et al.*, 2009) et permet d'obtenir des résultats satisfaisants. Li *et al.* (2012); Li et Rabitz (2012) proposent quant à eux l'algorithme D-MORPH qui diffère des approches présentées ci-dessus. La base polynomiale utilisée n'est pas construite spécifiquement en fonctions des distributions des entrées comme c'est la cas dans l'approche par polynômes de chaos et HOGS. La COH est prise en compte via l'ajout d'un terme de pénalisation au moment d'estimer la projection du modèle. Dans la pratique, il s'agit d'une transformation linéaire de l'estimateur des moindres carrés généralisé. Cette méthode est également efficace sous l'hypothèse de sparsité. Dans cet article, nous nous intéresserons à l'algorithme HOGS avec et sans hypothèse de sparsité. Dans le cadre de l'estimation paramétrique, Zou et Hastie (2005) ont introduit la méthode d'estimation *elastic net* (EN) permettant de faire un compromis entre le lasso et la régression l_2 (*ridge*). Cette approche permet notamment de dépasser certaines limitations du lasso. En effet, ce dernier ne peut sélectionner qu'un nombre maximal de variables explicatives égal au nombre d'observations. Par ailleurs, le lasso perd en efficacité dans le cas d'entrées corrélées. Nous testerons si EN couplé à HOGS peut améliorer les résultats sous hypothèse de sparsité.

Dans cet article, nous présenterons tout d'abord la définition et les aspects théoriques de la généralisation des indices de Sobol ainsi que de la décomposition HSHG. Nous présenterons ensuite l'algorithme HOGS ainsi que les différents estimateurs paramétriques étudiés (OLS, lasso, ridge, EN). Une analyse numérique sera conduite pour différents cas d'étude permettant d'aborder les aspects théoriques et numériques de HOGS. Nous étudierons tout d'abord plusieurs modèles pour lesquels des résultats théoriques existent (polynômes quadratiques), puis nous considèrerons un modèle non-linéaire (Ishigami), et enfin un modèle plus complexe (Log-Normal Allocation and Senescence - LNAS) consacré à la dynamique de la croissance des plantes (Cournède

 ISSN: 2102-6238

et al., 2013). Ce dernier modèle est un cas d'étude intéressant pour ces nouvelles méthodes d'analyse de sensibilité à cause de la forte dépendance entre les paramètres. Nous proposerons une méthodologie pour déterminer la configuration de HOGS et nous discuterons de l'interprétation des indices de Sobol généralisés.

2. Indices de Sobol généralisés

Introduisons quelques notations et hypothèses. Considérons un modèle f comptant p paramètres incertains notés $(x_i)_{i\in[\![1,p]\!]}$. Soit ν une mesure σ-finie sur $(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p))$ et soit P_x la loi de probabilité de $(x_i)_{i\in[\![1,p]\!]}$ telle que $P_x \ll \nu$ et notons p_x la densité de P_x relativement à ν.

Supposons que le modèle $f \in H = L^2_{\mathbb{R}}(\mathbb{R}^p, \mathscr{B}(\mathbb{R}^p), P_x)$ où H est muni du produit scalaire

$$\forall f,g \in H,\ \langle f,g\rangle = \mathbb{E}[f(x)g(x)] = \int f(x)g(x)dP_x. \tag{1}$$

Les indices introduits par Sobol (1993) reposent sur la décomposition fonctionnelle d'Hoeffding (1948). Elle donne une représentation du modèle sous la forme d'une somme dont les termes sont des fonctions correspondant aux différents degrés d'interactions entre les $(x_i)_{i\in[\![1,p]\!]}$. Elle se note

$$\begin{aligned} f(x_1,\ldots,x_p) &= f_\emptyset + \sum_{i=1}^{p} f_i(x_i) + \sum_{\substack{i,j=1\\ i\neq j}}^{p} f_{i,j}(x_i,x_j) + \ldots + f_{1,\ldots,p}(x_1,\ldots,x_p), \\ &= f_\emptyset + \sum_{u\in S^*} f_u(x_u), \end{aligned} \tag{2}$$

où u est un multi-indice d'éléments de $[\![1,p]\!]$, S est la collection des sous-ensembles de $[\![1,p]\!]$ et $S^* = S\backslash\{\emptyset\}$. Nous noterons dans la suite $|u| = card(u)$.

Dans le cas indépendant, cette décomposition correspond à l'ANOVA et Hoeffding (1948) assure que cette représentation est unique dans la mesure où elle satisfait la COH,

$$\forall u \in S^*,\ \forall v \in S,\ v \subset u,\ \mathbb{E}\left[f_u(x_u)f_v(x_v)\right] = 0. \tag{3}$$

En particulier, $f_\emptyset = \mathbb{E}[f(x)]$ et pour tout $u \in S^*$, $\mathbb{E}[f_u(x_u)] = 0$.

A partir de la décomposition de la variance de $f(x)$,

$$\mathbb{V}(f(x)) = \mathbb{E}\left[(f(x)-f_\emptyset)^2\right] = \sum_{u\in S^*} \mathbb{E}\left[f_u(x_u)^2\right] = \sum_{u\in S^*} \mathbb{V}(f_u(x_u)),$$

Sobol (1993) introduit les indices de sensibilité S_u suivants,

$$1 = \sum_{u\in S^*} \frac{\mathbb{V}(f_u(x_u))}{\mathbb{V}(f(x))} = \sum_{u\in S^*} S_u.$$

Dans le cas dépendant, Chastaing *et al.* (2012) démontre dans la continuité des travaux de Stone (1994), Hooker (2007) et Li et Rabitz (2012), que la décomposition HSHG donnée par l'équation (2) est unique sous la COH s'il existe une constante $M \in]0,1]$ vérifiant pour tout $u \in S, p_x \geq M\, p_{x_u} p_{x_{-u}}$.

 ISSN: 2102-6238

Nous rappelons ici les éléments théoriques de ce résultat. Pour tout $u \in S^*$, on notera H_u le sous-espace vectoriel (s.-e. v.) de H tel que tous les éléments de H_u ne dépendent que de x_u. $H_\emptyset$ désigne l'ensemble des constantes. Nous considérons alors la famille de s.-e.v. $(H_u^0)_{u\in S}$ telle que $H_\emptyset^0 = H_\emptyset$, satisfaisant la propriété d'orthogonalité hiérarchique

$$\forall u \in S^*, \quad H_u^0 = \left\{ h_u \in H_u \mid \langle h_u, h_v \rangle = 0,\ \forall v \subset u,\ \forall h_v \in H_v^0 \right\}. \tag{4}$$

Enfin, $H^0 = \left\{ h(x) = \sum_{u\in S} h_u(x_u), h_u \in H_u^0 \right\}$. Les conditions d'existence et d'unicité données par Chastaing *et al.* (2012) assurent que $H = \bigoplus_{u\in S} H_u^0$ et pour tout $u \in S$, $f_u \in H_u^0$.

La décomposition obtenue dans la cas dépendant mène à une nouvelle décomposition de la variance de $f(x)$ (SCSA) introduite par Li *et al.* (2010),

$$\mathbb{V}(f(x)) = \mathrm{Cov}\left(f(x) - f_\emptyset, \sum_{u\in S^*} f_u(x_u) \right) = \sum_{u\in S^*} \mathbb{V}(f_u(x_u)) + \sum_{u\in S^*} \sum_{\substack{v\in S^* \\ u\cap v \neq \{u,v\}}} \mathrm{Cov}(f_u(x_u), f_v(x_v)).$$

Remarquons que cette décomposition est homogène à celle obtenue dans le cas indépendant, les termes covariés étant alors nuls. Les trois indices de sensibilité généralisés associés à x_u (*contributions totale*, *structurelle* et *corrélative*) sont donnés par

$$S_u = \frac{\mathrm{Cov}\left(f(x), f_u(x_u)\right)}{\mathbb{V}(f(x))}, \tag{5}$$

$$S_u^s = \frac{\mathbb{V}\left(f_u(x_u)\right)}{\mathbb{V}(f(x))}, \tag{6}$$

$$S_u^c = \sum_{\substack{v\in S^* \\ u\cap v \neq \{u,v\}}} \frac{\mathrm{Cov}\left(f_u(x_u), f_v(x_v)\right)}{\mathbb{V}(f(x))}. \tag{7}$$

Ils satisfont les égalités $S_u = S_u^s + S_u^c$ et $\sum_u S_u = 1$. Cette dernière relation a été suggérée par Li et Rabitz (2014) comme critère de consistance des algorithmes numériques (voir section 3.3). Par ailleurs, il est également possible de définir les indices totaux,

$$ST_u = \sum_{v/u\subset v} S_v, \tag{8}$$

$$ST_u^s = \sum_{v/u\subset v} S_v^s, \tag{9}$$

$$ST_u^c = \sum_{v/u\subset v} S_v^c. \tag{10}$$

Ces indices généralisés diffèrent dans leurs propriétés par rapport au cas classique. Dans le cas indépendant, les indices S_u sont compris entre 0 et 1. Cette propriété n'est plus satisfaite dans le cas généralisé. Seuls les indices S_u^s sont positifs. Il en découle que l'interprétation de ces indices est moins directe et doit être étudiée plus précisément.

L'interprétation des indices de Sobol dans le cas indépendant est triviale dans la mesure où celle-ci s'appuie sur la relation d'ordre sur $\mathbb{R}$ et permet de hiérarchiser l'importance des paramètres quant à l'incertitude de sortie du modèle. La dissociation entre la part structurelle et

 ISSN: 2102-6238

la part corrélative fait perdre cette relation d'ordre naturelle qui existait dans le cas indépendant. Un exemple simple permet d'illustrer ce point. Soient deux indices de Sobol généralisés S_i et S_j associés aux paramètres x_i et x_j tels que $S_i = S_j = 0$. Dans le cas indépendant, cela signifie que l'influence de ces deux paramètres est nulle. Dans notre situation de dépendance, il est possible que $S_i = S_i^s = S_i^c = 0$ ce qui indique que le paramètre x_i n'a vraiment aucune influence sur l'incertitude. En revanche, il est aussi tout à fait possible que $S_j^s = -S_j^c \gg 0$ ce qui implique que l'influence directe de x_j est forte mais est totalement tempérée par sa dépendance aux autres paramètres. Plus généralement et contrairement à la situation sans dépendance, il est intéressant de garder à l'esprit qu'un paramètre à priori non influent peut l'être par le biais de la dépendance.

Finalement, du point de vue de l'incertitude de la variable d'intérêt, les indices S_u permettent toujours de comprendre quels sont les indices les plus influents. Inspirés par l'interprétation proposée par Chastaing *et al.* (2015), nous proposons plutôt d'interpréter les indices totaux pour u tel que $|u| = 1$. Cinq cas de figure sont identifiés :

1. $ST_i > ST_i^c > ST_i^s > 0$: la part d'incertitude associée à x_i est significative et majoritairement due à ses corrélations avec les autres entrées,
2. $ST_i > ST_i^s > ST_i^c > 0$: la part d'incertitude associée à x_i est majoritairement due à l'influence directe de x_i et est amplifiée par les dépendances,
3. $ST_i^s > ST_i > 0 > ST_i^c$: la part d'incertitude associée à x_i est significative mais est tempérée par la dépendance avec les autres paramètres,
4. $ST_i = 0$ et $ST_i^s = -ST_i^c$: les éléments donnés précédemment illustrent la complexité de l'interprétation de ce cas de figure, l'influence directe de x_i est tempérée par sa dépendance aux autres paramètres,
5. $ST_i^s > 0 > ST_i > ST_i^c$: le paramètre x_i a essentiellement pour effet de réduire la contribution des autres paramètres sur l'incertitude et il n'a pas d'influence directe sur la variance totale.

Nous tenterons par la suite de commenter ces éléments d'interprétation lors de l'application au modèle LNAS (section 4.4).

3. Méthodologie

Dans cet article, nous étudierons l'algorithme HOGS développé par Chastaing (2013), Champion *et al.* (2015) et Chastaing *et al.* (2015). Il construit une base fonctionnelle de dimension finie satisfaisant la COH (eq.3) et permettant de projeter chaque f_u sur une base de dimension finie.

3.1. Algorithme "Hierarchically Orthogonal Gram-Schmidt" (HOGS)

L'algorithme HOGS requiert plusieurs paramètres : la *dimension des bases d'ordre 1* notée $L \in \mathbb{N}^*$, le *nombre d'échantillons* noté $n_s \in \mathbb{N}^*$ et le *degré d'interaction maximal* noté $d \in [\![2, p]\!]$.

Dans la suite, le produit scalaire est remplacé par sa version empirique

$$\langle f, g \rangle_n = \sum_{s=1}^{n_s} f(x^s) g(x^s),$$

 ISSN: 2102-6238

où $(x^s)_{s\in[\![1,n_s]\!]}$ est un échantillon de n_s réalisations de x suivant P_x. Insistons ici sur le fait que la détermination de la décomposition est intrinsèquement liée à la loi P_x et ne peut être estimée qu'à partir d'un échantillonnage représentatif de P_x.

Chaque terme f_u de la décomposition de Hoeffding–Sobol appartient à H_u^0 pour tout $u \in S$. Comme $\dim(H_u^0) = +\infty$, il est nécessaire de se ramener à un problème de dimension finie. Dans un premier temps, HOGS considère les espaces H_i^0 pour tout $i \in [\![1,p]\!]$ correspondant aux termes fonctionnels d'ordre 1 ($|u| = 1$) par l'introduction de p s-e.v. $H_i^{0,L}$ de dimension L. Dans notre étude, nous choisissons des bases polynomiales dont le degré est au moins égal à 1 et inférieur ou égal à L. Les termes constants sont exclus de ces bases compte tenu du fait que $\mathbb{E}[f_i] = 0$, i.e. $H_u^0 \perp H_\emptyset$. Ces bases ont été choisies car nous étudierons par la suite plusieurs modèles polynomiaux. Les bases associées aux espaces d'ordre supérieurs ($|u| > 1$) sont obtenues par construction en imposant la COH. L'algorithme fonctionne de la manière suivante :

1. Initialisation : pour tout $i \in [\![1,p]\!]$, soit une base $(\phi_l^i)_{l\in[\![1,L]\!]}$ de $H_i^{0,L}$.
2. Il est possible de fixer le degré maximal d d'interactions entre variables. Alors, pour tout $u \in S$ tel que $2 \le |u| \le d$, on construit une base $(\phi_l^u)_{l\in[\![1,L]\!]^{|u|}}$ à $L^{|u|}$ éléments. A chaque multi-indice $l = (l_1,\dots,l_{|u|}) \in [\![1,L]\!]^{|u|}$ est associé un terme de la base de la forme

$$\phi_l^u(X_u) = \prod_{i=1}^{|u|} \phi_{l_i}^{u_i}(x_{u_i}) + \sum_{\substack{v\subset u,\\ v\in S^*}} \left[\sum_{k\in[\![1,L]\!]^{|v|}} \lambda_{l,k}^v \phi_k^v(x_v) \right] + C_l^u, \tag{11}$$

où les constantes $\lambda_{l,k}^v$ et C_l^u sont déterminées par la résolution de

$$\begin{cases} \langle \phi_l^u, \phi_k^v \rangle_n = 0, \forall v \subset u,\ \forall k \in [\![1,L]\!]^{|v|}, \\ \langle \phi_l^u, 1 \rangle_n = 0. \end{cases}$$

Le nombre de fonctions de base déterminées par l'algorithme n_b est donné par

$$n_b = \sum_{j=1}^{d} \binom{p}{j} L^j. \tag{12}$$

3.2. *Estimation paramétrique*

Rappelons que la construction de la base fonctionnelle ne dépend pas du modèle étudié. Celui-ci n'intervient que lors de la projection sur la base. Chaque terme f_u de la décomposition pour $u \in S^* / |u| \le d$ sera approché par une combinaison linéaire de fonctions de base de la forme $f_u \simeq \tilde{f}_u = \sum_l \beta_l^u \phi_l^u(x_u)$ où les termes β_l^u sont à estimer. Comme $f_\emptyset = \mathbb{E}[f(x)]$, notons $y = f(x) - f_\emptyset$. Ainsi le problème d'estimation revient à optimiser l'approximation

$$y = f(x) - f_\emptyset \simeq \sum_{u\in S^*/|u|\le d} \sum_l \beta_l^u \phi_l^u(x_u).$$

L'estimation des moindres carrés revient à résoudre

$$\min_{(\beta_l^u)} \mathbb{E}\left[\left(y - \sum_{u\in S^*/|u|\le d} \sum_l \beta_l^u \phi_l^u(x_u) \right)^2 \right],$$

 ISSN: 2102-6238

et sous forme matricielle discrète

$$\min_{\mathbb{B}} \|\mathbb{Y} - \Phi\mathbb{B}\|_{n_s}^2,$$

en posant $\mathbb{Y} = \left(y^1, ..., y^{n_s}\right)^\top$, $\Phi = \left(\phi_l^u(x_u^s)\right)$ et $\mathbb{B} = \left(\beta_l^u\right)^\top$.

Dans cet article, nous nous intéressons également aux régressions pénalisées dans la cadre de la sparsité. Ceci consiste à amender l'estimateur des moindres carrés par un terme de pénalité défini positif $J(\mathbb{B})$

$$\min_{\mathbb{B}} \|\mathbb{Y} - \Phi\mathbb{B}\|_{n_s}^2 + \lambda J(\mathbb{B}),$$

où λ est un paramètre positif d'optimisation de la régression. Nous proposerons ici trois pénalités distinctes. Le *lasso* (l_1) correspond à $J(\mathbb{B}) = \|\mathbb{B}\|_1$, la régression *ridge* (l_2) à $J(\mathbb{B}) = \|\mathbb{B}\|_2$ et leur compromis *elastic net* (EN) $J(\mathbb{B}) = \mu \|\mathbb{B}\|_1 + (1-\mu)\|\mathbb{B}\|_2$ où $\mu \in [0,1]$. Le lasso et le ridge sont des cas particuliers d'EN ($\mu = 1$ et $\mu = 0$).

3.3. *Pertinence de la surface de réponse*

Comme mentionné précédemment, Li et Rabitz (2014) ont proposé d'utiliser la relation $\sum_u S_u = 1$ pour estimer la pertinence de la surface de réponse obtenue après projection. Si l'on note $\tilde{f}_u$ la projection de f_u, le calcul de cette somme revient à calculer

$$\begin{aligned}
\sum_{u \in S^*} S_u &= \sum_{u \in S^*} \frac{\mathrm{Cov}(f(x), \tilde{f}_u(x))}{\mathbb{V}(f(x))}, \\
&= \frac{\mathrm{Cov}(f(x), \sum_{u \in S^*} \tilde{f}_u(x))}{\mathbb{V}(f(x))}, \\
&= \frac{\mathrm{Cov}(f(x), \tilde{f}(x))}{\mathbb{V}(f(x))}.
\end{aligned}$$

Il ressort de cette relation que plus la surface de réponse $\tilde{f}$ est semblable à f, plus la valeur de la somme est proche de 1. Nous proposons d'appeler cette somme *indice de consistance* et nous le noterons i_c par la suite.

4. Analyse numérique

L'implémentation de l'algorithme HOGS ainsi que des modèles étudiés dans cet article a été effectuée au sein de la plate-forme informatique PyGMAlion-Julia inspirée de la version C++ de PyGMAlion [1] (Cournède *et al.*, 2013). Le langage informatique Julia© [2] est un langage libre et multi-plate-forme (Bezanson *et al.*, 2014). Le Mésocentre de CentraleSupélec [3] a été utilisé pour conduire l'ensemble des tests numériques de cet article, en mode séquentiel.

La génération des nombres aléatoires dans Julia fait appel à la librairie Mersenne Twister (Matsumoto et Nishimura, 1998). Les distributions multivariées utilisées dans cet article suivent

[1] www.digiplante.mas.ecp.fr/software/
[2] www.julialang.org
[3] www.mesocentre.ecp.fr/meso/presentation/materiels

 ISSN: 2102-6238

des loi multivariées gaussiennes dont l'implémentation est incluse dans Julia. Précisons que la condition d'unicité de la décomposition HSHG n'est pas assurée pour ce type de distributions dans les conditions du théorème d'unicité proposé par Chastaing *et al.* (2015). Leur sélection a néanmoins été motivée par le fait que les polynômes quadratiques admettent des valeurs analytiques pour les indices de Sobol généralisés lorsque leurs entrées suivent une loi multivariée gaussienne (Li et Rabitz, 2014) et serviront de référence dans la suite de l'étude.

La résolution numérique des différents problèmes de régressions pénalisées (lasso, ridge et EN) s'appuie sur les travaux de Zou et Hastie (2005) consacrés à EN. Ces travaux sont implémentés dans la librairie GLMNet (Friedman *et al.*, 2010) disponible librement pour le langage Julia [4]. GLMNet permet d'estimer, à μ fixé, le meilleur paramètre λ pour la régression pénalisée. Ce dernier est alors obtenu par validation croisée tel que décrit par Zou et Hastie (2005). La recherche du meilleur compromis μ entre la pénalisation l_1 et l_2 n'est pas automatique et se fait "à la main" par une exploration discrète de l'intervalle $[0,1]$. Le meilleur couple (λ,μ) au sens de la validation croisée est choisi comme paramétrage de EN. Dans cette article, nous avons choisi un maillage régulier de pas 0.1 pour μ. La répétition nécessaire de l'algorithme de recherche de λ engendre un surcoût évident de temps de calcul.

Plusieurs aspects de la méthode HOGS seront éprouvés. Tout d'abord, nous évaluerons sa convergence suivant le nombre d'échantillons dans le cas d'un polynôme quadratique à trois entrées. Dans ce cas simple ($p=3$), la configuration optimale à priori de l'algorithme HOGS est connue : $d=2$ et $L=2$ soit $n_b=18$. Dans un second temps, nous testerons un second polynôme quadratique à $p=10$ entrées. Dans ce cas, la configuration optimale est également connue : $d=2$ et $L=2$ soit $n_b=200$. Nous comparerons dans ce cas l'efficience du lasso et de EN dans le cas où $n_s < n_b$. Le troisième cas d'étude considérera le modèle d'Ishigami. Ce modèle est fortement non-linéaire et nous dicuterons du choix de la dimension de l'interpolation polynomiale ainsi que de la surface de réponse dans le cas indépendant. Enfin, la méthode sera appliquée au modèle LNAS (Cournède *et al.*, 2013), modèle dynamique de la croissance des plantes, afin d'illustrer les résultats obtenus précédemment pour un modèle complexe.

4.1. Cas d'étude 1 : polynôme quadratique à 3 variables (p = 3)

Le premier modèle considéré est un polynôme quadratique défini dans $\mathbb{R}^3$ et à valeurs réelles

$$f(x) = 6 + 4x_1 + 4x_2 + 2x_3 + 2x_1^2 + 6x_1x_2 + 2x_2^2 + 6x_1x_3 + 2x_3^2. \tag{13}$$

Nous supposerons que x suit une loi multivariée gaussienne centrée en $0_{\mathbb{R}^3}$ et de matrice de variance-covariance de la forme

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \rho_{12}\sigma_1\sigma_2 & 0 \\ \rho_{12}\sigma_1\sigma_2 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{bmatrix}$$

avec $\sigma_1 = 0.2$, $\sigma_2 = 0.2$, $\sigma_3 = 0.18$ et $\rho_{12} = 0.6$.

L'étude de ce modèle simple présente plusieurs intérêts. Tout d'abord, Li et Rabitz (2014) donnent une valeur exacte des différents indices de Sobol pour ce modèle. Ce point de référence

[4] www.github.com/simonster/GLMNet.jl

 ISSN: 2102-6238

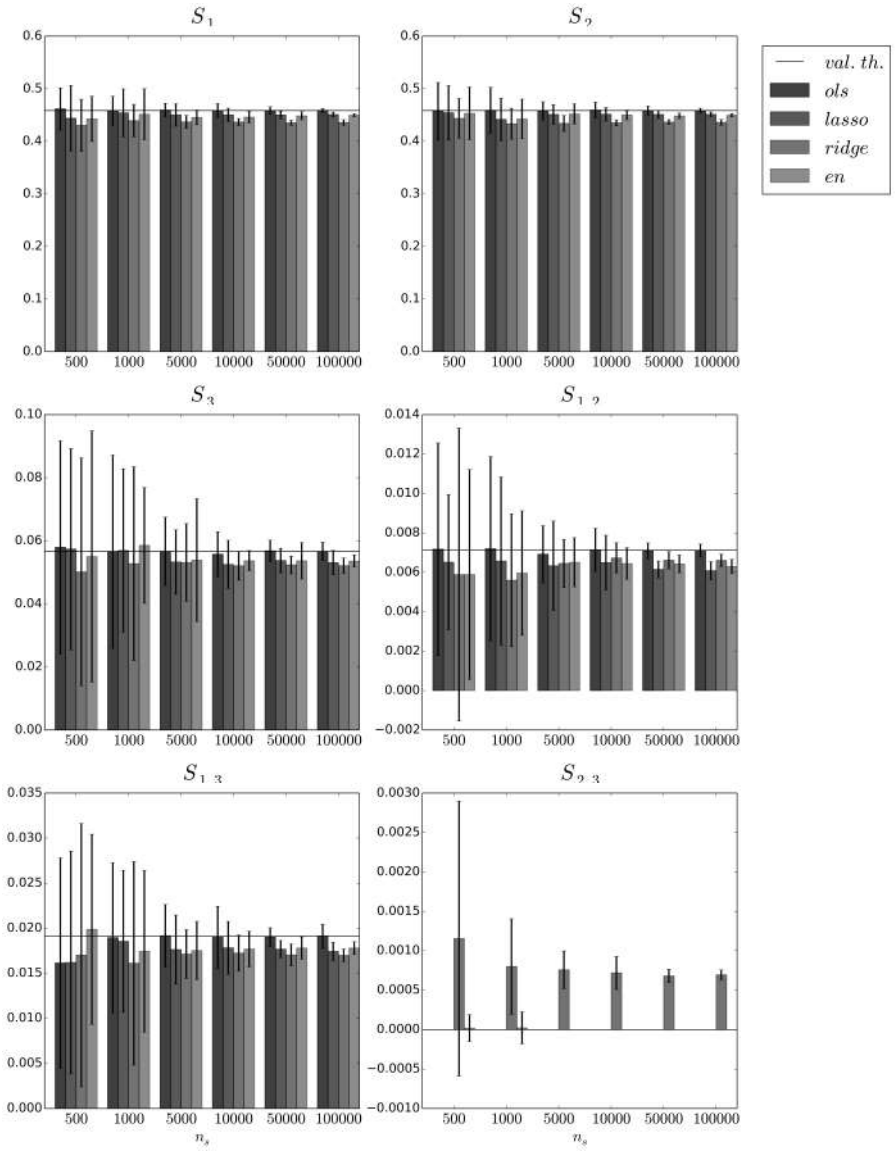

FIGURE 1. *Convergence des 4 variantes de l'algorithme HOGS vers les indices S_u pour $d = 2, L = 2, n_b = 18$ dans le cas d'étude 1. Chaque sous-figure est consacrée à un indice S_u donné, noté au dessus de la sous-figure. Sa valeur théorique, déterminée par les travaux de Li et Rabitz (2014), est donnée par la droite horizontale dans chacune des sous-figures. Les valeurs en abscisse correspondent aux différentes valeurs d'échantillons n_s. Les histogrammes associés aux différentes méthodes d'estimation paramétrique (OLS, lasso, ridge, EN) présentent les résultats obtenus en moyenne par les 10 réplications de chaque expérience. Les écarts-types obtenus sur ces réplications sont indiqués sur chaque histogramme de manière symétrique autour de la moyenne. Dans le cas de $S_{2,3}$, qui est nul en théorie, les valeurs pour OLS et lasso sont quasi-nulles et ne sont pas visibles sur le graphique quelle que soit la taille des échantillons.*

permettra tout d'abord de vérifier que l'implémentation de la méthode HOGS a été effectuée correctement au sein de la plate-forme PyGMAlion-Julia. Ensuite, il permettra de présenter l'évo-

 ISSN: 2102-6238

lution de la convergence de l'algorithme en fonction du nombre d'échantillons et en fonction des différentes méthodes d'estimations paramétriques. Bien que dans tous les cas étudiés le problème d'estimation ne soit pas en condition de sparsité ($n_b < n_s$), nous présentons tout de même les résultats obtenus avec les régressions pénalisées afin de vérifier la bonne implémentation de ces méthodes. Enfin, dans ce cas polynomial, le choix de la configuration de HOGS semble évident. En effet, le nombre d'interaction maximal et le degré polynomial sont explicites ce qui conduit à fixer $d = 2$ et $L = 2$. Cependant, pour se convaincre de la consistance de la décomposition fonctionnelle sous-jacente à la construction de la base de projection, nous avons tout de même testé des configurations augmentées de l'algorithme.

Plusieurs degrés d'interaction ($d = 2, 3$), degrés d'interpolation ($L = 2, 3, 5$), tailles d'échantillons ($n_s = 500, 1000, 5000, 10000, 50000, 100000$) et méthodes d'estimation paramétrique (OLS, lasso, ridge, EN) ont été testés. Chaque expérience a été répliquée 10 fois et les résultats présentés le sont en moyennes et en écarts-types.

Premièrement, l'augmentation des configurations au-delà de la configuration optimale est inutile. Une observation détaillée montre que les coefficients de projection associés à des degrés polynomiaux et à des interactions inexistants dans le modèle sont nuls (lasso, EN) ou infinitésimaux (OLS, ridge). Dans un soucis de concision, nous ne présenterons dans la suite que les résultats obtenus pour la configuration optimale (fig. 1).

Nous constatons la convergence en moyenne des différentes méthodes à des vitesses similaires. Remarquons de plus que dans les cas augmentés, la convergence de l'algorithme est dégradée à mesure que n_b augmente.

Dans l'ensemble des cas, HOGS-OLS est la méthode la plus précise alors que la régression ridge est la moins performante. Les régressions pénalisées lasso et EN sont équivalentes dans l'ensemble, la précision des indices obtenus variant suivant les cas. Par ailleurs, les régressions pénalisées convergent vers des résultats biaisés. Néanmoins, ce biais assez faible n'altère pas l'étude qualitative de la sensibilité du modèle pour le modélisateur.

Illustrant sur ce cas simple que la régression ridge est la moins performante, nous ne présenterons pas dans la suite les résultats obtenus par cette méthode qui seront systématiquement plus faibles.

4.2. Cas d'étude 2 : polynôme quadratique à 10 variables (p = 10)

Le modèle considéré maintenant est un polynôme quadratique défini sur $\mathbb{R}^{10}$ et à valeurs réelles

$$\begin{aligned} f(x) = 6 + 4x_1 + 4x_2 + 2x_3 + 2x_4 + 2x_5 + 4x_6 + 4x_7 + 4x_{10} + 2x_2^2 + 2x_3^2 + 3x_4^2 + x_5^2 + x_6^2 + \\ 2x_7^2 + x_8^2 + x_9^2 + x_{10}^2 + 6x_1x_2 + 12x_1x_9 + 8x_2x_3 + 2x_2x_4 + 10x_2x_7 + 4x_3x_5 + 2x_4x_6 + \\ 4x_4x_9 + 2x_5x_7 + 2x_5x_8 + 2x_5x_{10} + 6x_6x_8 + 4x_6x_{10} + 4x_7x_9 + 6x_8x_{10} + 4x_9x_{10}. \end{aligned} \tag{14}$$

Nous supposerons que x suit une loi multivariée gaussienne centrée en $0_{\mathbb{R}^{10}}$ et de matrice de variance-covariance de la forme $\Sigma = (\sigma_i \rho_{i,j} \sigma_j)_{i,j \in [\![1,10]\!]}$ avec $\sigma_1 = 0.2$, $\sigma_2 = 0.2$, $\sigma_3 = 0.18$, $\sigma_4 = 0.2$, $\sigma_5 = 0.2$, $\sigma_6 = 0.5$, $\sigma_7 = 0.2$, $\sigma_8 = 0.2$, $\sigma_9 = 0.18$, $\sigma_{10} = 0.2$ et $\rho_{ii} = 1$ pour tout $i \in [\![1, 10]\!]$, $\rho_{12} = 0.6$, $\rho_{23} = 0.3$, $\rho_{34} = 0.5$, $\rho_{45} = 0.6$, $\rho_{56} = 0.2$, $\rho_{67} = 0.1$, $\rho_{78} = 0.3$, $\rho_{89} = 0.2$, $\rho_{910} = 0.1$ et $\rho_{ij} = 0$ sinon. Comme montré dans le cas d'étude précédent, ce type de modèle

 ISSN: 2102-6238

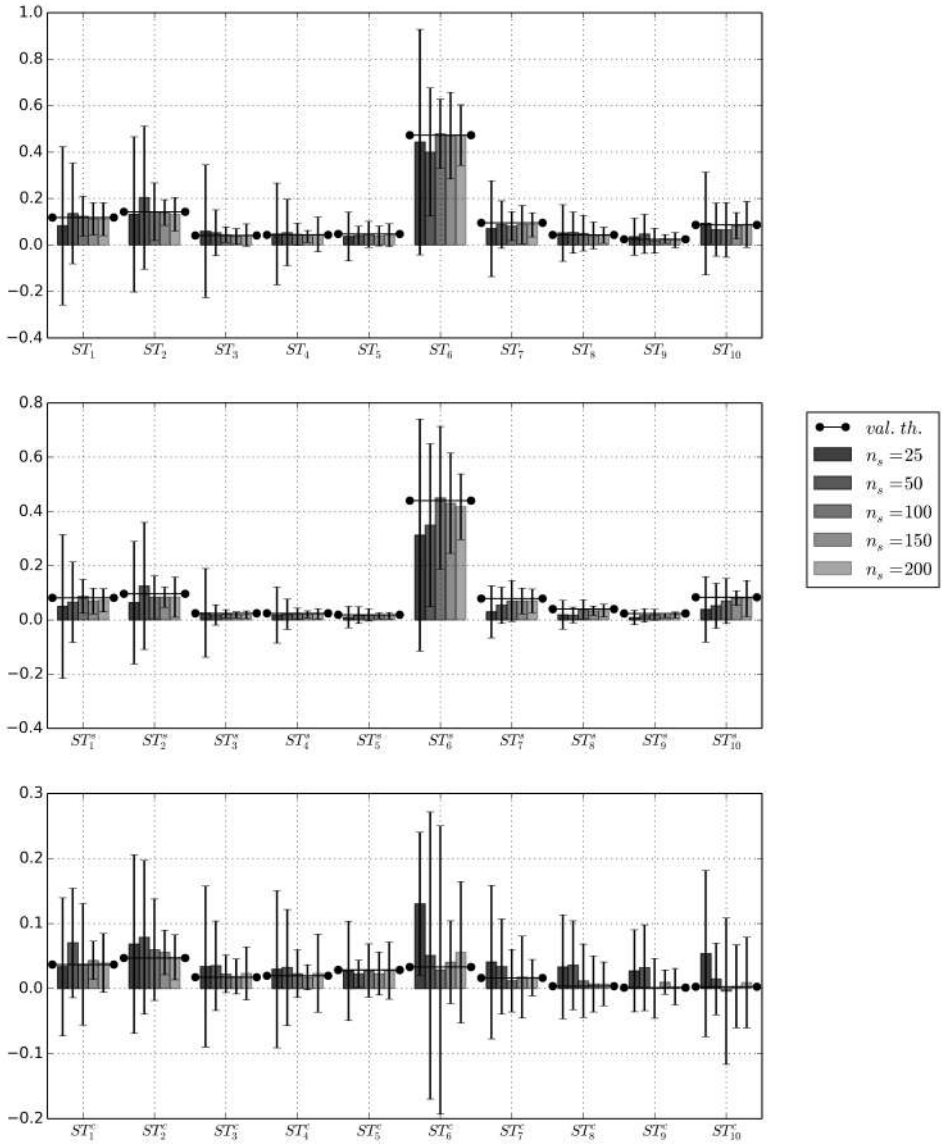

FIGURE 2. *Indices de Sobol totaux obtenus par la méthode HOGS-lasso ($d = 2$, $L = 2$, $n_b = 200$) dans le cas d'étude 2. La première sous-figure est consacrée aux indices totaux ST_i pour chaque x_i. Les deux sous-figures suivantes donnent la part structurelle ST_i^s de ces indices puis leur part corrélée ST_i^c. Les valeurs théoriques de chaque indice, déterminées par les travaux de Li et Rabitz (2014), sont données par un segment horizontal. Les différents histogrammes, associés aux différentes configurations de HOGS-lasso, présentent les résultats obtenus en moyenne par les 10 réplications de chaque expérience. Les écarts-types obtenus sur ces réplications sont indiqués sur chaque histogramme de manière symétrique autour de la moyenne.*

procure immédiatement une configuration idéale de l'algorithme ($d = 2$, $L = 2$, $n_b = 200$) ainsi qu'une valeur exacte pour les différents indices de Sobol généralisés (Li et Rabitz, 2014).

Le modèle étudié ici permet d'aborder la question de l'estimation paramétrique sous hypothèse

 ISSN: 2102-6238

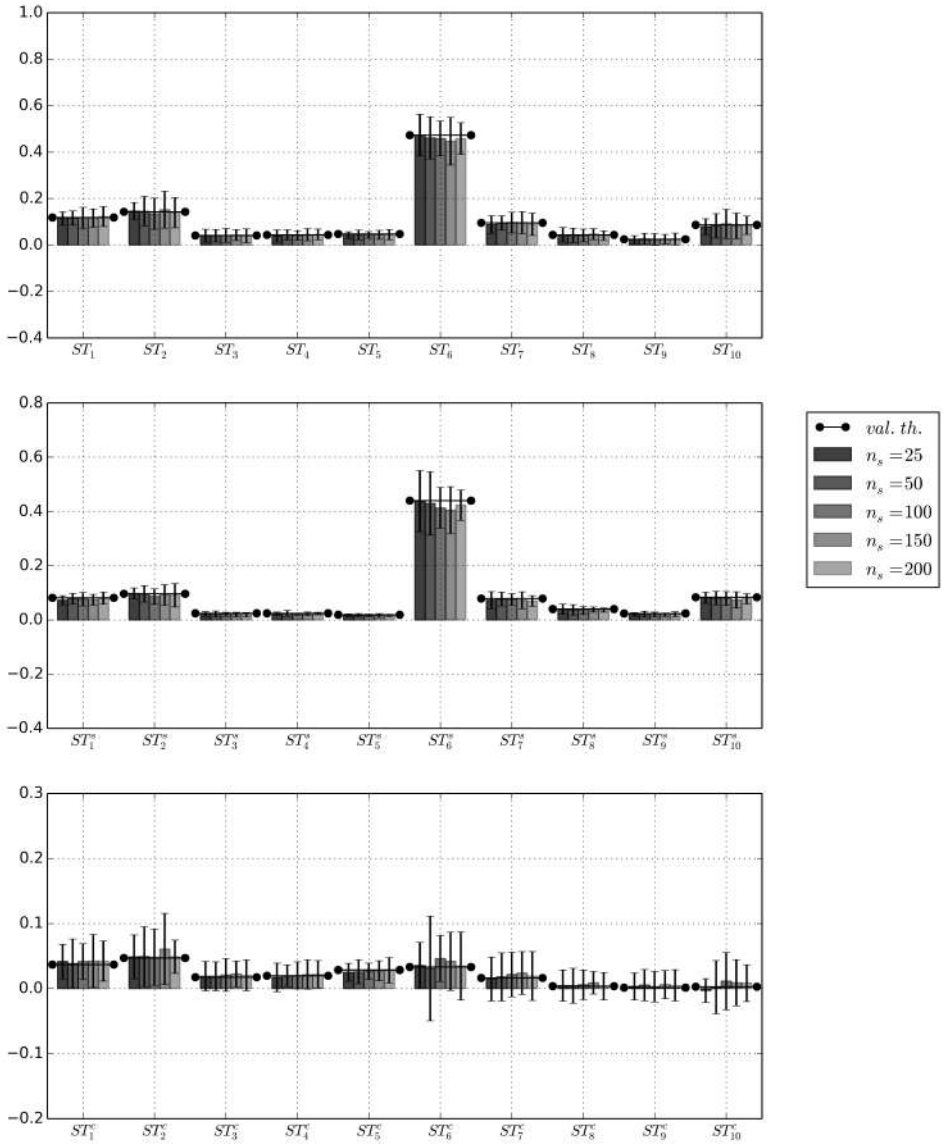

FIGURE 3. *Indices de Sobol totaux obtenus par la méthode HOGS-EN ($d = 2$, $L = 2$, $n_b = 200$) dans le cas d'étude 2. La première sous-figure est consacrée aux indices totaux ST_i pour chaque x_i. Les deux sous-figures suivantes donnent la part structurelle ST_i^s de ces indices puis leur part corrélée ST_i^c. Les valeurs théoriques de chaque indice, déterminées par les travaux de Li et Rabitz (2014), sont données par un segment horizontal. Les différents histogrammes, associés aux différentes configurations de HOGS-EN, présentent les résultats obtenus en moyenne par les 10 réplications de chaque expérience. Les écarts-types obtenus sur ces réplications sont indiqués sur chaque histogramme de manière symétrique autour de la moyenne.*

de sparsité. Le modèle compte 8 monômes de degré 1, 9 monômes de degré 2 et 16 doubles produits de degré 1 soit au total 33 termes non constants. Comme la construction de la base de projection créée par l'algorithme HOGS est polynomiale, nous nous attendons à ce que le nombre de termes de bases significatifs dans l'HDMR soit de l'ordre de 33 parmi les 200 termes de la

 ISSN: 2102-6238

base. Nous proposons de comparer lasso et EN pour des tailles d'échantillons très petites, pouvant illuster la situation où le temps nécessaire à une simulation du modèle étudié est important.

Les valeurs suivantes de n_s seront testées : 25, 50, 100, 150, 200. Ces faibles valeurs auront une double influence sur le comportement de HOGS : le produit scalaire empirique utilisé pour la détermination de la base de projection sera influencé ainsi que l'estimateur paramétrique choisi. Les résultats sont présentés aux figures 2 et 3.

De manière générale, l'estimateur EN est plus précis et plus stable que le lasso (écarts-types plus faibles). Il permet d'obtenir une bonne approximation des indices de Sobol généralisés. Le lasso quant à lui est moins performant à fortiori lorsque le nombre d'échantillons diminue. Dans la situation de sparsité forte ($n_s = 25$), EN permet d'obtenir des résultats satisfaisants alors que le lasso est clairement moins performant. Ceci vient du fait que le lasso ne peut pas sélectionner plus de 25 termes dans la base ce qui est insuffisant pour estimer les 33 termes non-constants du modèle. Nous conseillons donc de privilégier l'utilisation d'EN dans les cas de forte sparsité plutôt que le lasso, surtout s'il est difficile d'anticiper à priori la forme de la surface de réponse.

4.3. Cas d'étude 3 : fonction d'Ishigami (p = 3)

Nous étudions ici la fonction test d'Ishigami

$$f(x) = \sin(x_1) + 7\sin(x_2)^2 + 0.1x_3^4\sin(x_1). \tag{15}$$

Nous supposerons comme la plupart des études consacrées à ce modèle, que chaque composante de x suit une loi uniforme sur l'intervalle $]-\pi,\pi[$.

Nous nous placerons ici dans le cas indépendant et nous nous intéresserons à la pertinence du méta-modèle estimé par HOGS. Concernant la configuration, le degré d'interaction maximal d à choisir est clairement égal à 2 vu la forme du modèle. En revanche, contrairement aux exemples précédents, il n'existe pas de choix optimal à priori pour L. L'objectif de ce cas d'étude est de tirer bénéfice de la non-linéarité de f afin de déterminer l'influence de L sur l'indice de consistance (voir 3.3). Nous nous placerons dans un cas d'étude avec un grand nombre d'échantillons ($n_s = 100000$) afin de bien fixer l'étude sur le degré L.

OLS étant la meilleure estimation d'après les exemples précédents, nous continuerons à tester ses performances. De plus, afin d'observer plus précisément le comportement de la surface de réponse, nous proposons d'utiliser le lasso pour sélectionner les termes de la base de projection les plus significatifs. La pertinence de l'utilisation de termes de base dont le degré est très élevé pourra alors être discutée.

TABLEAU 1. *Comparaison de l'indice de consistance i_c de la surface de réponse pour differents degrés d'interpolation polynomiale pour OLS et lasso dans le cas d'étude 3. Les deux méthodes ont une efficience semblable.*

L	3	5	7	9
i_c-OLS	0.56	0.86	0.99	1.00
i_c-lasso	0.56	0.86	0.99	0.98

Le tableau 1 montre l'influence du degré L sur l'indice de consistance de la surface de réponse que ce soit dans le cas de la régression OLS ou avec le lasso. L'augmentation de la qualité de la

 ISSN: 2102-6238

surface de réponse ne dépend pas de l'estimateur paramétrique utilisé et la convergence de i_c vers 1 en fonction de L est semblable pour les deux estimateurs dans ce cas où n_s est grand.

Par ailleurs, un examen détaillé des résultats obtenus avec le lasso dans le cas où $L = 9$, donne un éclairage important sur les termes de base sélectionnés et donc significatifs. Sur les dix réplications, en plus de la fonction de base constante associée à l'espérance du modèle, le lasso a retenu systématiquement les fonctions de bases associées aux multi-indices suivants :

- pour x_1 : (1,0,0), (3,0,0), (5,0,0),
- pour x_2 : (0,2,0), (0,4,0), (0,6,0), (0,8,0),
- pour x_1, x_3 : (1,0,2), (1,0,4), (3,0,2), (3,0,4), (5,0,2), (5,0,4).

Il ressort de cette sélection que le degré polynomial ne dépasse jamais la valeur de $L = 9$ y compris pour les termes quadratiques. Ce résultat semble indiquer que l'utilisation de termes quadratiques d'ordre plus élevés n'est pas pertinente. Cette remarque permet de suggérer une modification importante de la procédure HOGS. En effet, plutôt que de construire une base inutilement grande pour l'étude des interactions d'ordre élevé, le degré maximal des fonctions de base pourrait être limité à L. Ceci aurait alors pour conséquence de réduire le nombre de fonctions à construire et à estimer et éviterait à HOGS de créer des situations de sparsité artificielle. Considérant la formule (12), nous obtiendrions par construction un nouveau nombre $\hat{n}_b$ de fonctions tel que

$$\hat{n}_b = L \sum_{j=1}^{\min(d,L,p)} \binom{p}{j}. \tag{16}$$

Il convient de remarquer qu'avec une telle modification, le nombre d'intéractions entre variables d serait nécessairement limité à la valeur de L. À titre indicatif, dans la configuration utilisée ici ($p = 3$, $d = 2$, $L = 9$), $n_b = 270$ et $\hat{n}_b = 54$. Une telle réduction de la base est considérable et mérite d'être étudiée dans des travaux ultérieurs.

4.4. Cas d'étude 4 : modèle LNAS

Le modèle LNAS (Log-Normal Allocation and Senescence) est un modèle de la croissance des plantes, qui décrit dynamiquement les processus de production de biomasse par photosynthèse, d'allocation de biomasse entre les différents types d'organes de la plante et la sénescence foliaire. Nous étudions ici la version simple dédiée à la betterave décrite dans (Cournède *et al.*, 2013; Chen et Cournède, 2014), avec seulement deux compartiments, les feuilles et la racine.

Dans ce type de modèle, certains processus élémentaires sont représentés par des fonctions empiriques, comme par exemple des sigmoïdes, des fonctions de Michaelis–Menten ou encore des fonctions de répartition de la loi log-normale dans le cas du modèle LNAS. Ces fonctions empiriques mettent en jeu plusieurs paramètres, et une modification de la dynamique d'un processus se répercute généralement sur plusieurs paramètres à la fois, ce qui entraîne une forte dépendance entre ces différents paramètres. Cette dépendance pourrait d'ailleurs être expliquée à des niveaux plus élémentaires encore, par le déterminisme génétique des différents paramètres mis en jeu dans les modèles de croissance de plantes (Tardieu, 2003) et le phénomène de pléiotropie, les gènes pouvant avoir de l'influence sur plusieurs caractères phénotypiques ou sur plusieurs paramètres dans les modèles (Letort *et al.*, 2008). L'analyse de sensibilité globale par indices de Sobol commence à être largement utilisée pour l'étude des modèles de croissance de plantes

 ISSN: 2102-6238

et de cultures (voir par exemple Monod *et al.* (2006); Lamboni *et al.* (2011); Wu *et al.* (2012)), car elle offre un diagnostic intéressant pour mieux comprendre les interactions entre processus dans les modèles mais également pour leur paramétrisation. Cependant, cette étude est toujours conduite en supposant les entrées indépendantes. L'utilisation d'indices de Sobol généralisés et de l'algorithme HOGS prend donc tout son sens pour ce type de modèle.

Nous étudions ici la dynamique de la sensibilité à 10 paramètres de la *biomasse foliaire sèche* au cours du temps. Son évolution est présentée dans la figure 4. Elle est étudiée sur l'intervalle de temps $[\![25, 160]\!]$ avec un pas régulier de 5 jours.

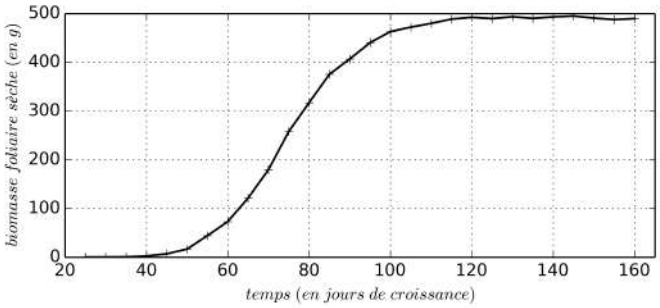

FIGURE 4. *Dynamique de la variable* biomasse foliaire sèche *du modèle LNAS obtenue avec les valeurs moyennes des paramètres étudiés.*

Les distributions de probabilités proposées sont arbitraires mais sont néanmoins inspirées des travaux consacrés précédemment au modèle LNAS. Considérant que ce modèle est composé de quatres modules, nous proposons de doter les paramètres de chaque module d'une dépendance et de laisser ces quatres groupes de paramètres indépendants entre eux.

Un premier groupe de paramètres déterminant la dynamique de sénescence foliaire $(tt_{sen}, \mu_{sen}, s_{sen})^\top$ suit une loi multivariée gaussienne $\mathcal{N}_3(\mu_3, \Sigma_3)$ où $\mu_3 = (644,\ 2400,\ 4520)^\top$ et $\Sigma_3 = \sigma_3^\top \rho_3 \sigma_3$ avec $\sigma_3 = (32.2,\ 120,\ 226)^\top$ et

$$\rho_3 = \begin{bmatrix} 1 & 0,5 & -0.5 \\ 0.5 & 1 & 0.2 \\ -0.5 & 0.2 & 1 \end{bmatrix}.$$

Un second groupe de paramètres correspondants à la dynamique d'allocation entre les deux compartiments foliaire et racinaire $(\mu_{alloc}, s_{alloc}, s_{init}, s_{end})^\top$ suit une loi multivariée gaussienne $\mathcal{N}_4(\mu_4, \Sigma_4)$ où $\mu_4 = (550,\ 300,\ 0.7,\ 0.15)^\top$ et $\Sigma_4 = \sigma_4^\top \rho_4 \sigma_4$ avec $\sigma_4 = (27.5,\ 15,\ 0.035,\ 0,075)^\top$ et

$$\rho_4 = \begin{bmatrix} 1 & 0,2 & 0 & 0 \\ 0.2 & 1 & 0.5 & -0.5 \\ 0 & 0.5 & 1 & -0.5 \\ 0 & -0.5 & -0.5 & 1 \end{bmatrix}.$$

Les 3 paramètres $rue \sim \mathcal{N}(3.6,\ 0.15)$, $e \sim \mathcal{N}(60,\ 3)$, $k_b \sim \mathcal{N}(0.7,\ 0.035)$ correspondant respectivement à l'efficience de conversion de l'énergie lumineuse, la masse surfacique et le

 ISSN: 2102-6238

coefficient d'interception de la loi de Beer–Lambert sont indépendants des autres.

Nous utiliserons ici une configuration de HOGS assez petite ($d = 2$, $L = 5$, $n_b = 1175$) afin d'évaluer l'indice de consistance pour les différents régresseurs et différentes tailles d'échantillons résumés dans la table 2. Le choix de cette configuration permet également de se placer dans des durées d'exécution acceptables.

TABLEAU 2. *Synthèse des performances temporelles pour différentes configurations de HOGS appliquées au modèle LNAS (cas d'étude 4). La recherche à chaque pas de temps du meilleur paramétrage* μ *de EN entraîne un surcoût en temps de calcul comparativement au lasso.*

n_s	EN	lasso	OLS
100	77 s.	22 s.	-
500	292 s.	41 s.	-
1000	561 s.	64 s.	-
10000	-	-	103 s.
50000	-	-	517 s.

Bien que nous ayons conseillé précédemment l'utilisation d'EN comparativement au lasso, il apparaît ici que son coût en terme de temps de calcul est très important à fortiori lorsque l'analyse du modèle porte sur plusieurs états successifs du modèle. Ce point devra être remis en question au regard de la proposition d'amendement de la procédure HOGS proposée dans la section précédente. En effet, si la taille de la base de projection est fortement réduite, l'hypothèse de sparsité forte sera rarement rencontrée dans la pratique. Le choix de l'utilisation d'EN plutôt que du lasso relève donc d'un compromis entre sparsité, ressources numériques et temps de calcul.

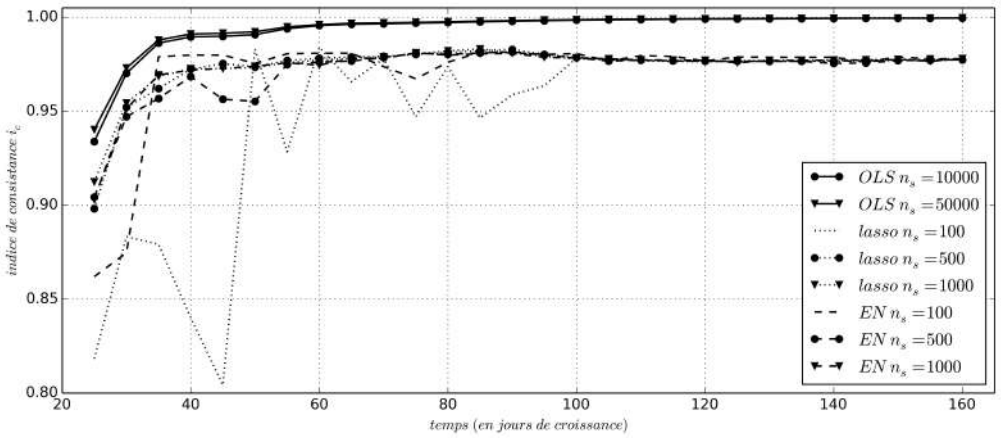

FIGURE 5. *Dynamique de l'indice de consistance* i_c *associés à la variable* biomasse foliaire sèche *du modèle LNAS pour différentes associations de méthodes de régression et de tailles d'échantillons et avec* $d = 2$, $L = 5$, $n_b = 1175$ *fixés (cas d'étude 4).*

L'influence des différents régresseurs sur l'indice de consistance est présentée dans la figure 5. Quels que soient les échantillons et le régresseur choisis, les premiers temps étudiés ont un indice de consistance plus faible. Ce phénomène est très probablement lié à un effet de seuil

 ISSN: 2102-6238

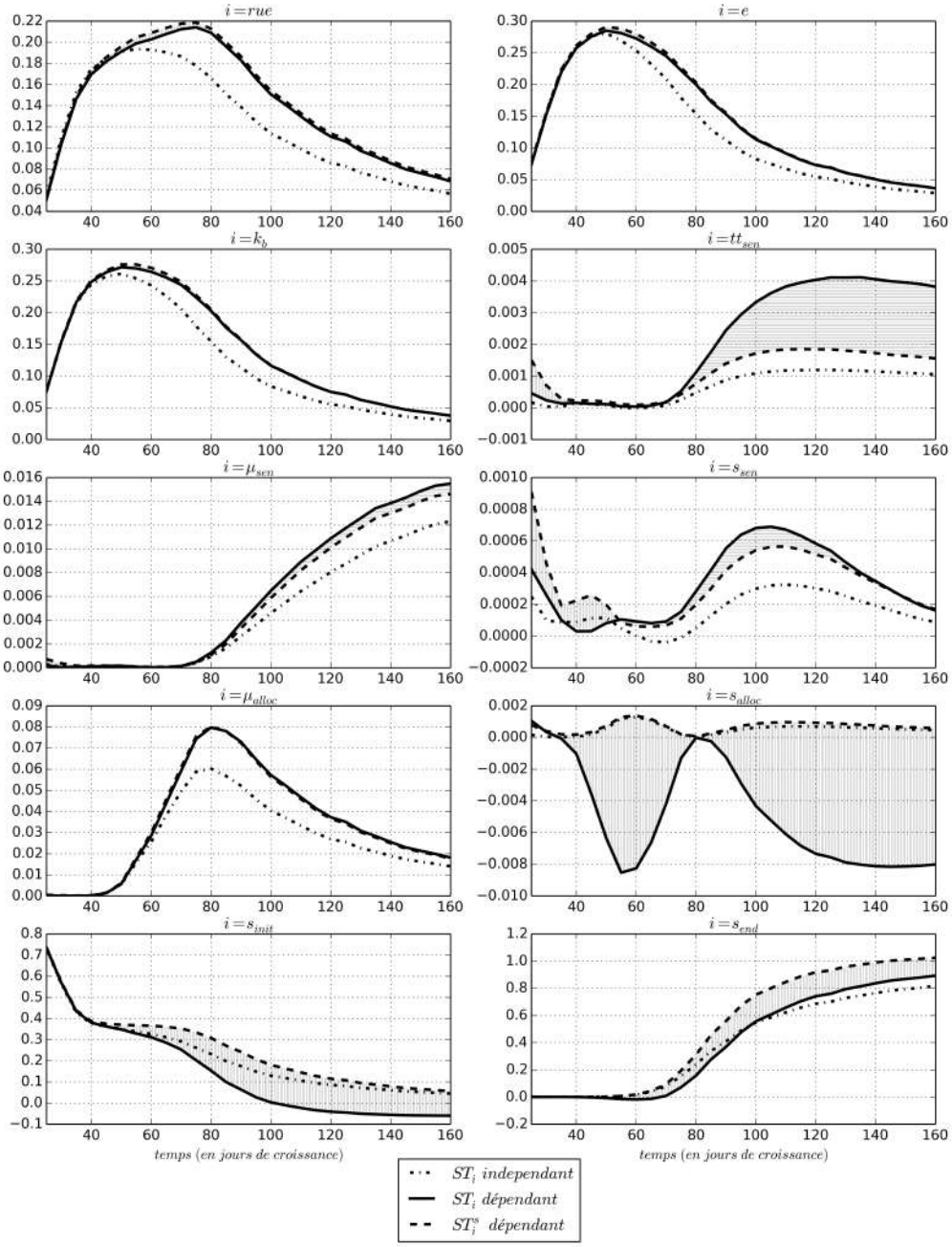

FIGURE 6. *Dynamique des indices de Sobol totaux ST_i et des indices totaux structurels ST_i^s associés à la variable* biomasse foliaire sèche *du modèle LNAS pour chacun des paramètres i (cas d'étude 4). Les indices obtenus dans le cas indépendant sont également présentés. Ils ont étés déterminés en supposant que les paramètres ont les mêmes lois marginales gaussienne que dans le cas dépendant. L'impact des corrélations est donné par les zones hachurées : i) horizontalement pour une augmentation d'influence et ii) verticalement pour une réduction d'influence . Ces résultats ont été obtenus par HOGS-OLS avec $n_s = 50000$ échantillons.*

 ISSN: 2102-6238

qui sera discuté plus loin. Considérons pour le moment les temps supérieurs à 40. Dans ce cas d'étude, OLS est la méthode de régression offrant la meilleure consistance comparativement aux deux régresseurs pénalisés. En étudiant les résultats plus précisément pour $n_s = 500$ et $n_s = 1000$, il apparaît que peu de termes significatifs sont choisis par le lasso et EN dans la base fonctionnelle. Nous sommes donc à nouveau dans une situation de sparsité artificielle similaire à celle rencontrée dans le cas d'étude précédent. Remarquons qu'en reprenant la formule (16) et la configuration utilisée ici, l'amendement de la méthode conduirait à $\hat{n}_b = 275$. Cette valeur est largement inférieure aux deux tailles d'échantillons proposées ici. En revanche, dans le cas où $n_s = 100$, la qualité de la méta-modélisation est fortement dégradée pour le lasso qui montre une forte instabilité. Dans cette situation, EN a un meilleur comportement et ses performances sont équivalentes à celles obtenues avec davantage d'échantillons. Nous retrouvons donc les constatations de la section précédente qui incitent à l'amendement de la méthode HOGS et à privilégier l'utilisation d'EN dans les situations de forte sparsité.

Outre les aspects de convergence des estimateurs pénalisés, l'indice de consistance permet également d'évaluer si le degré maximal d'interaction d choisi est pertinent. En effet, vu les cas d'étude précédents, il semble raisonnable de considérer que l'indice de consistance obtenu avec HOGS-OLS et $n_s = 100000$ est très proche de sa valeur théorique. Aussi, si le degré polynomial L est élevé comme dans le cas d'étude précédent et si l'écart entre la valeur de i_c et 1 est significatif, il est raisonnable d'attribuer cette différence à l'absence de termes d'interaction d'ordre élevé dans la base de projection. Leur estimation peut s'avérer nécessaire et l'amendement proposé pour la méthode HOGS précédemment est d'autant plus pertinent qu'il allégerait l'estimation de ces termes.

Revenons maintenant aux premiers temps de l'analyse (<40) où l'indice de consistance se dégrade significativement. Dans cette période, la variable *biomasse foliaire sèche* est proche de 0 (fig. 4). Par construction du modèle LNAS, l'augmentation de la biomasse n'est possible qu'à partir d'une date seuil fixée et la variance de la *biomasse foliaire sèche* est par suite nécessairement nulle avant cette date. Le modèle étant continu ainsi que les distributions des paramètres, l'augmentation la variance au cours du temps est également continue. Aussi, plus l'analyse de la variance est effectuée dans un voisinage de la date seuil, plus les différentes approximations faites par la méthode (erreur machine, échantillonnage Monte–Carlo) sont significatives relativement à la variance à étudier. Dans ce cas particulier, la perte de consistance n'est donc pas à attribuer à un mauvais choix de configuration mais plutôt aux limites numériques. Il est par ailleurs important de souligner que l'étude d'une variance très faible n'a pas forcément beaucoup d'intérêt.

L'analyse comparée dans les cas indépendant et dépendant des indices de sensibilité obtenus avec le meilleur indice de consistance (HOGS-OLS et $n_s = 50000$) est donnée par la figure 6. Tout d'abord, les résultats montrent que les paramètres tt_{sen}, s_{sen} et s_{alloc} ont une influence minime que ce soit dans les cas indépendant et dépendant. Considérons maintenant les résultats obtenus dans le cas dépendant. Nous retrouvons que les paramètres *rue*, e et k_b supposés indépendants le sont effectivement (en effet $ST_i = ST_i^s$). Notons que le paramètre μ_{alloc} qui est faiblement dépendant (corrélé à s_{alloc}) se comporte comme s'il était indépendant. Ceci s'explique aussi par le fait que s_{alloc} n'a presque pas d'influence. Le même phénomène s'observe pour μ_{sen} bien qu'il soit fortement corrélé à tt_{sen} et s_{sen} qui sont très peu influents.

 ISSN: 2102-6238

Finalement, dès lors que les paramètres peu influents sont exclus, nous nous retrouvons dans une situation où cinq paramètres (rue, e, k_b, μ_{sen}, μ_{alloc}) peuvent être considérés comme étant indépendants et deux paramètres sont corrélés (s_{init}, s_{end}). Après le 40^{eme} jour de croissance, l'introduction de la dépendance a augmenté l'influence des cinq paramètres indépendants. Ceci s'explique par la forte réduction de l'indice total associé à s_{init} par rapport au cas indépendant et qui est due à l'introduction des corrélations. Dans ce cas d'étude, nous observons que l'introduction des corrélations a pour effet de réduire l'influence d'un groupe de paramètres et par suite, d'augmenter l'influence des autres. Ceci est un résultat significatif et encourageant la poursuite de l'application de HOGS à différents modèles de croissance des plantes.

5. Conclusion et perspectives

Cette étude a permis de mettre en évidence plusieurs propriétés numériques de HOGS. Du point de vue des performances, les temps de calculs observés et la meilleure qualité des estimations proposées par HOGS-OLS suggèrent de privilégier son utilisation. La régression ridge est quant à elle à proscrire.

L'indice de consistance est un bon indicateur de la qualité de la surface de réponse estimée par HOGS et l'influence de la configuration de l'algorithme sur son comportement a clairement été mise en évidence surtout lorsque le modèle n'est pas linéaire.

Une constatation importante de cet article est que HOGS génère une quantité très importante de fonctions de base ce qui a pour conséquence de créer des situations où la sparsité est artificielle. Dans ce cas de figure, le lasso et EN ont des performances identiques en matière d'estimation des indices de Sobol. En revanche, EN a un coût plus élevé en temps de calcul. Du point de vue de HOGS, la limitation du degré maximal des polynômes pour estimer les fonctions d'ordre supérieur à 1 permettrait de réduire la sparsité artificielle. L'étude menée dans le cas du modèle d'Ishigami semble indiquer que cette solution est pertinente et mérite une étude approfondie. Les travaux ultérieurs que nous conduirons tiendront compte de ce constat.

Concernant les estimateurs pénalisés, les estimations obtenues avec EN sont clairement meilleures dans le cas où la sparsité réelle est avérée (cas d'étude 2 et 4) mais requièrent plus de ressources numériques. Ainsi, il est préférable d'utiliser tout de même le lasso dans un premier temps et de s'appuyer sur l'indice de consistance qui est un bon indicateur de la qualité de la surface de réponse estimée.

Finalement, nous constatons dans tous les cas d'étude l'importance de bien choisir la configuration de HOGS. Lors de l'étude d'un modèle quelconque avec un grand nombre de paramètres, l'utilisation d'une configuration faible (d et L petits) et de HOGS-lasso peut permettre d'identifier les entrées les plus influentes et de réduire la dimension du problème en excluant les paramètres sans influence. Suivant la stabilité de l'indice de consistance obtenu avec le lasso, EN pourra être privilégié. Une réduction suffisante permettrait d'appliquer HOGS-OLS à un groupe réduit de paramètres afin d'obtenir des résultats plus fins. Ainsi, nous suggérons d'utiliser la méthode HOGS de manière graduelle, en plusieurs étapes si nécessaire. Cette approche permet de faire un compromis entre les ressources numériques et la qualité de l'analyse de sensibilité.

Il ressort de cette étude que l'utilisation et l'interprétation de la méthode HOGS n'est pas triviale. Néanmoins, l'indice de consistance, une bonne connaissance à priori du modèle et la méthodologie proposée ci-dessus permettent de guider le modélisateur dans la configuration de

 ISSN: 2102-6238

HOGS. Cette méthode offre des résultats et des perspectives encourageantes pour l'étude de nombreux modèles, comme par exemple les modèles de croissance de plantes dont la dépendance entre paramètres est un élément essentiel.

Remerciements

Les auteurs souhaitent tout particulièrement remercier les rapporteurs dont les diverses remarques ont permis une amélioration significative du contenu de cet article. Elles ont notamment permis d'approfondir l'analyse de la méthode HOGS et de proposer un amendement de cette dernière.

Références

Bezanson, J., Edelman, A., Karpinski, S. et Shah, V. B. (2014). Julia : A fresh approach to numerical computing. *CoRR*, abs/1411.1607.

Caniou, Y. (2012). *Global sensitivity analysis for nested and multiscale modelling*. Thèse de doctorat, Université Blaise Pascal - Clermont-Ferrand II.

Champion, M., Chastaing, G., Gadat, S. et Prieur, C. (2015). L2-Boosting for sensitivity analysis with dependent inputs. 25(4):1477–1502. Statistica Sinica.

Chastaing, G. (2013). *Indices de Sobol généralisés pour variables dépendantes*. Thèse de doctorat, Université de Grenoble.

Chastaing, G., Gamboa, F. et Prieur, C. (2015). Generalized sobol sensitivity indices for dependent variables : numerical methods. *Journal of Statistical Computation and Simulation*, 85(7):1306–1333.

Chastaing, G., Gamboa, F., Prieur, C. *et al.* (2012). Generalized hoeffding-sobol decomposition for dependent variables-application to sensitivity analysis. *Electronic Journal of Statistics*, 6:2420–2448.

Chen, Y. et Cournède, P.-H. (2014). Data assimilation to reduce uncertainty of crop model prediction with convolution particle filtering. *Ecological Modelling*, 290:165–177.

Cournède, P.-H., Chen, Y., Wu, Q., Baey, C. et Bayol, B. (2013). Development and evaluation of plant growth models : Methodology and implementation in the pygmalion platform. *Mathematical Modelling of Natural Phenomena*, 8(4):112–130.

Friedman, J., Hastie, T. et Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(February):1–22.

Hastie, T., Tibshirani, R. et Friedman, J. (2009). *The elements of statistical learning : data mining, inference and prediction*. Springer, 2 édition.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325.

Hooker, G. (2007). Generalized Functional ANOVA Diagnostics for High-Dimensional Functions of Dependent Variables. *Journal of Computational and Graphical Statistics*, 16(November):709–732.

Kucherenko, S., Tarantola, S. et Annoni, P. (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183(4):937–946.

Lamboni, M., Monod, H. et Makowski, D. (2011). Multivariate sensitivity analysis to measure global contribution of input factors in dynamic models. *Reliability Engineering & System Safety*, 96(4):450–459.

Lecoeur, J., Poiré-Lassus, R., Christophe, A., Pallas, B., Casadebaig, P., Debaeke, P., Vear, F. et Guilioni, L. (2011). Quantifying physiological determinants of genetic variation for yield potential in sunflower. sunflo : a model-based analysis. *Functional Plant Biology*, (38):246–259.

Letort, V., Mahe, P., Cournède, P.-H., de Reffye, P. et Courtois, B. (2008). Quantitative genetics and functional-structural plant growth models : Simulation of quantitative trait loci detection for model parameters and application to potential yield optimization. *Annals of Botany*, 101(8):951–963.

Li, G. et Rabitz, H. (2012). General formulation of HDMR component functions with independent and correlated variables. *Journal of Mathematical Chemistry*, 50(1):99–130.

Li, G. et Rabitz, H. (2014). Analytical HDMR formulas for functions expressed as quadratic polynomials with a multivariate normal distribution. *Journal of Mathematical Chemistry*, 52(8):2052–2073.

 ISSN: 2102-6238

Li, G., Rabitz, H., Yelvington, P. E., Oluwole, O. O., Bacon, F., Kolb, C. E. et Schoendorf, J. (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *The journal of physical chemistry. A*, 114(19):6022–32.

Li, G., Rey-de Castro, R. et Rabitz, H. (2012). D-MORPH regression for modeling with fewer unknown parameters than observation data. *Journal of Mathematical Chemistry*, 50(7):1747–1764.

Li, G., Rosenthal, C. et Rabitz, H. (2001). High Dimensional Model Representations. *The Journal of Physical Chemistry A*, 105(33):7765–7777.

Matsumoto, M. et Nishimura, T. (1998). Mersenne twister : A 623-dimensionally equidistributed uniform pseudo-random number generator. *ACM Trans. Model. Comput. Simul.*, 8(1):3–30.

Monod, H., Naud, C. et Makowski, D. (2006). Uncertainty and sensitivity analysis for crop models. *Working with dynamic crop models : Evaluation, analysis, parameterization, and applications*, 4:55–100.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M. et Tarantola, S. (2008). *Global sensitivity analysis : the primer*. John Wiley & Sons.

Sobol, I. (1993). Sensitivity analysis for non-linear mathematical models. *Mathematical Modeling and Computational Experiment*, 1:407–414.

Stone, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, 22(1):118–184.

Tardieu, F. (2003). Virtual plants : modelling as a tool for the genomics of tolerance to water deficit. *Trends in Plant Science*, 8(1):9–14.

Wu, Q.-L., Cournède, P.-H. et Mathieu, A. (2012). An efficient computational method for global sensitivity analysis and its application to tree growth modelling. *Reliability Engineering & System Safety*, 107:35 – 43.

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

 ISSN: 2102-6238