

Indices de sensibilité, sélection de paramètres et erreur quadratique de prédiction : des liaisons dangereuses ?

Title: Sensitivity indices, parameter selection and squared error of prediction : dangerous liaisons?

Matieyendou Lamboni¹, David Makowski² et Hervé Monod³

Résumé : Lorsqu'un modèle contient un grand nombre de paramètres, l'analyse de sensibilité globale est souvent utilisée pour sélectionner les paramètres à estimer parmi ceux identifiés comme les plus influents. Une telle procédure de sélection est basée sur des données simulées et se distingue de la procédure de validation du modèle qui est basée sur des données réelles. Néanmoins, ces deux procédures sont liées dans leurs objectifs et il est intéressant d'évaluer les bénéfices de la sélection par analyse de sensibilité à l'aide des critères EQMP (Erreur Quadratique Moyenne de Prédiction) et EQM (Erreur Quadratique Moyenne). Dans cet article, nous formalisons d'abord la démarche consistant à sélectionner les paramètres à estimer par analyse de sensibilité et à fixer les autres paramètres à leur valeur nominale. Dans le cadre du modèle linéaire, nous explicitons ensuite, de façon exacte, les liens formels qui existent entre les indices de sensibilité des paramètres et la qualité de prédiction mesurée par les critères d'évaluation EQM et EQMP. Nous complétons ces résultats par des simulations pour étudier l'impact sur la qualité prédictive du modèle, d'une part du plan d'expériences (variables d'entrées du modèle), d'autre part du point où l'analyse de sensibilité est effectuée. Dans ces simulations, l'approche de sélection de paramètres par analyse de sensibilité est comparée à la méthode LASSO qui sert de référence pour sélectionner des modèles creux. Les résultats montrent qu'estimer les paramètres les plus influents contribue à la réduction de l'EQM et de l'EQMP mais que cette réduction n'est pas systématique. En effet, la relation entre l'EQMP et les indices de sensibilité est complexe et elle dépend fortement du plan d'expériences. Par exemple, seul un plan d'expériences orthogonal garantit une réduction systématique de l'EQMP. De plus, les résultats dépendent des points supports de l'analyse de sensibilité. La performance de la sélection des paramètres par l'analyse de sensibilité est équivalente à celle de LASSO en termes de l'EQMP si nous disposons *a priori* des connaissances pertinentes sur le degré d'incertitudes des différents paramètres pour conduire l'analyse de sensibilité. Les conséquences pratiques des résultats font l'objet d'une discussion en fin d'article.

Abstract: When a model contains a large number of parameters, sensitivity analysis is often used to select the parameters to be estimated among those identified as the most influential. This selection procedure is based on simulated data and is different from the model validation procedure that is based on real data. Nevertheless, these two processes are interrelated in their objectives and it is interesting to quantify the benefit of this practice in terms of MSE (Mean

¹ ANSES-DER ; 27-31, avenue du Général Leclerc, BP 19 - 94701 Maisons-Alfort Cedex, France.

E-mail : matieyendou.lamboni@gmail.com

² INRA, UMR 211 INRA AgroParisTech, BP 01, F78850, Thiverval-Grignon, France.

E-mail : david.makowski@grignon.inra.fr

³ INRA, UR 341 MIA-Jouy, Domaine de Vilvert, F78352 Jouy-en-Josas, France.

E-mail : herve.monod@jouy.inra.fr

Square Error of Prediction) and MSE (Mean Square Error) criteria. In this paper, we investigate the relationship between the model validation criteria and the sensitivity indices. We first formalize the process of selecting the parameters to be estimated by sensitivity analysis and of fixing other parameters at their nominal value. Under the linear model, we show an explicit relationship between the sensitivity indices of model parameters and the model quality criteria such as MSE and MSEP. We also study the impact on prediction quality of both the design of experiments (input variables of the model) and the point where sensitivity analysis is performed. In these simulations, we compare the procedure of parameters selection by sensitivity indices and the LASSO method well suited for sparse model. The results show that estimating the most influent parameters reduces the MSE and the MSEP all things being equal. However this reduction is not systematic. Indeed, the relationship between MSEP and sensitivity indices is complex and depends heavily on experimental design. For example, only an orthogonal experimental design ensures a systematic reduction of MSEP. Moreover, the results depend on the support points of sensitivity analysis. The performance of the parameters selection by sensitivity analysis is equivalent to that of LASSO in terms of MSEP if we have relevant prior knowledge on the degree of uncertainty in different parameters to perform the sensitivity analysis. The practical implications of the results are discussed at the end of the paper.

Mots-clés : analyse de sensibilité globale, EQM, EQMP, LASSO, plan d'expériences, sélection des paramètres

Keywords: global sensitivity analysis, MSE, MSEP, LASSO, design of experiments, parameter selection

Classification AMS 2000 : 62F07, 62G05, 62K20

1. Introduction

De plus en plus fréquemment, des modèles quantitatifs comportant de nombreux paramètres et variables d'entrée sont utilisés pour anticiper des phénomènes et aider à la décision. Dans ces situations, les modélisateurs se trouvent confrontés à de multiples sources d'incertitude notamment l'incertitude sur les entrées, sur les paramètres, et sur les conditions de fonctionnement du modèle (de Rocquigny, 2006 [8], [9]). La prise en compte de ces différentes sources d'incertitudes et de variabilité est indispensable, que ce soit lors de la mise au point du modèle ou lors de son exploitation pour la prédiction, la préconisation, la gestion des risques ou la prise de décision.

L'estimation des paramètres est une étape cruciale dans le processus de modélisation puisqu'elle permet de réduire l'incertitude sur les paramètres et influence ainsi la performance du modèle (Butterbach-Bach *et al.*, 2004 [6]). Une des caractéristiques de la modélisation des phénomènes industriels, hydrologiques, agronomiques et autres réside dans le fait que le nombre d de paramètres incertains est souvent très élevé et ne s'accompagne pas d'un nombre n d'observations proportionnel, pour des raisons de coûts et de difficultés de mesures (Brun *et al.*, 2001 [5]; Bechini *et al.*, 2006 [3]; de Rocquigny, 2006 [8], [9]). Dans cette configuration ($n \ll d$), l'estimation des paramètres sans réduction de dimension n'est pas performante (Fort *et al.*, 2005 [11]; Saporta, 2006 [23]). Dans le cas simple d'un modèle linéaire, les techniques de réduction de dimension ont fait l'objet de nombreux développements récents tels que la méthode "partial least squares" (PLS) (Frank et Friedman, 1993 [12]); la méthode LASSO (Tibshirani, 1996 [26]); plus généralement la méthode LARS (Efron *et al.*, 2004 [10]; Zou et Hastie, 2005 [30]); plus toute la gamme de régressions régularisées ou pénalisées (Hoerl et Kennard, 1970 [13]; Frank et Friedman, 1993

[12]). Mais ces méthodes ne sont pas adaptées à des modèles complexes, dynamiques, et non linéaires, coûteux en temps de simulation. De même, une approche bayésienne est difficile à mettre en œuvre sur de tels modèles (voir par exemple Brun *et al.*, 2001 [5], Wallach *et al.*, 2001 [27], Wallach *et al.*, 2002 [28]). Même si ce domaine évolue rapidement, l'approche bayésienne est restreinte aux modèles à faible coût de simulations avec un nombre de paramètres relativement petit. Il est aussi connu que l'estimation bayésienne en présence d'un petit nombre d'observations peut fournir des distributions *a posteriori* peu précises (Robert, 2006 [21]).

En présence d'un modèle complexe ou d'un code de calcul, les modélisateurs sont souvent conduits à estimer un sous-groupe de paramètres sélectionnés selon des critères plus ou moins objectifs (de Rocquigny, 2006 [8], [9]). L'analyse de sensibilité (Saltelli *et al.*, 2000 [7] ; Monod *et al.*, 2006 [18]) est une des techniques les plus utilisées pour effectuer une telle sélection (Perrin *et al.*, (2001) [19] ; Wallach *et al.*, 2002 [28] ; Makowski *et al.*, 2006b [17] ; Lamboni *et al.*, 2009 [16]). La procédure de sélection des paramètres les plus influents par analyse de sensibilité est basée sur des données simulées. Elle est réalisée indépendamment de la procédure de validation du modèle qui est basée sur les données réelles. Néanmoins, ces deux procédures sont liées dans leurs objectifs et il est pertinent d'évaluer l'intérêt de la sélection par analyse de sensibilité avec des critères de qualité de modèle tels que l'EQMP (Erreur Quadratique Moyenne de Prédiction) et l'EQM (Erreur Quadratique Moyenne).

Cet article a pour ambition d'évaluer de façon rigoureuse la qualité de modèles dont seuls les paramètres les plus influents sont estimés, après sélection par analyse de sensibilité. Nous nous plaçons dans le cadre du modèle linéaire, tout en nous intéressant aux implications pour des situations plus générales. Dans la Section 2, nous formalisons le cadre générique de la modélisation complexe et sur-paramétrée. Dans la Section 3, nous explicitons de façon exacte, dans le cas du modèle linéaire, les liens formels qui existent entre les indices de sensibilité et la qualité d'estimation et de prédiction mesurée par les critères EQM et EQMP. Ces résultats sont complétés par des simulations pour étudier l'impact sur la qualité prédictive du modèle, d'une part du plan d'expériences (variables d'entrées du modèle), d'autre part du point où l'analyse de sensibilité est effectuée (Section 4). Dans ces simulations, l'approche de sélection de paramètres par analyse de sensibilité est comparée à la méthode LASSO qui sert de référence pour sélectionner des modèles creux. Enfin la Section 5 est consacrée à la discussion des résultats.

2. Formalisation de la sélection de paramètres par analyse de sensibilité

Nous commençons par définir un cadre formel pour décrire la démarche d'un modélisateur qui sélectionne les paramètres à estimer par analyse de sensibilité. Après avoir défini cinq postulats, nous précisons les deux conditions sous lesquelles les résultats des sections suivantes sont obtenus.

2.1. Postulats sur le modèle

Nous considérons une situation où le phénomène que l'on veut prédire et le phénomène observé sont en étroite concordance.

P1 (vrai modèle) le phénomène à prédire est représenté par la fonction de réponse suivante, considérée comme le “vrai” modèle :

$$m^* = f(\mathbf{x}, \beta^*), \quad (1)$$

où \mathbf{x} est le vecteur de variables d'entrée et β^* le vecteur de paramètres de dimension d . Nous nous plaçons dans un cadre fréquentiste, selon lequel β^* représente une valeur précise mais inconnue des paramètres β_j , pour $j \in \{1, 2, \dots, d\}$.

P2 (modèle statistique) le phénomène à prédire est observable selon le modèle statistique :

$$y = f(\mathbf{x}, \beta^*) + \varepsilon,$$

où y est une observation du phénomène au point \mathbf{x} et ε représente une erreur d'observation. Les erreurs d'observation sont supposées indépendantes et identiquement distribuées, centrées et de variance σ_ε^2 .

2.2. Postulats sur la démarche du modélisateur

Les trois postulats suivants décrivent la démarche classique dans le cas où le modèle représente fidèlement le phénomène étudié et où l'on sélectionne les paramètres à estimer par analyse de sensibilité.

P3 (informations a priori du modélisateur) le modélisateur connaît la fonction $f(.,.)$ du modèle (1) mais il ignore la vraie valeur β_j^* des paramètres β_j , pour $j \in \{1, 2, \dots, d\}$. Grâce à des experts du domaine, il dispose néanmoins d'une distribution de probabilité *a priori* sur chaque β_j . La distribution *a priori* de β_j est caractérisée par son espérance ou valeur nominale b_j et par sa variance s_j^2 reflétant le degré d'incertitude exprimé par l'expert. Pour le modélisateur, les valeurs b_j sont fixées. Mais elles sont considérées aussi, dans notre étude, comme des réalisations de variables aléatoires reflétant une “population” virtuelle de situations d'expertise analogues.

P4 (objectifs du modélisateur) le modélisateur doit réaliser une prédiction \hat{m}° de la réponse du modèle en un point particulier des variables d'entrée que nous notons \mathbf{x}° .

P5 (stratégie du modélisateur) le modélisateur dispose d'un jeu de données (\mathbf{x}_i, y_i) , avec $i \in \{1, 2, \dots, n\}$. Ce jeu étant considéré insuffisant pour estimer tous les paramètres inconnus, il utilise les indices de sensibilité globale de premier ordre des paramètres β_j (définis plus bas), mesurés en \mathbf{x}° , comme critère objectif pour sélectionner les paramètres clés à estimer. Il fixe les autres paramètres à leurs valeurs nominales b_j .

2.3. Conditions supplémentaires

Deux conditions sont nécessaires pour obtenir les principaux résultats des sections suivantes. L'une porte sur le modèle, et l'autre sur les incertitudes.

C1 (linéarité du modèle vrai) la fonction de réponse du postulat P1 est un modèle linéaire multiple défini par :

$$f(\mathbf{x}, \boldsymbol{\beta}) = \mathbf{x}'\boldsymbol{\beta}, \quad (2)$$

avec $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)'$ le vecteur de d paramètres inconnus et $\mathbf{x} = (x_1, \dots, x_d)'$ le vecteur de variables d'entrée.

C2 (qualité de l'information a priori) le degré d'incertitude exprimé par l'expert sur chaque paramètre β_j , défini dans le postulat P3, vérifie

$$s_j^2 = \mathbb{E}(b_j - \beta_j^*)^2,$$

où $\mathbb{E}(b_j - \beta_j^*)^2$ est l'écart quadratique moyen entre la valeur nominale b_j fournie par l'expert et la vraie valeur de β_j . L'espérance est ici définie par rapport à la distribution de b_j dans la population de situations d'expertise. En d'autres termes, on suppose que l'expert évalue de façon non biaisée le manque de précision s_j^2 de chacune de ses valeurs nominales b_j , au sens de l'écart quadratique moyen.

2.4. Commentaires sur la condition C1 de linéarité

Dans la suite de l'article, le modèle $f(.,.)$ est un modèle linéaire de régression multiple (condition C1), ce qui permet d'avoir des expressions explicites des estimateurs des paramètres, des indices de sensibilité et des critères de qualité du modèle. Notre but principal est d'établir des relations formelles entre les indices de sensibilité et les critères de validation de modèle, notamment l'EQM et l'EQMP. Le lien avec des modèles plus complexes peut s'envisager en considérant le modèle (2) comme un méta-modèle linéaire ou comme une linéarisation par développement limité d'un modèle non linéaire $g(.,.)$ plus complexe. Considérons en effet le modèle

$$z = g(\mathbf{u}; \boldsymbol{\theta})$$

et soit \mathbf{t} un vecteur de valeurs nominales des paramètres $\boldsymbol{\theta}$. Alors un développement limité d'ordre 1 autour de \mathbf{t} conduit au modèle linéaire par rapport aux paramètres

$$m = \boldsymbol{\beta}'\mathbf{x},$$

avec

$$\begin{cases} m &= g(\mathbf{u}; \boldsymbol{\theta}) - g(\mathbf{u}; \mathbf{t}) \\ \boldsymbol{\beta} &= \boldsymbol{\theta} - \mathbf{t} \\ x_j &= \left(\frac{\partial g(\mathbf{u}; \boldsymbol{\theta})}{\partial \theta_j} \right)_{\mathbf{t}}. \end{cases}$$

D'après ce développement, la valeur nominale qui s'impose pour chaque paramètre β_j est $b_j = 0$, puisque la valeur la plus plausible de θ pour l'expert est \mathbf{t} . C'est cette valeur nominale qui sera utilisée dans l'étude par simulations de la Section 4.

2.5. Comparaison avec le cadre bayésien

La stratégie décrite dans Brun *et al.* (2002) [4] et formalisée par les postulats P1-P5 est largement utilisée dans la littérature pour résoudre le problème d'estimation des modèles sur-paramétrés. Dans des travaux tels que ceux de Kennedy *et al.* (2001) [14], le modèle du phénomène d'intérêt, le modèle développé par le modélisateur, et le modèle des observations peuvent être différents. Ici par contre, les trois modèles sont supposés aussi cohérents que possible (postulats P1, P2 et P3). Nous nous plaçons volontairement dans un cadre où le modèle est très bien adapté au phénomène étudié.

La sélection et l'estimation de paramètres formalisées par les postulats P1-P5 s'apparentent à de l'inférence bayésienne, puisqu'on associe dans les deux cas une distribution *a priori* au vecteur de paramètres β . Mais dans P1-P5, la distribution *a priori* n'est utilisée que pour sélectionner les paramètres à estimer, par analyse de sensibilité indépendante des données. Puis les données sont utilisées pour estimer les paramètres sélectionnés, sans tenir compte de leur distribution *a priori*. L'estimation bayésienne (Robert, 2006 [21]), par contre, exploite conjointement l'information *a priori* et l'information apportée par les données, en appliquant la formule de Bayes pour obtenir la distribution *a posteriori* des paramètres et, éventuellement, leur sélection.

3. Liens entre la qualité de prédiction et les indices

3.1. Indices de sensibilité

Nous utilisons la définition probabiliste des indices de sensibilité globale (Sobol, 1993 [24]; Saltelli *et al.*, 2008 [22]), basée sur la décomposition de la variance des sorties du modèle. Pour des raisons de simplicité, les indices ne sont pas normalisés ici par la variance marginale des sorties. Conformément à la notion de *factor priorization* (Saltelli *et al.*, 2008 [22]), ce sont les indices de sensibilité du 1er ordre qui sont utilisés pour sélectionner les paramètres à estimer.

Définition 3.1. Pour la fonction réponse (1) et pour $j \in \{1, 2, \dots, d\}$, l'indice de sensibilité globale de 1er ordre du paramètre β_j au point $\mathbf{x}^\circ = (x_1^\circ, x_2^\circ, \dots, x_d^\circ)'$ des variables d'entrée, est défini par

$$\mathbb{I}\mathbb{S}_{\beta_j} = \text{Var}_{\text{prior}} [\mathbb{E}_{\text{prior}}(f(\mathbf{x}^\circ, \beta) | \beta_j)],$$

où $\mathbb{E}_{\text{prior}}$ et $\text{Var}_{\text{prior}}$ désignent l'espérance et la variance pour la distribution *a priori* des paramètres β . Dans cette distribution, les paramètres β_j sont supposés indépendants.

Plusieurs méthodes d'estimation des indices de sensibilité sont décrites dans Saltelli *et al.*, 2008 [22]. Ces méthodes nécessitent des simulations du modèle, associées à des tirages des valeurs des paramètres inconnus β dans leur loi de distribution *a priori*. Dans le cas du modèle linéaire, cependant, les indices se calculent analytiquement.

Proposition 3.1. *Dans le cas particulier du modèle linéaire (2) de la condition C1, les indices de sensibilité vérifient :*

$$\mathbb{I}\mathbb{S}_{\beta_j} = s_j^2 x_j^{\circ 2}. \quad (3)$$

Démonstration. On a

$$\mathbb{E}_{\text{prior}}(f(\mathbf{x}^\circ, \beta) | \beta_j) = x_j^\circ \beta_j + \sum_{k \neq j} x_k^\circ b_k,$$

d'où

$$\begin{aligned} \mathbb{I}\mathbb{S}_{\beta_j} &= \text{Var}_{\text{prior}} \left(x_j^\circ \beta_j + \sum_{k \neq j} x_k^\circ b_k \right) \\ &= x_j^{\circ 2} \text{Var}_{\text{prior}}(\beta_j) \\ &= s_j^2 x_j^{\circ 2}. \end{aligned}$$

□

L'indépendance entre les distributions *a priori* des β_j n'est pas nécessaire pour démontrer la Proposition 3.1. Par contre, elle est nécessaire pour que la somme des indices de sensibilité (interactions comprises) soit égale à $\text{Var}_{\text{prior}}(f(\mathbf{x}^\circ, \beta))$. C'est seulement sous cette condition que les indices sont interprétables sans ambiguïté comme des parts de variance de la sortie.

Les indices de sensibilité de la Proposition 3.1 sont des fonctions croissantes de s_j^2 et $|x_j^\circ|$. Les degrés d'incertitudes s_j^2 jouent un rôle déterminant sur les indices de sensibilité et donc sur le choix des paramètres à estimer. Plus l'incertitude sur la valeur d'un paramètre est jugée importante, plus il est susceptible d'être associé au sous-groupe de paramètres les plus influents. Les indices de sensibilité dépendent également du point de prédiction x° visé par le modélisateur d'après le postulat P4.

3.2. Estimation

Selon le postulat P5, le modélisateur dispose de données observées lui permettant d'estimer le sous-ensemble des paramètres qu'il a sélectionné.

$$\text{On note } \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \text{ et } \mathcal{X} = \begin{pmatrix} \mathbf{x}'_1 \\ \mathbf{x}'_2 \\ \vdots \\ \mathbf{x}'_n \end{pmatrix},$$

avec \mathbf{y} le vecteur des observations, $\boldsymbol{\varepsilon}$ le vecteur des erreurs, \mathcal{X} la matrice des variables d'entrée, et \mathbf{x}' la transposée de \mathbf{x} . Sous les postulats P1 et P2, ces quantités vérifient

$$\mathbf{y} = f(\mathcal{X}, \boldsymbol{\beta}^*) + \boldsymbol{\varepsilon}.$$

et, sous la condition C1,

$$\mathbf{y} = \mathcal{X} \boldsymbol{\beta}^* + \boldsymbol{\varepsilon}.$$

D'après P3, le modélisateur connaît le modèle mais ignore la valeur $\boldsymbol{\beta}^*$. Pour des raisons de parcimonie, il choisit de fixer une partie des paramètres et de n'estimer que ceux dont il juge l'influence sur les prédictions prépondérante (postulat P5). Nous utiliserons dans toute la suite l'indice "e" pour désigner les quantités relatives aux paramètres estimés et l'indice "f" pour les termes liés aux paramètres fixés. Supposons sans perte de généralité que les $q < d$ premiers paramètres $\boldsymbol{\beta}_e = (\beta_1, \beta_2, \dots, \beta_q)'$ sont ceux que l'on choisit d'estimer à partir des observations, et que les paramètres $\boldsymbol{\beta}_f = (\beta_{q+1}, \beta_{q+2}, \dots, \beta_d)'$ sont ceux que l'on fixe à leurs valeurs nominales $\mathbf{b}_f = (b_j)_{j=q+1, q+2, \dots, d}$. Nous avons alors les notations suivantes :

$$\begin{aligned} \boldsymbol{\beta} &= (\boldsymbol{\beta}'_e, \boldsymbol{\beta}'_f)', \\ \mathbf{x} &= (\mathbf{x}'_e, \mathbf{x}'_f)', \\ \mathbf{x}^\circ &= (\mathbf{x}^\circ_e, \mathbf{x}^\circ_f)', \\ \mathcal{X} &= (\mathcal{X}_e, \mathcal{X}_f). \end{aligned}$$

L'estimation consiste à chercher les valeurs du vecteur de paramètres $\boldsymbol{\beta}_e$ afin que les sorties de la fonction réponse se rapprochent au mieux des observations. En prenant comme critère à minimiser la fonction de perte $\|\mathbf{y} - \mathcal{X} \boldsymbol{\beta}\|^2$ qui mesure l'écart quadratique entre les observations et les sorties de la fonction réponse, l'estimation s'écrit comme un problème d'optimisation :

$$\hat{\boldsymbol{\beta}}_e = \arg \min_{\boldsymbol{\beta}_e} \|\mathbf{y} - (\mathcal{X}_e \boldsymbol{\beta}_e + \mathcal{X}_f \mathbf{b}_f)\|^2.$$

Sous l'hypothèse que la matrice \mathcal{X}_e est de plein rang, $\hat{\boldsymbol{\beta}}_e$ s'obtient par la méthode des moindres carrés ordinaires (MCO) et vérifie :

$$\hat{\boldsymbol{\beta}}_e = (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathcal{X}_e' (\mathbf{y} - \mathcal{X}_f \mathbf{b}_f).$$

Les différentes hypothèses et propriétés des estimateurs des moindres carrés ordinaires sont exposées dans Azaïs et Bardet (2005) [2].

3.3. Qualité d'estimation

La qualité d'estimation des paramètres $\boldsymbol{\beta}_e$, obtenue en estimant $\boldsymbol{\beta}_e$ à partir des données et $\boldsymbol{\beta}_f$ par \mathbf{b}_f , est ici mesurée par l'EQM, c'est-à-dire l'écart quadratique moyen entre les paramètres

estimés et leurs vraies valeurs. Nous procédons en deux temps. Dans le Lemme 3.1, la qualité d'estimation est déterminée conditionnellement au choix de la partition (β'_e, β'_f) et à la valeur de \mathbf{b}_f utilisée comme estimation *a priori* de β_f . Puis dans la Proposition 3.2, c'est la qualité d'estimation marginale qui est considérée, par rapport à la distribution de \mathbf{b}_f dans la population virtuelle de situations d'expertise évoquée dans le postulat P3.

On note $\text{Tr}(M)$ la trace d'une matrice carrée M .

Lemme 3.1. *Sous le modèle linéaire de la condition C1, l'erreur quadratique moyenne de l'estimateur $\hat{\beta}_e$ conditionnelle à $\{\hat{\beta}_f = \mathbf{b}_f\}$, notée $\text{EQM}[\hat{\beta}_e | \hat{\beta}_f = \mathbf{b}_f]$, vérifie :*

$$\text{EQM}[\hat{\beta}_e | \hat{\beta}_f = \mathbf{b}_f] = \sigma_\varepsilon^2 \text{Tr} \left[(\mathcal{X}'_e \mathcal{X}_e)^{-1} \right] + \|\mathbf{A}(\mathbf{b}_f - \beta_f^*)\|^2, \quad (4)$$

avec $\mathbf{A} = (\mathcal{X}'_e \mathcal{X}_e)^{-1} \mathcal{X}'_e \mathcal{X}_f$, matrice d'ordre $(q, d - q)$.

Démonstration. Voir Annexe 6.1. □

Si l'on considère la population virtuelle de situations d'expertise évoquée dans le postulat P3 et la condition C2, l'EQM marginal représente l'erreur quadratique moyenne liée à la stratégie suivie. La Proposition 3.2 donne cette espérance en fonction des indices de sensibilité.

Proposition 3.2. *Sous les conditions C1 et C2, le critère EQM vérifie :*

$$\text{EQM}[\hat{\beta}_e] = \sigma_\varepsilon^2 \text{Tr} \left[(\mathcal{X}'_e \mathcal{X}_e)^{-1} \right] + \sum_{j=q+1}^d \gamma_j \text{IS}_{\beta_j}, \quad (5)$$

où $\gamma_j = \frac{1}{(x_j^\circ)^2} \sum_{i=1}^q a_{ij}^2$, $\mathbf{A} = (a_{ij})$ a été définie dans le Lemme 3.1, et x° désigne le point d'intérêt.

Démonstration. Voir Annexe 6.2. □

L'équation (5) fait apparaître des pondérations γ_j des indices de sensibilité qui dépendent du jeu de données. Remarquons que \mathbf{A} mesure le degré d'orthogonalité entre les variables explicatives associées aux paramètres estimés et celles associées aux paramètres non estimés. Ces points seront développés dans les Sections 3.5 et 3.6.

3.4. Qualité de prédiction

Selon le postulat P4, le modélisateur doit prédire le phénomène au point \mathbf{x}° c'est à dire au point où l'analyse de sensibilité a été effectuée. Il est en effet important de faire la prédiction dans les mêmes conditions que l'analyse de sensibilité, du fait que les indices de sensibilité sont dépendants des variables d'entrée \mathbf{x} du modèle. La qualité de la prédiction est ici évaluée par le critère EQMIP, c'est-à-dire l'écart quadratique moyen entre $\hat{m}^\circ = f(\mathbf{x}^\circ, \hat{\beta})$ et $m^* = f(\mathbf{x}^\circ, \beta^*)$.

Dans le Lemme 3.2, on s'intéresse tout d'abord à la qualité de prédiction conditionnelle au choix de la partition (β_e, β_f) et à la valeur de \mathbf{b}_f .

Lemme 3.2. *Sous le modèle linéaire de la condition C1, l'erreur de prédiction en \mathbf{x}° conditionnelle à $\{\hat{\beta}_f = \mathbf{b}_f\}$, notée $\text{EQMP}[\hat{m}^\circ | \hat{\beta}_f = \mathbf{b}_f]$, vérifie :*

$$\text{EQMP} \left[\hat{m}^\circ | \hat{\beta}_f = \mathbf{b}_f \right] = \sigma_\varepsilon^2 \mathbf{x}_e^{\circ'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^\circ + [\mathbf{w}' (\mathbf{b}_f - \beta_f^*)]^2 \quad (6)$$

où $\mathbf{w} = \mathbf{x}_f^\circ - \mathbf{A}' \mathbf{x}_e^\circ$, vecteur de longueur $d - q$, avec \mathbf{A} définie dans le Lemme 3.1.

Démonstration. Voir Annexe 6.3. □

Comme pour l'EQM, si l'on considère la population virtuelle de situations d'expertise évoquée dans le postulat P3 et la condition C2, l' $\text{EQMP}[\hat{m}^\circ | \hat{\beta}_f = \mathbf{b}_f]$ devient une variable aléatoire dont l'espérance représente l'erreur quadratique moyenne de prédiction en \mathbf{x}° , pour la stratégie de sélection des paramètres par analyse de sensibilité. La Proposition 3.3 permet de déterminer cette espérance en fonction des indices de sensibilité.

Proposition 3.3. *Sous les conditions C1 et C2, le critère EQMP vérifie :*

$$\text{EQMP}[\hat{m}^\circ] = \sigma_\varepsilon^2 \mathbf{x}_e^{\circ'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^\circ + \sum_{j=q+1}^d \lambda_j \text{IS}_{\beta_j}, \quad (7)$$

avec $\lambda_j = w_j^2 / (x_j^\circ)^2$, $\mathbf{w} = \mathbf{x}_f^\circ - \mathbf{A}' \mathbf{x}_e^\circ$, \mathbf{A} définie dans le Lemme 3.1, et \mathbf{x}° le point d'intérêt.

Démonstration. Voir Annexe 6.4. □

Remarque. *Le Lemme 3.2 et la Proposition 3.3 se généralisent en fait à toute combinaison linéaire $\mathbf{c}_f' \beta_f + \mathbf{c}_e' \beta_e$ des paramètres. Ils sont énoncés dans le cas particulier où $\mathbf{c}_f = \mathbf{x}_f^\circ$ et $\mathbf{c}_e = \mathbf{x}_e^\circ$.*

3.5. Commentaires

Lorsque certains paramètres du modèle sont fixés à des valeurs nominales, l'erreur quadratique de prédiction se décompose en trois termes. Les deux premiers apparaissent dans l'équation (7) de la Proposition 3.3 :

- le premier est un terme d'erreur dû à la variance d'estimation des paramètres, et il dépend donc du choix des paramètres à estimer. L'analyse de sensibilité ne fournit aucune information ni aucun contrôle sur ce terme, qui relève d'une analyse du jeu de données non prévue dans nos postulats ;
- le second terme est une fonction croissante des indices de sensibilité des paramètres non estimés, toutes choses égales par ailleurs. A première vue, il conforte la pratique qui consiste à estimer les paramètres dont les indices de sensibilité sont les plus élevés. Mais la situation est plus complexe car les indices sont pondérés par des coefficients λ_j qui dépendent du jeu de données et du point d'intérêt. Nous détaillerons l'influence de ces coefficients dans la Section 3.6.

Un troisième terme, noté $\mathbf{w}'\Delta\mathbf{w}$ dans l'Annexe 4, est absent de l'équation (7) car il est nul sous la condition C2. Ce troisième terme est lié aux écarts $s_j^2 - \mathbb{E}(b_j - \beta_j^*)^2$, pour $j \in \{1, 2, \dots, d\}$. Il apparaît si les incertitudes *a priori* sur les paramètres sont biaisées. Plus ces biais sont importants, moins les indices de sensibilité sont un indicateur fiable pour sélectionner les paramètres à estimer, surtout si ces biais modifient le classement des indices. En pratique, la sélection des paramètres par analyse de sensibilité (postulat P5) ne peut être envisagée que si la condition C2 (qualité de l'information *a priori*) est implicitement supposée valide, c'est-à-dire si l'avis des experts sur les degrés d'incertitudes est jugé suffisamment fiable.

3.6. Cas particuliers

Le second terme des erreurs moyennes d'estimation et de prédiction, dans les équations (5) et (7) respectivement, dépend non seulement des indices de sensibilité mais aussi du jeu de données et des coordonnées du point d'intérêt. Les poids γ_j de la Proposition (3.2) dépendent en particulier de la matrice $\mathbf{A} = (\mathcal{X}'_e \mathcal{X}_e)^{-1} \mathcal{X}'_e \mathcal{X}_f$. Et les poids λ_j , $j \in \{q+1, \dots, d\}$ de la Proposition (3.3) vérifient

$$\lambda_j = \left(1 - \frac{1}{x_j^o} \mathcal{X}'_j \mathcal{X}_e (\mathcal{X}'_e \mathcal{X}_e)^{-1} \mathbf{x}_e^o \right)^2.$$

La matrice \mathbf{A} dépend de la corrélation entre les variables explicatives associées aux paramètres estimés et celles associées aux paramètres non estimés. Dans le cas général où les colonnes de \mathcal{X}_f et \mathcal{X}_e sont corrélées, la relation entre l'EQMP et les indices de sensibilité est complexe. Pour mieux la comprendre et l'interpréter nous distinguons donc plusieurs cas particuliers.

3.6.1. Cas 1 : orthogonalité

Si les colonnes de la matrice \mathcal{X} sont mutuellement orthogonales ou, de façon moins contraignante, si les colonnes de \mathcal{X}_e sont orthogonales à celles de \mathcal{X}_f alors la matrice \mathbf{A} est nulle. Par conséquent, l'erreur d'estimation des paramètres β_e donnée dans l'équation (4) est réduite au minimum, et elle est indépendante des indices de sensibilité des paramètres fixés. Les coefficients λ vérifient $\lambda_j = 1$, $\forall j \in \{q+1, \dots, d\}$. Dans l'équation (7), le même poids est donc affecté aux indices de sensibilité de tous les paramètres fixés. Si les colonnes de \mathcal{X} sont non seulement orthogonales mais aussi de même norme, le premier terme de l'équation (7) ne dépend pas du choix des paramètres et il faut estimer les paramètres les plus influents pour réduire l'EQMP.

3.6.2. Cas 2 : confusion totale d'effets

Si les colonnes i de \mathcal{X}_e ($i \leq q$) et j de \mathcal{X}_f ($j > q$) sont colinéaires, on a $\mathcal{X}_j = c \mathcal{X}_i$, avec $c \in \mathbb{R}$. La combinaison linéaire $\beta_i + c \beta_j$ est la seule fonction éventuellement estimable de ces deux

paramètres. Dans la terminologie des plans factoriels (Azaïs et Bardet [2] ; Kobilinsky [15]), les paramètres β_i et β_j sont dits totalement confondus. Si, de plus, \mathcal{X}_i et \mathcal{X}_j sont orthogonales aux autres colonnes de \mathcal{X} , on a alors $\gamma_j = (c/x_j^\circ)^2$ et $\lambda_j = 1 - cx_i^\circ/x_j^\circ$.

L'indice de sensibilité \mathbb{S}_{β_j} contribue à l'EQM de façon proportionnelle au rapport $(c/x_j^\circ)^2$. Il contribue également à l'EQMIP, sauf pour les points d'intérêt qui satisfont la même contrainte de colinéarité que ceux du plan, c'est-à-dire $x_j^\circ = cx_i^\circ$. Pour ces points, l'EQMIP est identique que l'on choisisse d'estimer β_i et de fixer β_j , ou bien l'inverse. Elle est donc insensible au rapport entre les indices de sensibilité de ces deux paramètres : la confusion d'effets rend ces deux paramètres complètement redondants.

3.6.3. Cas 3 : $d = 2$ et $q = 1$

Considérons un modèle de régression avec deux variables explicatives ($d = 2$) et supposons que le premier paramètre est estimé à partir des données et que le second est fixé à sa valeur nominale. Les quantités \mathbf{A} , γ , \mathbf{w} et λ valent alors :

$$\begin{aligned} \mathbf{A} &= \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} \\ \gamma &= \left(\frac{\sum_{i=1}^n x_{i1}x_{i2}}{x_2^{\circ 2} \sum_{i=1}^n x_{i1}^2} \right)^2 \\ \mathbf{w} &= x_1^\circ \frac{\sum_{i=1}^n x_{i1}x_{i2}}{\sum_{i=1}^n x_{i1}^2} - x_2^\circ \\ \lambda &= \left(1 - \frac{x_1^\circ \sum_{i=1}^n x_{i1}x_{i2}}{x_2^\circ \sum_{i=1}^n x_{i1}^2} \right)^2. \end{aligned}$$

Cet exemple simple permet d'interpréter ce que représentent \mathbf{A} et \mathbf{w} . En posant τ_1 (resp. τ_2) l'écart type de la variable explicative X_1 (resp. X_2) et ρ le coefficient de corrélation entre les deux variables explicatives, on obtient $\mathbf{A} = \rho \frac{\tau_1}{\tau_2}$. En régression normalisée avec des variables explicatives de même variance ($\tau_1 = \tau_2$), \mathbf{A} est égal au coefficient de corrélation entre les deux variables explicatives ($\mathbf{A} = \rho$). De plus, on a $\gamma = \rho^2 x_2^{\circ -4}$, $\mathbf{w} = \rho x_1^\circ - x_2^\circ$ et $\lambda = \left(1 - \frac{x_1^\circ}{x_2^\circ} \rho \right)^2$. Dans le cas où les deux variables explicatives sont orthogonales ($\rho = 0$), l'EQMIP le plus faible est bien obtenu lorsque le paramètre à estimer est celui dont l'indice de sensibilité est le plus élevé. Dans le cas où la corrélation est élevée en valeur absolue, le meilleur choix dépend complètement des valeurs x_1° et x_2° .

4. Étude par simulations

En dehors des cas extrêmes de l'orthogonalité et de la confusion des effets, la relation entre les indices de sensibilité et l'EQMIP dépend des pondérations λ associées aux différents paramètres

fixés. Ces pondérations dépendent des corrélations entre les variables explicatives, qui changent d'un jeu de données à un autre.

En pratique, nous disposons souvent d'un jeu de données corrélées. Pour évaluer l'influence des pondérations λ sur l'EQMP selon l'équation (7), une étude par simulations a été réalisée dans des conditions où les variables explicatives sont partiellement corrélées. Une étude comparative des valeurs des pondérations λ et des indices de sensibilité à travers ces simulations permet de vérifier sur un exemple si l'estimation d'un paramètre ayant un faible indice de sensibilité pourrait réduire beaucoup plus l'EQMP que ne le laisse entendre la Proposition 3.3. Ces simulations visent aussi à comparer la stratégie décrite par les postulats P1-P5 à la méthode LASSO de sélection et estimation des paramètres, sous les conditions C1 et C2. Pour effectuer cette comparaison, nous utilisons l'un des modèles proposé par Tibshirani (1996) [26] pour illustrer la méthode LASSO. Nous nous plaçons dans les mêmes conditions que cet article pour générer les données sauf qu'en accord avec le postulat P5, on choisit des tailles d'échantillon très petites par rapport au nombre de paramètres. Ceci ne remet pas en cause l'utilisation de la méthode LASSO comme référence, du fait que cette méthode s'applique pour des tailles très petites par rapport au nombre de paramètres. Ces conditions sont détaillées dans la Section 4.1.

4.1. *Modèle et données simulées*

Nous considérons la modélisation et la démarche d'estimation décrites par les postulats P1-P5 et les conditions C1-C2. Les variables explicatives et les observations sont simulées selon Tibshirani (1996) [26]. La taille des observations est fixée à $n = 20$ conformément au postulat P5.

Les réalisations des huit variables explicatives (colonnes de la matrice \mathcal{X}) sont générées suivant une loi multivariée centrée et de matrice de variance-covariance définie par

$$\text{Cov}(X_{j_1}, X_{j_2}) = \sigma_x \times \rho^{|j_1 - j_2|}.$$

C'est la covariance d'un processus AR(1). Dans le but de pouvoir introduire de petites corrélations entre les variables explicatives, ρ est fixé à 0.5 et $\sigma_x = 3$. La plus forte corrélation entre variables X_{j_1} et X_{j_2} se produit lorsque $|j_1 - j_2| = 1$. Les vraies valeurs des paramètres sont $\beta^* = [3, 1.5, 0, 0, 2, 0, 0, 0]'$.

Les observations \mathbf{y} sont générées suivant l'équation :

$$\mathbf{y} = \mathcal{X}\beta^* + \varepsilon,$$

avec ε un vecteur gaussien de moyenne nulle et de matrice de variance-covariance $\Sigma_\varepsilon = 3 \times \mathbf{I}$ avec \mathbf{I} la matrice identité.

4.2. Points de prédiction

Les résultats théoriques ayant montré que le point de prédiction \mathbf{x}° du postulat P4 intervient dans la qualité des prédictions, quatre points de prédiction différents $\mathbf{x}^{\circ(1)}$, $\mathbf{x}^{\circ(2)}$, $\mathbf{x}^{\circ(3)}$, $\mathbf{x}^{\circ(4)}$ sont utilisés pour comparer les méthodes. Les coordonnées de ces quatre points (Table 1) ont été tirés aléatoirement et indépendamment selon la loi normale $\mathcal{N}(0,3)$, afin que les points \mathbf{x}° soient dans le même domaine que les variables explicatives du jeu de données. Nous nous plaçons donc dans le cadre d'une interpolation, en considérant que les corrélations entre variables explicatives concernent uniquement le jeu de données.

Points	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\mathbf{x}^{\circ(1)}$	2.743	-0.671	2.976	4.301	0.991	3.439	-3.261	3.145
$\mathbf{x}^{\circ(2)}$	1.131	0.905	-3.294	-3.391	-8.389	2.161	2.817	-0.688
$\mathbf{x}^{\circ(3)}$	0.538	-0.267	1.026	0.178	0.970	-1.398	-0.889	0.981
$\mathbf{x}^{\circ(4)}$	-1.879	0.551	-2.507	4.786	0.989	-2.461	1.462	2.215

TABLEAU 1. Les quatre points de prédiction considérés pour le calcul des indices de sensibilité et de l'EQMP.

4.3. Méthodes d'analyse simulées

Les indices de sensibilité sont calculés selon l'équation (3), avec des valeurs nominales fixées à $b_j = 0$ et des degrés d'incertitude sur chaque paramètre β_j fixés à $s_j = |\beta_j^*| + s_0$, où $s_0 = 0,005$ représente un degré d'incertitude minimal sur les paramètres pour éviter la situation irréaliste où les paramètres nuls seraient considérés comme connus exactement. Ces valeurs sont, à s_0 près, conformes à la condition C2, selon laquelle les modélisateurs évaluent l'incertitude de façon adéquate.

Dans la procédure de sélection des paramètres à estimer par les indices de sensibilité, il est indispensable de se donner un seuil pour déterminer le sous-groupe de paramètres à estimer. En effet, les indices de sensibilité fournissent uniquement un classement des paramètres. De même, pour la méthode LARS (LASSO) (Zou *et al.*, 2005 [30]), il est nécessaire de fixer la pénalité pour sélectionner les paramètres. Dans cette étude, nous fixons les différents seuils ou pénalités par la validation croisée (Allen, 1971 [1] ; Stone, 1974 [25] ; Yang, 2007 [29]). Pour chaque valeur du seuil, l'échantillon est divisé en 5 groupes et nous utilisons 4 groupes pour l'estimation et le dernier groupe pour la prédiction. Nous répétons cette procédure sur les 5 combinaisons de 4 groupes possibles et nous calculons l'EQMP moyen. A la fin de la procédure, nous retenons la valeur du seuil qui minimise l'EQMP moyen estimé par cette procédure. La même partition en 5 groupes est utilisée pour la méthode LASSO et pour la méthode basée sur les indices de sensibilité.

Tous les résultats des simulations sont obtenus en utilisant le logiciel statistique R (Venables and Ripley, 2003 ; R Development Core Team, 2007 [20]).

4.4. Résultats

Afin que les résultats des simulations soient fiables, 10000 échantillons de taille $n = 20$ sont générés. Pour chaque échantillon, nous disposons de 20 observations pour estimer 8 paramètres. Sur chacune des simulations, nous appliquons trois méthodes d'estimation : moindres carrés ordinaires (MCO), LASSO et enfin sélection des paramètres les plus influents par analyse de sensibilité suivie de leur estimation par la méthode des moindres carrés ordinaires (AS+MCO). Les résultats figurent dans les Tables 2 et 3 pour chaque point de prédiction considéré.

Points		β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8
$\mathbf{x}^{(1)}$	\mathbb{IS} (%)	88.4	2.7	0.2	0.3	7.7	0.2	0.2	0.2
	λ_{mean}	-	1.0	0.9	0.6	0.002	0.7	1.2	1.1
	λ_{max}	-	34.4	6.1	3.5	3.9	5.5	12.4	13.4
	Nb fixé (%)	0	9.2	97.2	68.2	0.3	81.4	88.4	93.8
	$\hat{\beta}_{AS+MCO}$	3.04	1.42	-0.003	0.004	1.99	-0.005	0.001	0.001
$\mathbf{x}^{(2)}$	\mathbb{IS} (%)	2.6	0.8	0.04	0.04	96.4	0.02	0.03	0.001
	λ_{mean}	-	0.5	0.6	0.2	-	8.5	3.2	11.2
	λ_{max}	-	83.0	14.8	6.3	-	38.0	22.1	307.2
	Nb fixé (%)	0	9.5	81.0	66.8	0	93.6	88.9	97.1
	$\hat{\beta}_{AS+MCO}$	3.04	1.399	0.005	-0.001	2.000	-0.001	0.002	-0.0008
$\mathbf{x}^{(3)}$	\mathbb{IS} (%)	30.1	3.7	0.2	0.005	65.4	0.3	0.1	0.2
	λ_{mean}	-	0.6	0.8	8.3	-	1.3	1.3	1.1
	λ_{max}	-	30.0	7.9	269.8	-	4.5	10.6	16.2
	Nb fixé (%)	0	9.4	81.5	97.4	0	67.4	93.0	89.0
	$\hat{\beta}_{AS+MCO}$	3.03	1.41	0.001	-0.001	2.007	0.006	-0.001	0.0003
$\mathbf{x}^{(4)}$	\mathbb{IS} (%)	80.2	3.4	0.2	0.9	14.8	0.2	0.1	0.2
	λ_{mean}	-	0.7	1.4	0.5	0.005	1.4	1.4	1.1
	λ_{max}	-	34.6	12.2	2.7	4.0	16.4	74.6	34.1
	Nb fixé (%)	0	9.7	81.0	67.5	0.3	88.6	97.1	93.6
	$\hat{\beta}_{AS+MCO}$	3.039	1.41	0.005	0.003	2.00	0.007	-0.0013	-0.002

TABLEAU 2. Indices de sensibilité et résultats de simulations pour la méthode AS+MCO, pour les huit paramètres et les quatre points de prédiction. \mathbb{IS} : indice de sensibilité ; λ_{mean} et λ_{max} : valeur moyenne et valeur maximum des pondérations de l'indice dans la formule de l'EQMP, sur 10000 simulations ; Nb fixé (%) : pourcentage de simulations où le paramètre a été fixé et non estimé à partir des données ; $\hat{\beta}_{AS+MCO}$: valeur moyenne de l'estimation de β par la méthode AS+MCO.

Les valeurs, notamment les indices de sensibilité (Table 2) changent en fonction du point de prédiction et nous notons même une inversion d'ordre de classification des paramètres β_1 et β_5 . Ces deux paramètres (β_1, β_5) ont les plus grands indices et sont quasi-systématiquement sélectionnés comme paramètres à estimer tandis que le troisième paramètre le plus influent (β_2) n'est sélectionné que dans 90% des simulations. Ces trois paramètres sont les trois plus importants quel que soit le point de prédiction considéré et leur estimation conduit à des EQMP faibles.

Les paramètres $\beta_3, \beta_6, \beta_7, \beta_8$ sont fixés à leur valeur nominale dans plus de 80 % des 10000 simulations réalisées et ceci pour les quatre points de prédiction considérés. Les différentes valeurs

Points de prédiction	Méthodes d'estimation	EQMP	Meilleure méthode (%)
$\mathbf{x}^{o(1)}$	MCO	34.3	15.4
	LASSO	11.4	39.1
	AS+MCO	11.7	46.2
$\mathbf{x}^{o(2)}$	MCO	44.3	24.7
	LASSO	38.9	28.8
	AS+MCO	26.3	47.8
$\mathbf{x}^{o(3)}$	MCO	3.4	18.2
	LASSO	1.6	34.7
	AS+MCO	1.4	48.0
$\mathbf{x}^{o(4)}$	MCO	26.9	13.9
	LASSO	8.0	44.4
	AS+MCO	10.4	42.3

TABLEAU 3. Valeurs moyennes de l'EQMP pour trois méthodes d'estimation et proportions des cas où la méthode a la plus petite valeur de l'EQMP sur 10000 simulations.

de pondérations λ sont très contrastées pour ces quatre paramètres dans le cas particulier $\mathbf{x}^{o(2)}$. En moyenne, la pondération λ_8 correspondant au paramètre β_8 est 20 fois plus importante que celle de λ_3 et 50 fois plus importante que celle de λ_4 . Ces différences importantes entre les valeurs de λ peuvent dégrader la réduction de l'EQMP lorsque les paramètres à estimer sont sélectionnés par analyse de sensibilité.

La moyenne des estimations des paramètres par la stratégie AS+MCO est néanmoins très proche des vraies valeurs des paramètres.

Les pourcentages obtenus par la méthode MCO (Table 3) indiquent que, dans 15% à 25% des simulations, la sélection des paramètres à estimer par analyse de sensibilité ou LASSO n'a pas réduit l'EQMP par rapport à l'estimation des huit paramètres. Ce résultat est probablement lié à des confusions d'effets ou, pour AS+MCO, à de fortes variations des valeurs des pondérations λ .

Les stratégies LASSO, AS+MCO sont au même niveau de performances en termes de l'EQMP. Chacune de ces deux stratégies conduit à l'EQMP le plus faible dans environ 40% de toutes les simulations réalisées. La performance de la méthode LASSO sur le modèle considéré n'est pas étonnante du fait que la méthode LASSO s'adapte bien aux modèles creux. Les performances similaires entre les méthodes LASSO et AS+MCO soulignent l'intérêt de l'approche AS+MCO dans ce cas de figure. Notons qu'il n'y a pas de différence entre les valeurs de l'EQMP des méthodes LASSO et AS+MCO et ceci, quelque soit le point où la prédiction est faite. Par contre la stratégie AS+MCO est meilleure que l'estimation brute MCO et l'EQMP de la méthode MCO est 2 fois plus grand que celui de la stratégie AS+MCO : AS+MCO assure un gain important en terme de l'EQMP par rapport à MCO.

5. Discussion

Lorsque la fonction de réponse à calibrer est sur-paramétrée par rapport aux observations disponibles, une pratique courante consiste à sélectionner les paramètres clés à estimer en se basant sur les indices de sensibilité globale. Dans cet article, nous avons formalisé les différents concepts utilisés par les modélisateurs qui appliquent cette pratique. A l'aide de la décomposition de l'EQM et de l'EQMP, nous avons établi une relation formelle entre les qualités du modèle et les indices de sensibilité dans le cas particulier d'un modèle linéaire. Toutes choses égales par ailleurs, estimer les paramètres les plus influents contribue à la réduction de l'EQM et de l'EQMP.

Cependant, la relation entre les indices de sensibilité et l'EQMP est complexe même pour le modèle linéaire considéré. La sélection des principaux paramètres à estimer par le biais des indices de sensibilité et la fixation des autres paramètres ne réduisent pas l'EQM et l'EQMP de manière systématique. L'EQMP par exemple se décompose en trois termes d'erreur (deux si l'on évalue correctement l'incertitude), dont l'un seulement est fonction des indices de sensibilité. De plus la relation entre l'EQMP et les indices de sensibilité dépend directement des variables explicatives à travers les pondérations des indices. Ces pondérations, qui dépendent des données disponibles, peuvent compromettre la réduction de l'EQMP lorsque l'on sélectionne les paramètres clés à estimer à l'aide des indices, sauf dans le cas particulier où les variables explicatives sont orthogonales.

La comparaison de la stratégie AS+MCO et de la méthode LASSO montre une performance équivalente entre les deux méthodes de sélection de paramètres en terme de qualité prédictive du modèle sur un modèle très bien adapté à la méthode LASSO. Cette égalité de performance n'est évidemment possible que si nous disposons *a priori* des connaissances pertinentes sur le degré d'incertitude sur les différents paramètres pour conduire l'analyse de sensibilité. Contrairement à la procédure de sélection LASSO, la méthode AS+MCO peut s'appliquer aussi bien sur des modèles creux que sur des modèles non creux qui sont fréquemment rencontrés, par exemple, en modélisation agronomique et environnementale.

Le gain en EQMP des modèles sur-paramétrés est discutable lorsque les paramètres à estimer sont sélectionnés uniquement à l'aide d'indices de sensibilité. En particulier, le jeu de données disponible a une forte influence sur les conséquences du choix des paramètres à estimer, alors que l'analyse de sensibilité est indépendante de ce jeu de données. Il serait donc intéressant de définir de nouveaux indices qui prennent en compte les pondérations et la précision des estimateurs, qui figurent dans la relation entre l'EQMP et les indices. Ainsi, l'estimation des paramètres jugés influents grâce à ces nouveaux indices pourrait contribuer à améliorer la qualité prédictive du modèle.

Les résultats obtenus dans cet article portent sur le modèle linéaire. Néanmoins ils sont également utiles pour réfléchir sur une pratique couramment appliquée à des modèles non linéaires plus complexes. Le modèle linéaire peut en effet être considéré comme une approximation par linéarisation grâce au développement limité de Taylor ou par construction d'un meta-modèle linéaire. La complexité de la relation établie entre indices de sensibilité et EQMP doit attirer l'attention du modélisateur sur la nécessité de disposer d'information *a priori* de grande qualité et de niveau d'incertitude bien évalué. Elle montre également le rôle essentiel joué par le jeu de données disponibles.

Remerciements

Cet article a bénéficié de la participation des auteurs au réseau MEXICO (reseau-mexico.fr) et au GDR MASCOT-NUM (www.gdr-mascotnum.fr), tous deux consacrés aux méthodes d'analyse de sensibilité et d'exploration numérique des modèles. Nous remercions les deux relecteurs pour leur lecture détaillée et leurs commentaires très constructifs.

6. Annexes

Les preuves présentées dans les Annexes 6.2 et 6.4 donnent des résultats plus généraux que le texte principal. Elles fournissent en effet une expression de l'EQM et l'EQMP sous les postulats P1-P5 et la condition C1, alors que les Propositions 3.2 et 3.3 sont établies sous la condition supplémentaire C2.

6.1. Démonstration du Lemme 3.1

Démonstration. Selon les postulats P1-P5 et la condition C1, le modèle de régression s'écrit

$$\mathbf{y} = \mathcal{X}_e \beta_e + \mathcal{X}_f \beta_f + \varepsilon,$$

avec β_e^* et β_f^* les vraies valeurs des paramètres. L'estimateur des moindres carrés ordinaires (MCO) de β_e conditionnellement à l'estimation *a priori* de β_f par \mathbf{b}_f est donné par

$$\hat{\beta}_e = (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathcal{X}_e' (\mathbf{y} - \mathcal{X}_f \mathbf{b}_f).$$

Le biais d'estimation de β_e dû à l'estimation *a priori* de β_f par \mathbf{b}_f vaut :

$$\begin{aligned} \mathbb{B}[\hat{\beta}_e | \hat{\beta}_f = \mathbf{b}_f] &= \mathbb{E}[\hat{\beta}_e | \hat{\beta}_f = \mathbf{b}_f] - \beta_e^* \\ &= (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathcal{X}_e' \mathcal{X}_f (\beta_f^* - \mathbf{b}_f) \\ &= \mathbf{A} (\beta_f^* - \mathbf{b}_f), \end{aligned}$$

avec $\mathbf{A} = (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathcal{X}_e' \mathcal{X}_f$ une matrice d'ordre $(q, d - q)$ dont on note a_{ij} l'élément en ligne i et colonne j .

L'erreur quadratique moyenne (EQM) de l'estimateur $\hat{\beta}_e$ vaut :

$$\begin{aligned} \text{EQM} [\hat{\beta}_e | \hat{\beta}_f = \mathbf{b}_f] &= \mathbb{E}_{\mathbf{b}_f} \left[\left\| \hat{\beta}_e - \beta_e^* \right\|^2 \right] \\ &= \mathbb{E}_{\mathbf{b}_f} \left[\left\| \hat{\beta}_e - \mathbb{E}_{\mathbf{b}_f} (\hat{\beta}_e) \right\|^2 + \left\| \mathbb{E}_{\mathbf{b}_f} (\hat{\beta}_e) - \beta_e^* \right\|^2 \right] \\ &= \text{Tr} \left\{ \mathbb{E}_{\mathbf{b}_f} \left[\left(\hat{\beta}_e - \mathbb{E}_{\mathbf{b}_f} (\hat{\beta}_e) \right) \left(\hat{\beta}_e - \mathbb{E}_{\mathbf{b}_f} (\hat{\beta}_e) \right)' \right] \right\} + \left\| \mathbf{A} (\beta_f^* - \mathbf{b}_f) \right\|^2 \\ &= \text{Tr} \left[\text{Cov}_{\mathbf{b}_f} (\hat{\beta}_e) \right] + \left\| \mathbf{A} (\beta_f^* - \mathbf{b}_f) \right\|^2 \\ &= \sigma_\varepsilon^2 \text{Tr} \left[(\mathcal{X}_e' \mathcal{X}_e)^{-1} \right] + \left\| \mathbf{A} (\beta_f^* - \mathbf{b}_f) \right\|^2, \end{aligned}$$

où $\mathbb{E}_{\mathbf{b}_f}$ désigne l'espérance conditionnelle à $\hat{\beta}_f = \mathbf{b}_f$. □

6.2. Démonstration de la Proposition 3.2

Démonstration. On considère maintenant chaque valeur nominale b_j comme une variable aléatoire par rapport à une situation virtuelle de situations d'expertise. On suppose par contre que les degrés d'incertitude s_j^2 sont fixes. Cette condition, nécessairement vérifiée sous la condition C2, assure que le classement des paramètres par leurs indices de sensibilité est fixe. Dans les résultats suivants, les espérances sont conditionnelles à la partition des paramètres entre β_e et β_f .

On note $Q = \mathbb{E}(\beta_f^* - \mathbf{b}_f)(\beta_f^* - \mathbf{b}_f)'$ la matrice des écarts quadratiques moyens de \mathbf{b}_f , Σ_f la matrice diagonale des s_j^2 et $\Delta = Q - \Sigma_f$. On obtient

$$\begin{aligned}
 \text{EQM}[\hat{\beta}_e] &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \text{Tr}\left\{\mathbb{E}\left[(\beta_f^* - \mathbf{b}_f)(\beta_f^* - \mathbf{b}_f)'\right] \mathbf{A}' \mathbf{A}\right\} \\
 &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \text{Tr}\{Q \mathbf{A}' \mathbf{A}\} \\
 &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \text{Tr}(\mathbf{A} \Sigma_f \mathbf{A}') + \text{Tr}[\mathbf{A} \Delta \mathbf{A}'] \\
 &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \sum_{i=1}^q \sum_{j=q+1}^d a_{ij}^2 s_j^2 + \text{Tr}[\mathbf{A} \Delta \mathbf{A}'] \\
 &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \sum_{i=1}^q \sum_{j=q+1}^d \frac{a_{ij}^2}{(x_j^\circ)^2} \mathbb{I} \mathbb{S}_{\beta_j} + \text{Tr}[\mathbf{A} \Delta \mathbf{A}'] \\
 &= \sigma_\varepsilon^2 \text{Tr}[(\mathcal{X}_e' \mathcal{X}_e)^{-1}] + \sum_{j=q+1}^d \gamma_j \mathbb{I} \mathbb{S}_{\beta_j} + \text{Tr}[\mathbf{A} \Delta \mathbf{A}'].
 \end{aligned}$$

Sous la condition C2, $\Delta = 0$ et l'on retrouve les résultats présentés dans le texte. \square

6.3. Démonstration du Lemme 3.2

Démonstration. La prédiction \hat{m}° au point x° conditionnellement à $\hat{\beta}_f = \mathbf{b}_f$ est définie par

$$\hat{m}^\circ = \mathbf{x}_e^\circ' \hat{\beta}_e + \mathbf{x}_f^\circ' \mathbf{b}_f,$$

et le biais de prédiction conditionnel vaut

$$\begin{aligned}
 \text{Biais}[\hat{m}^\circ | \hat{\beta}_f = \mathbf{b}_f] &= \mathbb{E}_{\mathbf{b}_f}[\hat{m}^\circ] - m^{\circ*} \\
 &= [\mathbf{x}_e^\circ' \mathbf{A} - \mathbf{x}_f^\circ'] (\beta_f^* - \mathbf{b}_f),
 \end{aligned}$$

où m^{o*} est la vraie valeur du phénomène que nous cherchons à prédire. Dans le cas particulier du modèle linéaire, l'EQMP s'écrit :

$$\begin{aligned}
 \text{EQMP} \left[\hat{m}^o \mid \hat{\beta}_f = \mathbf{b}_f \right] &= \mathbb{E}_{\mathbf{b}_f} (\hat{m}^o - m^{o*})^2 \\
 &= \text{Var}_{\mathbf{b}_f} [\hat{m}^o] + \text{Biais} \left(\hat{m}^o \mid \hat{\beta}_f = \mathbf{b}_f \right)^2 \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \left\{ [\mathbf{x}_e^{o'} \mathbf{A} - \mathbf{x}_f^{o'}] (\beta_f^* - \mathbf{b}_f) \right\}^2 \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + [\mathbf{w}' (\beta_f^* - \mathbf{b}_f)]^2,
 \end{aligned}$$

avec $\mathbf{w} = [\mathbf{x}_e^{o'} \mathbf{A} - \mathbf{x}_f^{o'}]'$.

□

6.4. Démonstration de la Proposition 3.3

Démonstration. Avec les mêmes hypothèses que dans l'Annexe 2, on obtient

$$\begin{aligned}
 \text{EQMP} [\hat{m}^o (\mathbf{B}_f)] &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \mathbb{E} [\mathbf{w}' (\beta_f^* - \mathbf{b}_f)]^2 \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \mathbf{w}' \mathbf{Q} \mathbf{w} \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \mathbf{w}' \Sigma_f \mathbf{w} + \mathbf{w}' \Delta \mathbf{w} \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \sum_{j=q+1}^p w_j^2 s_j^2 + \mathbf{w}' \Delta \mathbf{w} \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \sum_{j=q+1}^p \frac{w_j^2}{(x_j^o)^2} \mathbb{I} \mathbb{S}_{\beta_j} + \mathbf{w}' \Delta \mathbf{w} \\
 &= \sigma_\varepsilon^2 \mathbf{x}_e^{o'} (\mathcal{X}_e' \mathcal{X}_e)^{-1} \mathbf{x}_e^o + \sum_{j=q+1}^p \lambda_j \mathbb{I} \mathbb{S}_{\beta_j} + \mathbf{w}' \Delta \mathbf{w},
 \end{aligned}$$

avec $\lambda_j = \frac{w_j^2}{(x_j^o)^2}$.

Sous la condition C2, $\Delta = 0$ et l'on retrouve les résultats présentés dans le texte.

□

Références

- [1] D. M. ALLEN : Mean square error of prediction as a criterion for selecting variables. *Technometrics*, 13:469–475, 1971.
- [2] J.-M. AZAÏS et J.-M. BARDET : *Le modèle linéaire par l'exemple : régression, analyse de la variance et plans d'expériences illustrés par R, SAS et Splus*. Dunod, Paris, 2005.
- [3] L. BECHINI, S. BOCCHI, T. MAGGIORE et R. CONFALONIERI : Parameterization of a crop growth and development simulation model at sub model component level. An example for winter wheat (*Triticum aestivum L.*). *Environmental Modelling & Software*, 21:1042–1054, 2006.
- [4] R. BRUN, M. KUHNI, H. SIEGRIST, W. GUJER et P. REICHERT : Practical identifiability of ASM2d parameters - systematic selection and tuning of parameter subsets. *Water Research*, 36:4113–4127, 2002.
- [5] R. BRUN, P. REICHERT et R. KUNSCH H. : Practical identifiability of large environmental simulation models. *Water Resources Research*, 37:1015–1030, 2001.
- [6] K. BUTTERBACH-BAHL, M. KESIK, P. MIEHLE, H. PAPEN et C. LI : Quantifying the regional source strength of n-trace gases across agricultural and forest ecosystems with process based models. *Plant and Soil*, 260:311–329, 2004.
- [7] K. CHAN, S. TARANTOLA, A. SALTELLI et I. M. SOBOL : Variance-based methods. In A. SALTELLI, K. CHAN et E. M. SCOTT, éditeurs : *Sensitivity Analysis*, Probability and Statistics, chapitre 8. Wiley, 2000.
- [8] E. de ROCQUIGNY : La maîtrise des incertitudes dans un contexte industriel. 1ère partie : une approche méthodologique globale basée sur des exemples. *Journal de la Société Française de Statistique*, 147:33–71, 2006.
- [9] E. de ROCQUIGNY : La maîtrise des incertitudes dans un contexte industriel. 2ème partie : revue des méthodes de modélisation statistique, physique et numérique. *Journal de la Société Française de Statistique*, 147:73–106, 2006.
- [10] B. EFRON, T. HASTIE, I. JOHNSTONE et R. TIBSHIRANI : Least angle regression. *Annals of Statistics*, 32:407–451, 2004.
- [11] G. FORT, S. LAMBERT-LACROIX et J. PEYRE : Réduction de la dimension dans les modèles linéaires généralisés : application à la classification supervisée de données issues des biopuces. *Journal de la Société Française de Statistique*, 146:117–152, 2005.
- [12] I.E. FRANK et J. H. FRIEDMAN : A statistical view of some chemometrics regression tools (with discussion). *Technometrics*, 35:109–148, 1993.
- [13] A.E. HOERL et R.W. KENNARD : Ridge regression based estimation for nonorthogonal problems. *Technometrics*, 12:55–67, 1970.
- [14] M.C. KENNEDY et A. O'HAGAN : Bayesian calibration of computer models. *Journal of the Royal Statistical Society*, 63:425–464, 2001.
- [15] A. KOBILINSKY : Les plans factoriels. In J.-J. DROESBEKE, J. FINE et G. SAPORTA, éditeurs : *Plans d'expériences. Applications à l'entreprise*, pages 69–209. Technip, Paris, 1997.
- [16] M. LAMBONI, D. MAKOWSKI, S. LEHUGER, B. GABRIELLE et H. MONOD : Multivariate global sensitivity analysis for dynamic crop models. *Field Crops Research*, 113:312–320, 2009.
- [17] D. MAKOWSKI, C. NAUD, M.H. JEUFFROY, A. BARBOTTIN et H. MONOD : Global sensitivity analysis for calculating the contribution of genetic parameters to the variance of crop model prediction. *Reliability Engineering and System Safety*, 91:1142–1147, 2006.

- [18] H. MONOD, C. NAUD et D. MAKOWSKI : Uncertainty and sensitivity analysis for crop models. In D. WALLACH, D. MAKOWSKI et J. JONES, éditeurs : *Working with Dynamic Crop Models*, pages 55–100. Elsevier, Amsterdam, 2006.
- [19] C. PERRIN, C. MICHEL et V. ANDREASSIAN : Does a large number of parameters enhance model performance ? Comparative assessment of common catchment model structure on 429 catchments. *Journal of Hydrology*, 242:275–301, 2001.
- [20] R DEVELOPMENT CORE TEAM : *R : a language and environment for statistical computing*. R Foundation for Statistical Computing, Austria, 2008.
- [21] C. P. ROBERT : *Le choix bayésien, principes et pratique*. Statistique et Probabilités Appliquées. Springer, 2006.
- [22] A. SALTELLI, M. RATTO, T. ANDRES, F. CAMPOLONGO, J. CARIBONI, D. GATELLI, M. SAISANA et S. TARANTOLA : *Global Sensitivity Analysis : The Primer*. Wiley, 2008.
- [23] G. SAPORTA : *Probabilité, Analyse des Données et Statistique*. Technip, 2nd édition, 2006.
- [24] I.M. SOBOL : Sensitivity analysis for non-linear mathematical model. *Mathematical Modelling and Computational Experiments*, 1:407–414, 1993.
- [25] M. STONE : Cross-validated choice and assessment of statistical predictions (with discussion). *Journal of the Royal Statistical Society B*, 36:111–147, 1974.
- [26] R. TIBSHIRANI : Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society B*, 58:267–288, 1996.
- [27] D. WALLACH, B. GOFFINET, J.-E. BERGEZ, P. DEBAEKE, D. LEENHARDT et J.-N. AUBERTOT : Parameter estimation for crop models : a new approach and application to a corn model. *Agronomy Journal*, 93:757–766, 2001.
- [28] D. WALLACH, B. GOFFINET, J.-E. BERGEZ, P. DEBAEKE, D. LEENHARDT et J.-N. AUBERTOT : The effect of parameter uncertainty on a model with adjusted parameters. *Agronomie*, 22:159–170, 2002.
- [29] Y. YANG : Consistency of cross validation for comparing regression procedures. *The Annals of Statistics*, 35:2450–2473, 2007.
- [30] H. ZOU et T. HASTIE : Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society B*, 67:301–320, 2005.