Agrégation d'informations et alternative au krigeage en environnement aléatoire

Pierre Ribereau, Didier Rullière

Agrégation d'informations et alternative au krigeage en environnement aléatoire*

Didier Rullière† Pierre Ribereau‡

10 mars 2011

Résumé

Nous définissons une notion d'information d'une variable aléatoire, qui permet en particulier de modéliser les erreurs d'estimation. Nous bâtissons un opérateur d'agrégation d'informations, qui reste cohérent avec la réduction des erreurs d'estimation lors de la réitération d'observations d'une même variable aléatoire. Nous évoquons alors les problèmes liés à l'usage du krigeage en environnement bruité, et proposons l'usage de l'opérateur d'agrégation comme alternative à la construction de surfaces de réponse et à l'estimation de la variance de krigeage. Un algorithme d'optimisation utilisant ces surfaces de réponse alternatives est proposé. Quelques applications numériques illustrent enfin l'intérêt de ces surfaces de réponse et de l'algorithme proposé.

mots-clés : Agrégation d'informations, krigeage, optimisation globale, fonction objectif bruitée, potentiel, fonction d'information.

1 Introduction

Considérons le champ aléatoire suivant, défini sur un domaine $\Theta \subset \mathbb{R}^d$, $d \in \mathbb{N}^*$:

$$F(\theta) = f(\theta) + \epsilon(\theta)\,, \quad \theta \in \Theta\,,$$

où $f(\theta)$, $\theta \in \Theta$ est une fonction déterministe réelle, continue mais non nécessairement dérivable, et où $\{\epsilon(\theta)\}_{\theta\in\Theta}$ est un bruit que l'on cherche à éliminer, d'espérance nulle, éventuellement non homogène, mais sans structure de dépendance : les éléments distincts de $\{\epsilon(\theta)\}_{\theta\in\Theta}$ sont mutuellement indépendants.

Pour chaque point $\theta \in \Theta$, on suppose qu'un observateur ne peut pas disposer directement de la valeur de $f(\theta)$, mais seulement de tirages de la variable aléatoire $F(\theta)$.

Krigeage en environnement aléatoire Dans la théorie du krigeage, on se place sous l'hypothèse supplémentaire que $\{f(\theta)\}_\theta$ est une trajectoire déterministe générée par un champ gaussien $\{X(\theta)\}_{\theta\in\Theta}$. Considérons un ensemble de n_T points déjà observés $T = \{\theta_1, \ldots, \theta_{n_T}\}$, $n_T \geqslant 1$. En chaque point $\theta \in T$, on suppose que l'on dispose d'un échantillon de n_θ tirages de $F(\theta)$: $\{F_1(\theta), \ldots, F_{n_\theta}(\theta)\}$. Considérons tout d'abord les surfaces de réponse qui tentent de prédire $F(\theta)$ sur des sites θ inexplorés. Un technique possible pour construire de telles surfaces de réponse est ce que nous appellerons par la suite le *krigeage avec effet pépite*. Ce type de krigeage considère bien le bruit ϵ, et tente d'exprimer un prédicteur $\tilde{F}$ de F comme une combinaison linéaire des valeurs observées aux points de T. La présence d'un bruit

*Ce travail a bénéficié d'un soutien partiel du projet ANR ANR-08-BLAN-0314-01 et du projet MIRACCLE-GICC.

†Université de Lyon, F-69622, Lyon, France; Université Lyon 1, Laboratoire SAF, EA 2429, Institut de Science Financière et d'Assurances, 50 Avenue Tony Garnier, F-69007 Lyon, France. didier.rulliere@univ-lyon1.fr

‡Université de Montpellier, équipe Proba-Stat, CC 051, pl. E.Bataillon, 34000 Montpellier. pierre.ribereau@univ-montp2.fr

supplémentaire induit un accroissement ponctuel de la variance sur les seuls points d'observation, et cause ainsi une discontinuité dans la fonction de covariance, usuellement qualifiée d'*effet pépite*.

$$\mathrm{Cov}\left(F(\theta_1), F(\theta_2)\right) \quad = \quad \mathrm{Cov}\left(X(\theta_1), X(\theta_2)\right) + \ \mathbb{1}_{\theta_1=\theta_2}\mathrm{Cov}\left(\epsilon(\theta_1), \epsilon(\theta_2)\right), \quad \theta_1, \theta_2 \in \Theta\,.$$

Des exemples dans lesquels les prédictions par krigeage basées sur ce modèle sont meilleures que les prédictions issues de régression sont proposés par [34]. La recherche d'un prédicteur de F est un domaine de recherche populaire [voir par exemple 21]. Dans ce type de recherche, il est tenu compte du bruit entachant la fonction f, mais on ne cherche pas explicitement à éliminer ce bruit. A notre connaissance, la littérature de krigeage est plus réduite lorsqu'il s'agit de prédire f et non F sur des sites inexplorés en réduisant explicitement le bruit pesant sur f. Il s'agit dans ce cas d'une prédiction en présence d'observations incomplètes, parce que l'on estime alors uniquement une distribution possible de f à chaque point d'observation, et non une valeur déterministe fixée.

Raisons d'une alternative Lorsque f est estimée à l'aide de simulations stochastiques, le simulateur renvoie bien des valeurs $F(\theta)$ centrées sur $f(\theta)$ et la technique de krigeage à effet pépite peut être utilisée ($\theta \in \Theta$). Néanmoins, dans beaucoup de situations, on cherche en effet à prédire f et non F. Lorsque deux sites d'exploration sont très proches, la variance de la prédiction de f doit être réduite sur chacun des sites. Par ailleurs, le poids affecté à une observation très bruitée doit intuitivement être très réduit. Dans la mesure où le krigeage propose une interpolation de F et non de f, on peut se demander si le krigeage conduit bien à ce type de comportement, d'autant que l'amplitude du bruit peut être radicalement différente d'un site à un autre. Ces questions et la question du choix des effets pépite sont discutées dans [28]. En outre, à supposer que le krigeage soit retenu, des problèmes peuvent subsister :

- Le krigeage implique l'inversion de matrices qui peuvent être très grandes lorsque l'on dispose d'un nombre élevé d'observations. Cette inversion peut en conséquence être très lente et coûteuse. Même si l'on considère les seules observations suffisamment proches d'un point donné, les procédures d'optimisation peuvent conduire à une grande densité d'observations aux alentours d'un point jugé intéressant (cela constitue même un objectif des procédures d'optimisation).
- La fonction de covariance est inconnue en pratique. Bien qu'elle puisse être estimée, elle peut être entachée d'erreurs d'estimation importantes. En outre, lors de procédures d'optimisation, lorsque la densité des points explorés augmente autour d'une solution potentielle, la corrélation devient très forte et la nécessité de prendre en compte les corrélations croisées semble moins évidente.
- Enfin, le non-respect des hypothèses prises de stationnarité peuvent parfois rendre le krigeage moins performant que des interpolations plus simples.

Pour ces raisons, nous envisagerons dans les sections suivantes une alternative au krigeage.

Contraintes pour les alternatives La recherche d'un optimum global de $f(\theta)$ va conduire l'observateur à opérer des tirages de $F(\theta)$ en différents points de Θ. Afin de déterminer où explorer la fonction, il apparaît en premier lieu nécessaire de disposer d'intervalles de confiance pour la valeur de f en des points θ non encore explorés. Dans un souci global d'économie du nombre de tirages, il va s'agir d'exploiter au mieux les informations obtenues grâce aux précédents tirages de F.

Cette question introduit la problématique plus générale de l'agrégation des informations. Supposons que l'on dispose de n_1 observations de F en un point θ_1, et n_2 observations en un point θ_2. Il semble logique d'exiger, lorsque θ_1 et θ_2 sont presque confondus, que cet ensemble d'observations soit presque équivalent à $n_1 + n_2$ observations en θ_1. La formalisation de cette équivalence nous conduira à définir précisément une *information*, et la façon dont l'on agrège les informations.

Structure du document Dans une section 2, nous étudierons l'information définie pour une unique variable aléatoire réelle. Nous utiliserons ensuite ces définitions et propriétés sur des ensembles de variables aléatoires réelles (par exemple issues de champs gaussiens). Nous définirons alors un opérateur

de dégradation d'information, qui permettra notamment de prendre en compte les incertitudes liées aux distances spatiales entre deux observations. Dans une section 3, nous proposerons l'utilisation d'informations agrégées pour la construction de surfaces de réponse alternatives. Dans une section 4, nous bâtirons un algorithme d'optimisation très simple exploitant les propriétés de ces surfaces de réponses alternatives. Enfin, dans une section 5, des applications numériques illustreront la construction de surfaces de réponse et le comportement de l'algorithme d'optimisation proposé.

2 Agrégation d'informations

2.1 Notion d'information

Nous allons évoquer ici des considérations élémentaires sur les erreurs d'estimation, afin de formaliser la notion d'agrégation d'information. L'information la plus complète caractérisant une variable aléatoire est sa loi. Néanmoins, en présence d'un nombre limité d'observations, nous allons représenter la connaissance d'une variable aléatoire au moyen de deux réels, qui quantifierons par la suite une moyenne et une variance. Cette représentation s'avérera notamment intéressante lorsque nous considérerons des champs gaussiens. La notion d'information développée ici n'est pas a priori liée à la théorie de l'information de Shannon ou à l'entropie. Nous avons néanmoins utilisé le même vocable d'*information*.

Définition 1 (Information) *Nous appellerons information une quantité : $I = (m, \sigma^2) \in \mathbb{R} \times \bar{\mathbb{R}}^+$. Nous dirons que l'information est certaine si $\sigma^2 = 0$. Nous dirons que l'information est nulle si $\sigma^2 = +\infty$.*

2.2 Informations et erreurs d'estimation

Nous allons dans un premier temps essayer de représenter l'incertitude liée à l'estimation d'une variable aléatoire réelle X.

Définition 2 (Information d'estimation) *Soit un ensemble de tirages $G = \{X_1, \ldots, X_n\}$ d'une variable aléatoire générique X, $n \in \mathbb{N}^*$ (on suppose que les tirages correspondent à une réalisation d'une suite iid de variables aléatoires de même loi que X). On suppose également que les deux premiers moments de X sont finis et que la variance σ_X^2 de X est connue.*
Nous dirons que $I = \mathscr{I}(G, \sigma_X)$ est l'information d'estimation, en connaissance de G et de σ_X, lorsque :

$$\mathscr{I}(G, \sigma_X) = (m, \sigma^2) \quad avec \quad \left\{ \begin{array}{rcl} m & = & \frac{1}{n} \sum_{X_i \in G} X_i, \\ \sigma^2 & = & \frac{\sigma_X^2}{n} \end{array} \right.$$

Cette définition permet de représenter de façon logique l'incertitude liée à l'estimation de la variable aléatoire X :

- la première composante m de l'information représente la moyenne empirique de X en connaissance de G.
- La seconde composante σ^2 de l'information représente la variance de cette moyenne empirique en connaissance de G et de σ_X. Cette seconde composante quantifie donc l'erreur d'estimation commise lors de l'estimation de la moyenne de X.

Soient I_1, I_2 deux informations : $I_1 = (m_1, \sigma_1^2)$ et $I_2 = (m_2, \sigma_2^2)$. Nous allons essayer de spécifier les contraintes souhaitables pour qu'une information $I = (m, \sigma^2)$ représente l'agrégation des deux informations I_1 et I_2. La définition précédente va nous permettre de formaliser l'agrégation d'information : l'agrégation des informations d'estimation issues de deux ensembles de tirages G_1 et G_2 devra simplement correspondre à l'information issue de l'ensemble des tirages $G_1 \cup G_2$.

Définition 3 (Agrégation cohérente) *Soient deux ensembles de tirages G_1 et G_2 d'une variable aléatoire générique X dont les deux premiers moments sont finis, et dont la variance $\sigma_X^2 > 0$ est connue. Soient I_1 et I_2 les informations d'estimation en connaissance respectivement de G_1 et de G_2 :*

$$I_1 = \mathscr{I}(G_1, \sigma_X) \text{ et } I_2 = \mathscr{I}(G_2, \sigma_X),$$

Nous dirons qu'une information I est l'agrégation cohérente des informations I_1 et I_2 si I est l'information d'estimation en connaissance de $G_1 \cup G_2$:

$$I = \mathscr{I}(G_1 \cup G_2, \sigma_X)$$

Proposition 1 (Condition d'agrégation cohérente) *Soient deux ensembles de tirages G_1 et G_2 d'une variable aléatoire générique X dont les deux premiers moments sont finis, et dont on connaît la variance $\sigma_X^2 > 0$. Soient I_1 et I_2 les informations d'estimation en connaissance respectivement de G_1 et de G_2 :*

$$I_1 = \mathscr{I}(G_1, \sigma_X) \text{ et } I_2 = \mathscr{I}(G_2, \sigma_X),$$

Soit I l'information résultant de l'agrégation cohérente de I_1 et I_2. Si n_1 et n_2 désignent les cardinaux respectifs de G_1 et G_2, alors I est telle que :

$$I = (m, \sigma^2) \quad avec \quad \left\{ \begin{array}{lcl} m & = & \frac{n_1}{n_1+n_2} m_1 + \frac{n_2}{n_1+n_2} m_2 \\ \sigma^2 & = & \frac{\sigma_X^2}{n_1+n_2} \end{array} \right. \tag{1}$$

Proof: L'information I étant l'agrégation cohérente de I_1 et I_2, on a $m = \frac{1}{n_1+n_2} \sum_{X_i \in G_1 \cup G_2} X_i$ et $\sigma^2 = \frac{\sigma_X^2}{n_1+n_2}$. Par définition de I_1 et I_2, on en déduit pour la moyenne $m = \frac{1}{n_1+n_2}(n_1 m_1 + n_2 m_2)$. □

On aimerait par ailleurs ne faire dépendre (m, σ^2) que de (m_1, σ_1^2) et (m_2, σ_2^2). Il faut donc chercher à exprimer (m, σ^2) indépendamment de n_1 et n_2.

2.3 Définition et propriétés de l'agrégation

Définissons tout d'abord l'opérateur d'agrégation que nous allons utiliser.

Définition 4 (Opérateur d'agrégation) *Soient deux informations (m_1, σ_1^2) et (m_2, σ_2^2), avec $\sigma_1 \sigma_2 > 0$. On définit l'opérateur d'agrégation $\oplus$ tel que :*

$$\begin{array}{l} (m_1, \sigma_1^2) \oplus (m_2, \sigma_2^2) = (m, \sigma^2)\,, \\ avec \left\{ \begin{array}{lcl} m & = & \frac{\sigma_2^2}{\sigma_1^2+\sigma_2^2} m_1 + \frac{\sigma_1^2}{\sigma_1^2+\sigma_2^2} m_2\,, \\ \sigma^2 & = & \frac{\sigma_1^2 \sigma_2^2}{\sigma_1^2+\sigma_2^2}\,, \end{array} \right. \end{array} \tag{2}$$

Proposition 2 (Cohérence de l'opérateur d'agrégation) *Lorsque I_1 et I_2 sont les informations d'estimation de deux ensembles de tirages G_1 et G_2, alors l'information $I = I_1 \oplus I_2$ est l'agrégation cohérente de I_1 et I_2.*

Proof: Soient deux ensembles de tirages G_1 et G_2 d'une variable aléatoire générique X dont les deux premiers moments sont finis, et dont on connaît la variance $\sigma_X^2 > 0$. Soient n_1 et n_2 les cardinaux respectifs de G_1 et G_2, $n_1, n_2 \in \mathbb{N}^*$. Supposons que $I_1 = (m_1, \sigma_1^2)$ et $I_2 = (m_2, \sigma_2^2)$ sont les informations d'estimation en connaissance respectivement de G_1 et de G_2 :

$$I_1 = \mathscr{I}(G_1, \sigma_X) \text{ et } I_2 = \mathscr{I}(G_2, \sigma_X),$$

Notons $I = (m, \sigma) = I_1 \oplus I_2$. Il s'agit de montrer que

$$I_1 \oplus I_2 = \mathscr{I}(G_1 \cup G_2, \sigma_X)\,.$$

Pour la variance, $\sigma^2 = \sigma_X^2/(n_1+n_2)$. Pour les ensembles d'observation G_1 et G_2, l'erreur d'estimation de la moyenne de X est respectivement quantifiée par $\sigma_1^2 = \frac{\sigma_X^2}{n_1}$ et $\sigma_2^2 = \frac{\sigma_X^2}{n_2}$. De par la définition de $\oplus$, la variance σ^2 dans l'information I est $\sigma^2 = \sigma_1^2\sigma_2^2/(\sigma_1^2+\sigma_2^2)$, de sorte que $1/\sigma^2 = \frac{n_1+n_2}{\sigma_X^2}$ et finalement l'équation 1 est bien respectée pour la variance :

$$\sigma^2 \;=\; \frac{\sigma_X^2}{n_1+n_2}\,.$$

Pour les moyennes, les moyennes empiriques dans I_1 et I_2 sont respectivement : $m_1 = \frac{1}{n_1}\sum_{X_i\in G_1} X_i$ et $m_2 = \frac{1}{n_2}\sum_{X_i\in G_2} X_i$. De par la définition de $\oplus$, on établit facilement $m = \frac{\sigma^2}{\sigma_1^2}m_1 + \frac{\sigma^2}{\sigma_2^2}m_2$. Or $\sigma_1^2 = \sigma_X^2/n_1$, $\sigma_2^2 = \sigma_X^2/n_2$ et $\sigma^2 = \sigma_X^2/(n_1+n_2)$. On vérifie alors que $m = \frac{n_1}{n_1+n_2}m_1 + \frac{n_2}{n_1+n_2}m_2$, et finalement l'équation 1 est bien respectée pour la moyenne.

$$m = \frac{1}{n_1+n_2}\sum_{X_i\in G_1\cup G_2} X_i$$

Finalement, les conditions de la proposition 1 sont respectées, et l'on a donc $I_1 \oplus I_2 = \mathscr{I}(G_1\cup G_2, \sigma_X)$. □

On montre également que l'agrégation d'informations ne dépend pas de l'ordre d'agrégation.

Proposition 3 (commutativité) *L'opérateur $\oplus$ est commutatif : soient deux informations non certaines I_1 et I_2 :*

$$I_1 \oplus I_2 = I_2 \oplus I_1$$

Proof: Se déduit directement par symétrie des expressions définissant l'opérateur $\oplus$. □

Ensuite, les conséquence d'ajout d'informations sont logiques : l'agrégation d'information conduit notamment à une diminution d'incertitude.

Proposition 4 (Domaine de l'information agrégée) *Soient deux informations non certaines $I_1 = (m_1,\sigma_1)$, $I_2 = (m_2,\sigma_2)$. Soit $I = (m,\sigma) = I_1 \oplus I_2$.*

$$\begin{aligned} m &\in [m_1, m_2]\,, \\ \sigma &\leqslant \min\{\sigma_1,\sigma_2\}\,. \end{aligned}$$

Proof: Se déduit facilement de la définition de l'opérateur $\oplus$. □

Le fait que l'agrégation conduise à une moyenne m appartenant à $[m_1,m_2]$ peut revêtir un intérêt lorsque les quantités à agréger sont bornées (par exemple lorsque l'on estime des probabilités).

Enfin, l'ajout d'une information infiniment incertaine correspond à n'ajouter aucune information, et une information certaine annihile les effets des informations incertaines :

Proposition 5 (valeurs particulières d'agrégation)

$$\begin{aligned} (m_1,\sigma_1^2)\oplus(m_2,+\infty) &= (m_1,\sigma_1^2) \\ (m_1,\sigma_1^2)\oplus(m_2,0) &= (m_2,0)\,,\ \textit{si } \sigma_1{>}0\,, \\ (m_1,0)\oplus(m_2,0) &= \left(\frac{m_1+m_2}{2},0\right)\ \textit{par convention}\,. \end{aligned}$$

Proof: Se déduit facilement de la définition de l'opérateur $\oplus$. En principe, aucun modèle ne devrait conduire à agréger deux informations certaines contradictoires $(m_1,0)$ et $(m_2,0)$ avec $m_1 \neq m_2$. La dernière convention est choisie pour parer à une telle éventualité : elle est cohérente avec le cas $m_1 = m_2$, et revient à imaginer dans le cas contraire qu'une source d'incertitude à été négligée, frappant de la même façon m_1 et m_2, par exemple une erreur de précision arithmétique causant indûment

$m_1 \neq m_2$. Cette convention permet d'éviter la distinction de cas particuliers lors de l'implémentation pratique de $\oplus$, car elle correspond à l'agrégation d'informations auxquelles on a rajouté une source négligeable d'incertitude ϵ :

$$(m, \sigma^2) = \lim_{\epsilon \to 0} (m_1, \sigma_1^2 + \epsilon^2) \oplus (m_2, \sigma_2^2 + \epsilon^2)$$

□

Proposition 6 (interpolation) *Considérons deux variables aléatoires $X_1 = m_1 + E_1$ et $X_2 = m_2 + E_2$, où m_1 et m_2 sont déterministes, E_1 et E_2 sont deux bruits centrés de variances respectives σ_1^2 et σ_2^2. Notons $I_1 = (m_1, \sigma_1^2)$ et $I_2 = (m_2, \sigma_2^2)$ les informations correspondantes. Notons $I = (m, \sigma^2)$, alors :*

$$I = I_1 \oplus I_2 \Leftrightarrow \left\{ \begin{array}{l} m = \mathrm{E}\left[\widehat{X}\right] \\ \sigma^2 = \mathrm{V}\left[\widehat{X}\right] \end{array} \right. \text{ avec } \widehat{X} = \frac{\sigma_2^2}{\sigma_1^2 + \sigma_2^2} X_1 + \frac{\sigma_1^2}{\sigma_1^2 + \sigma_2^2} X_2 \,.$$

Proof: Découle directement du calcul de $\mathrm{E}\left[\widehat{X}\right]$ et $\mathrm{V}\left[\widehat{X}\right]$, qui permet de retrouver la définition de l'opérateur $\oplus$. □

Cette propriété permet de voir l'agrégation d'information comme une interpolation dont les coefficients sont proportionnels au degré de certitude de chaque information.

Ce type d'interpolation peut être rapproché de la théorie de la crédibilité, bien que les objectifs de cette théorie diffèrent légèrement [8]

Définition 5 (Agrégation totale d'un ensemble d'information) *Supposons que l'on ait un ensemble d'informations $\{I_1, \dots I_n\}$, $n \geqslant 2$. Nous appellerons l'information totale de cet ensemble la quantité :*

$$I = \bigoplus_{i=1}^{n} I_i = I_1 \oplus \cdots \oplus I_n \,.$$

Remarque 1 (Allure de l'agrégation totale) *Si pour tout $j \in \{1, \dots, n\}$, $I_j = (m_j, \sigma_j^2)$, $n \geqslant 2$, alors :*

$$I_1 \oplus \cdots \oplus I_n = (M_n, V_n), \text{ avec } \left\{ \begin{array}{lcl} M_n & = & \frac{P_n}{S_n} \sum_{i=1}^{n} \frac{m_i}{\sigma_i^2} \,, \\ \sigma^2 & = & \frac{P_n}{S_n} \,, \end{array} \right.$$

$$\text{où } \left\{ \begin{array}{lcl} P_n & = & \prod_{i=1}^{n} \sigma_i^2 \,, \\ S_n & = & \sum_{1 \leqslant j_1 < \cdots < j_n \leqslant n-1} \sigma_{j_1}^2 \dots \sigma_{j_{n-1}}^2 \,. \end{array} \right.$$

Ce résultat se montre aisément par récurrence, en utilisant le fait que $P_n + \sigma_n^2 S_n = S_{n+1}$.

Cette dernière formule permet de mieux faire le lien avec d'autres techniques d'interpolation, et de visualiser l'allure de l'interpolation réalisée. Toutefois, lors de l'implémentation de l'agrégation de plusieurs points, il est naturellement plus rapide de calculer l'information agrégée en utilisant itérativement l'opérateur $\oplus$ tel qu'il est présenté dans la définition 4.

2.4 Dégradation d'information

L'intérêt de l'opérateur $\oplus$ précédemment défini est qu'il permet l'agrégation d'informations sans connaissance ni de l'ensemble des observations G, ni de la variance supposée connue σ_X. Il devient alors possible, par exemple, de tenir compte d'informations entachées d'autres sources d'incertitudes que l'estimation : erreurs liées à des informations incomplètes, à une source supplémentaire de bruit, à une distance spatiale, etc.

Nous allons ici envisager un mécanisme de dégradation d'information. En pratique, une information peut en effet être dégradée pour plusieurs raisons :

– distance spatiale : une observation d'une fonction aléatoire a pu être menée en un point distinct de celui que l'on aimerait explorer. Pour autant, si le point observé est proche, une information est quand même partiellement exploitable.
– distance temporelle : des observations d'une variable aléatoire ont pu être menées suffisamment loin dans le passé, et loi de la variable aléatoire considérée peut avoir évolué au fil du temps. L'allure de ce qu'était la loi de cette variable aléatoire est partiellement informative.
– variation de contexte : assimilable en un sens à une distance spatiale et temporelle, mais parfois plus difficilement paramétrable, le contexte d'observation d'une variable aléatoire peut avoir évolué. Ainsi, dans un environnement économique globalement changé, les conclusions d'hier peuvent n'être que partiellement informatives.
– prudence de l'observateur : l'observateur peut délibérément atténuer la confiance qu'il accorde à certaines observations, dans un souci prudentiel. Cela peut être requis d'un point de vue réglementaire, ou pour tenter de prendre en compte des variations de contexte qui n'ont pu être intégrées au modèle.
– imprécisions numériques : des sources de bruit supplémentaires peuvent apparaître naturellement, sans être a priori intégrées au modèle. Il peut s'agir par exemple d'une imprécision numérique.
– etc.

Nous nous intéresserons ici à une dégradation d'une information I par modification du deuxième paramètre de celle-ci, correspondant à l'erreur d'estimation pour une information d'estimation. L'information dégradée sera notée $D_h(I)$:

Définition 6 (information dégradée) *Nous appellerons D_hI l'information dégradée de l'information I par un bruit $h \in \mathbb{R}^+$, telle que :*

$$\begin{aligned} I &= (m, \sigma^2)\,, \\ D_h(I) &= (m, \sigma^2 + h^2)\,, \end{aligned}$$

Lorsque $h = 0$, l'information est inchangée. Lorsque $h = +\infty$, l'information devient nulle.

Remarque 2 *La dégradation d'une information correspond à l'ajout d'un bruit indépendant, d'espérance nulle et de variance h^2.*

Imaginons un vecteur gaussien (X_1, X_2) avec $\rho = \mathrm{Cov}(X_1, X_2)$, $m_i = \mathrm{E}[X_i]$, $\sigma_i = \mathrm{V}[X_i]$, $i \in \{1, 2\}$. La loi conditionnelle de X_1 sachant X_2 est bien connue :

$$X_1 \Big| X_2 \sim \mathcal{N}\left(m_1 + \frac{\rho}{\sigma_2}(X_2 - m_2), \sigma_1^2 - \rho^2/\sigma_2^2\right)\,.$$

Supposons que $X_1 = G(\theta_1)$ et $X_2 = G(\theta_2)$, où G est un champ gaussien, et notons d la distance séparant X_1 et X_2. Imaginons que l'on observe $X_2 = x_2$, et que l'on suppose que G est un champ gaussien stationnaire isotrope, de noyau de covariance défini par la fonction $k(d)$, de sorte que $\sigma_1 = \sigma_2 = \sigma$ et $\rho = k(d)$. La variance conditionnelle de X_1 sachant X_2 est donnée par $\sigma^2 - k^2(d)/\sigma^2$.

Imaginons désormais la situation où l'on considère l'information certaine $I_2 = (x_2, 0)$ correspondant à X_2 au point θ_2. La projection de cette information au point θ_1 revient à déterminer l'information dégradée $I_{\theta_2}(\theta_1) = D_h(I_2) = (x_2, h^2(d))$.

L'information $I_{\theta_2}(\theta_1)$ aura la même variance que $G(\theta_1)$ sachant $G(\theta_2)$ si :

$$h^2(d) = \sigma^2 - k^2(d)/\sigma^2$$

En notant $\gamma(d) = \sigma^2 - k(d)$, on établit alors le résultat suivant.

Proposition 7 (dégradation d'information et processus gaussien) *Considérons un processus gaussien stationnaire isotrope G, de variogramme $\gamma(d)$. La dégradation d'information liée à la connaissance de G en un unique point situé à une distance d est telle que :*

$$h^2(d) = \gamma(d)\left(2 - \frac{\gamma(d)}{\sigma^2}\right)\,.$$

Proposition 8 (lien avec variogramme exponentiel généralisé) *Pour un processus gaussien stationnaire isotrope de variogramme exponentiel généralisé,*

$$\gamma(d) = \sigma^2 \left(1 - e^{-(d/w)^p}\right) ,$$

avec $\sigma > 0$, $w > 0$ et $p \in]0, 2]$, la dégradation d'information à une distance d correspond à l'ajout d'un bruit $h(d)$ tel que :

$$h^2(d) = \sigma^2 \left(1 - e^{-2(d/w)^p}\right) .$$

En particulier, pour ce même variogramme exponentiel généralisé, lorsque $(d/w)^p$ est petit, $h^2(d) \simeq 2\sigma^2(d/w)^p$ et $h(d) \simeq \sigma\sqrt{2/w^p}d^{p/2}$.

3 Surfaces de réponse par agrégation d'informations

Supposons qu'un observateur observe des réalisations d'une variable aléatoire $F(\theta)$, d'espérance $f(\theta)$. Comment bâtir un prédicteur de f en dehors des points d'observation ? Comment représenter la connaissance qu'a l'observateur de $f(\theta)$?

L'idée est d'ajouter une source d'incertitude à la deuxième composante d'une information avant de l'agréger : vue du point θ, l'information disponible en un point θ' n'est que partiellement informative. Cela revient à dire qu'une information θ' vue de θ est dégradée, par exemple en fonction de la distance $d_{\theta,\theta'}$ de θ à θ' :

L'ampleur de dégradation peut être quantifiée par une fonction $h(\theta, \theta')$. Néanmoins, par souci de concision, nous supposerons ici que la dégradation ne dépend ici que de la distance considérée $d_{\theta,\theta'}$. La dégradation sera donc une fonction $h(d)$ croissante en fonction de la distance d considérée, telle que $h(0) = 0$ et $h(+\infty) = +\infty$.

Définition 7 (information projetée) *L'information disponible en un point θ compte tenu d'une information $I(\theta') = (m', \sigma'^2)$ en un point θ' sera notée $I_{\theta'}(\theta)$ et nommée information de θ' projetée en θ :*

$$\begin{aligned} I_{\theta'}(\theta) &= (m', \sigma'^2 + h^2(d_{\theta,\theta'})) , \\ \text{avec } I(\theta') &= (m', \sigma'^2) . \end{aligned}$$

où $d_{\theta,\theta'}$ représente une distance entre θ et θ'.

Remarque 3 (agrégation d'une information distante) *Soient deux points θ et θ'. Supposons qu'une information $I(\theta) = (m, \sigma^2)$ est disponible en θ, et que l'on agrège $I_{\theta'}(\theta)$. D'après la définition de l'information projetée, en utilisant la définition de l'opérateur $\oplus$, on a, en notant $h_d^2 = h^2(d_{\theta,\theta'})$:*

$$\begin{aligned} & I(\theta) \oplus I_{\theta'}(\theta) = (m'', \theta'') , \\ & \text{avec } \left\{ \begin{array}{lcl} m'' & = & \frac{\sigma'^2 + h_d^2}{\sigma^2 + \sigma'^2 + h_d^2} m + \frac{\sigma^2}{\sigma^2 + \sigma'^2 + h_d^2} m' , \\ \sigma''^2 & = & \frac{\sigma^2\sigma'^2 + \sigma^2 h_d^2}{\sigma^2 + \sigma'^2 + h_d^2} . \end{array} \right. \end{aligned}$$

En particulier, on peut noter pour les moyennes :

$$m'' = \alpha_{d,\sigma,\sigma'} m + (1 - \alpha_{d,\sigma,\sigma'}) m' , \quad \text{avec} \quad \alpha_{d,\sigma,\sigma'} = \frac{\sigma'^2 + h_d^2}{\sigma^2 + \sigma'^2 + h_d^2} .$$

Définition 8 (information totale en un point) *Supposons que l'on explore un ensemble fini de n points $T = \{\theta_1, \theta_2, \ldots \theta_n\}$, et que l'on dispose en chaque point exploré $\theta_i \in T$ d'une information I_{θ_i},*

$n \in \mathbb{N}^*$. Nous appellerons information totale au point θ l'agrégation $I_T(\theta)$ de toutes les informations disponibles, dégradées en fonction de la distance :

$$I_T(\theta) \quad = \quad \bigoplus_{\theta_i \in T} I_{\theta_i}(\theta)\,.$$

avec $I_{\theta_i}(\theta)$ l'information obtenue au point θ grace à l'observation θ_i.

Remarque 4 *Le choix d'une fonction h telle que $h(d) = +\infty$ pour tout $d > d_0$ donné, $d_0 \in \mathbb{R}^+$, conduit à ne prendre en compte pour θ que les informations disponibles sur des points situés à une distance inférieure ou égale à d_0 : seuls les explorations des points situés dans une sphère $S_{d_0}(\theta)$, de rayon d_0 et centrée sur θ, seront utilisées pour l'information $I_T(\theta)$. Dans ce cas, pour tout point θ, $I_T(\theta) = I_{T \cap S_{d_0}(\theta)}(\theta)$.*

Définition 9 (Fonction d'information) *Soit une fonction réelle aléatoire $F(\theta)$, $\theta \in \mathbb{R}^+$. Supposons que l'on observe un ensemble de réalisations indépendantes de F en différents points déjà explorés d'un ensemble T. Nous définirons la fonction d'information $\tilde{f}_T$ de F la fonction aléatoire telle que : $\forall \theta \in \Theta$, $\tilde{f}_T(\theta)$ est une variable aléatoire normale ayant pour espérance et pour variance les deux composantes respectives de $I_T(\theta)$:*

$$\tilde{f}_T(\theta) \sim \mathcal{N}(m, \sigma^2)\,, \quad \textit{avec} \quad (m, \sigma^2) = I_T(\theta)\,.$$

La structure de dépendance entre les $\tilde{f}(\theta)$ n'est pas ici précisée. Dans les cas où elle s'avère nécessaire, elle doit être construite de façon à être aussi proche que possible de celle des $X(\theta)$, processus gaussien dont on suppose que f est une trajectoire (sans considération de bruit ϵ ni d'effet pépite correspondant). En effet, $\tilde{f}_T$ est un prédicteur de f au vu de l'ensemble d'observations aux points de T. En un point θ non observé, conditionnellement à $\tilde{f}_T(\theta) = x$, $x \in \mathbb{R}$, on doit avoir $\tilde{f}_T(\theta + h)$ proche de x lorsque $d(x, x + h)$ tend vers 0.

La fonction d'information associe à chaque $\theta \in \Theta$ un prédicteur aléatoire $\tilde{f}_T(\theta)$. Ce prédicteur est la représentation de $f(\theta) = \mathrm{E}\left[F(\theta)\right]$ par l'observateur, en connaissance des observations T.

- Lorsque $f(\theta)$ est connu, $\tilde{f}_T(\theta) = f(\theta)$.
- Lorsque θ est l'unique point exploré de T, $\tilde{f}_T(\theta)$ est une v.a. de loi normale, centrée sur la valeur estimée $\widehat{f}(\theta)$ (moyenne empirique des observations), et d'écart-type identique à celui de cette moyenne empirique.
- Un intervalle de confiance $I_r = [\widehat{f}(\theta) - r, \widehat{f}(\theta) + r]$ peut être construit pour f en choisissant r tel que $\mathrm{P}\left[\tilde{f}_T(\theta) \in I_r\right] > 1 - \epsilon$, $\epsilon \in]0, 1[$.

L'avantage notable de cette approche est qu'en un point θ, l'intervalle de confiance obtenu peut être beaucoup moins étendu qu'en utilisant des statistiques isolées au seul point θ. La prise en compte des informations aux points adjacents est susceptible de beaucoup améliorer l'estimation, si toutefois les paramètres de l'hypothèse de dégradation spatiale sont convenablement choisis.

Lorsque deux points observations θ_1 et θ_2 sont très proches, l'observation de n réalisations de $F(\theta_1)$ d'une part et de n réalisations de $F(\theta_2)$ d'autre part revient bien à considérer ces $2n$ réalisations au seul point θ_1.

L'intérêt majeur de l'opérateur $\oplus$ présenté est qu'il permettra d'agréger des informations dégradées, là où il serait délicat d'envisager des erreurs d'estimations. Ainsi, si l'on considère un point θ non exploré mais dans un voisinage de points d'exploration, le nombre d'observations à l'origine de l'information dégradée en θ est perdu. Il serait envisageable de comptabiliser chaque observation comme une observation partielle (comme une demi-observation par exemple), mais le lien avec l'ajout d'un bruit est alors moins évident.

4 Optimisation et inversion de quantiles

Supposons que l'on cherche à minimiser la fonction objectif $f(\theta)$ en déterminant le minimum m^* de cette fonction ainsi que les minimiseurs de Θ :

$$m^* = \inf_{\theta \in \Theta} \{f(\theta)\} \ .$$

De très nombreux algorithmes d'optimisation existent. Dans le cas d'une fonction objectif f non bruitée et de la recherche d'un optimum local, on peut consulter par exemple les méthodes de descente de gradient, de Newton-Raphson, de Hooke et Jeeves, la méthode de [25], ou des méthodes spécifiquement adaptées à certaines formes de f, comme lorsque f est convexe. S'il s'agit de la recherche d'un optimum global, les recherches sont plus récentes et l'on pourra se référer par exemple à [14]. L'optimisation Lipschitzienne, l'algorithme de Schubert [cf. 31] ou l'algorithme DIRECT [cf. 17, 18] sont des algorithmes de type Branch and Bound qui s'intègrent dans ce cadre [cf. 23, 15, 11, 12, 35, pour une revue de différentes techniques].

Dans le cas d'une fonction objectif bruitée, différents algorithmes stochastiques visent à rechercher un optimum local ou des racines [cf. 20, 5, 33, 29]. Enfin, dans le cas de la recherche d'un minimum global d'une fonction bruitée non nécessairement convexe, des méthodes de recuit simulé [cf 1, 6] ou des méthodes génétiques [cf. 2, 24] sont envisageables, bien que le bruit rende le problème plus difficile [cf 9]. Quelques méthodes dérivées d'approches Branch-and-Bound existent [cf. par exemple 26]. Une classification d'algorithmes basés sur des surfaces de réponse peut être trouvée dans [19]. Des méthodes utilisant l'entropie de Shannon (IAGO) ou des algorithmes dérivées de EGO sont développées dans [3, 27] ainsi que dans les autres présentations du *Workshop on Noisy Kriging-based Optimization* (Berne, novembre 2010).

Dans [30] un algorithme d'optimisation dérivé d'une approche de type Branch-and-Bound se fonde sur l'estimation d'une quantité nommée "potentiel" et calculable pour chaque point candidat. Cette quantité traduit la probabilité que le prédicteur de la fonction f conduise, en un point donné, à un optimum meilleur que celui observé. Nous reprenons ici l'idée d'un potentiel, en montrant que l'on peut bâtir un potentiel calculé à partir d'agrégation d'informations (cf définition 10). En outre, plutôt que de partitionner la zone de recherche, nous supposerons que l'on se donne à chaque étape un ensemble de points candidats, ce qui conduira à un algorithme dont l'implémentation est beaucoup plus simple.

Il est de nombreuses situations où l'on cherche à déterminer l'ensemble des points conduisant à une valeur de f appartenant à un intervalle donné J :

$$\mathcal{S}(J) = \{\theta \in \Theta, f(\theta) \in J\} \ , \quad J \subset \mathbb{R} \, .$$

Les intervalles J considérés peuvent revêtir de nombreuses formes :

- Considérons la variable aléatoire $Q = f(U)$, où U est une variable aléatoire de loi connue sur le domaine Θ. La recherche des quantiles de Q revient à rechercher les points conduisant f à appartenir à $J_q = \inf\{x \in \mathbb{R}, \, \mathrm{P}\,[Q > x] \geqslant q\}$. Il s'agit alors de déterminer $\mathcal{S}(J_q)$, $q \in [0, 1]$.
  Ce type d'application revêt une grande importance dans le cadre de l'assurance et de la finance, notamment lorsque l'application de règles prudentielles impose aux organismes financiers ou assureurs de disposer d'une réserve d'argent suffisante pour ne pas être ruiné dans un certain pourcentage de scénarios.
- Les problèmes classiques d'optimisation globale reviennent quant-à-eux à déterminer $\mathcal{S}(J)$, avec $J =]-\infty, m^*]$, où m^* désigne la valeur minimale, supposée finie de f : $m^* = \inf\{f(\theta), \, \theta \in \Theta\}$.
  Là encore, ce type d'application permet par exemple d'optimiser un indicateur de gain ou un indicateur de risque dans un cadre d'allocation optimale ou de réassurance optimale. Plus généralement, de très nombreux problèmes d'ingénierie et de recherche opérationnelle conduisent à ce type d'optimisation.
- Enfin, la recherche des points conduisant f à appartenir à $J = [-\eta, \eta]$, $\eta \in \mathbb{R}^+$, revient à la recherche de racines de f.

A titre d'exemple, lorsque $J =]-\infty, m^*]$, l'algorithme 1 vise à obtenir un estimateur de la valeur supposée unique de l'optimum global $m^* = \inf_{\theta \in T} f(\theta)$, ainsi que des zones de confiance pour les minimiseurs supposés.

Cet algorithme s'appuie sur la notion de potentiel d'un point que nous définissons de la façon suivante :

Définition 10 (potentiel) *Nous définissons le potentiel d'un point θ, en connaissance d'un échantillon T et d'un intervalle J comme :*

$$\beta_{T,J}(\theta) = \mathrm{P}\left[\tilde{f}_T(\theta) \in J\right].$$

Par convention, en l'absence d'information tout point est susceptible de conduire à $f \in J$: $\beta_{\emptyset,J}(\theta) = 1$.

En un point $\theta \in \Theta$, une fois défini $\tilde{f}_T$ le prédicteur (aléatoire) de la fonction f (déterministe mais parfois bruitée), il est possible de définir différentes quantités indiquant si $\tilde{f}_T$ est susceptible d'appartenir à un intervalle cible J au point θ. Dans le cas de l'optimisation, pour $J =]-\infty, m]$, la définition d'un potentiel $\beta_{T,m}(\theta)$ se rapproche de l'usage classique d'un *Expected Improvement* $EI_{T,m}(\theta)$ [cf. 18, pour une surface de réponse $\tilde{f}_T$ construite différemment] :

$$\beta_{T,m}(\theta) = \mathrm{P}\left[\tilde{f}_T(\theta) < m\right] \quad \text{et} \quad EI_{T,m}(\theta) = \mathrm{E}\left[\left(m - \tilde{f}_T(\theta)\right)_+\right] = \int_{-\infty}^{m} \beta_{T,s}(\theta) ds\,.$$

On peut ainsi voir l'expected improvement comme un potentiel agrégé. Le débat de l'usage de l'une ou l'autre de ces quantités rejoint le débat du choix de mesures de risques (*Value-at-risk* ou *Tail-Value-at-risk*, cf. [10]), l'expected improvement possédant par exemple une propriété de sous-additivité. Ce choix peut aussi être vu comme le choix d'une moyenne conditionnelle ou d'une médiane conditionnelle, cette dernière étant plus robuste au sens de Hampel et Huber [cf. 13, 16].

Définition 11 (zone de confiance) *Nous définissons la zone de confiance pour l'intervalle J en connaissance de l'échantillon T et du seuil s comme :*

$$\mathrm{SS}_{T,s}(J) = \{\theta \in \Theta, \beta_{T,J}(\theta) \geqslant s\}\,.$$

En supposant $\tilde{f}_T(\theta)$ de loi normale pour tout $\theta \in \Theta$, alors l'information $I_T(\theta) = (m, \sigma)$ disponible en θ donne la moyenne et la variance de cette loi. Dans le cas gaussien, cette information est donc suffisante pour déterminer le potentiel

Proposition 9 (valeur du potentiel dans le cas gaussien) *Notons Φ la fonction de répartition d'une loi normale centrée réduite. $\tilde{f}_T(\theta)$ étant supposée de loi normale pour tout $\theta \in \Theta$, alors pour $J = [a, b]$, $a, b \in \mathbb{R}$:*

$$\begin{aligned} \beta_{T,J}(\theta) &= \Phi\left(\frac{b-m}{\sigma}\right) - \Phi\left(\frac{a-m}{\sigma}\right), \\ \textit{avec } (m, \sigma) &= I_T(\theta)\,. \end{aligned}$$

L'idée de l'algorithme 1 est de piocher un ensemble de points candidats pour l'exploration, de calculer leur potentiel, et choisir le point à explorer en fonction de son potentiel.

On pourrait ne piocher qu'un point à chaque étape, et effectuer une méthode des rejets pour n'explorer le point que si un tirage uniforme sur $[0, 1]$ est inférieur à son potentiel. En grande dimension, le nombre de points pouvant être rejeté risque d'être important, et on ne contrôlerait pas ainsi le nombre de tentatives avant de valider un point pour l'exploration. C'est la raison pour laquelle nous proposons de sélectionner un point pour l'exploration parmi un ensemble de candidats Θ_e de taille n_e à fixer.

Supposons que l'on dispose de n_e candidats à départager, pour lesquels n_e potentiels ont été calculés. On peut imaginer sélectionner le point de potentiel maximal parmi ces candidats (choix que

nous qualifierons de *priorité absolue*) ou de choisir un point avec une probabilité proportionnelle à son potentiel (choix que nous qualifierons de *priorité relative*). Dans le cas d'une priorité absolue, le choix de $n_e = 1$ conduit à une exploration uniforme de la zone de départ, tandis que $n_e \to +\infty$ conduirait à choisir un des points de potentiel maximal sur la zone de départ (par exemple celui ayant conduit à une valeur optimale). Un compromis entre ces deux solutions semble souhaitable. Nous retiendrons donc le choix d'une priorité relative : dans ce cas, le choix de $n_e = 1$ conduit à une exploration uniforme de la zone de départ, tandis que $n_e \to +\infty$ a moins d'incidence sur l'exploration que le choix d'une priorité absolue.

L'algorithme 1 permet ainsi de construire itérativement l'ensemble T_j des points observés à l'étape j et la succession des estimateurs J_j de l'intervalle cible J. La constante n représente le nombre d'étapes de l'algorithme.

Algorithme 1 algorithme d'exploration par potentiel

Entrée: n, n_e, n_0, fonction h, zone de départ Θ

$T_0 = \emptyset$

$J_0 = \mathbb{R}$

pour $j = 0$ à $n - 1$

échantillon uniforme sur Θ

générer un échantillon $\Theta_e = \{\theta_1, \dots, \theta_{n_e}\}$ pioché de façon uniforme sur Z_0

Calcul des potentiels

pour tout $\theta \in \Theta_e$, déterminer $\beta_{T_j,J_j}(\theta)$

Exploration d'un point de Θ_e

piocher un point θ^+ de Θ_e en fonction des $\left\{\beta_{T_j,J_j}(\theta)\right\}_{\theta\in\Theta_e}$

$T_{j+1} = T_j \cup \{\theta^+\}$

générer un échantillon de taille n_0 de tirages de $F(\theta^+)$

calculer $I(\theta^+)$ et estimer J_{j+1}

fin pour

Sortie: J_n

Sortie: $\forall\theta \in T_n$, $I_{T_n}(\theta)$, $\beta_{T_n,J_n}(\theta)$

Du fait de $T_0 = \emptyset$ et de $\beta_{\emptyset,J}(\theta) = 1$, le premier point θ^+ est sélectionné sans préférence parmi les n_e candidats. Il est naturellement possible de placer l'exploration de ce premier point en initialisation de l'algorithme. Il s'agit d'ailleurs d'une propriété intéressante de cet algorithme, qui ne requiert pas d'exploration préalable de la fonction objectif et permet donc d'utiliser $T = \emptyset$.

Le détail de l'estimation des J_{j+1} à l'étape j en fonction des points déjà explorés $\theta_1, \dots, \theta_j$ n'est pas ici fourni dans la mesure ou cette estimation peut différer en fonction de la signification de l'intervalle cible J. Lorsqu'il s'agit de la recherche d'un minimum m^*, on peut prendre par exemple $J_{j+1} =]-\infty, m_j^*]$, avec :

$$m_j^* = \min_{\theta\in T}\left\{m_{T_j}(\theta) + \lambda\sigma_{T_j}(\theta)\right\},$$
$$\text{où } \left(m_{T_j}(\theta), \sigma_{T_j}(\theta)\right) = I_{T_j}(\theta),$$

où λ est une constante numérique détaillée dans la partie application numérique, et où T représente un ensemble de points sur lequel est conduite cette estimation, par exemple l'ensemble des points déjà explorés $\theta_1, \dots, \theta_j$. L'idée de la constante λ est d'intégrer une certaine prudence dans cette estimation : le minimum estimé sera supposé inférieur à $m_i + \lambda\sigma_i$ avec une probabilité suffisante, et les points conduisant à un minimum potentiellement inférieur à ce seuil relevé seront jugés intéressants. Cette constante λ permet ainsi d'éviter les situations où, du fait d'une grande volatilité, on a estimé en θ_i un minimum $m^*(\theta_i)$ beaucoup plus petit que l'optimum m^*. Cette dernière situation est à éviter

dans car elle exclut l'exploration de beaucoup de points pour lesquels f est pourtant proche du vrai m^*.

Le tirage des candidats de façon uniforme peut se faire, lorsque Θ est représenté par une union finie de simplexe, à l'aide de la méthode de Kraemer, détaillée dans [32] (en finance, par exemple, la recherche de pourcentages d'allocation d'actifs conduit à une recherche à mener dans le simplexe orthogonal unité). Si de nombreuses zones sont a priori exclues de l'analyse, et si l'on désire limiter le nombre n_e de candidats, il est possible de modifier ce choix de distribution uniforme initiale.

Par ailleurs, on peut imaginer une évolution de la fonction h en fonction de l'étape en cours, la connaissance sur la régularité de h s'affinant au fur et à mesure de l'exploration de la fonction bruitée. Des éléments sur l'estimation de la fonction h peuvent être trouvés dans [30].

5 Illustrations numériques

5.1 Surfaces de réponse en dimension 1

Dans ce paragraphe, nous comparons des surfaces de réponse obtenues par krigeage ou par agrégation d'information, en présence d'un bruit non homogène et d'ampleur importante.

Dans une première application, nous essayons d'estimer la probabilité p_θ de réalisation d'une variable aléatoire de type Bernoulli, pour un paramètre $\theta \in \mathbb{R}$. Cela correspond à la situation classique où l'on essaye d'estimer une probabilité au vu de réalisations appartenant à $\{0,1\}$. Il ne s'agira pas de trouver le meilleur modèle pour faire ce type de prédiction, mais d'analyser sur cet exemple simple le comportement de l'agrégation d'information et du krigeage.

On considère donc le modèle suivant :

$$F(\theta) = p_\theta + \epsilon(\theta)\,,$$

où $F(\theta)$ est une variable aléatoire de type Bernoulli de paramètre p_θ, donc d'espérance $\mathrm{E}\left[F(\theta)\right] = p_\theta$. Les bruits distincts de $\{\epsilon(\theta)\}_{\theta\in\Theta}$ sont donc d'espérance nulle, et considérés mutuellement indépendants.

La probabilité p_θ varie pour différents paramètres $\theta \in \mathbb{R}$ en entrée, et la connaissance d'estimations de p_θ en d'autres points d'un ensemble déjà exploré T constitue naturellement une information intéressante. Nous avons considéré dans nos applications l'ensemble $\Theta = [0, \nu]$, $\nu \in \mathbb{R}$, et

$$p_\theta = \frac{1}{2}\left(1 + \cos(\theta)\right)\,.$$

On note $T \subset \Theta$ l'ensemble des points déjà explorés : en chaque point exploré $\theta \in T$, on dispose de $n_\theta \geqslant 2$ observations. Ces observations forment une réalisation $(F_1(\theta), \ldots, F_{n_\theta}(\theta))$ d'une suite de variables aléatoires iid de loi Bernoulli de paramètre p_θ. On dispose donc de l'estimateur $\widehat{p}_\theta$ de p_θ et de la variance d'estimation $\sigma_e^2(\theta) = \mathrm{V}\left[\widehat{p}_\theta\right]$:

$$\widehat{p}(\theta) = \frac{1}{n_\theta}\sum_{j=1}^{n_\theta} F_j(\theta)\,,\quad \widehat{\sigma}^2(\theta) = \frac{1}{n_\theta - 1}\sum_{j=1}^{n_\theta}(F_j(\theta) - \widehat{p}(\theta))^2 \quad \text{et} \quad \widehat{\sigma}_e^2(\theta) = \frac{1}{n_\theta}\widehat{\sigma}^2(\theta)\,.$$

En dehors des points $\theta \in T$ d'observations, un prédicteur $\tilde{p}(\theta)$ est construit, associé à une variance de prédiction $\tilde{\sigma}(\theta)$. Ce prédicteur peut être obtenu soit par agrégation d'information, soit par krigeage. Les figures 1 et 2 montrent l'allure des prédicteurs obtenus. Dans toutes ces illustrations, nous avons considéré le cas où $\nu = 10$, où l'ensemble T des sites déjà explorés est constitué de $n_T = 25$ réalisations de variables aléatoires uniformes sur Θ (et mutuellement indépendantes). En chaque site exploré $\theta \in T$, on estime p_θ à partir d'un échantillon de n_θ valeurs (avec $n_\theta = 20$ ou $n_\theta = 5$). Dans ces deux figures, le krigeage se base sur une structure de covariance de type gaussienne dont les paramètres ont été estimés par la méthode des moindres carrés pondérés en utilisant le logiciel `R`. L'utilisation d'autres méthodes d'estimation comme le maximum de vraisemblance ou par moindres carrés classiques, ne modifie pas sensiblement la surface de réponse.

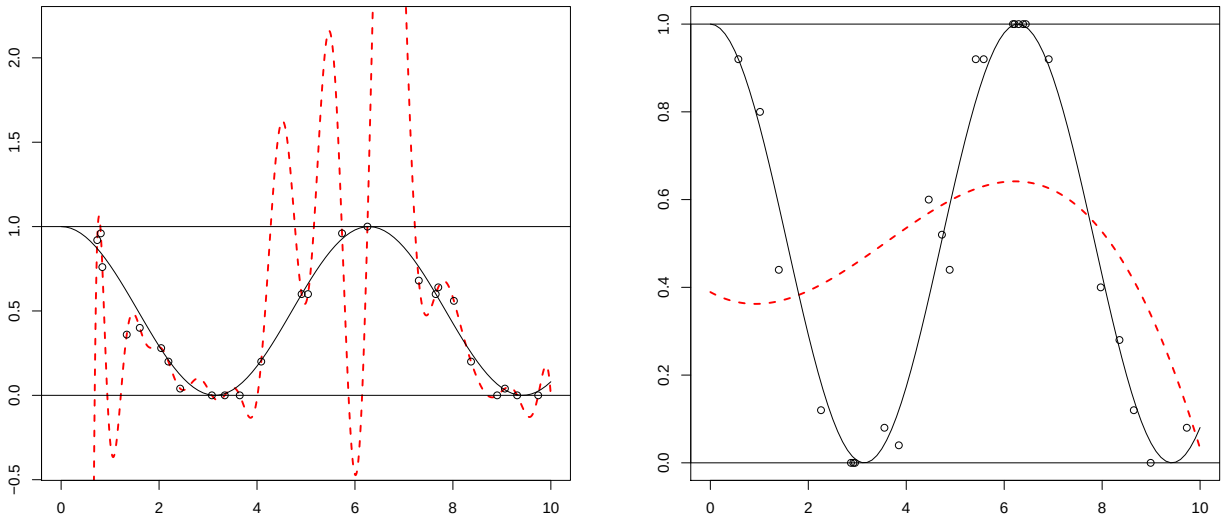

FIG. 1 – Allure de prédicteurs $\tilde{p}_\theta$ de p_θ obtenus par krigeage, sans effet pépite (à gauche), ou avec effet pépite stationnaire (à droite). Les cercles représentent les estimations bruitées de p_θ.

La figure 1 montre deux écueils à éviter lorsqu'on utilise le krigeage en environnement bruité : la non-considération d'effet pépite force le krigeage à passer par tous les points observés, et peut ainsi conduire à assimiler des variations proches liées au bruit à une très brusque variation de la fonction objectif. A contrario, la considération d'un effet pépite stationnaire conduit dans ce modèle, pour certaines simulations, à une sur-évaluation du bruit aux sommets des crêtes, et à une interpolation d'amplitude beaucoup trop faible.

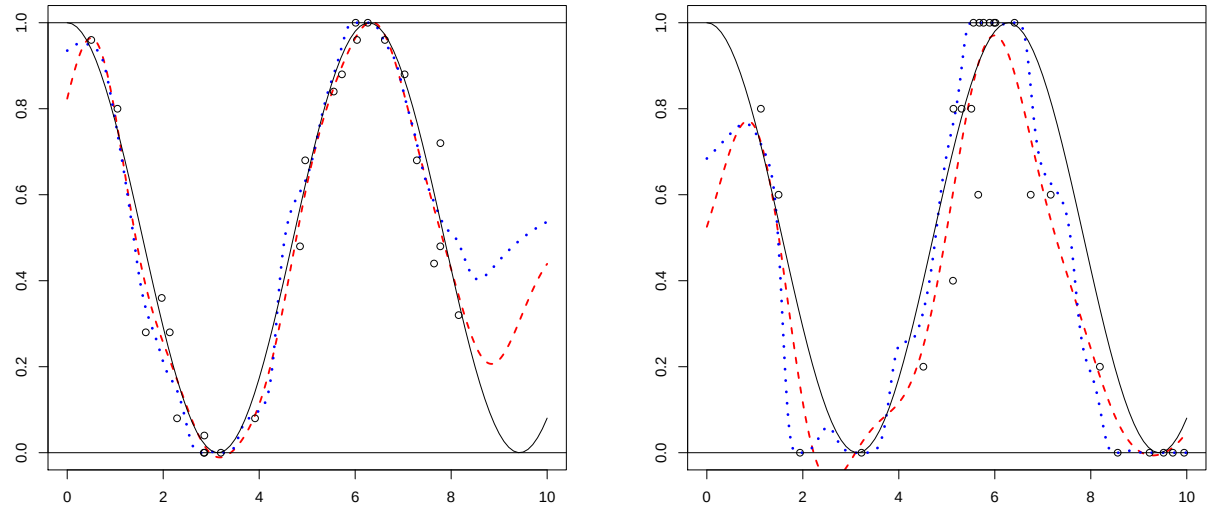

FIG. 2 – Allure des prédicteurs $\tilde{p}_\theta$ de p_θ obtenus par agrégation d'information avec $h(d) = \sigma_K d^2$ (en pointillés), ou par krigeage (tirets), les cercles représentent les estimations bruitées de p_θ. A gauche : $n_\theta = 20$, à droite : $n_\theta = 5$.

La figure 2 montre deux prédicteurs : l'un obtenu avec un krigeage avec effet pépite et matrice de covariance gaussienne estimée (moindre carrés), l'autre avec l'agrégation d'information, avec une fonction de dégradation d'information très simple : $h(d) = \sigma_K d^2$.

En pratique, l'estimation du paramètre σ_K de la fonction de dégradation h a été conduite de la façon suivante : en supposant que l'on dispose d'une unique information certaine $(p_{\theta'}, 0)$ en un point θ', l'information projetée en θ est $(p_{\theta'}, h^2(d_{\theta,\theta'}))$, où $d_{\theta,\theta'}$ représente une distance entre θ et θ'. Cela revient à supposer qu'en l'absence de bruit, le prédicteur $\tilde{p}_\theta$ de p_θ est distribué selon une loi normale d'écart-type $h(d_{\theta,\theta'}) = \sigma_K d_{\theta,\theta'}$, et que $(\tilde{p}_\theta - p_{\theta'})/d_{\theta,\theta'}$ est de loi normale centrée et d'écart-type σ_K. Pour U_1 et U_2 variables aléatoires uniformes indépendantes sur Θ, nous avons calculé la variable aléatoire $IR = (p_{U_1} - p_{U_2})/d(U_1, U_2)$, et pris σ_K égal à l'écart-type empirique d'un échantillon issu de IR.

Sur les parties gauche et droite de la figure 2, krigeage comme agrégation d'information s'écartent naturellement de la cible, dans la mesure ou l'on ne dispose plus d'observations pour l'estimer. En présence de peu d'observations pour estimer p_θ, par exemple lorsque $n_\theta = 5$, l'estimateur $\widehat{p}_\theta$ subit un bruit important, et les prédicteurs sont plus éloignés de la cible, tant pour le krigeage que pour l'agrégation d'incertitudes. Dans ces figures, en dépit des paramètres supplémentaires et des inversions de matrices qu'utilise le krigeage, les résultats issus de l'agrégation d'information semblent acceptables, sans toutefois deviner exactement la position de p_θ du fait de l'incidence des erreurs d'estimation.

5.2 Surfaces de réponse en dimension 2

La zone Θ initiale est ici supposée être un simplexe, situation qui survient notamment lors de la recherche de poids d'allocation d'actifs en finance, les pourcentages d'allocation se sommant à un [cf. 30].

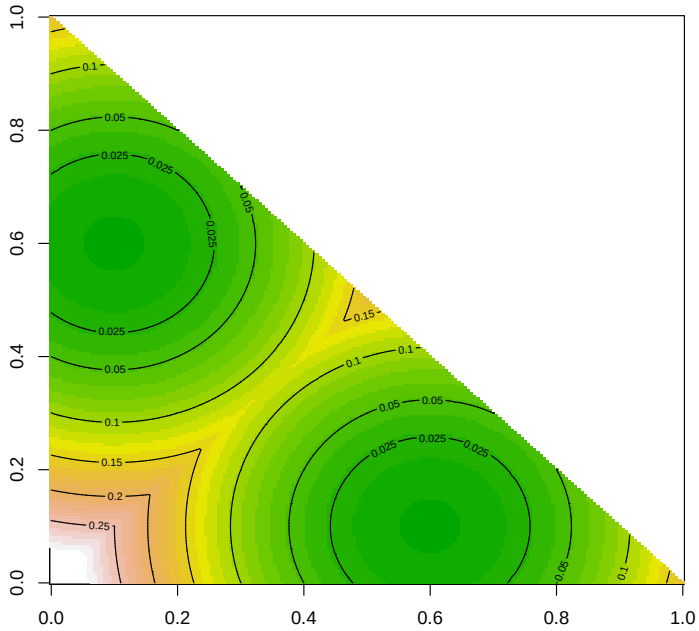

FIG. 3 – Courbes de niveaux de la fonction $f(\theta)$ en l'absence de bruit ($\sigma_B = 0$).

Nous prenons ici en dimension $d = 2$:

$$\begin{aligned} f(\theta) &= (\min(x,y) - 0.1)^2 + (\max(x,y) - 0.6)^2 \,, \\ F(\theta) &= f(\theta) + \sigma_B (U - 0.5) \,, \\ \theta &= (x,y) \,, \end{aligned}$$

où U est une variable aléatoire de loi uniforme sur $[0,1]$. La fonction f possède deux minima, l'un en $\theta_1^* = (0.1, 0.6)$, l'autre en $\theta_2^* = (0.6, 0.1)$. La valeur de f en ces deux points est $f(\theta_1^*) = f(\theta_2^*) = 0$. Afin d'imaginer les variations possibles de cette fonction, la fonction f atteint son maximum sur Θ en $\theta = (0,0)$ et l'on a alors $f(\theta) = 0.37$. Nous utiliserons par la suite un bruit d'amplitude $\sigma_B = 0.1$, qui est donc assez élevé au vu de cette amplitude de 0.37 de variation de f. Les courbes de niveau de la fonction f non bruitée sont présentées dans la figure 3.

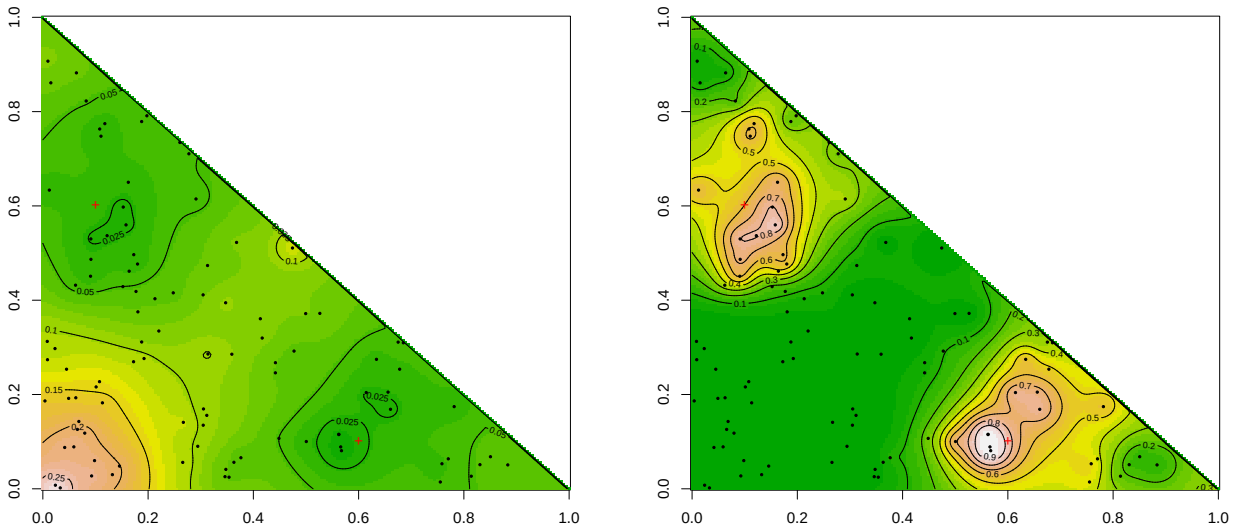

FIG. 4 – Courbes de niveaux estimées de la fonction f, en présence d'un bruit $\sigma_B = 0.1$, par agrégation d'information avec $h(d) = d^2$ (à gauche), potentiel correspondant (à droite).

Dans les figures 4 et 5, différentes surfaces de réponse ont été produites à partir d'un ensemble T de $n_T = 100$ points déjà explorés, représentés par des points noirs. En chaque point observé $\theta \in T$, les estimations de $f(\theta)$ et de l'erreur d'estimation $\sigma_e(\theta)$ ont été menées à partir de $n_\theta = 10$ réalisations de $F(\theta)$. Dans la figure 4, les courbes de niveaux obtenues par agrégation d'information montrent l'interpolation réalisée de f en présence d'un bruit $\sigma_B = 0.1$. Les potentiels correspondant donnent la valeur de $\mathrm{P}\left[\tilde{f}_T(\theta) < m_T + 2\sigma_T\right]$, où (m_T, σ_T) correspond à l'information totale au point minimiseur estimé. La figure 5 donne les mêmes courbes obtenues par krigeage avec une covariance gaussienne et un effet pépite, estimés par la méthode des moindres carrés pondérés.

Si le krigeage se comporte ici très bien, l'agrégation d'information reste également intéressante, malgré l'absence d'inversion de matrice et un nombre de paramètres ici très limité.

5.3 Optimisation

L'algorithme d'optimisation proposé dans la section 4 repose sur l'idée d'une exploration privilégiée des zones de fort potentiel, à l'instar de l'algorithme de type Branch-and-Bound développé dans [30]. Nous reprenons ici la même fonction test que celle présentée en 5.2.

La fonction de dégradation utilisée pour cet exemple est la fonction suivante :

$$h(d) = \sigma_K d^\alpha \,,$$

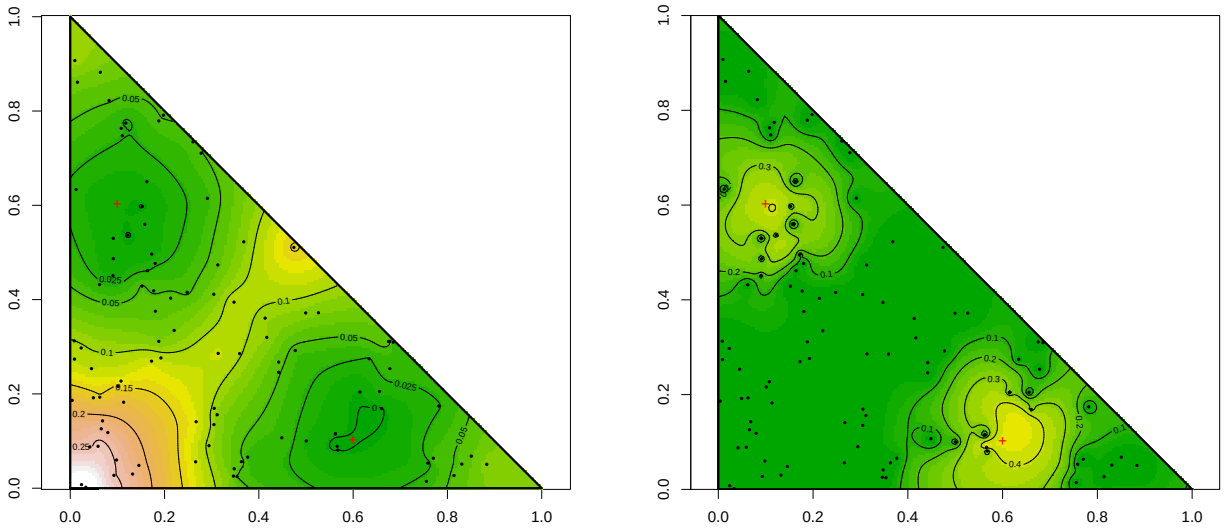

FIG. 5 – Courbes de niveaux estimées de la fonction f, en présence d'un bruit $\sigma_B = 0.1$, par krigeage (à gauche), potentiel correspondant (à droite).

où (σ_K, α) sont des constantes liées à la régularité de la fonction objectif. Des propositions pour l'estimation des constantes en question, basées sur des techniques de maximisation de vraisemblance, peuvent être trouvées dans [30]. Dans les applications numériques présentées, nous avons utilisé par exemple $(\sigma_K, \alpha) = (0.6, 1.2)$. Ces quantités peuvent être mises en correspondance avec un variogramme exponentiel généralisé (cf. Prop 8), mais correspondent ici à une fonction un peu plus régulière que celle générée par un processus gaussien de noyau de covariance gaussien. Lors de chaque exploration d'un nouveau point θ, n_0 tirages de $F(\theta)$ sont opérés afin d'estimer la variance, supposée inconnue, de $F(\theta)$.

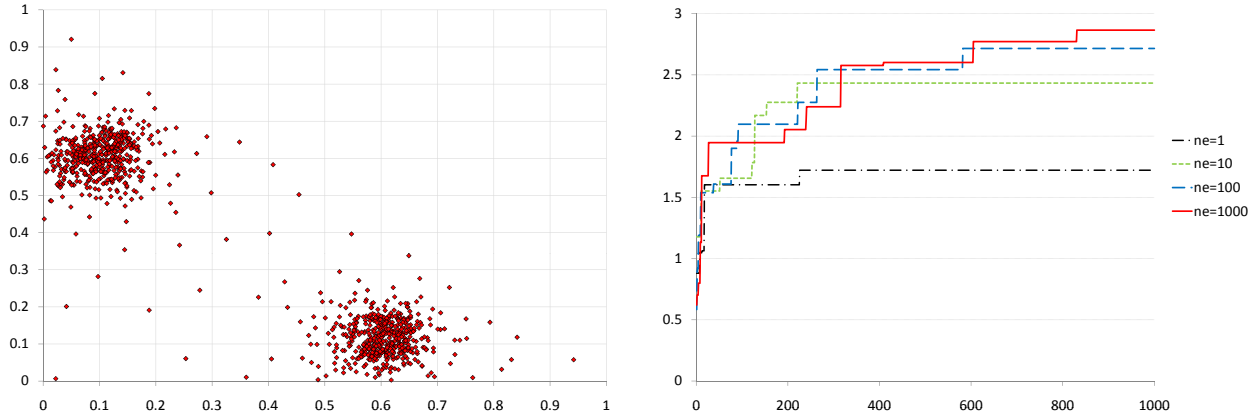

FIG. 6 – Position des 1000 premiers points d'exploration obtenus pour $n_e = 100$, par agréagation d'information (gauche) et valeur absolue du logarithme base 10 de la distance minimale à un optimum, en fonction du nombre de points explorés (droite). $\lambda = 2$, $n_0 = 10$, $\sigma_B = 0.1$.

La figure 6 montre un comportement logique de l'algorithme, qui privilégie l'exploration de chacune des deux zones susceptibles de contenir un optimum. Sur la partie droite de la figure 6 apparaît la valeur absolue du logarithme base 10 de la distance minimale entre les points explorés et une des deux

solutions. Cette quantité donne une indication sur le nombre de décimales de la distance à une solution : ainsi, avec un unique candidat $n_e = 1$, la distance à la solution est de l'ordre de 10^{-2}, tandis qu'avec $n_e = 1000$ candidats, elle est plutôt de l'ordre de 10^{-3}. Le choix $n_e = 1$ correspond à une exploration uniforme du domaine de recherche. Comme on peut le constater, pour $n_e > 1$ l'algorithme explore plus rapidement des points à faible distance d'un optimum que lors d'une exploration uniforme : sur ces trajectoires une centaine d'observations choisies avec $n_e = 100$ sont suffisantes pour obtenir une meilleure précision qu'avec un millier d'observations uniformément réparties ($n_e = 1$).

Pour un nombre d'étapes fixé, du fait de la nature aléatoire de ces trajectoires, le choix d'un n_e supérieur ne garantit pas que le plus proche point exploré sera meilleur. Par contre, le nombre d'évaluation de F ne dépend pas de n_e : le calcul du potentiel des n_e candidats ne requiert aucune exploration supplémentaire, et dans le cas où l'évaluation de F est particulièrement coûteuse, le choix d'un n_e élevé modifiera peu le temps d'exécution de l'algorithme.

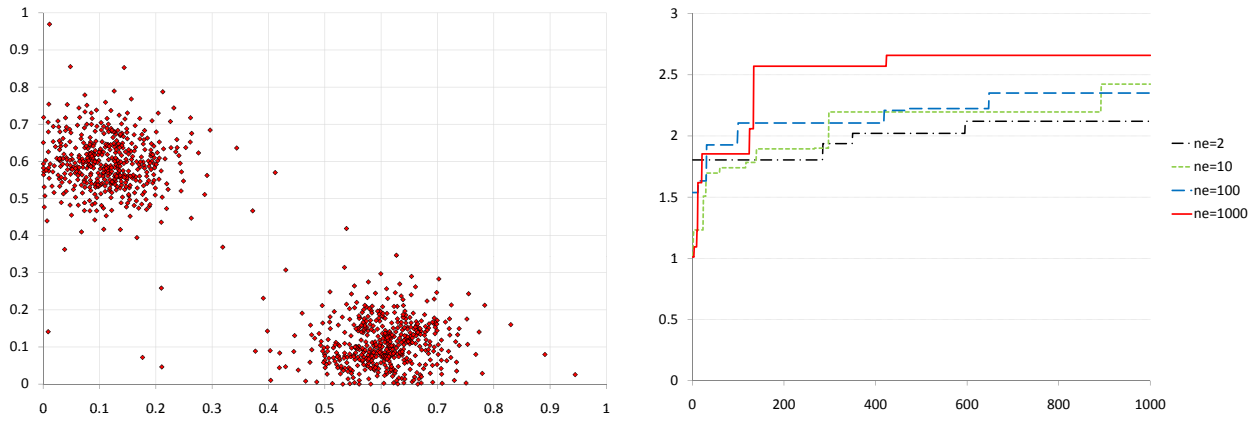

FIG. 7 – Position des 1000 premiers points d'exploration obtenus pour $n_e = 1000$, par une version adaptée de l'algorithme EGO (gauche) et valeur absolue correspondante du logarithme base 10 de la distance minimale à un optimum, en fonction du nombre de points explorés (droite). $\lambda = 2$, $n_0 = 10$, $\sigma_B = 0.1$.

La figure 7 montre le comportement d'une version adaptée de l'algorithme EGO sur ce même type de données. Cette version d'EGO s'appuie sur l'algorithme 1, en calculant un Expected Improvement [cf. 18, 22] au lieu du potentiel, en en choisissant comme futur point d'exploration celui (parmi les candidats de Θ_e) maximisant cet Expected Improvement. A chaque étape j de l'algorithme, la covariance exponentielle avec effet pépite est estimée (par la méthode des moindres carrés pondérés). Le comportement est très voisin de celui de l'algorithme présenté. L'algorithme semble légèrement plus sensible à l'augmentation du nombre de points candidats n_e, le passage à $n_e = 1000$ candidats conduisant à une amélioration significative. Là encore, les mille inversions de matrices comportant jusqu'à un million de termes n'apportent pas, en environnement bruité, de précision supplémentaire sur le résultat final.

Avec cette même fonction objectif, il apparaît par ailleurs que l'usage de l'algorithme d'optimisation locale stochastique de Kiefer-Wolfowitz-Blum, lorsqu'il est initialisé avec une suite de valeurs par défaut non spécifiquement optimisée, conduit à une précision inférieure, de l'ordre de $3.3 \cdot 10^{-2}$, même après 2000 explorations [cf. 30, pour plus de détails].

Enfin, l'algorithme présenté n'est pas limité à la dimension $d = 2$. Toutefois, comme cela est évoqué dans [4] et [30], ce type d'algorithme est vite piégé par la dimension de l'espace à explorer : lorsque la dimension augmente, l'information recueillie en chaque point exploré devient insuffisante et le gain de performance par rapport à une exploration uniforme du domaine s'estompe.

6 Conclusion

Nous avons proposé dans cette étude une définition de l'information d'une variable aléatoire et nous avons vu comment agréger des informations de façon à tenir compte de la réduction du bruit lorsque l'on réitère des tirages d'une même variable aléatoire.

La méthode proposée présente l'avantage de réduire simplement les erreurs d'estimation, notamment lorsque la fonction objectif est très bruitée. Le remplacement de l'hypothèse de processus gaussien sous-jacent (utilisée dans les modèles de krigeage) par une hypothèse de dégradation d'information permet l'utilisation d'une fonction de dégradation d'information, d'usage plus simple qu'une fonction de covariance, dans la mesure où l'agrégation d'information ne requiert pas d'inversion de matrices. Cette hypothèse peut être débattue, mais l'abandon d'inversion de matrice devient numériquement incontournable lorsque le nombre de points exploré est élevé.

La variance du prédicteur obtenu s'obtient naturellement, par construction. Malgré des résultats perfectibles lorsque les sites d'observations sont très distants, l'intérêt majeur de l'agrégation d'information est la prise en compte de la réduction de l'erreur d'estimation lorsque les points observés se rapprochent, situation survenant naturellement lors de la recherche d'optimum.

De nombreuses pistes et extensions peuvent être apportés au travail présenté ici :

- Les estimations des erreurs d'estimation en chaque point exploré reposent sur la réitération de n_0 tirages, la question du choix de n_0 reste ici en suspens.
- L'usage d'un ensemble de candidats choisis de façon uniforme peut conduire à un choix de n_e élevé afin de garantir l'exploration de chaque zone potentiellement intéressante. La modification de la loi de la position d'un candidat a priori peut constituer une piste intéressante.
- L'analyse détaillée des performances comparées des différents algorithmes d'optimisation globale d'une fonction bruitée permettrait de mieux quantifier l'intérêt de la prise en compte de la réduction du bruit d'estimation au fil des tirages. Le grand nombre d'algorithmes et de fonctions tests peut néanmoins conduire à une étude assez longue.

Références

[1] Aarts, E.H.L., Laarhoven V. (1985) *Statistical cooling : a general approach to combinatorial optimization problems,* Philips J.Res, 40 (4), 193-226.

[2] Alliot, J.M. (1996) *Techniques d'optimisation stochastique appliquées aux problèmes du contrôle aérien.* INPT, Habilitation à Diriger des Recherches.

[3] Bect, J. (2010), *IAGO for global optimisation with noisy evaluations,* Workshop on Noisy Kriging-based Optimization, (NKO Workshop), Bern, 22-24 nov. 2010. Slides available at `http://www.imsv.unibe.ch/content/continuingeducation/nko_workshop/program/index_ger.html`.

[4] Bellman, R.E. (1957) *Dynamic Programming.* Princeton University Press, Princeton, NJ.

[5] Blum, J.R. (1954) *Multidimensional stochastic approximation methods.* Annals of Mathematical Statistics, 25, 737-744.

[6] Branke, J., Meisel, S., Schmidt, C. (2008) *Simulated annealing in the presence of noise.* Journal of Heuristics, vol. 14, n° 6, pp.627-654.

[7] Broadie, M., Cicek, D.M., Zeevi, A. (2009) *An adaptative multidimensional version of the kiefer-Wolfowitz stochastic approximation algorithm,* Proceeding of the 2009 Winter Simulation Conference. M.D. Rossetti, R.R. Hill, B. Johansson, A. Dunkin and R.G. Ingalls, eds.

[8] Bühlmann, H., Gisler, A. (2005), *A course in credibility theory and its applications,* Springer.

[9] Bulger, D.W., Romeijn, H.E. (2005) *Optimising noisy objective functions,* Journal of Global Optimization, 31 : 599-600.

[10] Dhaene, J., Vanduffel, S., Kaas, R., Tang, Q., Vyncke (2006), *Risk measures and comonotonicity : a review.* Stochastic models, 22 :573–606.

[11] Emmerich, M.T.M. (2005) *Single and Multi-objective evolutionary design optimization assisted by Gaussian Random Field Metamodels.* Dissertation zur Erlangung des Grades eines Doktors der Naturwissenschaften der Universität Dortmund, Dortmund.

[12] Ginsbourger, D. (2009) *Multiples métamodèles pour l'approximation et l'optimisation de fonctions numériques multivariables,* Thèse de doctorat de mathématiques appliquées, Ecole nationale supérieure des mines de Saint-Etienne, n° 519MA.

[13] Hampel, F.R., Ronchetti, E.M., Rousseeuw, P.J., Stahel, W.A. (1986), *Robust Statistics. The Approach Based on Influence Functions,* Wiley Series in Probability and Statistics, Wiley, New York.

[14] Hansen, E.R. (1979) *Global optimization using interval analysis : the one dimensional case,* JOTA 29 :331-344.

[15] Horst, R., Parlados, P.M. (1995) *Handbook of Global Optimization,* Kluwer Academic Publishers, Dordrecht Boston London.

[16] Huber, P.J. (1981), *Robust Statistics,* Wiley Series in Probability and Statistics, Wiley, New York.

[17] Jones, D.R., Pertunen, C.D., Stuckman (1993) *Lipschitzian optimization without the Lipschitz constant.* Journal of Optimization Theory and Applications, 79(1), 157-181.

[18] Jones, D.R., Schonlau, M., Welch, W.J. (1998) *Efficient global optimization of expensive black-box functions,* Journal of Global Optimization, 13, 455-492.

[19] Jones, D.R. (2001) *A taxonomy of global optimization methods based on response surface (2001).* Journal of Global Optimization, 21 :345-383.

[20] Kiefer, J., Wolfowitz, J. (1952) *Stochastic estimation of the maximum of a regression function.* Annals of Mathematical Statistics, 23, 462-466.

[21] Kleijnen, J.P.C. (2009) *Kriging metamodeling in simulation : A review,* European Journal of Operational Research, 192, 707–716.

[22] Jack P.C. Kleijnen, J.P.C., van Beers, W., van Nieuwenhuyse, I. (2010) *Expected improvement in efficient global optimization through bootstrapped kriging .*

[23] Lawler, E.L., Wood, D.E. (1966) *Branch and Bound methods : a survey,* Operations Research, Vol. 14, n° 4, pp 699-719.

[24] Mathias, K., Whitley, D., Kusuma, A., Stork, C. (1996) An empirical evaluation of genetic algorithms on noisy objective functions.

[25] Nelder, J., Mead, R. (1965) *A simplex method for function minimization,* Computer Journal, vol. 7, n° 4, p.308-313.

[26] Norkin, V., Pflug, G.Ch., Ruszczyński, A. (1996) *A branch and bound method for stochastic global optimization,* Mathematical Programming, vol 83, n° 1-3, pp 452-450.

[27] Picheny, V., Ginsourger, D., Richet, Y. (2010), *Optimization of Noisy Computer Experiments with Tunable Precision,* Workshop on Noisy Kriging-based Optimization, (NKO Workshop), Bern, 22-24 nov. 2010. Slides available at `http://www.imsv.unibe.ch/content/continuingeducation/nko_workshop/program/index_ger.html`.

[28] Ribereau, P., Rullière, D., *Krigeage à effet pépite pour simulations stochastiques.* Preprint available online at `http://hal.archives-ouvertes.fr/`

[29] Robbins, H., Monro, S. (1951) *A Stochastic approximation method.* Annals of Mathematical Statistics, 22, 400-407.

[30] Rullière, D., Faleh, A., Planchet, F. (2011) *Un algorithme d'optimisation par exploration sélective.* Preprint available at `http://hal.archives-ouvertes.fr/hal-00411406` .

[31] Schubert, B. (1972) *A sequential method seeking the global maximum of a function.* SIAM J. Numer. Anal., 9 :379-388.

[32] Smith, N.A., Tromble, R.W. (2004) *Sampling uniformly from the unit simplex,* Technical Report, Johns Hopkins University.

[33] Strugarek, C. (2006) *Approches variationnelles et autres contributions en optimisation stochastique.* ENPC, Thèse de doctorat.

[34] Van Beers, W., Kleijnen, J.P.C. (2003) *Kriging for interpolation in random simulations,* Journal of the Operational Research Society (54), 255–262.

[35] Villemonteix, J. (2009) *Optimisation de fonctions coûteuses,* Thèse de doctorat de physique, Université Paris Sud 11, Faculté des sciences d'Orsay, n° 9278.

Table des matières

version de ce document en date du 10 mars 2011, 15 :03.