Journal de la Société Française de Statistique
Vol. 156 No. 3 (2015)

Substitution de modèle et approche multifidélité en expérimentation numérique

Title: Surrogate modeling and multifidelity approach in computer experimentation

Matthias De Lozzo[1]

Résumé : Cet article présente une synthèse bibliographique sur la substitution de modèle en expérimentation numérique où l'objectif est d'approcher un simulateur numérique à partir de quelques unes de ses évaluations. Les principaux modèles de substitution y sont décrits : réseaux de neurones artificiels, modèles par processus gaussien, machines à vecteurs de support et polynômes de chaos. Des éléments d'apprentissage statistique sont par ailleurs exposés afin de choisir la complexité et les paramètres d'un modèle de substitution permettant une bonne approximation du simulateur numérique. Une ouverture à la modélisation multifidélité est proposée afin de tenir compte de sources d'observations complémentaires lorsque l'évaluation du simulateur est trop coûteuse.

Abstract: This article presents a review of research literature on surrogate modeling in the context of computer experimentation where the goal is to approach a numerical simulator from some evaluations. The main surrogate models are described: artificial neural networks, gaussian process models, support vector machines and polynomial chaos expansions. Elements of statistical learning are expounded in order to select the complexity and the parameters of a surrogate model which assure a good approximation of the numerical simulator. An extension to multifidelity modelization is also proposed so as to take into account complementary sources of observations when the simulator evaluation is too expensive.

Mots-clés : expérimentation numérique, apprentissage supervisé, modèle de substitution, multifidélité, régression hétéroscédastique, modèle par processus gaussien, synthèse bibliographique
Keywords: computer experiments, supervised learning, surrogate model, multifidelity, heteroscedastic regression, Gaussian process model, survey
Classification AMS 2000 : 35L05, 35L70

1. Introduction

Au fil des avancées scientifiques, la connaissance des phénomènes physiques n'a cessé de croître et d'enrichir des théories sans cesse plus pertinentes. Ainsi, les méthodes numériques de résolution des systèmes d'équations correspondants permettent régulièrement d'améliorer la représentation de la réalité ; c'est le cadre de l'expérimentation numérique (*computer experiments*). Néanmoins, les coûts de calcul liés à ces approches sont élevés, ces dernières reposant sur des modèles physico-numériques toujours plus complexes, des grilles de maillage toujours plus fines et des pas de temps toujours plus faibles. Conséquemment, les simulateurs numériques décrivant un phénomène physique en rendent de mieux en mieux compte tout en étant de plus en plus difficiles à manipuler. En particulier, leur utilisation devient inconcevable en optimisation (Bonnans et al.,

[1] CEA, DEN, DER, F-13108 Saint Paul Lez Durance, France.
E-mail : matthias.delozzo@cea.fr

 ISSN: 2102-6238

2006), propagation d'incertitudes, analyse de sensibilité (Faivre et al., 2013) ou encore pour des problèmes inverses (Tarantola, 2004), ces applications requérant un trop grand nombre d'appels à ces simulateurs. Dès lors, il devient nécessaire de remplacer ces modèles précis mais de coût calculatoire élevé par des approximateurs aux évaluations moins onéreuses.

Une première approche consiste à utiliser des modèles physiques réduits basés sur une linéarisation des équations par exemple ou sur une relaxation de certaines hypothèses. Il est aussi fréquent de remplacer un modèle à sortie tridimensionnelle par un modèle bidimensionnel, voire monodimensionnel. En thermique entre autres, la complexité d'un modèle physique peut être réduite en remplaçant sa formulation distribuée, *i.e.* définie en tout point de l'espace, par une formulation par bloc (*lumped model* en anglais), caractérisée par un nombre restreint de points. Un système d'équations différentielles ordinaires est alors associé au modèle réduit physique quand au modèle physique initial correspond un système d'équations aux dérivées partielles (Lasance, 2008).

En dehors des modèles physiques réduits, un simulateur numérique peut être remplacé par un modèle approché purement mathématique et rapide en temps d'exécution, appelé modèle de substitution ou métamodèle (Forrester et al., 2008). Parmi les plus répandus se trouvent les réseaux de neurones (Bishop, 1995), les modèles par processus gaussien (Rasmussen and Williams, 2005) ainsi que les développements par polynômes de chaos (Ghanem and Spanos, 1991). Un métamodèle est paramétré par apprentissage statistique (Hastie et al., 2009) d'un ensemble d'évaluations du simulateur numérique ; en particulier, les méthodes de sélection de modèles et de régularisation lui assurent une qualité d'approximation optimale sur l'espace d'étude de ses entrées. Cette qualité de substitution dépend du phénomène physique, de la famille de métamodèles, du rapport entre le nombre d'évaluations et celui de paramètres ainsi que du plan d'expériences (Dean et al., 2015) associé aux évaluations. En effet, la planification d'expériences ne saurait être séparée de l'étape de modélisation ; ainsi, l'ouvrage de Kleijnen (2015) présente des modèles polynomiaux d'ordre peu élevé, des modèles à surface de réponse ou encore un modèle de krigeage, tout en faisant le lien avec le choix du plan servant à leurs paramétrisations. Santner et al. (2003) traitent également de cette sélection en fonction des objectifs de recherche de l'expérimentateur (prédiction, analyse de sensibilité, ...) tandis que Fang et al. (2006) détaillent l'expérimentation numérique et ses finalités avant de présenter différents plans d'expériences puis des métamodèles plus ou moins élaborés. Des revues sur la planification et la métamodélisation des expériences numériques ont également été proposées par Simpson et al. (2001), Chen et al. (2006) et Forrester and Keane (2009).

Par ailleurs, lorsque le coût calculatoire du simulateur numérique contraint trop son nombre d'évaluations, il peut être opportun d'utiliser des observations émanant de sources complémentaires (*multi-sensor data fusion* ou *sensor fusion*) avec des coûts moins élevés et des représentations du phénomène physique différentes. Selon le domaine d'études, on parle de modélisation multifidélité (*multifidelity modeling*, Forrester et al., 2007), de modélisation à complexité variable (*variable-complexity modeling*), de fusion de données (*data fusion*, Mitchell, 2007), de fusionnement de données (*data merging*) ou encore de modélisation hétéroscédastique (*heteroscedastic modeling*, Gendre, 2008). Bien que les méthodes et terminologies employées diffèrent, l'objectif est commun : approcher la quantité physique d'intérêt à partir de sources d'information ayant différentes qualités de représentation de cette quantité.

 ISSN: 2102-6238

Dans ce papier, nous appliquons la modélisation multifidélité au cadre de la substitution de modèle. Dans le cas fréquent où les sources d'information sont des codes de calcul déterministes, les écarts de représentation entre les codes de calcul sont expliqués par des différences d'hypothèses physiques ou de paramétrisations numériques. Dans le cas où ces sources sont des codes de calcul stochastiques ou des séries de mesures, les écarts sont dus à des différences de niveaux de bruit d'observation. Ces deux considérations ont été étudiées dans la thèse de De Lozzo (2013) appliquée à l'ingénierie thermique. Ces travaux traitent notamment des problèmes de mélange de co-krigeages (Kennedy and O'Hagan, 2000) lorsque plusieurs simulateurs numériques de basse-fidélité complètent un simulateur précis, ainsi que de la modélisation hétéroscédastique pour des études comportant des séries d'observations associées à différents niveaux d'incertitude.

Cet article propose une synthèse sur la substitution de modèle en expérimentation numérique, avec une ouverture sur la modélisation multifidélité. Il s'adresse tout particulièrement aux ingénieurs et chercheurs souhaitant se familiariser avec ce domaine. La section 2 présente les principales familles de modèles de substitution tandis que la section 3 fournit des outils d'apprentissage statistique pour construire un métamodèle. Par la suite, la section 4 étend ces éléments au cadre de la modélisation multifidélité. Une conclusion de cette synthèse et des perspectives associées sont proposées dans la section 5.

2. Modèles de substitution pour les codes de calcul

Cette section introduit la notion de modèle de substitution (Forrester et al., 2008) pour les codes de calcul et présente quelques uns des modèles les plus utilisés : les réseaux de neurones artificiels, les régressions par machines à vecteurs de support, les développements par polynômes de chaos et les modèles par processus gaussien.

2.1. Substitution d'un code de calcul

L'essor de la puissance computationnelle a donné naissance aux sciences numériques (*computational sciences* ou *scientific computing* en anglais). Ces dernières étudient les phénomènes physiques au moyen de codes de calcul, c'est-à-dire de simulateurs numériques implémentés sur ordinateur et basés sur des équations de la physique et sur des méthodes de résolution numérique. L'observation de la sortie d'un code de calcul associée à une certaine configuration de ce modèle physico-numérique est appelée expérience simulée, expérience numérique ou *computer experiment* :

> *« Suppose that a mathematical theory exists, e.g, a set of differential equations, that relates the output of a complex physical process to a set of input variables. Suppose also that a numerical method exists for accurately solving the mathematical system. The presence of these two elements with appropriate computer hardware and software to implement the numerical methods allows one to conduct a* ***computer experiment*** *[...] »*, Santner et al. (2003).

Pour des soucis de clarté, ce papier se limite aux situations où la variable d'intérêt y et les variables physico-numériques $x_1, \ldots, x_d$ dont elle dépend sont des scalaires réels. Sauf mention contraire, le code de calcul associé est considéré comme déterministe ; ceci écarte la spécificité

 ISSN: 2102-6238

de modélisation des codes de calcul stochastiques pour lesquels deux évaluations associées à un même jeu $x_1,\ldots,x_d$ donnent deux résultats différents. Ce code de calcul déterministe est décrit mathématiquement par la fonction f associant au vecteur de paramètres d'entrée $\mathbf{x} = (x_1,...,x_d)^T \in \mathscr{X} \subset \mathbb{R}^d$ la sortie réelle $y \in \mathscr{Y} \subset \mathbb{R}$:

$$\begin{array}{rcl} f : \mathscr{X} \subset \mathbb{R}^d & \to & \mathscr{Y} \subset \mathbb{R} \\ \mathbf{x} & \mapsto & y := f(\mathbf{x}). \end{array}$$

D'un côté, la fonction f est considérée comme un modèle de référence vis-à-vis du problème étudié car le code de calcul repose sur des hypothèses physiques et des méthodes mathématiques éprouvées. D'un autre côté, la complexité de ces dernières rend les évaluations du simulateur numérique onéreuses en temps de calcul, ce qui en limite le nombre. Cette limitation est particulièrement contraignante pour des applications d'optimisation, de propagation d'incertitudes, d'analyse de sensibilité ou de problèmes inverses, requérant toutes une grande quantité d'observations de la fonction f. Il est donc souhaitable de substituer au modèle f un modèle mathématique $\hat{f}_\theta$ à temps de réponse rapide, dont la sortie sur $\mathscr{X}$ est suffisamment proche de celle de f :

$$\begin{array}{rcl} \hat{f}_\theta : \mathscr{X} \times \Theta \subset \mathbb{R}^d \times \mathbb{R}^p & \to & \mathscr{Y} \subset \mathbb{R} \\ (\mathbf{x};\theta) & \mapsto & \hat{y} := \hat{f}_\theta(\mathbf{x}). \end{array}$$

Le vecteur θ de $\Theta \subset \mathbb{R}^p$ regroupe les paramètres du modèle $\hat{f}_\theta$; de leurs valeurs dépend la qualité d'approximation de f par $\hat{f}_\theta$, et donc la faculté de $\hat{f}_\theta$ à bien représenter le phénomène physique.

Cette fonction $\hat{f}_\theta$ est appelée modèle de substitution (Forrester et al., 2008), modèle réduit, métamodèle ou encore surface de réponse (Myers et al., 2009), cette dernière terminologie étant issue du domaine de la fiabilité des structures. Dans ce papier, les expressions "métamodèle" et "modèle de substitution" sont privilégiées, la seconde étant la traduction de *surrogate model*, terme prépondérant dans la littérature anglophone.

Afin de rendre la sortie de $\hat{f}_\theta$ la plus proche possible de celle de f, les paramètres θ sont optimisés afin de réduire les écarts $\left(\left|y^{(i)} - \hat{f}_\theta\left(\mathbf{x}^{(i)}\right)\right|\right)_{1\leq i\leq n}$ associés à la base d'apprentissage $\mathscr{A} = \left(\mathbf{x}^{(i)}, y^{(i)}\right)_{1\leq i\leq n}$, où $y^{(i)} = f\left(\mathbf{x}^{(i)}\right)$. Le plan d'expériences correspondant $\mathbf{D} = \left(\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}\right)$ est classiquement obtenu par une méthode de *space-filling* ou par des tirages de Monte-Carlo (Dean et al., 2015). D'autre part, des techniques complémentaires visent à réduire l'écart entre f et $\hat{f}_\theta$ sur l'ensemble du domaine $\mathscr{X}$. Les méthodes d'apprentissage statistique permettent ainsi de construire un métamodèle proche du code de calcul ; les principaux éléments sont présentés dans la section 3.

2.2. *Réseaux de neurones artificiels*

Les réseaux de neurones artificiels considérés dans ce papier sont des perceptrons multicouches à une couche cachée et sont présentés dans les ouvrages de Dreyfus et al. (2008) et Bishop (1995), couvrant un plus large spectre de métamodèles neuronaux. Un perceptron multicouche à M couches cachées est une fonction mathématique non-linéaire appliquant N_1 transformations non-linéaires identiques à N_1 combinaisons linéaires des entrées, puis de façon itérative, N_i

 ISSN: 2102-6238

transformations non-linéaires identiques à N_i combinaisons linéaires des N_{i-1} transformations de la $i-1^{\text{ème}}$ couche cachée pour tout $i \in \{2,\dots,M\}$; enfin, la sortie de cette fonction est une combinaison linéaire des combinaisons de la $M^{\text{ème}}$ couche cachée. Ces réseaux de neurones sont dits à propagation en avant (*feed-forward*) car ces neurones ne comportent pas de cycle, contrairement aux réseaux de neurones récurrents. Les entrées, les transformations des couches cachées et la sortie sont respectivement appelées neurones d'entrée, cachés et de sortie ; les transformations non-linéaires sont appelées fonctions d'activation. Pour un grand nombre de cas d'étude, un perceptron multicouche à une couche cachée est suffisant ; c'est celui qui est présenté ici. Cependant, dans le cas de phénomènes physiques multi-échelles, l'augmentation du nombre de couches cachées peut considérablement améliorer l'approximation de f ; on parle de réseau de neurones profond (*deep neural network*).

Un perceptron multicouche à une couche cachée, appelé abusivement réseau de neurones artificiel dans la suite de ce papier, s'écrit :

$$\hat{f}_\theta(\mathbf{x}) = \sum_{i=1}^{N} w_i h\left(\sum_{j=1}^{d} w_{ij}x_j + w_{i0}\right) + w_0 \tag{1}$$

où w_{i0} est le $i^{\text{ème}}$ biais d'entrée, w_{ij} le poids de la $j^{\text{ème}}$ entrée associée au $i^{\text{ème}}$ neurone de la couche cachée, w_i le poids du $i^{\text{ème}}$ neurone de la couche cachée associé à la sortie et w_0 le biais de sortie. La fonction d'activation h du $i^{\text{ème}}$ neurone caché applique au potentiel $v = \sum_{j=1}^{d} w_{ij}x_j + w_{i0}$ une transformation non-linéaire. Il est courant de considérer la fonction sigmoïde $h(v) = 1/(1+\exp(-v))$ ou la fonction tangente hyperbolique $h(v) = 2/(1+\exp(-2v)) - 1$ dont les dérivées respectives s'obtiennent rapidement au moyen des expressions $h'(v) = h(v)(1-h(v))$ et $h'(v) = 4h(v)(1-h(v))$.

Réseaux de neurones artificiels à base radiale - Certaines modèles neuronaux utilisent des potentiels quadratiques plutôt que des potentiels linéaires comme dans l'équation (1). Ainsi, le métamodèle s'écrit $\hat{f}_\theta(\mathbf{x}) = \sum_{i=1}^{N} w_i \exp\left(-\frac{\sum_{j=1}^{d}(x_j - w_{ij})^2}{2\sigma_{ij}^2}\right) + w_0$ dans le cas d'une fonction d'activation Gaussienne. Cette fonction d'activation compare les valeurs prises par les entrées à une certaine référence avant de renvoyer en sortie une valeur d'autant plus faible que la distance séparant les entrées de cette référence est grande. Parmi ces réseaux de neurones artificiels à base radiale, on retrouve également les réseaux d'ondelette (Oussar et al., 1998).

Propriétés

Les perceptrons multicouches comportent des propriétés mathématiques intéressantes. Tout d'abord, Hornik (1991) a montré que toute fonction continue et bornée de $\mathbb{R}^d$ dans $\mathbb{R}$ peut être approchée avec une précision donnée par le réseau de neurones défini par l'expression (1) dès lors que la fonction d'activation h est continue, non-constante et bornée. Ce résultat avait été auparavant démontré par Cybenko (1989) pour une fonction d'activation de type sigmoïde. $\hat{f}_\theta$ est qualifié d'approximateur universel sur le sous-ensemble compact $\mathscr{X}$ de $\mathbb{R}^d$. De plus, si f est non-linéaire, Barron (1993) a montré qu'un métamodèle non-linéaire en les paramètres d'entrée est plus parcimonieux qu'un métamodèle linéaire. En d'autres termes, à erreur d'approximation fixée, le nombre de paramètres croît exponentiellement avec le nombre d'entrées pour un métamodèle

 ISSN: 2102-6238

linéaire tandis qu'il croît linéairement pour un métamodèle non-linéaire. Enfin, la fonction f peut être approchée avec une erreur $\|f-\hat{f}_\theta\|_\infty = \sup_{x\in\mathscr{X}} |f(x)-\hat{f}_\theta(x)|$ majorée par $\mathscr{O}\left(N^{-(\alpha+1)/d}\right)$ lorsque $f\in\mathscr{C}^\alpha$ (Attali and Pagès, 1997). On note ainsi qu'à nombre de neurones cachés donné, l'erreur d'approximation est d'autant plus faible que le nombre de paramètres d'entrée est faible ou que la fonction à approcher est régulière. Ces propriétés justifient le choix courant du perceptron multicouche à une couche cachée comme réseau de neurones artificiel, avec souvent une fonction d'activation sigmoïde.

Paramétrisation

La paramétrisation d'un perceptron multicouche comporte quelques étapes de pré-traitement. Les paramètres d'entrée sont d'abord ramenés dans un intervalle de faible amplitude (souvent $[-1,1]$) grâce à une normalisation géométrique ou à une normalisation statistique par centrage-réduction des entrées. Lors de l'initialisation de la phase d'optimisation de $\hat{f}_\theta$, les paramètres initiaux du réseau de neurones sont souvent considérés comme des réalisations de variables aléatoires uniformes de faibles amplitudes, avec des biais d'entrée égaux à zéro, de sorte que les sorties des neurones soient centrées. Néanmoins cette méthode conduit souvent à des minima locaux très éloignés du minimum global. On lui préfère alors l'algorithme de Nguyen and Widrow (1990) qui s'efforce à remplir l'espace des poids et des biais du réseau de neurones de façon plus parcimonieuse, de sorte qu'à chaque neurone caché soit associé un sous-ensemble de l'espace de sortie.

De là, diverses méthodes d'optimisation classiques ou dédiées existent selon la dérivabilité ou non-dérivabilité de ce risque, c'est-à-dire de la fonction d'activation sous-jacente h. Dans le cas où le risque est dérivable, il est courant d'utiliser une descente de gradient en utilisant l'algorithme de la rétropropagation du gradient (Rumelhart et al., 1986). La rétropropagation résiliente du gradient (Riedmiller and Braun, 1993) et son amélioration iRPROP+ (Igel and Hüsken, 2000) sont des versions de cette technique moins coûteuses en mémoire, basées sur les signes des dérivées partielles du risque empirique et non pas sur leurs valeurs. Ces méthodes d'optimisation d'ordre 1 sont rapides et peu gourmandes en mémoire de stockage. Néanmoins le risque empirique comporte un nombre de minima locaux trop important pour ces approches et sa minimisation conduit souvent à un métamodèle de mauvaise qualité de prédiction. L'utilisation d'une méthode d'ordre 1 est alors à déconseiller dès lors que le réseau comporte plusieurs centaines de paramètres. Une première alternative consiste à procéder à une optimisation *multi-start* en répétant plusieurs fois la descente de gradient à partir de paramètres initiaux différents. Cependant la complexité du problème est souvent telle qu'elle requiert l'utilisation de méthodes d'ordre 2 basées sur la matrice hessienne. Les plus connues sont les algorithmes BFGS (Broyden, 1970) et Levenberg-Marquardt (Levenberg, 1944; Marquardt, 1963). Dans le cas où le risque n'est pas dérivable, il est nécessaire de considérer des méthodes d'optimisation de type recuit simulé ou algorithmes évolutionnistes (Simon, 2013). Enfin, ces algorithmes d'optimisation peuvent être appliqués au risque empirique associé à la base d'apprentissage ou de façon itérative aux risques empiriques associés à des sous-ensembles emboîtées de cette base (*incremental training*), les paramètres optimaux d'une étape servant de paramètres initiaux à la suivante.

 ISSN: 2102-6238

2.3. *Modèle par processus gaussien*

Plus récemment, des modèles de substitution basés sur des processus gaussiens ont été étudiés en optimisation, en analyse de sensibilité ou encore en prédiction ; ils comprennent notamment le krigeage (*kriging en anglais*). Formalisé mathématiquement par Matheron (1963), ce dernier a été imaginé par Krige (1951) afin de modéliser la teneur en or du complexe récifal sud-africain de Witwatersrand par une moyenne basée sur la teneur en quelques sites et pondérées par la distance à ces points. Le krigeage est un modèle de géostatistique prenant ses entrées dans $\mathbb{R}^2$ ou $\mathbb{R}^3$ et dont l'estimation des paramètres repose sur une analyse variographique. Sacks et al. (1989) ont introduit ce modèle mathématique en expérimentation numérique. L'ouvrage de Rasmussen and Williams (2005) est une référence pour la métamodélisation au moyen de processus gaussiens tandis que pour le modèle de krigeage plus spécifiquement, on peut se référer à Stein (1999). Le krigeage étant un cas particulier du modèle par processus gaussien et par abus de langage en sciences numériques, nous confondrons par la suite krigeage et modèle par processus gaussien.

Modèle d'observation par processus gaussien

La modélisation par processus gaussien suppose que f est la réalisation d'un processus gaussien Z d'espérance $\mu(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\beta$ et de structure de covariance $\sigma^2 r(\mathbf{x},\tilde{\mathbf{x}}) + \lambda\delta_{\mathbf{x}\tilde{\mathbf{x}}}$ pour tous les $\mathbf{x}$ et $\tilde{\mathbf{x}}$ de $\mathscr{X}$, avec δ le symbole de Kronecker.

La tendance $\mu(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\beta$ de $Z(\mathbf{x})$ est une somme pondérée de n_β fonctions de régression $\mathbf{h}(\mathbf{x}) = \left(h_1(\mathbf{x}),...,h_{n_\beta}(\mathbf{x})\right)^T \in \mathbb{R}^{n_\beta}$. On se limite souvent au cas "$h(\mathbf{x}) = 1$ et β connu" (krigeage simple) et "$h(\mathbf{x}) = 1$ et β inconnu" (krigeage ordinaire) ; pour les autres situations (krigeage universel), $h(\mathbf{x})$ est souvent un vecteur de monômes, typiquement $(1\ x_1\ \ldots\ x_d)$.

La structure de covariance $\sigma^2 r(\mathbf{x},\tilde{\mathbf{x}})$ est le produit d'un niveau de bruit σ^2 et d'une fonction de corrélation $r(\mathbf{x},\tilde{\mathbf{x}}) = \mathrm{Corr}\,(Z(\mathbf{x}),Z(\tilde{\mathbf{x}}))$. La fonction $\sigma^2 r(\mathbf{x},\tilde{\mathbf{x}})$ est un noyau défini positif décrivant la structure du processus gaussien Z, c'est-à-dire la texture de ses instances ; elle caractérise la régularité de la fonction f à approcher. Pour une facilité d'exploitation, ce noyau est souvent considéré comme le produit tensorisé [1] de d noyaux de corrélation de dimension 1 : $r(\mathbf{x},\tilde{\mathbf{x}}) = \prod_{i=1}^d r_i(x_i,\tilde{x}_i)$. Dans le cadre stationnaire défini par $\forall h \in \mathbb{R}^d\ \mathrm{Corr}(Z(\mathbf{x}+\mathbf{h}),Z(\mathbf{x})) = \kappa(\|\mathbf{h}\|_2)$, le tableau 1 représente des exemples de noyaux de dimension 1 et stationnaires. La structure de covariance atteint son maximum σ^2 lorsque $\mathbf{x} = \tilde{\mathbf{x}}$.

TABLEAU 1. *Différents noyaux de corrélation fonction de la distance $h = |x - \tilde{x}|$ et d'un paramètre de portée ζ.*

Gaussien	exponentiel	Matérn 3/2	Matérn 5/2
$\exp\left(-\frac{1}{2}\left(\frac{h}{\zeta}\right)^2\right)$	$\exp\left(-\frac{h}{\zeta}\right)$	$\left(1+\sqrt{3}\frac{h}{\zeta}\right)\exp\left(-\sqrt{3}\frac{h}{\zeta}\right)$	$\left(1+\sqrt{5}\frac{h}{\zeta}+\frac{5}{3}\left(\frac{h}{\zeta}\right)^2\right)\exp\left(-\sqrt{5}\frac{h}{\zeta}\right)$

Sous l'hypothèse d'isotropie, l'échelle des fluctuations de f est identique dans toutes les directions. Pour tout $i \in \{1,\ldots,d\}$, le noyau élémentaire r_i est alors fonction d'un même hyperpa-

[1] Le noyau tensorisé est appelé noyau factorisé en géostatistique.

 ISSN: 2102-6238

ramètre $\zeta_i := \zeta > 0$, appelé *portée* ou *longueur de corrélation*. Cette portée ζ est homogène à la distance au-delà de laquelle deux réels x et $\tilde{x}$ cessent d'avoir une influence mutuelle. Une faible valeur de ζ_* est associée à de fortes variations de f selon l'entrée x_* et inversement, une grande valeur de ζ_* est associée à de faibles variations de f selon x_*.

Sous l'hypothèse d'anisotropie, la régularité dépend des paramètres d'entrée ; ceci se traduit par des hyperparamètres $\zeta_1, \ldots, \zeta_d$ distincts. Dans ce cas, il peut être pertinent de différencier l'anisotropie géométrique (Stein, 1999) de l'anisotropie zonale (Matheron, 1970). Dans le cas stationnaire, on considère pour cela le variogramme de Z défini pour tout $\mathbf{h} \in \mathbb{R}^d$ par $\gamma(\mathbf{h}) = \sigma^2(1 - \kappa(\|\mathbf{h}\|_2))$. Le terme σ^2 est alors nommé *palier* ; il représente la valeur limite du variogramme lorsque $\|\mathbf{h}\|_2$ tend vers l'infini. L'anisotropie géométrique revient à considérer une transformation linéaire du vecteur d'entrée $\mathbf{x}$. Dans ce cas, le variogramme γ n'est plus fonction de $\|\mathbf{h}\|_2$ mais de $\|M\mathbf{h}\|_2$, où M est une matrice définissant la transformation linéaire. Ainsi, le processus est anisotrope dans l'espace des paramètres d'entrée et isotrope dans une transformation linéaire de cet espace ; ceci se traduit par une rotation ou par un étirement des paramètres d'entrée. Dans le cas de l'anisotropie zonale, le variogramme comporte plusieurs paliers pour mieux caractériser des phénomènes locaux ; il se décompose alors en la somme de variogrammes sous-jacents représentant des anisotropies de support.

Modèle de prédiction par processus gaussien

Le modèle de substitution $\hat{f}_\theta$ associé au processus gaussien Z s'obtient en conditionnant ce dernier par la base d'apprentissage $\mathscr{A}$ et les paramètres $\theta = (\beta, \sigma^2, \lambda, \zeta)$. Ce métamodèle interpole les observations de la base d'apprentissage lorsque $\lambda = 0$. Précisément, $[Y|\mathscr{A}, \theta]$ est un processus gaussien défini pour tout $\mathbf{x}$ et tout $\tilde{\mathbf{x}}$ de $\mathscr{X}$ par son espérance :

$$\hat{f}_\theta(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\beta + \mathbf{r}(\mathbf{x},\mathbf{D})(\mathbf{R} + \lambda\mathbf{I}_n)^{-1}(\mathbf{y}^n - \mathbf{H}^T\hat{\beta}) = \mathbf{h}(\mathbf{x})^T\beta + \sum_{i=1}^{n} \alpha_i r\left(\mathbf{x}, \mathbf{x}^{(i)}\right),$$

avec $\alpha = (\mathbf{R} + \lambda\mathbf{I}_n)^{-1}(\mathbf{y}^n - \mathbf{H}^T\beta)$, et par sa structure de covariance :

$$s^2(\mathbf{x}, \tilde{\mathbf{x}}) = \sigma^2\left(r(\mathbf{x}, \tilde{\mathbf{x}}) - \mathbf{r}(\mathbf{x}, \mathbf{D})(\mathbf{R} + \lambda\mathbf{I}_n)^{-1}\mathbf{r}(\mathbf{D}, \tilde{\mathbf{x}})\right)$$

où $\mathbf{y}^n = \left(y^{(1)}, \ldots, y^{(n)}\right)^T$, $\mathbf{D} = \left(\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right)$, $\mathbf{r}(\mathbf{x}, \mathbf{D}) = \left(r\left(\mathbf{x}, \mathbf{x}^{(i)}\right), \ldots, r\left(\mathbf{x}, \mathbf{x}^{(n)}\right)\right) = \mathbf{r}(\mathbf{D}, \mathbf{x})^T$, $\mathbf{R} = \left(r\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)\right)_{1 \leq i,j \leq n}$, $\mathbf{H} = \left(\mathbf{h}\left(\mathbf{x}^{(1)}\right), \ldots, \mathbf{h}\left(\mathbf{x}^{(n)}\right)\right) \in \mathscr{M}_{n_\beta, n}(\mathbb{R})$ et $\mathbf{I}_n = (\delta_{ij})_{1 \leq i,j \leq n}$.

Les paramètres β, σ^2 et ζ sont en pratique remplacés par leurs estimateurs du maximum de vraisemblance (EMV) associés à $\mathscr{A}$. Ceux de β et σ^2 comportent des expressions analytiques tandis que celui de ζ requiert une méthode numérique d'optimisation non linéaire. Il est par exemple possible de coupler un algorithme évolutionnaire à la méthode BFGS (Le Gratiet, 2013) afin de minimiser l'opposé de la log-vraisemblance concentrée $\log(|\det(\mathbf{R})|) + (n - n_\beta)\log(\widehat{\sigma^2})$.

En considérant l'EMV de β et en fixant ζ, $\hat{f}_\theta$ est le meilleur prédicteur linéaire sans biais de f lorsque $\lambda = 0$; on parle d'estimateur BLUE (*Best Linear Unbiased Estimator*).

 ISSN: 2102-6238

Effet de pépite et régression

Le processus gaussien Z permet d'intégrer un bruit de mesure gaussien, centré et de variance λ lié aux observations de la base d'apprentissage $\mathscr{A}$. Or la plupart des modèles à approcher ne sont pas stochastiques mais déterministes. Dès lors, ce bruit de mesure n'a pas lieu d'être et le paramètre λ est fixé à 0, faisant de $\hat{f}_\theta$ un modèle de substitution interpolant.

Cependant lorsque le nombre d'évaluations n est faible, le choix d'un modèle régressant via une valeur non-nulle de λ peut être opportun en matière de régularité et de capacité de généralisation de $\hat{f}_\theta$. Ce choix, appelé "effet de pépite", peut être vu comme une pénalité ℓ_2 dont l'objectif est de régulariser la sortie de $\hat{f}_\theta$, sur le même principe que la régression par splines par exemple. De plus, en présence de points de $\mathscr{A}$ relativement proche sur $\mathscr{X}$, la matrice de covariance $\mathbf{R}$ peut être mal conditionnée ; l'effet de pépite est une solution pour pallier ce problème. Enfin, λ étant un paramètre du métamodèle, sa valeur peut être choisie via des méthodes classiques de sélection de modèles : minimisation d'une erreur de test, d'une erreur de validation croisée, ...

Mesure de confiance et planification adaptative

La variance conditionnelle $s^2(\mathbf{x}) = \mathbb{E}\left[\left(Z(\mathbf{x}) - \hat{f}_\theta(\mathbf{x})\right)^2 | \mathscr{A}, \theta\right]$ est par définition l'erreur quadratique moyenne de $\hat{f}_\theta(\mathbf{x})$ et sert souvent de mesure de l'erreur de $\hat{f}_\theta(\mathbf{x})$. Cette utilisation comme erreur de prédiction est l'un des principaux intérêts du krigeage. Néanmoins, cette considération repose sur l'hypothèse forte que f est une instance du processus gaussien Z. Par ailleurs, la variance de krigeage est optimiste car elle ne tient pas compte de l'incertitude associée aux estimateurs de θ. Pour pallier ceci, des techniques de *bootstrap* paramétrique ont été proposées (Den Hertog et al., 2005). L'estimateur de la variance de $[Z(\mathbf{x})|\mathscr{A}]$ s'écrit alors :

$$s^2_{\mathrm{BP}}(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\left(Z^{(b)}(\mathbf{x}) - \hat{f}_\theta^{(b)}(\mathbf{x})\right)^2, \tag{2}$$

où $Z^{(b)}$ représente la $b^{\text{ème}}$ réalisation du processus gaussien $[Z(\mathbf{x})|\mathscr{A}]$. Le nombre B de réplications *bootstrap* tend à réduire le biais de s^2_{BP}. Le terme $\hat{f}_\theta^{(b)}(\mathbf{x})$ quant à lui est l'espérance de $\left[Z(\mathbf{x})|\mathscr{A}, \hat{\theta}^{(b)}\right]$, avec $\hat{\theta}^{(b)}$ l'estimateur de θ obtenu à partir de $\mathscr{A}^{(b)} = \left(\mathbf{x}_i, Z^{(b)}(\mathbf{x}_i)\right)_{1\leq i\leq n}$.

D'autre part, il est courant d'associer à $\hat{f}_\theta(\mathbf{x})$ un intervalle de niveau de confiance à $1-\alpha$, $\alpha \in (0,1)$. Dans le cas de s^2 par exemple, cet intervalle s'écrit :

$$IC_{1-\alpha}(f(\mathbf{x})) = \left[\hat{f}_\theta(\mathbf{x}) - q_{1-\alpha/2}s(\mathbf{x}), \hat{f}_\theta(\mathbf{x}) + q_{1-\alpha/2}s(\mathbf{x})\right]$$

avec $q_{1-\alpha/2}$ le $1-\alpha/2$ quantile de la loi normale standard.

En planification adaptative, le terme s sert aussi d'indicateur de l'erreur ; en pratique, le métamodèle $\hat{f}_\theta$ est reconstruit via la base d'apprentissage $\mathscr{A}$ enrichie du point $(\mathbf{x}^*, f(\mathbf{x}^*))$ où $\mathbf{x}^*$ maximise s. Plusieurs itérations de ce procédé sont réalisées selon le budget de calcul et la précision de $\hat{f}_\theta$ souhaitée, comme l'illustre la figure 1. D'autres critères d'enrichissement de la base d'apprentissage permettent aussi d'améliorer la qualité d'approximation du métamodèle.

 ISSN: 2102-6238

Sacks et al. (1989) présentent la minimisation de l'IMSE (*Integrated Mean Squared Error*), de la MMSE (*Maximum Mean Squared Error*) et de l'entropie *a posteriori*. La MMSE correspond au maximum sur $\mathscr{X}$ de la variance du processus gaussien conditionnée par la base d'apprentissage et les nouveaux points candidats :

$$\text{MMSE} = \max_{x \in \mathscr{X}} \mathbb{E}\left[\left(Z(\mathbf{x}) - \hat{f}_\theta(\mathbf{x})\right)^2 | \mathscr{A}, \mathscr{D}^*, \theta\right]$$

où $\mathscr{D} = \left(\mathbf{x}^{*,(1)}, \ldots, \mathbf{x}^{*,(n_*)}\right)$ est l'ensemble des nouveaux points candidats. L'IMSE correspond quand à elle à la moyenne pondérée de cette variance :

$$\text{IMSE} = \int_{x \in \mathscr{X}} \mathbb{E}\left[\left(Z(\mathbf{x}) - \hat{f}_\theta(\mathbf{x})\right)^2 | \mathscr{A}, \mathscr{D}^*, \theta\right] W(\mathbf{x}) d\mathbf{x}$$

où $W(\mathbf{x})$ est la pondération, dont la dépendance en $\mathbf{x}$ est pertinente pour obtenir un métamodèle particulièrement précis dans une région cible de l'espace de sortie $\mathscr{Y}$ (Picheny et al., 2010). Enfin, l'entropie de Shanon *a posteriori* (Currin et al., 1988), c'est-à-dire l'entropie du vecteur aléatoire $[Z(\mathscr{D}_g) | \mathscr{A}, \mathscr{D}^*, \theta]$, s'écrit quant à elle :

$$H = 0.5 \ln\left\{\det\left[\text{Var}\left(Z(\mathscr{D}_g) | \mathscr{A}, \mathscr{D}^*, \theta\right)\right]\right\} + \text{constante}$$

où $\mathscr{D}_g$ est un ensemble discret d'éléments de $\mathscr{X}$. Choisir le plan $\mathscr{D}^*$ minimisant l'entropie *a posteriori* H revient à minimiser la quantité d'incertitude du processus gaussien conditionné par $\mathscr{A}$ et $\mathscr{D}^*$ en les points de généralisation $\mathscr{D}_g$.

2.4. *Régression par machines à vecteurs de support*

La Régression par machines à Vecteurs de Support (SVR) consiste à chercher une fonction $\hat{f}_\theta$ comportant une certaine platitude et pour tout $i \in \{1, \ldots, n\}$, une déviation $|y^{(i)} - \hat{f}\left(\mathbf{x}^{(i)}\right)|$ inférieure à un seuil $\varepsilon > 0$ (Vapnik et al., 1996; Scholkopf and Smola, 2001). L'erreur quadratique associée à $\hat{f}_\theta$ s'écrit :

$$L\left(\hat{f}_\theta; \mathscr{A}\right) = \frac{1}{n} \sum_{i=1}^{n} V\left(y^{(i)}, \hat{f}_\theta\left(\mathbf{x}^{(i)}\right), \varepsilon\right)$$

où $V\left(y^{(i)}, \hat{f}_\theta\left(\mathbf{x}^{(i)}\right), \varepsilon\right) = \begin{cases} 0 & \text{si } \left|y^{(i)} - \hat{f}_\theta\left(\mathbf{x}^{(i)}\right)\right| \leq \varepsilon, \\ \left(\left|y^{(i)} - \hat{f}_\theta\left(\mathbf{x}^{(i)}\right)\right| - \varepsilon\right)^2 & \text{sinon.} \end{cases}$

Ainsi, une prédiction $\hat{f}_\theta\left(\mathbf{x}^{(i)}\right)$ située à une distance de $y^{(i)}$ inférieure à ε ne contribue pas à l'erreur quadratique L, du fait de la marge d'insensibilité ε.

Dans le cadre linéaire, ce métamodèle s'écrit :

$$\hat{f}_\theta(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle_2 + b$$

où $\langle \cdot, \cdot \rangle_2$ est le produit scalaire sur $\mathscr{X}$, $\mathbf{w} \in \mathbb{R}^d$ sont les coefficients de régression et $b \in \mathbb{R}$ est le biais du métamodèle. Le problème d'optimisation associé s'écrit :

$$\underset{\mathbf{w}, b, \xi, \xi^*}{\text{argmin}} \frac{1}{2} \|\mathbf{w}\|_2^2 + C \sum_{i=1}^{n} (\xi_i + \xi_i^*)$$

 ISSN: 2102-6238

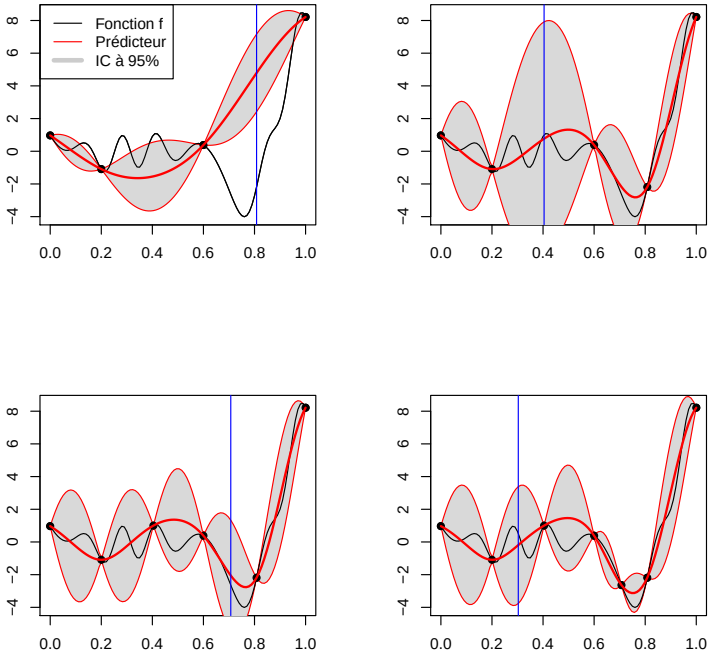

FIGURE 1. *Planification d'expériences adaptative pour la construction d'un modèle par processus gaussien : le plan d'expériences est complété à chaque étape par le point maximisant la variance avant de reconstruire le métamodèle.*

sous les contraintes $y^{(i)} - \hat{f}_\theta\left(\mathbf{x}^{(i)}\right) \leq \varepsilon + \xi_i$ et $\hat{f}_\theta\left(\mathbf{x}^{(i)}\right) - y^{(i)} \leq \varepsilon + \xi_i^*$, avec $\xi_i, \xi_i^* \geq 0$, pour tout $i \in \{1,\ldots,n\}$. Fixer les variables ξ et ξ^* à zéro suppose qu'il existe un vecteur $\mathbf{w}$ tel que pour tout $i \in \{1,\ldots,n\}$, l'observation $y^{(i)}$ se situe à une distance de $\hat{f}_\theta\left(\mathbf{x}^{(i)}\right)$ inférieure à ε, ce qui n'est pas toujours vrai. Ainsi, ces variables sont des variables de relâchement (*stack variables*) qui assurent la faisabilité des contraintes. L'hyperparamètre $C > 0$ permet de régler le compromis entre l'erreur quadratique L et la platitude du métamodèle $\hat{f}_\theta$.

La version duale de ce problème d'optimisation facilite la convergence vers le minimum :

$$(\hat{\alpha}, \hat{\alpha}^*) = \underset{\alpha,\alpha^*}{\operatorname{argmax}} -\frac{1}{2}\sum_{i,j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)}\rangle - \varepsilon\sum_{i=1}^{n}(\alpha_i + \alpha_i^*) + \sum_{i=1}^{n} y^{(i)}(\alpha_i - \alpha_i^*) \quad (3)$$

sous les contraintes $\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0$ et $\alpha_i, \alpha_i^* \in [0, C]$, $\forall i \in \{1,\ldots,n\}$. Le métamodèle s'écrit alors :

$$\hat{f}_\theta(\mathbf{x}) = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)\langle \mathbf{x}^{(i)}, \mathbf{x}\rangle_2 + b = \langle \hat{\mathbf{w}}, \mathbf{x}\rangle_2 + b$$

où $\hat{\mathbf{w}} = \sum_{i=1}^{n}(\hat{\alpha}_i - \hat{\alpha}_i^*)\mathbf{x}^{(i)}$. $\hat{f}_\theta$ est donc un développement en termes de $\mathbf{x}^{(1)},\ldots,\mathbf{x}^{(n)}$, appelés vecteurs de support (SVs). Sa complexité a l'avantage d'être indépendante du nombre d'entrées d

 ISSN: 2102-6238

et dépend seulement du nombre de SVs.

Dans le cadre où la fonction f est fortement non-linéaire, le métamodèle $\hat{f}_\theta = \langle \mathbf{w}, \cdot \rangle_2 + b$ peut ne plus être satisfaisant. Une méthode classique consiste alors à utiliser l'astuce du noyau, en plongeant l'espace des paramètres d'entrée $\mathscr{X}$ dans un espace des *features* de plus grande dimension $\tilde{\mathscr{X}} = \Phi(\mathscr{X})$, où Φ est une transformation non linéaire. Il vient alors $\hat{\mathbf{w}} = \sum_{i=1}^n (\hat{\alpha}_i, \hat{\alpha}_i^*) \Phi\left(\mathbf{x}^{(i)}\right)$ ainsi que $\hat{f}_\theta(\mathbf{x}) = \sum_{i=1}^n (\hat{\alpha}_i - \hat{\alpha}_i^*) \langle \Phi\left(\mathbf{x}^{(i)}\right), \Phi(\mathbf{x}) \rangle + b$. L'astuce du noyau consiste alors à considérer une transformation non-linéaire Φ telle que pour tout $\mathbf{x}, \tilde{\mathbf{x}} \in \mathscr{X}$, $\langle \Phi(\mathbf{x}), \Phi(\tilde{\mathbf{x}}) \rangle = k(\mathbf{x}, \tilde{\mathbf{x}})$ où k est une fonction à noyau. En pratique, la connaissance de Φ importe peu, seule celle de k étant nécessaire dans l'expression du modèle de substitution :

$$\hat{f}_\theta(\mathbf{x}) = \sum_{i=1}^n (\hat{\alpha}_i - \hat{\alpha}_i^*) k\left(\mathbf{x}^{(i)}, \mathbf{x}\right) + b$$

et dans celle du problème dual (3) reformulé où pour tout $i \in \{1, \ldots, n\}$, $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right)$ remplace $\langle \mathbf{x}^{(i)}, \mathbf{x}^{(j)} \rangle$. Les exemples classiques de fonctions à noyaux considérées sont celles à noyau linéaire $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \langle \mathbf{x}^{(i)}, \mathbf{x}^{(i)} \rangle$, à noyau polynomial $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \left(\langle \mathbf{x}^{(i)}, \mathbf{x}^{(i)} \rangle + 1\right)^\gamma$, $\gamma > 0$, et à noyau Gaussien $k\left(\mathbf{x}^{(i)}, \mathbf{x}^{(j)}\right) = \exp\left(-0.5\gamma \|\mathbf{x}^{(i)} - \mathbf{x}^{(j)}\|_2^2\right)$, $\gamma > 0$.

2.5. *Métamodèles par chaos polynomial*

On considère à présent une métamodélisation basée sur une décomposition en polynômes de chaos. Pour cela, le vecteur des entrées $\mathbf{x}$ est considéré comme la réalisation d'une variable aléatoire $\mathbf{X}$ dont les composantes $X_1, \ldots, X_d$ sont indépendantes. On note $\mu_{\mathbf{X}}(\mathbf{x}) = \prod_{i=1}^d \mu_{X_i}(x_i)$ la fonction de densité de probabilité de $\mathbf{X}$, $P_{\mathbf{X}}$ la mesure de probabilité correspondante telle que $P_{\mathbf{X}}(d\mathbf{x}) = \mu_{\mathbf{X}}(\mathbf{x}) d\mathbf{x}$ et $F_{\mathbf{X}}$ la fonction de distribution associée, avec $F_{X_1}, \ldots, F_{X_d}$ les distributions marginales. On suppose aussi que la variable aléatoire $f(\mathbf{X})$ est de carré intégrable. Pour tout $i \in \{1, \ldots, d\}$, on considère alors une famille de polynômes univariés orthonormaux $\{\pi_j^{(i)},\ j \in \mathbb{N}\}$. On introduit ensuite l'index multidimensionnel $\alpha = \{\alpha_0, \ldots, \alpha_d\}$ et on définit la famille de polynômes multivariés $\{\psi_\alpha : \mathbf{x} \mapsto \prod_{i=1}^d \pi_{\alpha_i}^{(i)}(x_i),\ \alpha \in \mathbb{N}^d\}$. Lorsque $X_1, \ldots, X_d$ sont des variables gaussiennes indépendantes, on obtient le développement en polynômes de chaos de Wiener-Hermite :

$$f(\mathbf{X}) = \sum_{\alpha \in \mathbb{N}^d} y_\alpha \psi_\alpha(\mathbf{X})$$

où les polynômes univariés sont les polynômes d'Hermite et où $\forall \alpha \in \mathbb{R}^2$, $y_\alpha \in \mathbb{R}$. Des extensions existent pour d'autres lois de probabilité, en se basant sur le schéma d'Askey pour les polynômes orthogonaux (Xiu and Karniadakis, 2002). On parle alors de développement généralisé en polynômes de chaos.

En pratique, on considère une version tronquée de la série $f(\mathbf{X}) = \sum_{\alpha \in \mathbb{N}^d} y_\alpha \psi_\alpha(\mathbf{X})$:

$$\hat{f}_\theta(\mathbf{X}) = \sum_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| \leq p}} y_\alpha \psi_\alpha(\mathbf{X})$$

 ISSN: 2102-6238

qui comporte $P = \begin{pmatrix} n+p \\ p \end{pmatrix}$ polynômes ψ_α de degré total $|\alpha| = \sum_{i=1}^{d} \alpha_i$ inférieur à un entier p à fixer. Ce nombre P croît rapidement en fonction du nombre d'entrée d et du degré total p. Pour ce qui est de l'estimation des paramètres $\theta = (y_\alpha)_{\substack{\alpha \in \mathbb{N}^d \\ |\alpha| \leq p}}$, deux types d'approches existent. La première se compose de méthodes intrusives de type Galerkin (Ghanem and Spanos, 1991) ; elles sont cependant onéreuses en temps de calcul et nécessitent des modifications du simulateur numérique (aspect intrusif). La seconde se compose de méthodes non-intrusives requérant uniquement des évaluations du code de calcul. Ce sont des dernières que nous présentons ici.

Une première méthode non-intrusive est l'approche par projection où pour tout $\alpha \in \mathbb{N}^d$ vérifiant $|\alpha| \leq p$, on a $y_\alpha = \mathbb{E}\left[\psi_\alpha(\mathbf{X})\hat{f}_\theta(\mathbf{X})\right]$ par orthonormalité. En utilisant une approximation de l'intégrale, y_α peut être estimé au moyen d'un plan d'expériences $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(d)}\right\}$; on obtient alors :

$$\hat{y}_\alpha = \sum_{i=1}^{N} w_i \psi_\alpha\left(\mathbf{x}^{(i)}\right) f\left(\mathbf{x}^{(i)}\right)$$

où $w_1, \ldots, w_d$ sont des poids de quadratures, égaux par exemple à N^{-1} dans le plus simple des cas.

Lorsque la matrice $\Phi = \left(\Phi_{\alpha_j}\left(\mathbf{x}^{(i)}\right)\right)_{1 \leq i \leq n 1 \leq j \leq P}$ est inversible, une autre approche repose sur la version empirique du problème de minimisation :

$$\underset{\mathbf{y} \in \mathbb{R}^P}{\operatorname{argmin}}\ \mathbb{E}\left[\left(\sum_{\alpha \in \mathbb{N}^d | \alpha| \leq p} y_\alpha \psi_\alpha(\mathbf{X}) - f(\mathbf{X})\right)^2\right]$$

dont la solution s'écrit $\hat{\mathbf{y}} = \left(\Phi^T\Phi\right)^{-1}\Phi\mathbf{y}^n$, avec $\mathbf{y}^n = \left(y^{(1)}, \ldots, y^{(n)}\right)^T$.

2.6. *Autres exemples de modèles de substitution*

Les précédents modèles de substitution font partie des plus répandus et comportent une abondante littérature théorique, méthodologique et applicative. On peut faire référence à l'article de Simpson et al. (2001) comparant notamment les réseaux de neurones aux modèles par processus gaussien. Alors que les premiers permettent d'approcher les modèles hautement non-linéaires ou comportant un grand nombre d'entrées, les seconds ne peuvent considérer qu'un faible nombre d'entrées et disposent d'une formulation plus complexe. Simpson et al. (2001) mettent par ailleurs en avant l'avantage des surfaces de réponses (Kleijnen, 2015) lorsque le code de calcul comporte moins d'une dizaine de paramètres, du fait de la richesse théorique l'accompagnant et de la facilité à approcher des simulateurs comportant une erreur aléatoire.

Par ailleurs, il existe également les réseaux de neurones à base radiale (RBF pour *Radial Basis Function Network* (Broomhead and Lowe, 1988), les moindres carrés mobiles (Levin, 1998) et les modèles à bases de splines comme MARS (*Multivariate Adaptative Regression Splines* (Friedman, 1991). Lorsque le comportement du modèle à remplacer est très hétérogène sur l'espace des paramètres d'entrées, il peut aussi être opportun de construire un modèle de substitution par zone de comportement homogène et d'agréger ensuite ces différents métamodèles ; on parle alors de mélanges d'experts (Bettebghor et al., 2011). Enfin, lorsque le code de calcul est trop complexe

 ISSN: 2102-6238

à approcher du fait de phénomènes multiéchelles, des extensions des précédents métamodèles par apprentissage profond (*deep learning*) peuvent donner de meilleurs résultats qu'avec leurs versions classiques ; on parle ainsi de *deep gaussian process* (Damianou and Lawrence, 2013).

3. Apprentissage statistique pour la construction d'un modèle de substitution

Les méthodes classiques d'apprentissage statistique permettent d'estimer les paramètres d'un modèle de substitution afin qu'il approche au mieux la sortie du simulateur numérique d'intérêt. L'ouvrage de Hastie et al. (2009) présente les principaux éléments mentionnés dans cette section. Une vue d'ensemble de la théorie de l'apprentissage statistique est également proposée dans l'article de Vapnik (1999), et de façon plus détaillée dans le livre de Vapnik (1995). Enfin, l'ouvrage de Dreyfus et al. (2008) aborde l'apprentissage statistique sous un angle méthodologique dans le cadre des réseaux de neurones et des machines à vecteurs de support.

3.1. Approximation sur une base d'apprentissage

Pour un modèle de substitution $\hat{f}_\theta$ donné, les paramètres $\theta \in \Theta$ sont choisis afin que $\hat{f}_\theta$ soit proche du modèle de référence f pour les n valeurs d'un plan d'expériences $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$ représentatif de $\mathscr{X}$. On considère pour cela la base d'apprentissage $\mathscr{A} = \left(\mathbf{x}^{(i)}, y^{(i)}\right)_{1 \leq i \leq n}$ où $y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \varepsilon^{(i)}$. Le terme $\varepsilon^{(i)}$ représente l'erreur d'évaluation du modèle f en $\mathbf{x}^{(i)}$, lorsque f est une expérience en laboratoire ou un code de calcul stochastique notamment. Dans ces cas, les réels $\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}$ sont souvent vus comme des réalisations indépendantes d'une variable aléatoire gaussienne centrée dont l'écart-type représente le degré d'incertitude liée à l'évaluation de f.

Disposant de la base d'apprentissage $\mathscr{A}$, on cherche les paramètres $\theta \in \Theta$ minimisant un critère d'ajustement fonction de $\mathscr{A}$. On utilise couramment le risque empirique :

$$\mathscr{R}_{\text{emp}}\left(\hat{f}_\theta; \mathscr{A}\right) = \frac{1}{n} \sum_{i=1}^{n} Q\left(y^{(i)}, \hat{y}^{(i)}\right)$$

où $\hat{y}^{(i)} = \hat{f}_\theta\left(\mathbf{x}^{(i)}\right)$ et où la fonction Q est nommée contraste. Dans le cadre de la régression, la quantité $Q\left(y^{(i)}, \hat{y}^{(i)}\right)$ est souvent égale à $\left(y^{(i)} - \hat{y}^{(i)}\right)^2$; le risque empirique $\mathscr{R}_{\text{emp}}$ s'appelle alors l'erreur des moindres carrés et Q le contraste des moindres carrés. Les paramètres $\hat{\theta}$ optimaux selon la base d'apprentissage $\mathscr{A}$ correspondent à l'argument minimum du risque empirique $\mathscr{R}_{\text{emp}}\left(\hat{f}_\theta; \mathscr{A}\right)$. On parle aussi de minimisation de l'erreur d'apprentissage.

Le métamodèle $\hat{f}_{\hat{\theta}}$ ainsi construit est un estimateur de f dont la précision au point $\mathbf{x}$ est d'autant plus élevée que $\mathbf{x}$ est proche d'un élément de $\left\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}\right\}$. L'enjeu est alors de garantir une faible erreur de prédiction non seulement aux points de cet ensemble mais aussi sur la totalité de l'espace $\mathscr{X}$. La minimisation de l'erreur de généralisation va correspondre à un choix optimal de la complexité du métamodèle, basée sur son architecture et sur l'amplitude de ses paramètres.

3.2. Sélection de modèle et régularisation

Pour un modèle de substitution $\hat{f}_\theta$ donné, on appelle complexité la dimension ou le volume de l'espace des entrées Θ. En pratique, on recherche dans un premier temps les paramètres $\theta_1, \theta_2, \ldots, \theta_M$

 ISSN: 2102-6238

minimisant l'erreur d'apprentissage sur des sous-domaines restreints $\Theta_1, \Theta_2, ..., \Theta_M \subset \Theta$ de dimension ou volume croissants. Ainsi pour une sous-famille de métamodèles donnée $\left(\hat{f}_\theta\right)_{\theta\in\Theta_m}$, $\hat{f}_{\theta_m}$ minimise l'erreur d'apprentissage sur $\mathscr{A}$. Dans un second temps, on souhaite retenir le métamodèle $\hat{f}_{\theta_{m_*}}$ minimisant le risque :

$$\mathscr{R}\left(\hat{f}_{\theta_m}\right) = \int_{\mathscr{X}\times\mathscr{Y}} Q(y, \hat{f}_{\theta_m}(\mathbf{x}))dF(\mathbf{x}, y),$$

F étant la fonction de répartition de $(\mathbf{x}, y)$. Cette quantité est appelée fonction de perte ou erreur de généralisation de $\hat{f}_\theta$.

Le choix d'un vecteur θ de grande taille ou avec des composantes de grandes amplitudes peut conduire le modèle de substitution à devenir un expert de la base $\mathscr{A}$ et du voisinage du plan d'expériences au détriment du reste du domaine $\mathscr{X}$. C'est ce que l'on appelle le surajustement, ou surapprentissage. Schématiquement, la décroissance permanente de l'erreur d'apprentissage liée à l'augmentation du nombre de paramètres entraîne tout d'abord une diminution de l'erreur de généralisation à mesure que l'architecture du métamodèle est enrichie, puis cette erreur de généralisation se met à croître. La complexité optimale du modèle de substitution $\hat{f}_{\theta_*}$ réalise alors le minimum de l'erreur de généralisation. Le problème de surapprentissage ici soulevé peut être relié au compromis biais-variance : à nombre d'observations fixé, l'augmentation de la complexité d'un modèle de substitution implique un accroissement de la variance des estimateurs de ses paramètres et une diminution du biais du modèle. Ainsi, le risque quadratique, décomposable en la somme de la variance du métamodèle et du carré de son biais, diminue à mesure que la complexité croît avant d'augmenter ; son minimum correspond au meilleur compromis biais-variance et le modèle de substitution associé $\hat{f}_{\theta_*}$ possède une des meilleurs capacités de généralisation possibles sachant $\mathscr{A}$.

La fonction de distribution F étant le plus souvent inconnue, l'impossibilité de calculer le risque $\mathscr{R}$ ne permet pas d'obtenir le modèle de substitution optimal $\hat{f}_{\theta_{m_*}}$. En pratique, on cherche alors à approcher $\mathscr{R}$ pour les différentes valeurs de θ. Lorsque $(\Theta_m)_{1\leq m\leq M}$ sont des sous-domaines de Θ de dimension croissante, des méthodes de sélection de modèles visent à approcher le métamodèle $\hat{f}_{\theta_{m_*}}$ minimisant une version approchée de $\mathscr{R}$. Lorsque $(\Theta_m)_{1\leq m\leq M}$ sont des sous-domaines de Θ de volume croissant, les méthodes de régularisation permettent d'effectuer un unique apprentissage où l'erreur d'apprentissage associée est pénalisée par le volume des paramètres.

3.2.1. Sélection de modèles

Erreur de test et méthode hold-out

La plus simple des approches consiste à compléter la base d'apprentissage $\mathscr{A}$ par une base de test $\mathscr{T}$ de n_t évaluations de f pour laquelle on note $\mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m}; \mathscr{T}\right)$ l'erreur de test du métamodèle $\hat{f}_{\theta_m}$. Minimiser l'erreur de généralisation revient alors à chercher la sous-famille $\left(\hat{f}_\theta\right)_{\theta\in\Theta_{\hat{m}}}$ dont le minimiseur de l'erreur d'apprentissage $\hat{f}_{\theta_{\hat{m}}}$ minimise l'erreur de test. Cette erreur de test est proche de celle de généralisation lorsque $\mathscr{T}$ est un échantillon représentatif de $\mathscr{X}\times\mathscr{Y}$ et que n_t est suffisamment élevé. Cette méthode nécessite donc une base de test ; or dans le contexte de la substitution de codes de calcul coûteux, un nouveau jeu d'observations $\mathscr{T}$ ne peut pas toujours être généré à partir de nouveaux appels à f. Conséquemment, $\mathscr{T}$ est souvent construit en

 ISSN: 2102-6238

partitionnant la base d'apprentissage initiale en un groupe de test et un groupe d'apprentissage, c'est-à-dire $(\mathscr{A}, \mathscr{T}) := \mathscr{A}$. On parle alors d'approche *hold-out* car on écarte des observations de l'apprentissage pour constituer une base de test. Typiquement le nombre d'observations de test n_t est égal à un tiers ou un cinquième de celui de départ. Enfin, lorsque le nombre d'observations d'apprentissage est initialement très faible, il est déconseillé d'en écarter de la sorte pour constituer une base de test ; d'autres estimateurs de l'erreur de généralisation basés sur un rééchantillonnage de la base d'apprentissage sont alors à privilégier.

Erreur de validation croisée

Plus complexe, la validation croisée *K-folds* permet d'approcher l'erreur de généralisation d'un estimateur $\hat{f}_\theta$ en s'appuyant uniquement sur la base d'apprentissage $\mathscr{A}$, contrairement à l'approche par erreur de test.

Pour cela, l'ensemble $\mathscr{A}$ est partitionné – de façon aléatoire ou déterministe – en K échantillons uniformes $\mathscr{A}^{(k)}$, $k \in \{1,...,K\}$, couramment appelés *folds*. La phase d'apprentissage donne ensuite lieu à K vecteurs de paramètres $\theta_m^{(k)}$, $k \in \{1,...,K\}$, où le $k^{\text{ème}}$ minimise sur Θ_m l'erreur d'apprentissage $\mathscr{R}_{\text{emp}}\left(\hat{f}_\theta; \mathscr{A} \setminus \mathscr{A}^{(k)}\right)$. L'erreur de généralisation est alors approchée par l'erreur de validation croisée (Kohavi et al., 1995) définie par :

$$\mathscr{R}_{\text{CV}}\left(\hat{f}_{\theta_m}; \mathscr{A}\right) = \frac{1}{n}\sum_{k=1}^{K} n_k \mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m^{(k)}}; \mathscr{A}^{(k)}\right)$$

où $n_k = \text{Card}\left(\mathscr{A}^{(k)}\right)$. Le modèle de substitution sélectionné s'écrit $\hat{f}_{\theta_{\hat{m}}}$ où $\hat{m}$ minimise $\mathscr{R}_{\text{CV}}(\hat{f}_{\theta_m}; \mathscr{A})$ sur $\{1,\ldots,M\}$ et où $\theta_{\hat{m}}$ minimise $\mathscr{R}_{\text{emp}}(\hat{f}_\theta; \mathscr{A})$ sur $\Theta_{\hat{m}}$.

Cette méthode requiert de choisir le nombre de *folds* K. Une valeur élevée entraîne une forte variance de l'estimateur $\mathscr{R}_{\text{CV}}$ et un coût important associé à son obtention ; cette version est donc à utiliser lorsque la taille de la base d'apprentissage $\mathscr{A}$ est faible. Dans le cas particulier où $K = n$, chaque *fold* se compose d'une seule observation : c'est l'approche *leave-one-out*. Inversement, une faible valeur de K implique une variance plus faible. Cette méthode est à utiliser de préférence lorsque le nombre d'observations n est grand. De façon courante, on utilise la validation croisée 5-*folds*, 10-*folds* ou *leave-one-out*. Enfin, on note que la validation croisée *K-folds* applique K fois l'approche *hold-out* en écartant à tour de rôle n/K observations de $\mathscr{A}$ pour en faire une base de test.

Erreurs bootstrap

Le *bootstrap* non paramétrique est une autre méthode de rééchantillonnage ; elle permet d'approcher la distribution de l'estimateur $\hat{f}_\theta$ lorsque la loi de $\mathscr{A}$ est inconnue (voir les articles de Efron, 1981 et Efron and Tibshirani, 1993).

Cette technique consiste à remplacer la distribution inconnue F de la variable aléatoire $(\mathbf{x}, y)$ par la distribution empirique F_n qui donne un poids de $1/n$ à chaque observation $\left(\mathbf{x}^{(i)}, y^{(i)}\right), i \in \{1,...,n\}$. Pour cela, n tirages avec remise selon F_n sont effectués dans $\mathscr{A}$ et les issues constituent un échantillon *bootstrap* $\mathscr{A}^{*1} = \left(\mathbf{x}^{(i),*1}, y^{(i),*1}\right)_{i\in\{1,\ldots,n\}}$. Répétés B fois, ces tirages donnent lieu à B bases d'apprentissage *bootstrap* $\left(\mathscr{A}^{*b}\right)_{1\leq b\leq B}$. Par la suite, ces ensembles servent à la

 ISSN: 2102-6238

construction de la collection de modèles de substitution *bootstrap* $\left(\hat{f}_{\hat{\theta}_m^{*b}}\right)_{1\leq b\leq B}$ où les paramètres $\hat{\theta}_m^{*b}$ réalisent le minimum de l'erreur d'apprentissage $\mathscr{R}_{\text{emp}}(\hat{f}_\theta;\mathscr{A}^{*b})$ sur Θ_m.

L'erreur de généralisation est alors estimée par l'erreur *bootstrap* définie par :

$$\mathscr{R}_{\text{boot}}(\hat{f}_{\theta_m};\mathscr{A}) = \frac{1}{B}\sum_{b=1}^{B}\mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m^{*b}};\mathscr{A}\right).$$

Le modèle de substitution sélectionné s'écrit $\hat{f}_{\theta_{\hat{m}}}$ où $\hat{m}$ minimise $\mathscr{R}_{\text{boot}}(\hat{f}_{\theta_m};\mathscr{A})$ sur $\{1,\ldots,M\}$ et où $\theta_{\hat{m}}$ minimise $\mathscr{R}_{\text{emp}}(\hat{f}_\theta;\mathscr{A})$ sur $\Theta_{\hat{m}}$. L'estimateur $\mathscr{R}_{\text{boot}}$ possède un biais qui décroît lorsque B et n augmentent. Ce biais rend cette erreur optimiste et s'explique par le fait que les observations apprises par un modèle *bootstrap* $\hat{f}_{\theta_m^{*b}}$ servent au calcul de son erreur de test $\mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m^{*b}};\mathscr{A}\right)$. Pour pallier ceci, l'erreur *out-of-bag* dissocie dans sa formulation les données d'apprentissage de celles de test, ce qui la rend pessimiste. Afin de trouver un équilibre entre l'excès d'optimisme de la première erreur et l'excès de pessimisme de la seconde, l'erreur $e.632+$ a été proposée, 0.632 étant le poids associé à l'erreur *out-of-bag* et 0.368 à l'erreur *bootstrap* (Efron, 1983).

Les erreurs de test ont le défaut de nécessiter un lot d'observations supplémentaire tandis que celles de validation croisée et *bootstrap* sont coûteuses en temps de calcul, car impliquant la construction de plusieurs métamodèles pour une complexité Θ_m fixée. Par ailleurs, l'obtention de ces dernières déstructure la qualité des plans d'expériences issus des rééchantillonnage ; elle pénalise ainsi les métamodèles construits avec les bases d'observations associées à ces plans et biaise les erreurs *bootstrap* et de validation croisée. Les méthodes suivantes ont l'avantage de ne pas comporter ces limites.

Critères pénalisés

Sous l'hypothèse que la base $\mathscr{A}$ provient du modèle d'observation bruité $Y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \varepsilon^{(i)}$, avec $\varepsilon^{(1)},\ldots,\varepsilon^{(n)} \overset{\text{i.i.d.}}{\sim} \mathscr{N}(0,\sigma^2)$, une méthode de sélection de modèles consiste à minimiser un critère pénalisé fonction du risque empirique de l'estimateur, de la dimension ou du volume de l'espace des paramètres d'entrée et du nombre d'observations (Massart, 2008). Ce critère est défini pour tout $m \in \{1,\ldots,M\}$ par :

$$\mathscr{R}_{\text{pen}}\left(\hat{f}_{\theta_m};\mathscr{A}\right) = \mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m};\mathscr{A}\right) + \text{pen}(\Theta_m,n)$$

Le terme de pénalité positif $\text{pen}(\Theta_m,n)$ croît avec m et décroît avec n. Il permet de corriger l'excès d'optimisme de la minimisation de la seule erreur d'apprentissage $\mathscr{R}_{\text{emp}}\left(\hat{f}_{\theta_m};\mathscr{A}\right)$.

Bien que l'hypothèse d'un modèle d'observation bruité n'est pas vérifiée dans le cadre des simulateurs numériques, l'utilisation de critères pénalisés est régulièrement considérée en méta-modélisation et produit souvent des résultats satisfaisants.

Tout d'abord, dans le cadre de la statistique asymptotique, un des premiers critères utilisés est le C_p de Mallows (Mallows, 1973) ; il mesure la qualité d'ajustement d'un modèle de régression

 ISSN: 2102-6238

dont les p_m paramètres réalisent le minimum de l'erreur des moindres carrés. Il est défini pour tout $m \in \{1,...,m\}$ par :

$$C_p\left(\hat{f}_{\theta_m};\mathscr{A}\right) = \mathscr{R}_{\text{emp}}(\hat{f}_{\theta_m};\mathscr{A}) + 2\frac{p_m}{n}\hat{\sigma}^2$$

où $\hat{\sigma}^2$ est un estimateur du bruit de mesure σ^2.

Défini pour tout $m \in \{1,...,m\}$ par :

$$AIC\left(\hat{f}_{\theta_m};\mathscr{A}\right) = -2\log\left(\mathscr{L}\left(\hat{f}_{\theta_m};\mathscr{A}\right)\right) + 2p_m$$

où le terme $\mathscr{L}\left(\hat{f}_{\theta_m};\mathscr{A}\right) = \frac{1}{(2\pi\sigma^2)^{n/2}}\exp\left(-\sum_{i=1}^{n}\frac{\left(y^{(i)}-\hat{f}_{\theta_m}\left(\mathbf{x}^{(i)}\right)\right)^2}{2\sigma^2}\right)$ est la vraisemblance associée à $\mathscr{A}$, le critère d'information d'Akaike (Akaike, 1973) est également fréquemment utilisé. Il se base sur une notion d'entropie, offrant ainsi un estimateur de la perte d'information associée au modèle de substitution $\hat{f}_{\theta_m}$ pour représenter la fonction f ; son maximiseur est le métamodèle le plus informatif et est noté $\hat{f}_{\theta_{\hat{m}}}$. Pour un faible nombre d'observations n, ce critère pénalisé s'écarte de l'erreur de généralisation. Il est possible de l'améliorer en lui ajoutant par exemple un terme supplémentaire fonction du nombre d'observations n : $AICc\left(\hat{f}_{\theta_m}\right) = AIC\left(\hat{f}_{\theta_m}\right) + \frac{2p_m(n_\theta+1)}{n-p_m-1}$.

Un dernier critère asymptotique classiquement utilisé est le critère bayésien non informatif (Schwarz, 1978) défini par :

$$BIC\left(\hat{f}_{\theta_m};\mathscr{A}\right) = -2\log(\mathscr{L}\left(\hat{f}_{\theta_m};\mathscr{A}\right) + \log(n)p_m.$$

Le critère BIC conduit à des modèles de substitution plus parcimonieux qu'avec le critère AIC dès lors que $n > e^2$. L'article de Lebarbier and Mary-Huard (2006) présente de façon détaillée les fondements des indices AIC et BIC et leur interprétation dans le cadre de la sélection de modèles.

Dans le cadre de la statistique non-asymptotique (Massart, 2008), les travaux récents de sélection de modèles par critère pénalisé visent notamment à chercher des pénalités optimales de la forme $\kappa \times \text{pen}_0(\Theta_m, n)$, produit d'une constante positive et d'une pénalité élémentaire. Ces derniers étendent au cadre de la théorie non-asymptotique les idées portées par des résultats asymptotiques couramment utilisés (Massart, 2007). Une pénalité de type $\kappa \times \text{pen}_0(\Theta_m, n)$ est dite optimale si l'estimateur sélectionné $\hat{f}_{\theta_{\hat{m}}}$ au risque empirique pénalisé par $\kappa \times \text{pen}_0(\Theta_m, n)$ satisfait une inégalité oracle de la forme :

$$l\left(f,\hat{f}_{\theta_{\hat{m}}}\right) \leq A_n l\left(f,\hat{f}_{\theta_{m_*}}\right) + \eta_n \tag{4}$$

où l'excès de risque l (ou perte relative) est une fonction positive et minimale en f, définie pour tout $\theta \in \Theta$ par $l\left(f,\hat{f}_\theta\right) = \mathscr{R}\left(\hat{f}_\theta\right) - \mathscr{R}(f)$. L'inégalité (4) est qualifiée d'oracle car elle majore l'excès de risque associé à l'estimateur sélectionné $\hat{f}_{\theta_{\hat{m}}}$ par celui associé à l'oracle $\hat{f}_{\theta_{m_*}}$ où m_* est défini par $m_* \in \underset{m\in\{1,...,M\}}{\text{argmin}}\ l\left(f,\hat{f}_{\theta_m}\right)$. Par ailleurs la constante A_n tend vers 1 lorsque n tend vers l'infini. D'autre part, le terme de résidu η_n est petit devant la perte relative $l\left(f,\hat{f}_{\theta_{\hat{m}}}\right)$ et tend à disparaître lorsque n croît. Ainsi pour un nombre suffisamment important d'observations, on obtient des inégalités fines assurant un choix de modèle de substitution dont le risque est proche

 ISSN: 2102-6238

de celui de l'oracle. Néanmoins, lorsque n est petit, ces relations sont toujours vérifiées mais le majorant peut être important.

L'inégalité oracle (4) est écrite sous une forme forte lorsqu'elle est valide avec une grande probabilité ou bien en espérance selon la base d'apprentissage $\mathscr{A}$. Parfois, de tels résultats théoriques sont difficiles à obtenir et il est alors d'usage de considérer la forme faible suivante, implication naturelle de la forme forte :

$$\mathbb{E}_{\mathscr{A}}\left[l\left(f,\hat{f}_{\theta_{\hat{m}}}\right)\right] \leq A_n \inf_{1\leq m\leq M} \mathbb{E}_{\mathscr{A}}\left[l\left(f,\hat{f}_{\theta_m}\right)\right] + \eta_n. \tag{5}$$

Lorsque le nombre d'observations n augmente, le rapport des excès de risque $l\left(f,\hat{f}_{\theta_{\hat{m}}}\right)/l\left(f,\hat{f}_{\theta_{m_*}}\right)$ tend vers 1, c'est-à-dire que le modèle sélectionné tend à produire un estimateur de risque minimal au sens de l'oracle. La forme faible (5) énonce la même conclusion non pas pour le rapport des risques mais pour le rapport du risque moyen de l'estimateur sélectionné par rapport au n-échantillon $\mathscr{A}$ sur le plus petit des risques moyens associés aux différentes complexités.

Différentes formes de pénalité $\text{pen}_0(\Theta_m,n)$ existent dans la littérature selon la famille de fonctions considérée ; le plus souvent elles sont proportionnelles au rapport de la dimension de Θ_m sur le nombre d'observations n. Un tableau de synthèse est présenté par Baudry et al. (2012) ; il comporte des formes de pénalité pour des problèmes classiques comme les modèles linéaires gaussiens, l'estimation de densité par histogramme, les modèles de mélange gaussien ou encore la régression par splines.

La constante optimale κ menant à l'inégalité oracle peut être approchée par une méthode de validation croisée ou bien par des méthodes plus récentes de saut de dimension et d'heuristique de pente (Arlot and Massart, 2009; Baudry et al., 2012).

Une dernière approche consiste à considérer non pas des pénalités proportionnelles à une forme prédéfinie comme $\kappa \times \text{pen}(\Theta_m,n)$ mais à utiliser des pénalités obtenues par rééchantillonnage de la base d'apprentissage (Arlot, 2010, 2009).

3.2.2. Régularisation

Les précédentes méthodes de sélection de modèles visent à sélectionner un métamodèle de complexité optimale au sein d'une collection donnée. Son nombre de paramètres ou leurs amplitudes doivent être suffisamment élevés afin d'approcher convenablement la fonction de référence f et suffisamment faibles pour ne pas être trop dépendant des données d'apprentissage.

Ces contraintes sur la complexité peuvent également être traitées par l'ajout d'un terme de régularisation au risque empirique. Plutôt que d'approcher la complexité optimale une fois des métamodèles construits à partir de $\mathscr{A}$ pour différents sous-domaines $\Theta_1,\ldots,\Theta_M$ de Θ, la régularisation vise à limiter l'amplitude des paramètres présents dans le métamodèle durant sa construction. Cette contrainte sur l'amplitude permet notamment d'établir des métamodèles comportant un nombre de paramètres important par rapport à celui d'observations, tout en garantissant une faible erreur de généralisation.

D'un point de vue terminologique, les méthodes de sélection de modèle opérant une fois les estimateurs de complexités différentes construits sont qualifiées de passives ; c'est le cas des précédents critères pénalisés. Les approches opérant sur le pouvoir de généralisation du métamodèle au cours de l'apprentissage de données sont quant à elles dites actives. En substitution

 ISSN: 2102-6238

de modèles, la méthode la plus utilisée est la régression *ridge* (Hoerl and Kennard, 1970),aussi appelée régularisation de Tikhonov (Tikhonov and Arsenin, 1977), basée sur une pénalisation ℓ_2, qui contribue à lisser la réponse du métamodèle. On retrouve également des pénalisations ℓ_1 de type *lasso* (Tibshirani, 1994) permettant de retirer les paramètres non significatifs d'un métamodèle ; les modèles de substitution ainsi obtenus sont qualifiés de creux (*sparse*). Cette approche est particulièrement utile pour les modèles riches en paramètres, comme les réseaux de neurones artificiels par exemple. Il est par ailleurs possible d'utiliser la pénalité *Elastic Net* (Zou and Hastie, 2005) qui combine les propriétés des deux précédentes.

Précisément, les régularisations *ridge* et *lasso* consistent à ajouter respectivement une pénalité ℓ_2 et ℓ_1 sur les paramètres θ du modèle de substitution :

$$\mathscr{R}_{\text{emp},\ell_i}\left(\hat{f}_\theta;\mathscr{A}\right) = \mathscr{R}_{\text{emp}}\left(\hat{f}_\theta;\mathscr{A}\right) + \|\Gamma\theta\|_i^2,\ i \in \{1,2\}$$

où Γ est une matrice de pondération, dite de Tikhonov pour la pondération *ridge*, souvent diagonale voire égale à la matrice identité. Ainsi on rencontre souvent le critère pénalisé $\mathscr{R}_{\text{emp},\ell_i}\left(\hat{f}_\theta;\mathscr{A}\right) = \mathscr{R}_{\text{emp}}\left(\hat{f}_\theta;\mathscr{A}\right) + \lambda\,\|\theta\|_i^2$, $i \in \{1,2\}$, où λ est une constante positive à calibrer, par validation croisée ou inférence bayésienne par exemple. La minimisation du risque empirique pénalisé se concentre d'autant plus sur la minimisation de la norme des paramètres au détriment de la quantité $\mathscr{R}_{\text{emp}}\left(\hat{f}_\theta;\mathscr{A}\right)$ que cette constante est grande.

La pénalisation *ridge* correspond au problème de minimisation du risque empirique sous la contrainte que l'amplitude des paramètres $\|\theta\|_2^2$ soit inférieure à une certaine valeur ; on parle en statistique de régression *ridge*. D'un point de vue bayésien (Goldstein, 2007), la pénalisation *ridge* dans un modèle linéaire revient à poser un *a priori* gaussien sur les paramètres. Le minimiseur du risque empirique pénalisé $\mathscr{R}_{\text{emp},\ell_2}$ correspond dans ce cas à l'espérance *a posteriori* de θ. Pour la pénalisation *lasso*, la même lecture peut être faite en considérant un *a priori* laplacien sur les paramètres du modèle linéaire.

Les éléments classiques d'apprentissage présentés dans ce chapitre sont utilisés pour la construction de modèles de substitution et particulièrement dans le choix de leur complexité architecturale. Dans le cas d'un réseau de neurones par exemple, ils permettent de sélectionner le nombre de neurones, de réguler l'amplitude des paramètres ou encore de supprimer ceux de faible significativité afin de garantir au métamodèle un bon pouvoir de généralisation. Dans le cas du modèle par processus gaussien d'autre part, ils servent à assurer cette capacité de prédiction sur l'ensemble du domaine $\mathscr{X}$ en sélectionnant les valeurs des différentes longueurs de portée ainsi que, cas échéant, la valeur de l'effet pépite pour un métamodèle régressant. Souvent, les différentes familles de modèles de substitution présentent des adaptations de ces techniques (voir l'ouvrage de Dreyfus et al., 2008 pour le cas des réseaux de neurones).

Après avoir présenté les principaux modèles de substitution et des outils permettant leur construction par apprentissage statistique, la section suivante étend la notion de métamodèle au cadre de la multifidélité où l'objectif est d'approcher un code de calcul à partir d'évaluations de ce dernier, complétées par des observations de modèles simplifiés.

 ISSN: 2102-6238

4. Modèles de substitution multifidélité

L'objet des sections précédentes consistait à approcher au moyen d'un modèle de substitution $\hat{f}_\theta$ un simulateur numérique f associant une sortie $y \in \mathscr{Y} \subset \mathbb{R}$ au vecteur des entrées $\mathbf{x} \in \mathscr{X} \subset \mathbb{R}^d$. Les paramètres $\theta \in \Theta$ minimisent une erreur d'apprentissage pénalisée par la complexité de Θ et associée au n-échantillon $\mathscr{A} = \left(\mathbf{x}^{(i)}, y^{(i)}\right)_{1 \leq i \leq n}$; $y^{(1)}, \ldots, y^{(n)}$ sont des réalisations de :

$$Y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \varepsilon^{(i)},\ i \in \{1, \cdots, n\} \tag{6}$$

avec $\varepsilon^{(1)}, ..., \varepsilon^{(n)}$ des variables aléatoires indépendantes et identiquement distribuées de moyenne nulle et de variance inconnue σ^2. Ce problème d'identification est dit homoscédastique car le niveau de bruit σ^2 est le même pour toutes les observations.

Du fait du compromis biais-variance, cette approximation est souvent de mauvaise qualité lorsque f est un code de calcul dont l'évaluation est onéreuse. En effet, ce coût calculatoire limite trop le nombre d'observations n par rapport au nombre de paramètres de $\hat{f}_\theta$ nécessaire pour la bonne prise en compte de la complexité du phénomène physique. Ne pouvant fournir une forte densité d'observations dans l'espace des paramètres d'entrée du modèle, le métamodèle $\hat{f}_\theta$ ne peut assurer une faible erreur de généralisation.

Il devient dès lors compliqué de lutter contre le dilemme du biais et de la variance en présence d'un simulateur numérique f coûteux en temps de calcul. Il peut alors être utile de compléter le modèle d'observation (6) par d'autres moins gourmands en temps d'exécution, permettant ainsi de construire des échantillons de tailles plus importantes. En thermique par exemple, un simulateur numérique peut s'accompagner de modèles réduits physiques ou de mesures expérimentales (De Lozzo, 2013). D'autre part, le maillage utilisé par le simulateur peut être dégradé de sorte à accélérer les évaluations du code de calcul. Un dernier point enfin est l'utilisation des résultats non convergés du simulateur (Picheny and Ginsbourger, 2013) ; ces derniers s'obtiennent en réduisant les nombres d'itérations dans les algorithmes de résolution des systèmes d'équations.

Cependant, ces modèles d'observation complémentaires sont d'une fidélité moindre, la fidélité d'un modèle d'observation correspondant ici à la confiance qu'on lui accorde pour représenter le phénomène d'intérêt ; ayant différents niveaux de fidélité pour approcher ce phénomène, on parle de modélisation multifidélité. Les modèles complétant le modèle de référence f peuvent être vus comme des versions dégradées de f soit parce qu'ils ne représentent pas f mais une fonction qui s'en approche, soit parce qu'ils observent f mais avec un bruit de mesure plus important que dans l'expression (6). L'exploration de ces sources de données multifidélité est l'objet de la présente section. Nous présentons dans un premier temps deux modèles-hypothèses permettant de les prendre en compte avant de présenter un état de l'art de différents outils existants. Nous concluons la section par un focus sur le modèle de co-krigeage basé sur la métamodélisation par processus gaussiens.

4.1. Modèles-hypothèses en simulation multifidélité

La prise en compte de différentes sources d'observations ne peut pas se limiter à l'utilisation d'un seul et unique type de modélisation, la notion de multifidélité variant selon la nature de ces sources. En présence de versions dégradées du simulateur numérique f ou de modèles physiques simplifiés,

 ISSN: 2102-6238

la fidélité correspond à une différence de comportement de la sortie par rapport au modèle de référence f sur l'ensemble du domaine $\mathscr{X}$, tandis que pour des séries de mesures expérimentales, elle correspond à une différence de bruit de mesure. À ces deux formes de multifidélité sont associées des hypothèses de modélisation différentes.

4.1.1. Hypothèse d'un bruit d'observation hétéroscédastique

On suppose que les modèles complémentaires représentent exactement le même phénomène physique f mais avec une erreur d'observation de niveau différent pour chacun d'eux liée à une imprécision de mesure dans le cas de mesures expérimentales par exemple ou à un bruit numérique. Cette approche est un premier pas vers la multifidélité car elle suppose que la différence de confiance entre les groupes d'observations n'est due qu'à un phénomène aléatoire additif. Concrètement, on dispose de K modèles d'observation avec des bruits différents :

$$\left\{\begin{array}{l} Y_K^{(i)} = f\left(\mathbf{x}_k^{(i)}\right) + \varepsilon_K^{(i)},\ 1 \leq i \leq n_K \\ Y_{K-1}^{(i)} = f\left(\mathbf{x}_{K-1}^{(i)}\right) + \varepsilon_{K-1}^{(i)},\ 1 \leq i \leq n_{K-1} \\ \vdots \\ Y_1^{(i)} = f\left(\mathbf{x}_1^{(i)}\right) + \varepsilon_1^{(i)},\ 1 \leq i \leq n_1 \end{array}\right.$$

où pour chaque niveau $k \in \{1,...,K\}$, les variables aléatoires $\varepsilon_k^{(1)},...,\varepsilon_k^{(n_k)}$ sont indépendantes et identiquement distribuées, d'espérance nulle et de variance σ_k^2. La base d'apprentissage associé à l'indice k s'écrit $\mathscr{A}_k = \left(\mathbf{x}_k^{(i)}, y_k^{(i)}\right)_{1 \leq i \leq n_k}$. Ces modèles sont ordonnés par niveau de bruit décroissant, *i.e.* $\sigma_1 > \ldots > \sigma_{K-1} > \sigma_K \geq 0$, où le $K^{\text{ème}}$ modèle d'observation correspond au modèle le plus fin et le plus coûteux, potentiellement non bruité, et le premier au modèle le plus grossier et le plus économique. Le nombre d'observations pouvant être obtenues avec chaque modèle satisfait alors en général les relations $n_K \leq n_{K-1} \leq ... \leq n_1$. Par identification au modèle de plus haute fidélité (6), on note $\sigma_K := \sigma$, $n_K := n$, $\mathscr{A}_K := \mathscr{A}$ ainsi que $\forall i \in \{1,...,n_K\},\ \varepsilon_K^{(i)} := \varepsilon^{(i)}$.

4.1.2. Hypothèse de tendances hétérogènes

On suppose à présent que bien que les K modèles cherchent à expliquer la même relation physique, ils n'offrent pas une même formulation de la tendance principale. Concrètement, on dispose de K modèles d'observation avec des tendances différentes :

$$\left\{\begin{array}{l} Y_K^{(i)} = f_K\left(\mathbf{x}_K^{(i)}\right) + \varepsilon_K^{(i)},\ 1 \leq i \leq n_K \\ Y_{K-1}^{(i)} = f_{K-1}\left(\mathbf{x}_{K-1}^{(i)}\right) + \varepsilon_{K-1}^{(i)},\ 1 \leq i \leq n_{K-1} \\ \vdots \\ Y_1^{(i)} = f_1\left(\mathbf{x}_1^{(i)}\right) + \varepsilon_1^{(i)},\ 1 \leq i \leq n_1 \end{array}\right.$$

où les variables aléatoires $\varepsilon_1^{(1)},...,\varepsilon_1^{(n_1)},...,\varepsilon_K^{(1)},...,\varepsilon_K^{(n_K)}$ sont indépendantes et identiquement distribuées, d'espérance nulle et de variance σ^2. La précision de ces modèles d'observation est

 ISSN: 2102-6238

croissante avec l'indice de fidélité $k \in \{1,...,K\}$ et donc, le nombre d'observations pouvant être obtenues avec chaque modèle est en général régi par les relations $n_1 \geq \ldots \geq n_{K-1} \geq n_K$. Par identification au modèle de plus haute fidélité (6), on note $f_K := f$, $\mathscr{A}_K := \mathscr{A}$, $n_K := n$ ainsi que $\forall i \in \{1,...,n_K\},\ \varepsilon_K^{(i)} := \varepsilon^{(i)}$.

Notons qu'il est aussi envisageable de procéder à une modélisation prenant en compte différentes tendances et différents niveaux de bruit. Ceci pourrait être justifié dans le cas où nous disposerions par exemple d'un modèle numérique fin, d'une version dégradée de ce dernier ainsi que de plusieurs séries de mesures du phénomène physique d'intérêt répondant à un même protocole expérimental mais réalisées par différents opérateurs.

Cas particulier à deux niveaux de fidélité

En général dans la littérature, on retrouve ces hypothèses pour $K = 2$, lorsqu'un modèle d'observation de basse-fidélité (BF) complète le modèle d'observation (6) de haute-fidélité (HF). Elles se déclinent sous les formes :

Hypothèse d'un bruit d'hétéroscédastique

$$\begin{cases} \text{HF} : Y_2^{(i)} = f\left(\mathbf{x}_2^{(i)}\right) + \varepsilon_2^{(i)},\ 1 \leq i \leq n_2 \leq n_1 \\ \text{BF} : Y_1^{(i)} = f\left(\mathbf{x}_1^{(i)}\right) + \varepsilon_1^{(i)},\ 1 \leq i \leq n_1 \\ \text{avec } \sigma_1 > \sigma_2 \geq 0 \text{ où } \mathbb{E}\left[\left(\varepsilon_k^{(1)}\right)^2\right] = \sigma_k^2,\ k \in \{1,2\} \end{cases}$$

et

Hypothèse de tendances hétérogènes

$$\begin{cases} \text{HF} : Y_2^{(i)} = f_2\left(\mathbf{x}_2^{(i)}\right) + \varepsilon_2^{(i)},\ 1 \leq i \leq n_2 \leq n_1 \\ \text{BF} : Y_1^{(i)} = f_1\left(\mathbf{x}_1^{(i)}\right) + \varepsilon_1^{(i)},\ 1 \leq i \leq n_1 \\ \text{avec fidélité}(f_2) > \text{fidélité}(f_1). \end{cases}$$

Les sections suivantes présentent des références bibliographiques répondant à ces deux hypothèses de modélisation de la fonction de référence f.

4.1.3. Processus de correction sur la sortie

Sous l'hypothèse de comportements distincts, il est courant d'approcher dans un premier temps le modèle d'observation de basse fidélité avant d'approcher dans un second temps la relation entre ce dernier et le modèle d'observation haute-fidélité. Autrement dit, on crée un estimateur $\hat{f}_{2,\theta}$ de f_2 sous la forme d'une fonction de lien $\hat{g}_{\text{OUT}}$ prenant en entrée le vecteur $\mathbf{x}$ et $\hat{f}_{1,\theta}(\mathbf{x})$, un estimateur de $f_1(\mathbf{x})$:

$$\hat{f}_{2,\theta}(\mathbf{x}) = \hat{g}_{\text{OUT}}\left(\mathbf{x}, \hat{f}_{1,\theta}(\mathbf{x})\right).$$

La fonction de lien $\hat{g}_{\text{OUT}}$ peut être vue comme un processus de correction sur la sortie basse-fidélité (Forrester and Keane, 2009). Typiquement on considère des métamodèles de la forme :

$$\hat{f}_{2,\theta}(\mathbf{x}) = \hat{\alpha}(\mathbf{x})\hat{f}_{1,\theta}(\mathbf{x}) + \hat{\beta}(\mathbf{x}) \tag{7}$$

 ISSN: 2102-6238

où les fonctions $\hat{\alpha}$ et $\hat{\beta}$ sont paramétrées à partir de la base d'apprentissage de haute-fidélité $\mathscr{A}_2$ associée au modèle f_2 tandis que le métamodèle $\hat{f}_{1,\theta}$ est paramétré à partir de la base d'apprentissage $\mathscr{A}_1$. Kennedy and O'Hagan (2000) ont proposé cette démarche en expérimentation numérique afin de prédire la sortie d'un code de calcul complexe lorsqu'un approximateur à évaluation rapide est disponible ; leur méthode repose sur des métamodèles par processus gaussiens et est détaillée à la section 4.2.

La réussite de cette approche basée sur un processus de correction repose sur l'hypothèse que, disposant de peu d'observations de haute-fidélité, il est plus facile d'approcher au moyen de ces dernières la relation entre le modèle d'observation de haute-fidélité f_2 et le modèle d'observation de basse-fidélité f_1, que le modèle de haute-fidélité f_2 directement.

Pour Sun et al. (2011) par exemple, le code grossier f_1 est remplacé par un modèle de régression linéaire basée sur des prédicteurs $p_1(\mathbf{x}),...,p_{n_p}(\mathbf{x})$ dont les paramètres $a_1(\mathbf{x}),...,a_{n_p}(\mathbf{x})$ dépendent de $\mathbf{x}$. Ils utilisent pour cela une méthode de moindres carrés glissants (MLS pour *moving least squares*) qui leur permet d'obtenir l'estimateur $\hat{f}_{1,\theta}(\mathbf{x}) = \hat{\mathbf{a}}^T(\mathbf{x})p(\mathbf{x})$ où le vecteur $\hat{\mathbf{a}}$ minimise le critère

$$J(\mathbf{a}) = \frac{1}{n_1}\sum_{i=1}^{n_1} w\left(\mathbf{x}-\mathbf{x}_1^{(i)}\right)\left(\mathbf{a}^T\left(\mathbf{x}_1^{(i)}\right)\mathbf{p}\left(\mathbf{x}_1^{(i)}\right) - f_1\left(\mathbf{x}_1^{(i)}\right)\right)^2.$$

Le terme $w\left(\mathbf{x}-\mathbf{x}_1^{(i)}\right)$ est une fonction de pondération tendant vers 0 en s'éloignant de $\mathbf{x}_1^{(i)}$ et $\mathbf{p}(\mathbf{x}) = \left(p_1(\mathbf{x}),...,p_{n_p}(\mathbf{x})\right)^T$ est le vecteur des n_p régresseurs qui peuvent être des monômes dans le cadre d'une approximation polynomiale.
Par la suite, ils supposent qu'il existe une relation liant le code grossier f_1 au code fin f_2 de la forme multiplicative $f_2(\mathbf{x}) = \alpha(\mathbf{x})f_1(\mathbf{x})$ ou bien de la forme additive $f_2(\mathbf{x}) = f_1(\mathbf{x}) + \beta(\mathbf{x})$.

Les inconnues α et β sont modélisées par des polynômes $\hat{\alpha}$ et $\hat{\beta}$ fonction de $\mathbf{x}$ et de degrés inférieurs ou égaux à m tels que :

$$\hat{\alpha} = \underset{Q\in\mathbb{R}_m[\mathbf{x}]}{\operatorname{argmin}} \frac{1}{n_2}\sum_{i=1}^{n_2}\left(y_2^{(i)} - Q\left(\mathbf{x}_2^{(i)}\right)\hat{f}_{1,\theta}\left(\mathbf{x}_2^{(i)}\right)\right)^2$$

et

$$\hat{\beta} = \underset{Q\in\mathbb{R}_m[\mathbf{x}]}{\operatorname{argmin}} \frac{1}{n_2}\sum_{i=1}^{n_2}\left(y_2^{(i)} - \left(\hat{f}_{1,\theta}\left(\mathbf{x}_2^{(i)}\right) + Q\left(\mathbf{x}_2^{(i)}\right)\right)\right)^2.$$

Ce modèle est employé avec réussite dans un cadre d'optimisation d'un processus de formage de tôle. Néanmoins, pour une relation plus complexe entre le modèle-basse fidélité et le modèle de haute-fidélité, il est parfois opportun de considérer à la fois le terme multiplicatif α et le terme additif β, ce qui correspond à la modélisation (7). Pour un système comportant un plus grand nombre de paramètres d'entrée, il peut par ailleurs s'avérer nécessaire de considérer des modèles de substitution plus complexes que des polynômes pour approcher α et β.

Ainsi, disposant déjà d'un estimateur $\hat{f}_{1,\theta}$ du modèle de basse-fidélité f_1, la méthode de Li and Wang (2011) vise à construire la fonction de lien $\hat{g}_{\text{OUT}}$ au moyen d'un terme multiplicatif $\alpha(\mathbf{x})$ et d'un terme additif $\beta(\mathbf{x})$. Pour cela, les auteurs utilisent un perceptron multicouche à une couche cachée et deux neurones de sortie : un jouant le rôle de $\alpha(\mathbf{x})$, l'autre celui de $\beta(\mathbf{x})$; la fonction d'activation des neurones cachés et de sortie est la tangente hyperbolique. Un coefficient

 ISSN: 2102-6238

w compris entre 0 et 1 permet par ailleurs de choisir si l'estimateur $\hat{f}_{2,\theta}$ du modèle de haute-fidélité est plutôt de la forme $\hat{f}_{1,\theta}$ corrigée par une tendance localisée $\beta(\mathbf{x})$ ou bien de la forme $\hat{f}_1$ corrigée par un facteur d'échelle localisé $\alpha(\mathbf{x})$; une fois ces différents termes estimés, il vient :

$$\hat{f}_{2,\theta}(\mathbf{x}) = w\hat{f}_{1,\theta}(\mathbf{x})\,\alpha(\mathbf{x}) + (1-w)\left(\hat{f}_{1,\theta}(\mathbf{x}) + \beta(\mathbf{x})\right)$$

où $w \in [0,1]$. Ce modèle est utilisé avec succès dans un processus d'optimisation globale de la forme du profil d'une aile d'avion. La combinaison de quelques simulations CFD (*Computational Fluid Dynamics*) coûteuses et d'un grand nombre de simulations d'un modèle de basse-fidélité donne ainsi de bons résultats tout en économisant significativement du temps de calcul.

Pour Kim et al. (2007) par ailleurs, la fonction $\hat{f}_{1,\theta}$ est un polynôme d'ordre 2 fonction des entrées x_1, x_2, ... et x_d calibré à partir d'un modèle physique simplifié et rapide en temps de calcul, tandis que la fonction de lien $\hat{g}_{\text{OUT}}$ est :

- soit un polynôme d'ordre 2 fonction de $\hat{f}_{1,\theta}$ uniquement : $\hat{f}_{2,\theta}(\mathbf{x}) = a_0 + a_1\hat{f}_{1,\theta}(\mathbf{x}) + a_2\left(\hat{f}_{1,\theta}(\mathbf{x})\right)^2$,
- soit un perceptron multicouche à deux couches cachées où les entrées de la première couche cachée sont $x_1 \ldots, x_d$ et où les entrées de la seconde sont les sorties de la première couche cachée ainsi que la sortie de $\hat{f}_{1,\theta}$. Kim et al. (2007) parlent alors de *knowledge-based artificial neural network* (KBNN), car le réseau de neurones s'appuie sur la connaissance d'un modèle physique simplifié. Ce métamodèle approchant le modèle de haute-fidélité peut être vu comme une boîte grise (Oussar and Dreyfus, 2001), représentation intermédiaire entre la boîte noire expliquant une sortie par des entrées en faisant une abstraction implicite de la physique et la boîte blanche reposant explicitement sur les équations décrivant cette dernière.

Sur un exemple industriel de processus de formage à chaud, Kim et al. (2007) montrent que cette "approche multifidélité hybride" donne de bien meilleurs résultats en prédiction que lorsque l'on construit uniquement un estimateur $\hat{f}_{1,\theta}$ du modèle d'observation basse-fidélité. Ceci est d'autant plus vrai lorsque $\hat{g}_{\text{OUT}}$ peut prendre une forme complexe, comme un perceptron multicouche par exemple.

Néanmoins, les méthodes développées par Li and Wang (2011) et Kim et al. (2007) présentent un inconvénient majeur : la relation entre le modèle de basse-fidélité et le modèle de haute-fidélité est assurée par un perceptron multicouche à deux couches cachées qui par nature comporte beaucoup de paramètres. Or, ces paramètres devant être fixés à partir des observations du modèle de haute-fidélité qui sont peu nombreuses, on est face à une situation de surparamétrisation. L'utilisation de réseaux de neurones pour modéliser le passage de la basse-fidélité à la haute-fidélité peut ainsi présenter des limites dans le cas où le nombre d'évaluations du modèle le plus fin est très faible relativement au nombre d'entrées. Il est ainsi fréquent de rencontrer des études où le code de calcul précis ne peut fournir que quelques dizaines de simulations, ces dernières nécessitant chacune plusieurs heures, voire journées, d'exécution. Pour cela sera abordé en fin de section le métamodèle co-krigeage (Kennedy and O'Hagan, 2000) ; ce modèle de substitution développé et appliqué depuis quelques années est un processus de correction sur la sortie étendant la notion de métamodèles par processus gaussiens et comportant un faible nombre de paramètres.

 ISSN: 2102-6238

4.1.4. *Processus de correction sur les entrées*

Une autre approche présentée par Forrester and Keane (2009) consiste non pas à appliquer un processus de correction sur la sortie mais sur le vecteur d'entrée : on parle alors de *space mapping* (Bandler et al., 2004). Ainsi, une fois l'estimateur $\hat{f}_{1,\theta}$ du modèle d'observation basse-fidélité obtenu, on cherche à construire une fonction de lien $\hat{g}_{\text{IN}}$ à partir des observations haute-fidélité telle que

$$\hat{f}_{2,\theta}(\mathbf{x}) = \hat{f}_{1,\theta}(\hat{g}_{\text{IN}}(\mathbf{x})).$$

La philosophie du *space mapping* consiste donc à déformer légèrement l'espace des entrées du modèle d'observation haute-fidélité afin de l'utiliser comme espace des entrées du modèle d'observation basse-fidélité et de faire coller la sortie de l'estimateur $\hat{f}_{1,\theta}$ sur celle de l'estimateur $\hat{f}_{2,\theta}$ pour les observations $\left(\mathbf{x}_2^{(1)}, y_2^{(1)}\right), \ldots, \left(\mathbf{x}_2^{(n_2)}, y_2^{(n_2)}\right)$. Lorsque ce raccord n'est pas satisfaisant, on peut ajouter un processus de correction sur la sortie :

$$\hat{f}_{2,\theta}(\mathbf{x}) = \hat{g}_{\text{OUT}}\left(\mathbf{x}, \hat{f}_{1,\theta}(\hat{g}_{\text{IN}}(\mathbf{x}))\right).$$

4.1.5. *Régression hétéroscédastique*

Le cadre classique du problème de régression consiste à chercher un estimateur de la fonction de référence f dans un espace de fonctions. La complexité de cet ensemble peut conduire à un surajustement. Des méthodes de sélection de modèles sont couramment utilisées afin d'éviter cette situation, en contraignant pour cela la complexité de l'estimateur. De cette manière, des sous-ensembles plus parcimonieux sont créés et le but est de trouver le meilleur estimateur dans le sous-espace optimal. Pour ce faire, une des principales approches est la pénalisation qui correspond à la minimisation d'un risque empirique complété par un terme additionnel appelé "pénalité", augmentant avec la dimension du modèle et diminuant avec l'augmentation du nombre d'observations (Barron et al., 1999). La plupart du temps, l'estimateur est une fonction d'observations indépendantes et identiquement distribuées ; c'est le cas de la régression homoscédastique où le niveau du bruit est uniforme sur l'espace des paramètres d'entrée. De plus, la pénalité est souvent proportionnelle à la dimension du modèle, notamment pour les critères bien connus que sont le C_p de Mallows (Mallows, 1973) ou la pénalité AIC (Akaike, 1973). Ce cadre de la régression homoscédastique est présenté plus en détails dans la section 3.

Lorsque l'hypothèse faite sur le type de multifidélité porte sur le bruit d'observation, ces considérations ne suffisent plus et il est nécessaire de passer à un cadre de régression hétéroscédastique. Il est pour cela courant de considérer le modèle de régression

$$Y^{(i)} = f\left(\mathbf{x}^{(i)}\right) + \sigma^{(i)}\varepsilon^{(i)},\ i = 1, \ldots, n,$$

où les variables $\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(n)}$ appartiennent à l'hypercube $[0,1]^d$ et où $\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}$ sont des variables aléatoires indépendantes et identiquement distribuées selon la loi normale standard. Chaque observation possède son propre niveau de bruit caractérisé par l'écart-type σ. L'intérêt

 ISSN: 2102-6238

est alors porté sur l'estimation de la fonction inconnue f. Pour cela, ce modèle est habituellement réécrit sous la forme matricielle

$$\mathbf{Y} = \mathbf{f} + \Sigma\varepsilon$$

où $\mathbf{Y} = \left(Y^{(1)}, \ldots, Y^{(n)}\right)^T$, $\mathbf{f} = \left(f\left(\mathbf{x}^{(1)}\right), \ldots, f\left(\mathbf{x}^{(n)}\right)\right)^T$ et $\varepsilon = \left(\varepsilon^{(1)}, \ldots, \varepsilon^{(n)}\right)^T$. Σ est une matrice diagonale, dont ses éléments diagonaux sont égaux à $\sigma^{(1)}, \ldots, \sigma^{(n)}$.

Partant de cette modélisation, le but recherché est l'estimation du vecteur $\mathbf{f}$. Les performances d'un estimateur $\hat{\mathbf{f}}_\theta$ de $\mathbf{f}$ sont pour cela évaluées au moyen de son risque quadratique $\mathbb{E}\left[\|\hat{\mathbf{f}}_\theta - \mathbf{f}\|_2^2\right]$, $\|.\|_2$ étant la norme euclidienne sur $\mathbb{R}^n$. À une collection d'espaces des paramètres donnée $(\Theta_m,\ m \in \mathscr{M})$ est alors associée une collection d'estimateurs $\left(\hat{\mathbf{f}}_{\theta_m}, m \in \mathscr{M}\right)$. Le meilleur élément de cette famille au sens de la métrique euclidienne est l'oracle $\hat{\mathbf{f}}_{\theta_{m_*}}$ pour un certain $m_* \in \mathscr{M}$:

$$\|\hat{\mathbf{f}}_{\theta_{m_*}} - \mathbf{f}\|_2^2 = \inf_{m \in \mathscr{M}} \|\hat{\mathbf{f}}_{\theta_m} - \mathbf{f}\|_2^2.$$

Les travaux existants dans la communauté statistique visent alors à construire un processus de sélection de modèles permettant de sélectionner dans la collection des estimateurs un élément $\hat{f}_{\theta_{\hat{m}}}$ dont les performances sont les plus proches possibles de celles de l'oracle $\hat{f}_{\theta_{m_*}}$.

La sélection de modèles dans ce cadre hétéroscédastique a été étudiée par plusieurs auteurs en procédant à différentes considérations sur la matrice Σ que certains ne prennent pas forcément diagonale. Arlot (2010) et Arlot and Massart (2009) ont montré que les pénalités basées sur la validation croisée et des méthodes de rééchantillonnage construites indépendamment de la forme de la collection de modèles satisfont des inégalités oracle pour la sélection de régressogrammes. Un inconvénient de cette procédure est le coût computationnel important qui augmente avec la taille de l'échantillon. De plus, Arlot (2010) a montré que les pénalités proportionnelles à la dimension sont sous-optimales lorsque les observations sont entachées d'un bruit hétéroscédastique et que la collection de modèles est composée de régressogrammes, c'est-à-dire que le risque de l'estimateur sélectionné est plus large que celui de l'oracle. Lorsque Σ est une matrice diagonale d'éléments diagonaux inconnus, quelques résultats ont été proposés dans le cas gaussien ; on peut se référer à Gendre (2012) et Comte and Rozenholc (2002) par exemple. Par ailleurs, Gendre (2008) traite du cas où la matrice Σ n'est pas supposée inversible mais où $\mathbf{f}$ appartient à l'image de Σ. Il obtient dans ce cas qu'en considérant les observations $\mathbf{Y} = \mathbf{f} + \Sigma\varepsilon$ où $\mathbf{f} \in \mathrm{Im}(\Sigma) \subset \mathbb{R}^n$, Σ une matrice quelconque de $\mathscr{M}_n(\mathbb{R})$ et ε un vecteur gaussien standard et en disposant d'une collection $(S_m, m \in \mathscr{M})$ de sous-espaces vectoriels de $\mathrm{Im}(\Sigma)$ où à chaque S_m est associé un projecteur orthogonal associé π_m, prendre une pénalité $\mathrm{pen}(m,n)$ proportionnelle à $n^{-1}\mathrm{Tr}(\Sigma^T \pi_m \Sigma)$ conduit à des inégalités oracles sous réserve que quelques contraintes sur Σ soient vérifiées. Cette forme de pénalité correspond à celle bien connue en régression homoscédastique lorsque Σ est une matrice diagonale dont les éléments sur la diagonale valent tous σ ; dans ce cas, en notant D_m la dimension de S_m, on obtient $\mathrm{pen}(m,n) \sim n^{-1}\sigma^2 D_m$. Ces résultats de sélection de modèles dans un contexte hétéroscédastique sont essentiellement théoriques et demandent à être éprouvés sur des exemples applicatifs.

 ISSN: 2102-6238

4.2. Modèle de co-krigeage

Parallèlement aux méthodes de correction sur la sortie abordée à la sous-section 4.1.3 et couramment utilisées en multifidélité pour des problèmes industriels, le co-krigeage (*cokriging* en anglais) est apparu il y a une quinzaine d'années et est depuis très employé. Ceci peut s'expliquer par le fait que les modèles par processus gaussien sous-jacents au co-krigeage comportent peu de paramètres, fournissent un estimateur de l'erreur de prédiction et généralisent un certain nombre de modèles de substitution comme le modèle linéaire gaussien ou les réseaux de neurones (Williams, 1998).

Le modèle de co-krigeage présenté par Kennedy and O'Hagan (2000) se base sur des processus gaussiens Z_2 et Z_1 dont f_2 et f_1 sont des instances respectives. Ces processus aléatoires obéissent à la relation :

$$Z_2(\mathbf{x}) = \rho Z_1(\mathbf{x}) + \delta(\mathbf{x}),\ \forall \mathbf{x} \in \mathscr{X}$$

où Z_1 et δ sont des processus gaussiens indépendants définis par :

$$Z_1(\mathbf{x}) \sim \mathscr{PG}\left(\mathbf{h}_1^T(\mathbf{x})\beta_1; \sigma_1^2 r_1(\mathbf{x},\tilde{\mathbf{x}})\right),\ \forall \mathbf{x},\tilde{\mathbf{x}} \in \mathscr{X}$$

et

$$\delta(\mathbf{x}) \sim \mathscr{PG}\left(\mathbf{h}_\delta^T(\mathbf{x})\beta_\delta; \sigma_\delta^2 r_\delta(\mathbf{x},\tilde{\mathbf{x}})\right),\ \forall \mathbf{x},\tilde{\mathbf{x}} \in \mathscr{X}.$$

Les processus δ et Z_1 sont considérés comme indépendants afin de satisfaire la propriété de Markov $\mathrm{cov}(Z_2(\mathbf{x}), Z_1(\tilde{\mathbf{x}})|Z_1(\mathbf{x})) = 0,\ \forall \mathbf{x} \neq \tilde{\mathbf{x}}$ (Kennedy and O'Hagan, 2000). Cette propriété signifie que lorsque la valeur du modèle de basse-fidélité f_1 est connue en un point $\mathbf{x}$, la connaissance du modèle de basse-fidélité en un point $\tilde{\mathbf{x}} \neq \mathbf{x}$ n'apporte pas plus d'information pour inférer la valeur du modèle de haute-fidélité f_2 au point $\mathbf{x}$.

Parfois, le coefficient d'ajustement ρ s'écrit comme une fonction $\rho(\mathbf{x})$ du vecteur des paramètres d'entrée $\mathbf{x}$ afin de mieux tenir compte d'un écart variable entre le modèle de basse-fidélité et celui de haute-fidélité. Il joue par ailleurs le rôle du terme α dans la fonction de lien présentée dans la sous-section 4.1.3 tandis que le processus gaussien δ remplace le terme β. Cette fonction de lien caractérisée ici par le couple (ρ, δ) comporte moins de paramètre que les polynômes ou réseaux de neurones utilisés dans la sous-section 4.1.3. Elle est donc particulièrement avantageuse lorsque le nombre d'observations haute-fidélité n_2 est faible (quelques dizaines).

Par ailleurs, les plans d'expériences sont emboîtés afin de faciliter les calculs, c'est-à-dire $\mathbf{D}_2 \subset \mathbf{D}_1$ avec $\mathbf{D}_i = \{\mathbf{x}_i^{(1)}, ..., \mathbf{x}_i^{(n_i)}\},\ i \in \{1,2\}$. Concrètement, ceci signifie que le code basse fidélité f_1 doit être évalué au minimum aux mêmes points que le code haute-fidélité f_2.

Partant de cette modélisation et en conditionnant Z_2 par les bases d'apprentissage $\mathscr{A}_1$ et $\mathscr{A}_2$, ainsi que par les paramètres $\theta = \left(\sigma_1^2, \sigma_\delta^2, \beta_1, \beta_\delta, \zeta_1, \zeta_\delta, \rho\right)$, on obtient selon Kennedy and O'Hagan (2000) :

$$[Z_2(\mathbf{x})|\mathscr{A}_1, \mathscr{A}_2, \theta] \sim \mathscr{N}\left(\hat{f}_{2,\theta}(\mathbf{x}), s_2^2(\mathbf{x})\right)$$

où $\hat{f}_{2,\theta}$ est le prédicteur de f_2 et $s_2(\mathbf{x})$ l'estimateur de son erreur. En notant $\beta^T = \left(\beta_1^T\ \beta_\delta^T\right)$ le vecteur des paramètres de prédiction et $\mathbf{y} = \left(y_1^{(1)}, ..., y_1^{(n_1)}, y_2^{(1)}, ..., y_2^{(n_2)}\right)^T$ celui des observations,

 ISSN: 2102-6238

les quantités $\hat{f}_{2,\theta}$ et $s^2(\mathbf{x})$ s'écrivent :

$$\begin{cases} \hat{f}_{2,\theta}(\mathbf{x}) = \mathbf{h}(\mathbf{x})^T\beta + \mathbf{c}(\mathbf{x})^T\mathbf{V}^{-1}(\mathbf{y} - \mathbf{H}\beta) \\ s^2(\mathbf{x}) = \sigma_\delta^2 + \rho^2\sigma_1^2 - \mathbf{c}(\mathbf{x})^T\mathbf{V}^{-1}\mathbf{c}(\mathbf{x}) \end{cases}$$

où les écritures des termes $\mathbf{h}$, $\mathbf{H}$, $\mathbf{V}$ et $\mathbf{c}$ détaillées par Kennedy and O'Hagan (2000) généralisent celles des termes intervenant dans le modèle de krigeage. Les paramètres de tendance ρ, β_1 et β_δ sont le plus souvent estimés par maximum de vraisemblance et les paramètres σ_δ^2 et σ_1^2 par maximisation de vraisemblance restreinte.

D'autre part, on peut associer au modèle de substitution $\hat{f}_{2,\theta}$ évalué au point $\mathbf{x}$ un intervalle de confiance de niveau $1-\alpha$:

$$\mathrm{IC}(f_{2,\theta}(\mathbf{x})) = \left[\hat{f}_{2,\theta}(\mathbf{x}) - q_{1-\alpha/2}s_2(\mathbf{x}); \hat{f}_{2,\theta}(\mathbf{x}) + q_{1-\alpha/2}s_2(\mathbf{x})\right]$$

où $q_{1-\alpha/2}$ est le $1-\alpha/2$ quantile de la loi normale standard. La figure 2 illustre ainsi l'estimation d'un modèle de haute-fidélité (HF) à partir d'évaluations de ce modèle et d'un modèle de haute-fidélité (HF), en fournissant un intervalle de confiance de niveau 95%.

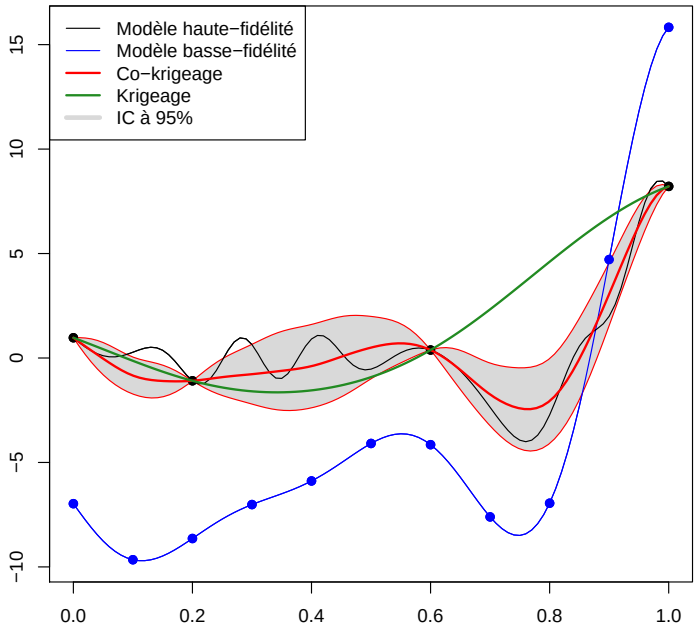

FIGURE 2. *Métamodèle de co-krigeage avec les fonctions proposées dans (Le Gratiet (2013)) et comparé à un métamodèle de krigeage dépendant des observations de haute-fidélité uniquement.*

Depuis le papier de Kennedy and O'Hagan (2000) établissant le modèle de co-krigeage, différentes contributions sont apparues pour la recherche des paramètres θ, l'application des méthodes de validation croisée, le développement en optimisation, la planification d'expériences adaptative, ...

Forrester et al. (2007) appliquent ainsi le co-krigeage à un problème d'optimisation globale en complétant le plan d'expériences avec le point $\mathbf{x}$ conduisant au maximum d'amélioration en espérance ; c'est la méthode de l'*expected improvement* :

$$\begin{aligned}\mathbb{E}[I(\mathbf{x})] &= \int_{-\infty}^{+\infty} \max\{\min(\mathbf{y}_2 - Z_2(\mathbf{x}),0)\}\Phi(Z_2(\mathbf{x}))dZ_2 \\ &= \begin{cases} (\min\{\mathbf{y}_2 - \hat{f}_{2,\theta}(\mathbf{x}))\Phi(W(\mathbf{x})) + s_2(\mathbf{x})\phi(W(\mathbf{x})) & \text{si } s > 0 \\ 0 & \text{sinon} \end{cases}\end{aligned}$$

où $W(\mathbf{x}) = \frac{\min\{\mathbf{y}_2\} - \hat{f}_{2,\theta}(\mathbf{x})}{s_2(\mathbf{x})}$ et où $\Phi(.)$ et $\phi(.)$ sont respectivement les fonctions de distribution et de densité de la loi normale standard.

Qian and Wu (2008) modélisent le coefficient d'ajustement ρ par un processus gaussien $\rho(\mathbf{x}) \sim \mathscr{PG}(\mu_\rho(\mathbf{x}), c_\rho(\mathbf{x},\tilde{\mathbf{x}}))$ et proposent une approche bayésienne afin de tenir compte de l'incertitude rencontrée dans l'estimation des différents paramètres. Ils s'intéressent ainsi à la loi de $[Z_2(\mathbf{x})|\mathscr{A}]$ plutôt qu'à celle de $[Z_2(\mathbf{x})|\mathscr{A},\theta]$. Ceci conduit à des variances *a posteriori* plus importantes et à des intervalles de confiance plus larges car considérant l'incertitude portée par les hyper-paramètres. Cette révision de la variance *a posteriori* est d'autant plus pertinente que la loi *a posteriori* des hyper-paramètres est faiblement concentrée autour de leurs modes. Toutefois, les lois *a posteriori* obtenues par Qian and Wu (2008) sont impropres et lors du calcul de $\hat{f}_{2,\theta}(\mathbf{x})$, ceci nécessite une approximation des intégrales par des méthodes de Monte Carlo, ce qui est coûteux en temps de calcul.

Le Gratiet (2013) propose une modélisation bayésienne alternative à celle de Qian and Wu (2008) aboutissant à une forme analytique de la moyenne et de la variance *a posteriori* de $Z_2(\mathbf{x})$ ne nécessitant pas ainsi d'évaluation numérique d'intégrales. Il y introduit par ailleurs une écriture récursive de la matrice de variance-covariance de l'ensemble des observations des deux niveaux de fidélité, réduisant le coût algorithmique de son inversion de $\mathscr{O}\left((n_1+n_2)^3\right)$ à $\mathscr{O}\left(n_1^3+n_2^3\right)$. Le Gratiet and Garnier (2014) proposent également une écriture alternative mais équivalente du modèle de Kennedy and O'Hagan (2000). Cette écriture a l'avantage de fournir une formulation récurrente de la moyenne et de la variance *a posteriori* de $Z_2(\mathbf{x})$ où l'on voit apparaître clairement que le co-krigeage correspond à un enchaînement de krigeages indépendants et que le prédicteur haute-fidélité $\hat{f}_{2,\theta}$ ne fait pas intervenir les variances σ_1^2 et σ_δ^2 :

$$\hat{f}_{2,\theta}(\mathbf{x}) = \rho(\mathbf{x})\hat{f}_{1,\theta}(\mathbf{x}) + \mathbf{h}_\delta^T(\mathbf{x})\hat{\beta}_\delta + r_\delta(\mathbf{x},\mathbf{D}_\delta)(\mathbf{R}_\delta)^{-1}\left(\mathbf{y}_2 - \rho\mathbf{y}_1 - \mathbf{H}_\delta\hat{\beta}_\delta\right),$$

$\hat{f}_{1,\theta}$ étant l'estimateur de f_1 associé au processus gaussien Z_1.

Les principales limites du co-krigeage sont les mêmes que celles rencontrées dans le krigeage. Notamment le nombre d'observations pour chaque niveau de fidélité est limité par la puissance de calcul à fournir lors de l'inversion de la matrice de variance-covariance de taille n dont le coût algorithmique est de l'ordre de n^3. Conséquemment il en est de même avec la dimension de

 ISSN: 2102-6238

l'espace des paramètres d'entrée puisque le nombre d'observations nécessaires pour maintenir une précision fixée augmente de manière exponentielle avec cette dimension. Par ailleurs, le choix des différentes structures de variance-covariance est arbitraire et il peut être nécessaire d'essayer plusieurs combinaisons afin de retenir celle qui généralise le mieux selon un critère à définir.

Néanmoins les méthodes à base de processus gaussiens ont l'avantage de pouvoir interpoler les observations de la base d'apprentissage, ce qui est particulièrement intéressant dans le cadre des codes de calcul qui sont des générateurs de données la plupart du temps déterministes. Par ailleurs afin d'accorder un peu plus de souplesse au métamodèle, il est possible d'introduire un effet de pépite dans chaque niveau du co-krigeage, ce qui produit une régression plutôt qu'une interpolation. De plus, ces méthodes possèdent peu de paramètres par rapport aux modélisations multifidélité à base de réseaux de neurones ce qui est un avantage dans notre situation où le nombre de simulations du code de calcul haute fidélité est faible. Enfin, le co-krigeage est un processus de correction sur la sortie comme ceux exprimés dans la section 4.1.3, mais à la différence de ces derniers, la fonction de lien requiert ici un nombre de paramètres à estimer moins élevé car elle repose sur un processus gaussien et non pas sur des polynômes ou des réseaux de neurones.

5. Conclusion et perspectives

L'objectif de cet article était de proposer une synthèse sur la substitution de modèle en expérimentation numérique ainsi qu'une ouverture sur leur déclinaison pour la modélisation multifidélité.

Après définition des notions d'expérience simulée et de substitution d'un code de calcul, les principaux modèles de substitution considérés en pratique ont été exposés : réseaux de neurones artificiels, modèles par processus gaussien, machines à vecteurs de support et développements en polynômes de chaos. Des références bibliographiques portant sur la comparaison de leurs performances ont complété cette présentation. Si les réseaux de neurones sont très répandus en ingénierie du fait de leur faculté à approcher des fonctions complexes comportant un grand nombre d'entrées, l'utilisation des modèles par processus gaussien est en plein essor depuis une décennie. En effet dans le cadre d'un nombre d'observations limité, leur construction est relativement rapide et leur formulation légère en nombre de paramètres ; ils s'accompagnent par ailleurs d'une évaluation locale de l'erreur.

Dans un autre temps, des éléments d'apprentissage statistique ont été exposés afin de choisir, pour un modèle de substitution donné, sa complexité architecturale et les valeurs de ses paramètres. Pour cela, différents estimateurs de l'erreur de généralisation, comme l'erreur de validation croisée ou de *bootstrap*, ont été abordés afin de sélectionner cette complexité au sein d'une collection de métamodèles minimisant l'erreur d'apprentissage sur différents sous-domaines des paramètres. D'autre part, des techniques de pénalisation de type AIC, BIC ou Cp de Mallows ont été présentées pour la sélection de modèles, ainsi que des critères pénalisés non-asymptotiques dont les travaux actuels sont essentiellement théoriques. Enfin, des approches de régularisation de type *ridge* et *lasso* ont été décrites, dont l'idée est de contraindre la dimension des paramètres durant la phase d'apprentissage afin de contrôler en même temps l'erreur de généralisation.

Dans un dernier temps a été abordée la métamodélisation multifidélité. Ce contexte apparaît dès lors que le coût d'évaluation du simulateur numérique est trop élevé pour garantir un nombre suffisamment élevé d'observations. Il repose sur l'existence de sources d'information complémentaires

 ISSN: 2102-6238

permettant de fournir de plus grands nombres d'observations mais comportant des niveaux de représentation de la réalité physique de moindre qualité. Les modèles à bruit hétéroscédastique ont été distingués des modèles à tendances hétérogènes. Le premier cas correspond plutôt à des séries d'observations du phénomène d'intérêt, chacune entachée d'une erreur numérique ou de mesure dotée d'une variance caractéristique ; des travaux en régression hétéroscédastique s'y rapportant ont été mentionnés. Le second cas quant à lui se rencontre lorsque les modèles complémentaires sont des versions dégradées du simulateur numérique de référence ; une distinction a été faite entre les méthodes utilisant un processus de correction sur la sortie et celles utilisant un processus de correction sur les entrées, de type *space-mapping*. Pour les processus de correction sur la sortie, le modèle de basse-fidélité est modélisé par un premier métamodèle, puis un second modèle de substitution vient approcher le lien entre les deux modèles. En particulier, un focus a été réalisé sur le métamodèle de co-krigeage couramment employé.

Par la suite, il pourrait être intéressant d'étudier les performances de prédiction des différents métamodèles en fonction des outils de sélection de modèles. D'autre part, une comparaison de ces métamodèles pourrait être menée selon leur utilisation : prédiction, analyse de sensibilité, problème inverse, ... Par ailleurs, plusieurs métamodèles comme les réseaux de neurones ou les modèles par processus gaussien comportent des méthodes de paramétrisation par inférence bayésienne ; une synthèse de ces approches et leur généralisation à d'autres modèles de substitution seraient une perspective intéressante. En modélisation multifidélité d'autre part, les méthodes de régression hétéroscédastique mériteraient d'être éprouvées sur des données réelles ; des études sur leur dépendance aux plans d'expériences et l'estimation des variances des différents modèles d'observation permettraient également d'élargir le champ d'application de la modélisation hétéroscédastique. Enfin, les simulateurs peuvent en pratique avoir en entrée ou en sortie des entiers, des variables catégorielles ou encore des variables fonctionnelles discrétisées. Les métamodèles présentés ne permettent pas dans leur formulation actuelle de les prendre en compte ; aussi, un état de l'art sur l'extension de ces modèles à ce type de variables serait pertinent.

Références

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, pages 267–281. Akademinai Kiado.

Arlot, S. (2009). Model selection by resampling penalization. *Electronic Journal of Statistics*, 3 :557–624.

Arlot, S. (2010). Choosing a penalty for model selection in heteroscedastic regression. *arXiv :0812.3141*.

Arlot, S. and Massart, P. (2009). Data-driven calibration of penalties for least-squares regression. *Journal of Machine Learning Research*, 10 :245–279.

Attali, J.-G. and Pagès, G. (1997). Approximations of functions by a multilayer perceptron : a new approach. *Neural Networks*, 10(6) :1069–1081.

Bandler, J. W., Cheng, Q. S., Dakroury, S. A., Mohamed, A. S., Bakr, M. H., Madsen, K., and Sondergaard, J. (2004). Space mapping : the state of the art. *Microwave Theory and Techniques, IEEE Transactions on*, 52(1) :337–361.

Barron, A., Birgé, L., and Massart, P. (1999). Risk bounds for model selection via penalization. *Probability theory and related fields*, 113(3) :301–413.

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *Information Theory, IEEE Transactions on*, 39(3) :930–945.

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics : overview and implementation. *Statistics and Computing*, 22(1) :455–470.

Bettebghor, D., Bartoli, N., Grihon, S., Morlier, J., and Samuelides, M. (2011). Surrogate modeling approximation

 ISSN: 2102-6238

using a mixture of experts based on em joint estimation. *Structural and Multidisciplinary Optimization*, 43(2) :243 – 259.

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press.

Bonnans, J. F., Gilbert, J. C., Lemaréchal, C., and Sagastizábal, C. A. (2006). *Numerical Optimization : Theoretical and Practical Aspects (Universitext)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Broomhead, D. S. and Lowe, D. (1988). Radial Basis Functions, Multi-Variable Functional Interpolation and Adaptive Networks. *Complex Systems*, 2 :321–355.

Broyden, C. G. (1970). The Convergence of a Class of Double-rank Minimization Algorithms 1. General Considerations. *IMA Journal of Applied Mathematics*, 6(1) :76–90.

Chen, V. C., Tsui, K.-L., Barton, R. R., and Meckesheimer, M. (2006). A review on design, modeling and applications of computer experiments. *IIE transactions*, 38(4) :273–291.

Comte, F. and Rozenholc, Y. (2002). Adaptive estimation of mean and volatility functions in (auto-)regressive models. *Stochastic Processes and their Applications*, 97(1) :111–145.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1988). A Bayesian approach to the design and analysis of computer experiments. Technical Report ORNL–6498, Oak Ridge Laboratory.

Cybenko, G. (1989). Approximations by superpositions of sigmoidal functions. *Mathematics of Control, Signals, and Systems*, 2(4) :303–314.

Damianou, A. C. and Lawrence, N. D. (2013). Deep gaussian processes. In *Proceedings of the Sixteenth International Conference on Artificial Intelligence and Statistics, AISTATS 2013, Scottsdale, AZ, USA, April 29 - May 1, 2013*, pages 207–215.

De Lozzo, M. (2013). *Modèles de substitution spatio-temporels et multifidélité : Application à l'ingénierie thermique*. PhD thesis, Université de Toulouse - Institut National des Sciences Appliquées de Toulouse.

Dean, A., Morris, M., Stufken, J., and Bingham, D. (2015). *Handbook of Design and Analysis of Experiments*. Chapman and Hall/CRC, 1st edition.

Den Hertog, D., Kleijnen, J., and Siem, A. (2005). The correct kriging variance estimated by bootstrapping. *Journal of the Operational Research Society*, 57(4) :400–409.

Dreyfus, G., Martinez, J.-M., Samuelides, M., Gordon, M. B., and Thiria, S. (2008). *Apprentissage statistique : Réseaux de neurones - Cartes topologiques - Machines à vecteurs supports*. Eyrolles.

Efron, B. (1981). Nonparametric estimates of standard error : The jackknife, the bootstrap and other methods. *Biometrika*, 68(3) :589–599.

Efron, B. (1983). Estimating the error rate of a prediction rule : Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382) :316–331.

Efron, B. and Tibshirani, R. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Faivre, R., Iooss, B., Mahévas, S., Makowski, D., and Monod, H., editors (2013). *Analyse de sensibilité et exploration de modèles : application aux sciences de la nature et de l'environnement*. Collection Savoir-faire. Éditions Quae.

Fang, K.-T., Li, R., and Sudjianto, A. (2006). *Design and modeling for computer experiments*. Chapman & Hall/CRC.

Forrester, A., Sobester, A., and Keane, A. (2008). *Engineering Design via Surrogate Modelling : A Practical Guide*. Wiley.

Forrester, A. I. and Keane, A. J. (2009). Recent advances in surrogate-based optimization. *Progress in Aerospace Sciences*, 45(1) :50 – 79.

Forrester, A. I. J., Sobester, A., and Keane, A. J. (2007). Multi-fidelity optimization via surrogate modelling. *Proceedings of Royal Society A*, 463 :3251–3269.

Friedman, J. H. (1991). Multivariate Adaptive Regression Splines. *The Annals of Statistics*, 19(1) :1–67.

Gendre, X. (2008). Simultaneous estimation of the mean and the variance in heteroscedastic gaussian regression. *Electron. J. Statist.*, 2 :1345–1372.

Gendre, X. (2012). Model selection and estimation of a component in additive regression. *ESAIM : Probability and Statistics*.

Ghanem, R. G. and Spanos, P. D. (1991). *Stochastic finite elements : a spectral approach*. Springer-Verlag, New York.

Goldstein, M. (2007). *Bayes Linear Statistics : Theory and Methods*. Wiley, Chichester.

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

Hoerl, A. E. and Kennard, R. W. (1970). Ridge Regression : Biased Estimation for Nonorthogonal Problems. *Technometrics*, 12(1) :55–67.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4(2) :251–257.

 ISSN: 2102-6238

Igel, C. and Hüsken, M. (2000). Improving the rprop learning algorithm. In Press, I. A., editor, *Proceedings of the Second International Symposium on Neural Computation, NC2000*, page pp. 115121.

Kennedy, M. C. and O'Hagan, A. (2000). Predicting the output from a complex computer code when fast approximations are available. *Biometrika*, 87(1) :1–13.

Kim, H. S., Ko, M., and Ni, J. (2007). A hybrid multi-fidelity approach to the optimal design of warm forming processes using a knowledge-based artificial neural network. *International Journal of Machine Tools and Manufacture*, 47(2) :211–222.

Kleijnen, J. (2015). *Design and Analysis of Simulation Experiments*. Springer.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, volume 2, pages 1137–1143.

Krige, D. G. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *Journal of the Chemical, Metallurgical and Mining Society*, 52 :119–139.

Lasance, C. J. M. (2008). Ten years of boundary-condition- independent compact thermal modeling of electronic parts : A review. *Heat Transfer Engineering*, 29(2) :149–168.

Le Gratiet, L. (2013). Bayesian analysis of hierarchical multi-fidelity codes. *SIAM/ASA Journal of Uncertainty Quantification*, pages 244–269.

Le Gratiet, L. and Garnier, J. (2014). Recursive co-kriging model for design of computer experiments with multiple levels of fidelity. *International Journal for Uncertainty Quantification*, 5(4) :365–386.

Lebarbier, E. and Mary-Huard, T. (2006). Une introduction au critère bic : fondements théoriques et interprétation. *Journal de la SFdS*, 147(1) :39–57.

Levenberg, K. (1944). A method for the solution of certain problems in least squares. *Quart. Applied Math.*, 2 :164–168.

Levin, D. (1998). The approximation power of moving least-squares. *Math. Comput.*, 67(224) :1517–1531.

Li, J. and Wang, H. (2011). Data fusion of multi-fidelity model and its application in low speed reflexed airfoil shape optimization. In *Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on*, pages 2910 –2913.

Mallows, C. L. (1973). Some Comments on CP. *Technometrics*, 15(4) :661–675.

Marquardt, D. W. (1963). An Algorithm for Least-Squares Estimation of Nonlinear Parameters. *SIAM Journal on Applied Mathematics*, 11(2) :431–441.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume Lecture Notes in Mathematics. Springer-Verlag.

Massart, P. (2008). Sélection de modèle : de la théorie à la pratique. *Journal de la SFDS 149*, 149(4) :5–28.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8) :1246–1266.

Matheron, G. (1970). La théorie des variables régionalisées et ses applications. In *Les Cahiers du Centre de Morphologie Mathématique*, volume 5. Fontainebleau, École Nationale Supérieure des Mines de Paris.

Mitchell, H. B. (2007). *Multi-Sensor Data Fusion : An Introduction*. Springer Publishing Company, Incorporated, 1st edition.

Myers, R., Montgomery, D., and Anderson-Cook, C. (2009). *Response Surface Methodology : Process and Product Optimization Using Designed Experiments*. Wiley Series in Probability and Statistics. Wiley, 3rd edition.

Nguyen, D. and Widrow, B. (1990). Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights. In *Proceedings of the International Joint Conference on Neural Networks*, volume 3, pages 21–26.

Oussar, Y. and Dreyfus, G. (2001). How to be a gray box : dynamic semi-physical modeling. *Neural Networks*, 14(9) :1161–1172.

Oussar, Y., Rivals, I., Personnaz, L., and Dreyfus, G. (1998). Training wavelet networks for nonlinear dynamic input-output modeling. *Neurocomputing*, 20(1-3) :173–188.

Picheny, V. and Ginsbourger, D. (2013). A nonstationary space-time gaussian process model for partially converged simulations. *SIAM/ASA Journal on Uncertainty Quantification*, 1(1) :57–78.

Picheny, V., Ginsbourger, D., Roustant, O., Haftka, R. T., and Kim, N. H. (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design*, 132(7).

Qian, P. Z. G. and Wu, C. F. J. (2008). Bayesian hierarchical modeling for integrating low-accuracy and high-accuracy experiments. *Technometrics*, 50(2) :192–204.

Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

Riedmiller, M. and Braun, H. (1993). A direct adaptive method for faster backpropagation learning : The rprop

 ISSN: 2102-6238

algorithm. In *IEEE International Conference on Neural Networks*, pages 586–591.
Rumelhart, D. E., Hinton, G. E., and Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, 323 :533–536.
Sacks, J., Welch, W. J., Mitchell, T. J., and Wynn, H. P. (1989). Design and analysis of computer experiments. *Statist. Sci.*, 4(4) :409–423.
Santner, T. J., B., W., and W., N. (2003). *The Design and Analysis of Computer Experiments*. Springer-Verlag.
Scholkopf, B. and Smola, A. J. (2001). *Learning with Kernels : Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA.
Schwarz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics*, 6(2) :461–464.
Simon, D. (2013). *Evolutionary Optimization Algorithms*. Wiley, 1 edition.
Simpson, T., Poplinski, J., Koch, P. N., and Allen, J. (2001). Metamodels for computer-based engineering design : Survey and recommendations. *Engineering with Computers*, 17(2) :129–150.
Stein, M. L. (1999). *Interpolation of Spatial Data : Some Theory for Kriging (Springer Series in Statistics)*. Springer.
Sun, G., Li, G., Zhou, S., Xu, W., Yang, X., and Li, Q. (2011). Multi-fidelity optimization for sheet metal forming process. *Structural and Multidisciplinary Optimization*, 44(1) :111–124.
Tarantola, A. (2004). *Inverse Problem Theory and Methods for Model Parameter Estimation*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.
Tibshirani, R. (1994). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58 :267–288.
Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solutions of Ill-Posed Problems*. Wiley.
Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.
Vapnik, V. (1999). An overview of statistical learning theory. *Neural Networks, IEEE Transactions on*, 10(5) :988–999.
Vapnik, V., Golowich, S. E., and Smola, A. J. (1996). Support vector method for function approximation, regression estimation and signal processing. In *Advances in Neural Information Processing Systems 9, NIPS, Denver, CO, USA, December 2-5, 1996*, pages 281–287.
Williams, C. K. I. (1998). Computation with infinite neural networks. *Neural Comput.*, 10(5) :1203–1216.
Xiu, D. and Karniadakis, G. E. (2002). The wiener–askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput.*, 24(2) :619–644.
Zou, H. and 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