Méthodes non linéaires pour des problèmes statistiques inverses

Pierre Barbillon, Gilles Celeux, Agnès Grimaud, Yannick Lefebvre, Etienne De Rocquigny

Méthodes non linéaires pour des problèmes statistiques inverses

Pierre Barbillon[1] & Gilles Celeux[1] & Agnès Grimaud[2] & Yannick Lefebvre[3] & Etienne De Rocquigny[4]

[1] *Université Paris-sud 11, Laboratoire de Mathématiques d'Orsay, Orsay Cedex, F-91405 INRIA Saclay, projet* SELECT

[2] *Institut de Mathématiques de Luminy Case 907 163 Avenue de Luminy 13288 Marseille cedex 9*

[3] *Schlumberger, 1 Cour du Triangle, 92936 La Defense Cedex*

[4] *Ecole Centrale Paris, Grande Voie des Vignes, 92290 Châtenay Malabry*

Résumé

Dans le cadre du traitement des incertitudes étudié ici, la variabilité intrinsèque des entrées d'un modèle physique est modélisée par une loi de probabilité multivariée. L'objectif est d'identifier cette loi de probabilité à partir d'observations des sorties du modèle. Afin de se limiter à un nombre d'appels raisonnable au code de calcul (souvent coûteux) du modèle physique dans l'algorithme d'inversion, une méthodologie d'approximation non linéaire faisant intervenir le krigeage et un algorithme EM stochastique est présentée. Elle est comparée à une méthode utilisant une approximation linéaire itérative sur la base de jeux de données simulées provenant d'un modèle de crues simplifié mais réaliste. Les cas où cette approche non linéaire est préférable seront mis en lumière.

Mots clés : Modélisation des incertitudes, Approximation non linéaire, Krigeage, Algorithme stochastique.

Abstract

In the uncertainty treatment framework considered in this paper, the intrinsic variability of the inputs of a physical simulation model is modelled by a multivariate probability distribution. The objective is to identify this probability distribution - the dispersion of which is independent of the sample size since intrinsic variability is at stake - based on observation of some model outputs. Moreover, in order to limit to a reasonable level the number of (usually burdensome) physical model runs inside the inversion algorithm, a non linear approximation methodology making use of Kriging and stochastic EM algorithm is presented. It is compared with iterated linear approximation on the basis of numerical experiments on simulated data sets coming from a simplified but realistic modelling of a

dyke overflow. Situations where this non linear approach is to be preferred to linearisation are highlighted.

Key words : Uncertainty Modelling, Non linear Approximation, Kriging, Stochastic Algorithm.

Résumé long

Le modèle considéré s'écrit :

$$Y_i = H(X_i, d_i) + U_i, \quad 1 \leq i \leq n \tag{1}$$

où

- $(Y_i) \in \mathbb{R}^p,\ 1 \leq i \leq n$, sont des vecteurs des données observées,
- H est une fonction connue de $\mathbb{R}^{(q+q_2)}$ dans $\mathbb{R}^p$. Elle est de type "boîte-noire", c'est-à-dire que nous ne disposons pas d'une formulation explicite de H. On peut l'évaluer en tout point par un code de calcul numérique exact mais cette évaluation est coûteuse. H doit être supposée injective afin de rendre le modèle (1) identifiable.
- $(X_i) \in \mathbb{R}^q,\ 1 \leq i \leq n$, désignent des vecteurs aléatoires non observés, supposés indépendants et identiquement distribués (i.i.d.) de loi normale $\mathcal{N}(\mu, C)$.
- $(d_i),\ 1 \leq i \leq n$, sont des variables observées liées aux conditions expérimentales, de dimension q_2.
- $(U_i),\ 1 \leq i \leq n$, sont des variables aléatoires qui représentent les erreurs de mesure. Elles sont supposées i.i.d. de distribution $\mathcal{N}(0, R)$, R ne doit pas être nécessairement supposé connu. Les variables aléatoires (X_i) et (U_i) sont supposées être indépendantes.

Notre but est d'estimer les paramètres (μ, C, R) à partir des observations $(Y_i, d_i), i = 1, \ldots, n$. Les $(X_i)_{1 \leq i \leq n}$ n'étant pas observés, cette structure de données manquantes mène à utiliser en général des algorithmes faisant appel un grand nombre de fois au modèle physique H. Mais comme nous l'avons précisé, H demande un temps de calcul important. Pour traiter ce problème, Celeux *et al.* (2009) ont proposé de linéariser la fonction H autour d'un point x_0 et d'employer un algorithme ECME pour estimer les paramètres (μ, C), R doit être fixé. Du fait de la linéarisation du modèle, la mise à jour des paramètres courants dans l'algorithme ECME est explicite. Ils proposent d'itérer la procédure de linéarisation afin d'améliorer la précision de l'algorithme et d'éviter une trop grande dépendance vis à vis du point x_0 choisi.
Si H peut être localement approchée de manière linéaire, cette méthode donne des

résultats satisfaisants. Par contre si H présente un comportement hautement non linéaire, il est risqué d'utiliser cet algorithme. C'est pourquoi dans (Barbillon *et al.*, 2009), nous proposons une méthode d'approximation de H par krigeage que nous intégrons dans un algorithme EM stochastique (SEM).
Nous présentons la $(k+1)^{\text{ème}}$ itération de l'algorithme SEM qui comporte trois étapes :

- Étape E : calcul de la densité conditionnelle $p(.|Y;\theta^{(k)})$ de $X^{(k)}$, où $\theta^{(k)}$ est la valeur courante du paramètre θ.
- Étape S (Stochastique) : restauration des données manquantes; des données complétes $Z^{(k)} = (Y, X^{(k)})$ sont générées en simulant $X^{(k)}$ suivant la loi conditionnelle $p(.|Y;\theta^{(k)})$.
- Étape M : L'estimateur mis à jour $\theta^{(k+1)}$ est l'estimateur du maximum de vraisemblance calculé sur les $Z^{(k)}$.

Cet algorithme génère une chaîne de Markov irréductible dont la loi stationnaire est concentrée autour des valeurs de l'estimateur du maximum de vraisemblance de θ (Nielssen, 2000). Une période de chauffe de longueur ℓ est nécessaire afin d'atteindre le régime stationnaire de la chaîne, ensuite la moyenne $\frac{1}{L-\ell}\sum_{k=\ell+1}^{L}\theta^{(k)}$ est calculée avec L assez grand pour obtenir un estimateur de θ. Pour le modèle (1), l'étape de simulation de SEM est problèmatique puisque la loi conditionnelle $(X|Y,\theta)$ n'est pas explicite. Un algorithme type MCMC (Markov Chain Monte Carlo) est nécessaire pour réaliser l'étape S. Pour la restauration de chacun des X_i $(1 \leq i \leq n)$, m itérations d'un algorithme de Metropolis-Hasting sont utilisées. Ainsi, à l'itération k, l'étape S s'écrit : Pour $1 \leq i \leq n$,

- Soit $X_{i,0} = X_i^{(k-1)}$.
- Pour $s = 1, ..., m$
  1. Simuler $\tilde{X}_{i,s}$ suivant la loi de proposition $q_{\theta_k}(X_{i,s-1}, .)$.
  2. $X_{i,s} = \tilde{X}_{i,s}$ avec probabilité

$$\alpha(X_{i,s-1}, \tilde{X}_{i,s}) = \min\left(1, \frac{p(\tilde{X}_{i,s}|Y_i;\theta^{(k)})q_{\theta_k}(\tilde{X}_{i,s}, X_{i,s-1})}{p(X_{i,s-1}|Y_i;\theta^{(k)})q_{\theta_k}(X_{i,s-1}, \tilde{X}_{i,s})}\right)$$

  et $X_{i,s} = X_{i,s-1}$ avec probabilité $1 - \alpha(X_{i,s-1}, \tilde{X}_{i,s})$.
- $X_i^{(k)} = X_{i,m}$.

Le calcul du taux d'acceptation $\alpha(.,.)$ dans une itération d'un algorithme de Metropolis-Hasting demande un appel à H. Si on suppose que m itérations permettent d'atteindre la distribution stationnaire de la chaîne, l'étape S dans une itération de l'algorithme SEM engendrerait alors $n \times m$ appels à H. Au total H serait calculée $L \times n \times m$ fois, ce qui

n'est pas envisageable puisque H est une fonction coûteuse. C'est pourquoi nous proposons comme alternative de remplacer H par une approximation non linéaire obtenue par krigeage (Cressie, 1990). Le krigeage permet à partir d'un nombre fixé d'évaluations de H sur un plan d'expérience numérique $D = \{x_1, \ldots, x_{N_{\max}}\}$ de fournir une approximation de H, notée $\hat{H}$ sur un domaine E choisi préalablement. Ainsi nous n'effectuerons en tout et pour tout que N_{max} appels au code de calcul de H. Le choix de E le domaine d'approximation est sensible puisqu'il doit être assez grand pour contenir avec grande probabilité la variable aléatoire X et pas trop étalé dans la mesure où la qualité de l'approximation dépend de la concentration des points de D dans E. Après s'être assuré par validation croisée de la qualité de l'approximation, nous remplaçons dans les calculs des taux d'acceptation tout appel à H coûteux par un appel à $\hat{H}$ d'évaluation quasi instantanée.

Nous comparons ensuite les méthodes ECME avec linéarisations itérées et SEM avec approximation par krigeage sur un modèle hydrodynamique simplifié mais réaliste. Ce modèle est lié au risque de dépassement d'une digue durant une crue. Les deux variables observées (le vecteur Y) sont la hauteur de la rivière au niveau de la digue et sa vitesse. La condition observée (d) est le débit de la rivière en amont et on cherche à estimer les paramètres des lois des variables non observées (vecteur X), le coefficient de friction du lit de la rivière et la côte de fond de la rivière au niveau de la digue. Les deux méthodes fournissent des estimations satisfaisantes. La méthode employant un algorithme ECME avec des linéarisations itérées est performante ici puisque la fonction H correspondant au modèle peut être assez finement approchée localement par une fonction linéaire.
Nous proposons aussi un exemple qui met en défaut cette méthode. Sur cet exemple, seule l'approche non linéaire utilisant l'algorithme SEM et une approximation par krigeage donne des estimateurs raisonnables car elle est plus flexible.

Bibliographie

[1] Barbillon, P., Celeux, G., Grimaud, A., Lefebvre, Y. and De Rocquigny, E. (2009). Non linear methods for inverse statistical problems. Rapport de recherche INRIA.

[2] Celeux, G., Grimaud, A., Lefebvre, Y. and De Rocquigny, E. (2009). Identifying variability in multivariate systems through linearised inverse methods. *Inverse Problems In Science & Engineering*, to appear.

[3] Cressie, N. (1990). The origins of kriging. *Mathematical Geology*, **22(2)**: 239-252.

[4] Koehler, J.R. and Owen, A.B. (1996). *Computer experiments*. Handbook of Statistics **Vol 13**, pp. 261-308, Elsevier Science, New York.

[5] Nielsen, S. F. (2000). The stochastic EM algorithm: estimation and asymptotic results *Bernoulli* **6**, 457-489.