Non linear methods for inverse statistical problems

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

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Non linear methods for inverse statistical problems

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

N° 7156

Decembre 2009

Thème COG

Rapport de recherche

ISSN 0249-6399 ISRN INRIA/RR--7156--FR+ENG

INSTITUT NATIONAL
DE RECHERCHE
EN INFORMATIQUE
ET EN AUTOMATIQUE

INRIA

centre de recherche
SACLAY - ÎLE-DE-FRANCE

Non linear methods for inverse statistical problems

Pierre Barbillon*, Gilles Celeux*, Agnès Grimaud† , Yannick Lefebvre‡ , Etienne De Rocquigny§

Thème COG — Systèmes cognitifs
Équipes-Projets SELECT

Rapport de recherche n° 7156 — Decembre 2009 — 20 pages

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

* Université Paris-sud 11, INRIA Saclay
† Institut Mathématique de Luminy
‡ Schlumberger
§ École Centrale Paris

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université
4, rue Jacques Monod, 91893 ORSAY Cedex
Téléphone : +33 1 72 92 59 00

Méthodes non linéaires pour des problèmes statistiques inverses

Résumé : Dans le cadre du traitement des incertitudes étudié dans cet article, 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 (sa dispersion est indépendante de la taille de l'échantillon puisque l'on traite de la variabilité intrinsèque) 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

1 Introduction

Probabilistic uncertainty treatment is gaining fast growing interest in the industrial field, as reviewed by Rocquigny (de) *et al.* (2008). In the energy sector, such uncertainty analyses are for instance carried out in environmental studies (flood protection, effluent control, etc.), or in nuclear safety studies involving large scientific computing (thermo-hydraulics, mechanics, neutronics etc.). Besides the uncertainty propagation challenges when dealing with complex and high CPU-time demanding physical models, one of the key issues regards the quantification of the sources of uncertainties. The problem is to choose reliable statistical models for the input variables such as uncertain physical properties of the materials or industrial process or natural random phenomena (wind, flood, temperature, etc.).

A key difficulty, traditionally encountered at this stage, is linked to the highly-limited sampling information directly available on uncertain input variables. An industrial case-study can largely benefit from integrate indirect information such as data on other more easily observable parameters linked to the uncertain variable of interest by a physical model. It demands methods using of probabilistic inverse methods since the recovering of indirect information involves generally the inversion of a physical model. Roughly speaking, this inversion transforms the information into a virtual sample of the variable of interest, before applying to it standard statistical estimation. Yet, it is mandatory to limit to a reasonable level the number of (usually large CPU-time consuming) physical model runs inside the inverse algorithms.

As in Celeux *et al.* (2009), this paper concentrates on the situation where there is an *irreducible* uncertainty or variability in the input parameters of a physical model. Observations are modelled with a vector of physical variables y that are connected to uncertain inputs x through a deterministic (and supposedly well-known) physical model $y = H(x, d)$. As a clear difference to classical parameter identification x is not supposed to have a fixed, albeit unknown physical value: It will be modelled with a random variable taking different realisations for each observation. The purpose is thus to estimate its probability distribution function instead of its point value. On the other hand, d stands for fixed inputs. A key difficulty is that the time needed to compute the physical function H is huge since H is often the result of a complex code. Thus, it is desirable or necessary to limit the number of calls to the H function. For this very reason, Celeux *et al.* (2009) investigated efficient estimation algorithms based on a linearisation of the model around a fixed value x_0 to estimate the parameters distributions in this context. But, the linearisation method has some drawbacks associated to the approximation error induced, and the potential difficulty in choosing an adequate linearisation point before identification. In this paper, we propose an alternative solution avoiding the linearisation of H by using a non linear approximation of the function H obtained through Kriging. The paper is organised as follows. In Section 2, the model is stated and the linear procedure of Celeux *et al.* (2009) is summarised. In Section 3, a stochastic procedure using a non linear approximation of H is presented. Section 4 is devoted to the presentation of numerical experiments for comparing the two approaches. A short discussion section ends the paper.

2 The model and its linear identification

The considered model takes the form

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

with the following features

- (Y_i) in $\mathbb{R}^p$ denotes the vector data,
- H denotes a known function from $\mathbb{R}^{(q+q_2)}$ to $\mathbb{R}^p$. The function H can be typically regarded as a "black box" and getting the output $H(x, d)$ from any input (x, d) is quite expensive. To ensure the identifiability of model (1), H is assumed to be injective.
- (X_i) in $\mathbb{R}^q$ denotes non observed random data, assumed to be independent and identically distributed (i.i.d.) with a Gaussian distribution $\mathcal{N}(\mu, C)$.
- (d_i) denotes observed variables related to the experimental conditions, with dimension q_2.
- (U_i) denotes measurement-model errors, assumed i.i.d. with distribution $\mathcal{N}(0, R)$, R being known or unknown. Variables (X_i) and (U_i) are assumed to be independent.

The aim is to estimate the parameters (μ, C, R) from the data $(Y_i, d_i), i = 1, \ldots, n$. Since the (X_i) are not observed, a good estimation of this missing structure model would require to compute the function H a lot of times. But, as written above, computing values of H is quite expensive.

A linearised method In order to limit to a reasonable amount of computation the number of calls to the function H to estimate the model parameters, a linear approximation of the model defined in (1) has been investigated in Celeux *et al.* (2009). In this approach the function H is linearised around a fixed value x_0 (chosen from expert informations). The approximated model is

$$Y_i = H(x_0, d_i) + J_H(x_0, d_i)(X_i - x_0) + U_i, \quad 1 \leq i \leq n \tag{2}$$

where $J_H(x_0, d_i)$ is the Jacobian matrix of the function H in x_0, with dimension $p \times q$.

In the following, for simplicity, the variance matrix R is assumed to be known to sketch the approach of Celeux *et al.* (2009). First the linear model (2) is supposed to be identifiable. It is ensured if and only if $\text{rank}(\mathbf{J_H}) = q$ with $\mathbf{J_H} = (J_H(x_0, d_1), ..., J_H(x_0, d_n))^T$.

The data (X_i) being non observed, the estimation problem is a missing data structure problem that can be solved with an EM-type algorithm (Dempster *et al.* 1977). The EM algorithm alternates two steps at iteration $(k + 1)$:

- E step (Expectation): It consists of computing $Q(\theta, \theta^{(k)}) = E[L(\theta, Z)|Y, \theta^{(k)}]$ where L is the completed loglikelihood.
- M step (Maximisation): $\theta^{(k+1)} = \arg\max_{\theta \in \Theta} Q(\theta, \theta^{(k)})$.

In the present context updating formulas for $\mu^{(k)}$ and $C^{(k)}$ in the M step are closed form.

A variant devoted to accelerate the EM algorithm, which is known to often encounter slow convergence situations, is the ECME (Expectation-Conditional Maximisation Either) algorithm of Liu and Rubin (1994). The M-step is replaced by CME-steps (Conditional Maximisation Either), maximising conditionally to some parameters, the Q-function or the actual observed loglikelihood, $\ln(L(\theta))$.

To compute $\theta^{(k+1)} = (\mu^{(k+1)}, C^{(k+1)})$ for model (2), the iteration $(k+1)$ of ECME is as follow: the E-step is the same as in EM and the M-step is replaced with two steps. The first CME step, to update the variance matrix C, is similar to the M step of EM with μ fixed to $\mu^{(k)}$. The second CME step, to update the parameter μ, maximises the incomplete-data loglikelihood over μ, assuming $C = C^{(k+1)}$ (see also De Crecy 1996). Introducing the notation: $h_i = H(x_0, d_i)$; $J_i = J_H(x_0, d_i)$,

$$A_i^{(k)} = Y_i - h_i - J_i(\mu^{(k)} - x_0), B_i^{(k)} = C^{(k)} J_i^T \text{ and } V_i^{(k)} = J_i C^{(k)} J_i^T + R,$$

the ECME updating equations for model (2) are

$$C^{(k+1)} = C^{(k)} + \frac{1}{n} \sum_{i=1}^{n} \left[(B_i^{(k)} (V_i^{(k)})^{-1} A_i^{(k)})(B_i^{(k)} (V_i^{(k)})^{-1} A_i^{(k)})^T - B_i^{(k)} (V_i^{(k)})^{-1} (B_i^{(k)})^T \right],$$

$$\mu^{(k+1)} - x_0 = \left(\sum_{i=1}^{n} J_i^T (V_i^{(k+1)})^{-1} J_i \right)^{-1} \left(\sum_{i=1}^{n} J_i^T (V_i^{(k+1)})^{-1} (Y_i - h_i) \right).$$

The EM and ECME algorithms have shown to work well in practice (Celeux *et al.*, (2009)). But the linearisation approach could be sensitive to the linearisation point x_0. To reduce its influence, a simple solution is to use an iterative linearisation of the physical model H, as now described:

- Initial Step: Starting from an initial linearisation point: $x_{\text{lin}} = x_0$; $(H(x_0, d_i))_i$ and $(J_H(x_0, d_i))_i$ are computed. Then the ECME algorithm, initiated at $\theta_{init} = (x_0, C_0)$, is run leading to the estimate $\hat{\theta}^{(1)}$.
- Step $l+1$: Let $x_{\text{lin}} = \hat{\mu}^{(l)}$. Then $(H(x_{\text{lin}}, d_i))$ and $(J_H(x_{\text{lin}}, d_i))$ are computed and the ECME algorithm initiated with $\theta_{init} = \hat{\theta}^{(l)}$, leads to the estimate $\hat{\theta}^{(l+1)}$.

This algorithm is run until some stopping criterion, as $\max_j \left(\frac{|\theta_j^{(l+1)} - \theta_j^{(l)}|}{|\theta_j^{(l)}|} \right) \leq \varepsilon$ with some fixed ε, is satisfied.

Remark: In the general case where d_i experimental conditions vary throughout the sample, changing the linearisation point requires n calls of H for $H(x_{\text{lin}}, d_i)$ plus $n \times q \times a$ calls of H for $J_H(x_{\text{lin}}, d_i)$ through finite differences where $a = 1$ to say $a = 5$ according to the roughness of H. This iterate linearisation is expected to perform well when the function H is not highly non linear. Otherwise alternative non linear approximations of H could be required.

3 Using a non linear approximation of the function H

In some cases, linear approximation of the function H could be unsatisfactory. But in such cases, the E and M steps in EM and ECME algorithms are difficult to implement. For instance, the conditional expectation function Q is not closed form. A possible answer is to use a stochastic version of the EM algorithm such as the SEM algorithm (Celeux and Diebolt, 1985, 1987) or the SAEM algorithm (Delyon *et al.*, 1999). However these algorithms which require to simulate the missing x_i according to their current conditional distribution at each iteration, need to call H some thousand times which is far too CPU time consuming. In practice, to save CPU running time, the number of calls to the function H is constrained to be smaller than a maximum value $N_{\max}$. Therefore, we propose a method coupling the SEM algorithm with a non linear approximation of H. Its principle is as follows: A set of points $D = \{(x_1, d_1), ..., (x_{N_{\max}}, d_{N_{\max}})\}$ with size $N_{\max}$ is chosen. Then H is computed at each point of D and will be not called again in the algorithm. Whenever H has to be computed at a point (x, d), the true value $H(x, d)$ is replaced by an approximation $\hat{H}(x, d)$, obtained with a barycentric interpolation or Kriging.

The considered model is the model (1):

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

In this section, the variance matrix R of the measurement model error can be assumed known or not. The aim is to estimate the parameter $\theta = (\mu, C, (R))$.

3.1 The SEM algorithm

The Stochastic EM (SEM) algorithm incorporates a simulation step between the E and M steps. Its $(k+1)$th iteration involves three steps:

- E step: Computation of the conditional density $p(.|Y; \theta^{(k)})$ of $X^{(k)}$, $\theta^{(k)}$ being the current fit of parameter θ.
- S step (Stochastic): It is a Restoration step: a completed sample $Z^{(k)} = (Y, X^{(k)})$ is generated by drawing $X^{(k)}$ from the conditional density $p(.|Y; \theta^{(k)})$.
- M step: The updated estimate $\theta^{(k+1)}$ is the maximum likelihood estimate computed on the basis of $Z^{(k)}$.

This SEM algorithm generates an irreducible Markov chain whose stationary distribution is concentrated around maximum likelihood estimate of θ (see Nielssen, 2000). To derive pointwise estimates from SEM, a warm-up step of length ℓ is required to reach the stationary regime of the generated Markov chain, then mean $\sum_{k=\ell+1}^{L} \theta^{(k)}$ is computed with L large enough to get an estimate of θ.

The SEM algorithm is now described for the model (1). The first task is to calculate the completed loglikelihood $L(\theta, Z) = \ln p(Y, X; \theta)$: We have $p(Y_i, X_i; \theta) = p(Y_i|X_i, \theta)p(X_i; \theta)$ with

$$p(Y_i|X_i, \theta) = (2\pi)^{-\frac{nd}{2}} |R|^{-\frac{n}{2}} \exp\left(-\frac{1}{2}\sum_{i=1}^{n}(Y_i - H(X_i, d_i))^T R^{-1}(Y_i - H(X_i, d_i))\right)$$

and

$$p(X_i;\theta) \quad = \quad (2\pi)^{-\frac{nq}{2}}|C|^{-\frac{n}{2}}\exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T C^{-1}(X_i-\mu)\right).$$

Thus

$$\ln p(Y,X;\theta) = -\frac{n}{2}\ln(|R|) - \frac{1}{2}\sum_{i=1}^{n}(Y_i - H(X_i,d_i))^T R^{-1}(Y_i - H(X_i,d_i))$$

$$-\frac{n}{2}\ln(|C|) - \frac{1}{2}\sum_{i=1}^{n}(X_i-\mu)^T C^{-1}(X_i-\mu) + Cste.$$

And, $\theta^{(k+1)}$ is obtained by solving the likelihood equations

$$\frac{\partial}{\partial R}\ln p(Y,X^{(k)};\theta) = \frac{\partial}{\partial \mu}\ln p(Y,X^{(k)};\theta) = \frac{\partial}{\partial C}\ln p(Y,X^{(k)};\theta) = 0.$$

This leads to the closed form formulas

$$R^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - H(X_i^{(k)},d_i))(Y_i - H(X_i^{(k)},d_i))^T$$

$$\mu^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}X_i^{(k)}$$

$$C^{(k+1)} = \frac{1}{n}\sum_{i=1}^{n}(X_i^{(k)} - \mu^{(k+1)})(X_i^{(k)} - \mu^{(k+1)})^T.$$

For model (1), the simulation step of SEM induces a difficulty since the conditional distribution of $(X|Y,\theta)$ is not directly available. A MCMC (Markov Chain Monte Carlo) algorithm is needed to perform the S step. At iteration k, the S step consists of m iterations of a Hastings-Metropolis algorithm. For $i = 1, ..., n$

- Let $X_{i,0} = X_i^{(k-1)}$.
- For $s = 1, ..., m$
  1. Generate $\tilde{X}_{i,s}$ from the proposal distribution $q_{\theta_k}(X_{i,s-1}, .)$.
  2. $X_{i,s} = \tilde{X}_{i,s}$ with probability

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

  and $X_{i,s} = X_{i,s-1}$ with probability $1 - \alpha(X_{i,s-1},\tilde{X}_{i,s})$.
- $X_i^{(k)} = X_{i,m}$.

Several proposal distributions taking into account assumptions made on the (X_i) distribution may be used. Here three proposals are alternately considered (see Kuhn, Lavielle 2004):

1. q_{θ_k} is the "prior" distribution of X_i at iteration k, that is the Gaussian distribution $\mathcal{N}(\mu_k, C_k)$. Then
$$\alpha(X_{i,s-1}, \tilde{X}_{i,s}) = \min\left(1, \frac{p(Y_i|\tilde{X}_{i,s}; \theta^{(k)})}{p(Y_i|X_{i,s-1}; \theta^{(k)})}\right).$$
2. q_{θ_k} is the multidimensional random walk with dimension q: $\mathcal{N}(X_{i,s-1}, \kappa C_k)$. Then
$$\alpha(X_{i,s-1}, \tilde{X}_{i,s}) = \min\left(1, \frac{p(Y_i, \tilde{X}_{i,s}; \theta^{(k)})}{p(Y_i, X_{i,s-1}; \theta^{(k)})}\right).$$
3. q_{θ_k} is the succession of q unidimensional Gaussian random walks $\mathcal{N}(X_{i,s-1}(l), \kappa C_k(l,l))$: each component of X is successively updated.

At iteration k, the S step consists of running m_1 iterations with proposal 1, m_2 iterations with proposal 2 and m_3 iterations with proposal 3, with $m_1 + m_2 + m_3 = m$. In proposal 2 and 3, κ has to be chosen between 0 and 1. κ is tuned so that the first iterations of the S step have acceptance rates between 0.3 and 0.6 to ensure that the Hasting-Metropolis chain well explores the possible values of X_i. In the following simulations, $\kappa = 0.1$ suits and $m_1 = 100$, $m_2 = 0$ and $m_3 = 100$ are set.

To compute the acceptance probabilities α, the function H is called m times, for each i. Hence for each iteration of the SEM algorithm, H is to be computed nm times. But recall that the number of calls to H is limited to at most $N_{\max}$. It means that in most situations the above described SEM algorithm is infeasible. To cope with this difficulty, we propose to first compute H on a set of $N_{\max}$ points. Then, H is replaced by an approximation $\hat{H}$ built from the $N_{\max}$ evaluations, in the SEM algorithm.

3.2 Kriging approximation of H

In this section, $H(z)$ could denote $H_i(z) = H(z, d_i)$ where $z \in \mathbb{R}^q$ as well as $H(z) = H(z_1, z_2)$ where $(z_1, z_2) \in \mathbb{R}^q \times \mathbb{R}^{q_2}$. That is to say that an approximation is made for each d_i (thus for each H_i) or a single approximation of H is made. This point is further discussed in the Remark (iii) of Section 3.2.3. It is considered that $z \in \mathbb{R}^Q$ where $Q = q$ or $Q = q + q_2$.

The approximation $\hat{H}$ could be a barycentric approximation derived from $N_{\max}$ exact values $H(z_1), \ldots, H(z_{N_{\max}})$ of H. The approximation is, for $z \notin D = \{z_1, \ldots, z_{N_{\max}}\}$
$$\hat{H}(z) = \sum_{j \in V_k(z)} \frac{\|z_j - z\|^{-1}}{\sum_j \|z_j - z\|^{-1}} H(z_j)$$
where $V_k(z)$ is the subset of the k nearest neighbours of z in D, for a fixed k. Preliminary numerical experiments (not reported here) show that this simple barycentric method could be not efficient enough and that Kriging that is now described is to be preferred.

Kriging is a method devoted to approximate a function $H : \Omega \mapsto \mathbb{R}$ where the input set $\Omega \subset \mathbb{R}^Q$ is a bounded hypercube. Our approximation will be warranted only on Ω. With no loss of generality, it is assumed that $\Omega = [0,\ 1]^Q$ for the clarity of exposition.

3.2.1 Choosing a design

The first concern is to select the set of points D where the function H is computed. This set will be called the design and has to be chosen carefully since the number of calls to H is limited to $N_{\max}$. In order to get an exploratory design, a Latin Hypercube Sampling (LHS)-*maximin* strategy is used. A design $D = \{z_1, \dots z_N\} \subset \Omega \subset [0,\ 1]^Q$ is a LHS if it is constructed as follows

$$z_i^j = \frac{\pi_j(i) - U_j^i}{N} \quad \forall\ 1 \leq i \leq N,\ \forall\ 1 \leq j \leq Q \tag{3}$$

where π_j are independent uniform random permutations of the integers 1 through N, and the U_j^i are independent $\mathcal{U}_{[0,1]}$ random variables independent of the π_js. A LHS guarantees good projection properties. The sample points are stratified on each of Q input axes and the projection of the design on any axe is well scattered. Therefore, it takes into account of the variability for all dimensions.

Then, in order to have good exploratory properties which means that the points are spread in the input set, the design D is said to be chosen to be *maximin*. A design D is *maximin* if the distance between the sites is maximum: D has to maximise

$$\delta_D = \min_{z_i, z_j \in D} \|z_i - z_j\|. \tag{4}$$

Morris and Mitchell (1995) provide an algorithm based on simulated annealing which can mix the two properties related to (3) and (4) to provide a LHS-*maximin* design.

3.2.2 Kriging predictor

It is assumed that $D = \{z_1, \dots z_N\}$ is a LHS-*maximin* design. The function H can be seen as the realisation of a Gaussian process Y

$$Y(z) = \sum_{i=1}^{p} \beta_i f_i(z) + G(z) = F(z)^T \boldsymbol{\beta} + G(z).$$

In this setting, the f_i are known regression functions, the β_i are unknown parameters to be estimated and G is a centered Gaussian process characterised by its covariance function $\mathrm{cov}(G(s), G(t)) = \sigma^2 K_\theta(s,t)$ where K_θ is a symmetric positive define kernel such that for all s, $K_\theta(s,s) = K_\theta(0,0) = 1$. The choice of the parameter θ allows to tune the regularity of the process G. For instance in the case of a Gaussian kernel where $\theta \in \mathbb{R}_+$ and $K_\theta(r,s) = e^{-\theta\|r-s\|^2}$, the larger θ is, the smoother the process G is. Therefore, the distribution of $Y_D = \{Y(z_1), \dots Y(z_N)\}$ is

$$p(Y_D) = \mathcal{N}(F_D\boldsymbol{\beta}, \sigma^2 \Sigma_{DD})$$

where $F_D = (F(z_1), \dots F(z_N))^T$ and $(\Sigma_{DD})_{1 \leq i,j \leq N} = K_\theta(z_i, z_j) = \mathrm{corr}(Y(z_i), Y(z_j))$. The conditional process knowing the vector Y_D, is a Gaussian process. The distribution of $Y(z_0)$, given Y_D, is $\mathcal{N}(\mu_{z_0|D}, \sigma_{z_0 z_0|D})$, with

$$\begin{aligned} \mu_{z_0|D} &= E(Y(z_0)|Y_D) = F(z_0)^T\boldsymbol{\beta} + \Sigma_{z_0 D}^T \Sigma_{DD}^{-1}(Y_D - F_D\boldsymbol{\beta}) \\ \sigma_{z_0 z_0|D} &= \mathrm{var}(Y(z_0)|Y_D) = \sigma^2\left(1 - \Sigma_{z_0 D}^T \Sigma_{DD}^{-1} \Sigma_{z_0 D}\right) \end{aligned}$$

where $\Sigma_{z_0 D} = (K_\theta(z_1, z_0), \dots K_\theta(z_N, z_0))^T$. The conditional mean $\mu_{z_0|D}$ can be used as a predictor of $H(z_0)$. Furthermore, $\boldsymbol{\beta}$, θ and σ^2 are estimated by maximising the likelihood. It leads to

$$\begin{aligned} \hat{\boldsymbol{\beta}} &= (F_D^T \Sigma_{DD}^{-1} F_D)^{-1} F_D^T \Sigma_{DD}^{-1} Y_D \\ \hat{\sigma}^2 &= \frac{1}{N}(Y_D - F_D\hat{\boldsymbol{\beta}})^T \Sigma_{DD}^{-1}(Y_D - F_D\hat{\boldsymbol{\beta}}). \end{aligned}$$

Both of these estimators depend on θ via Σ_{DD}. The maximisation in θ is not explicit and is made by minimising

$$\psi(\theta) = |\Sigma_{DD}|^{\frac{1}{N}} \sigma^2(\theta).$$

The Matlab toolbox DACE (Lophaven *et al.*, 2002) is used to compute all this parameters and to solve the optimisation problem in θ.

As a result, for all $z_0 \in \Omega$, the Kriging predictor of H is

$$\hat{H}(z_0) = F(z_0)^T \hat{\boldsymbol{\beta}} + \hat{\Sigma}_{z_0 D}^T \hat{\Sigma}_{DD}^{-1}(Y_D - F_D\hat{\boldsymbol{\beta}}),$$

where $\hat{\Sigma}$ stands for $\Sigma(\hat{\theta}, \hat{\sigma}^2)$. Moreover, this predictor is exact for any $z_0 = z_i$, and it is the best linear unbiased predictor of $Y(z_0)$ for all $z_0 \in \Omega$. A fully Bayesian method as described in Koehler and Owen (1996) is possible. In this framework, a Gaussian prior distribution is set on the parameters $(\beta_i)_{1 \le i \le p}$. If the prior distribution is diffuse enough, the posterior mean of $Y(z_0)$ (hence the predictor) tends to be the same than the maximum likelihood conditional mean of the Gaussian process.

3.2.3 Practical figures

(i) The choice of the input set Ω is sensitive. Ω has to be large enough to contain with a high probability the values of the random variable X and not too large in order to be efficient since the quality of Kriging depends on the design points concentration. The choice of Ω may rely on expert judgement. In practice, maximal plausible ranges for the x values are often known on a physical basis and are expected to be conservative though those ranges may exceed the likeliest (say 95%) range of true variability as the point in inverse statistical problems is precisely to identify the distribution. To prevent wrong results due to a poor approximation of H outside Ω, either the MCMC simulations are constrained to remain inside Ω or H is approximated thanks to a barycentric method outside Ω. It can lead to an adaptive scheme adapting the size of the domain according to early identification stages.

(ii) In order to compare Kriging to a barycentric approximation, a Monte Carlo half sampling strategy is used. The design D is randomly split in two equal parts M times: $D = (D_1^{(i)}, D_2^{(i)})_{1 \le i \le M}$. Then, for each i, the estimator is computed on the first part $D_1^{(i)}$, denoting $\hat{H}_{D_1^{(i)}}$, the relative prediction error computed on the other part $D_2^{(i)}$ is

$$ER(D_2^{(i)}|D_1^{(i)}) = \frac{2}{N} \sum_{z_j \in D_2^{(i)}} \left| \frac{H(z_j) - \hat{H}_{D_1^{(i)}}(z_j)}{H(z_j)} \right|.$$

Permuting the role of $D_1^{(i)}$ and $D_2^{(i)}$ leads to the error approximation

$$ER^{MC} = \frac{1}{2M}\sum_{i=1}^{M}\left(ER(D_2^{(i)}|D_1^{(i)}) + ER(D_1^{(i)}|D_2^{(i)})\right).$$

This Monte Carlo half sampling strategy is also a mean to choose the regression functions $(f_i)_{1\leq i\leq p}$ and the positive kernel K_θ for the Kriging predictor.

(iii) In order to decide if a single approximation suffices or if the approximation is to be made for each $H_i(= H(., d_i))$ the *maximin* distances (4) of the two strategies can be compared. For example, assuming there are ten different d_i for each i, these two strategies are respectively

1. Take a *maximin* design with 1000 points in $[0,1]^3$,
2. Take 10 (one for each d_i) *maximin* designs with 100 points in $[0,1]^2$.

For the first strategy, the *maximin* distance is denoted δ_1 and an upper bound can be found. Indeed, by comparing volumes

$$\begin{aligned} 1000\frac{4\pi}{3}\delta_1^3 &\leq 1 \\ \delta_1 &\leq \left(\frac{3}{4\pi\ 1000}\right)^{1/3} \approx 0.0620. \end{aligned}$$

Now, the minimal distance between the points of a regular grid of 100 points is $\frac{10}{9}$. Thus $\delta_2 \geq \frac{10}{9}$. As a consequence, the first strategy is to be preferred in this case since it allows for a better concentration of the design points favouring a good behaviour of Kriging (Schaback, 2007).

4 Numerical experiments

4.1 A flooding model

The model is related to the risk of dyke overflow during a flood event. It is a truly physics-based hydrodynamic model - even though quite simplified, as resulting from the well-known St-Venant equations in the one-dimensional case with a steady and uniform flow - that has been used as a benchmark in Rocquigny (de) (2009) or in Pasanisi *et al.* (2009). The available model computes the water level at the dyke position (Z_c) and the speed of the river (V) with respect to the observed flow of the river upstream of the dyke (Q), and non observed quantities: The river bed level at the dyke position (Z_v), and the value of Strickler coefficient K_s measuring the friction of the river bed, which is assumed to be homogeneous in this simplified model. Thus

$$\begin{pmatrix} Z_c \\ V \end{pmatrix} = H(Z_v, K_s; Q) + U \text{ with}$$

$$H(Z_v, K_s; Q) = \begin{pmatrix} Z_v + \left(\frac{\sqrt{L}}{B}\right)^{3/5} Q^{3/5} K_s^{-3/5} (Z_m - Z_v)^{-3/10} \\ B^{-2/5} L^{-3/10} Q^{2/5} K_s^{3/5} (Z_m - Z_v)^{3/10} \end{pmatrix}$$

where the values of the section length L and its width B are given and assumed to be fixed ($L = 5000, B = 300$). The river bed level beyond upstream (Z_m) has to be fixed to his mean value 55 in order to ensure identifiability. The ECME and SEM algorithms are tested in the case where:

- Q follows a Gumbel distribution with mode $a = 1013$ and scaling parameter $b = 458$. (Cumulative distribution function $F(q) = 1 - \exp[-\exp((q - a)/b)]$).
- K_s follows a normal distribution with mean $\mu_{K_s} = 30$ and standard deviation $\sigma_{K_s} = 7.5$.
- Z_v follows a normal distribution with mean $\mu_{Z_v} = 50$ and standard deviation $\sigma_{Z_v} = 1$.

The goal is to estimate properly the parameters of the normal distributions of the data K_s and Z_v which are not observed, while flow values Q are assumed to be measurable: indeed, while such flood flows are generally unpredictable, upstream hydrological observations generally issue credible estimates. The ECME algorithm is used with iterative linearisations of the function H. The SEM algorithm is used in the case where the real model H is computed and in the case where H is replaced by a Kriging approximation $\hat{H}$. They are called respectively "full SEM" and "Kriging SEM". 100 samples of $n = 50$ observations have been drawn to compare the parameter estimates given by these three algorithms. These estimates are compared to the ones which could have been obtained by maximising the likelihood if the non observed data were available.

The domain Ω where the Kriging approximation $\hat{H}$ of H is built, is chosen as $\Omega = [1,\ 65] \times [40,\ 54.9] \times [\min(Q_{obs}),\ \max(Q_{obs})]$ where $\min(Q_{obs})$, $\max(Q_{obs})$ are respectively the minimum and the maximum of the observations of Q. A smaller domain was early taken which have led to unsatisfying estimates with Kriging SEM. In the Kriging predictor, the regression functions are set to be linear and the kernel to be Gaussian i.e. $K_\theta(z, z') = \exp(-\theta\|z - z'\|_2^2)$. The initial values have been chosen as follows: for K_s, mean $\mu_{K_s}^{(0)} = 40$ and standard deviation $\sigma_{K_s}^{(0)} = 7.5$; for Z_v, mean $\mu_{Z_v}^{(0)} = 47$ and standard deviation $\sigma_{Z_v}^{(0)} = 3$. In ECME, the initial linearisation point is chosen to be $\mu^{(0)} = (\mu_{K_s}^{(0)}, \mu_{Z_v}^{(0)})$. The variance matrix of U is fixed to $R = \begin{pmatrix} 10^{-5} & 0 \\ 0 & 10^{-5} \end{pmatrix}$, and is supposed to be known.

Smooth histograms are plotted for the four parameters to be estimated in Figure 1.

Table 1 provides the mean and the standard error of the 100 computed estimates.

The methods give similar results. The model is simple and a local linear approximation of H is efficient, that is why the linearisations in ECME perform well. ECME algorithm needs between five and ten iterations of the linearisation process until the stopping criterion (set to 10^{-15}) is reached. For every iteration, $3n(= 150)$ calls to H are necessary. While only 100 calls to H are necessary to have a Kriging approximation with the Kriging SEM. The full SEM could not have worked if H were a real expensive black-box function since 50 iterations of S step are run. In each S step, there are 200 iterations of the Hasting-Metropolis

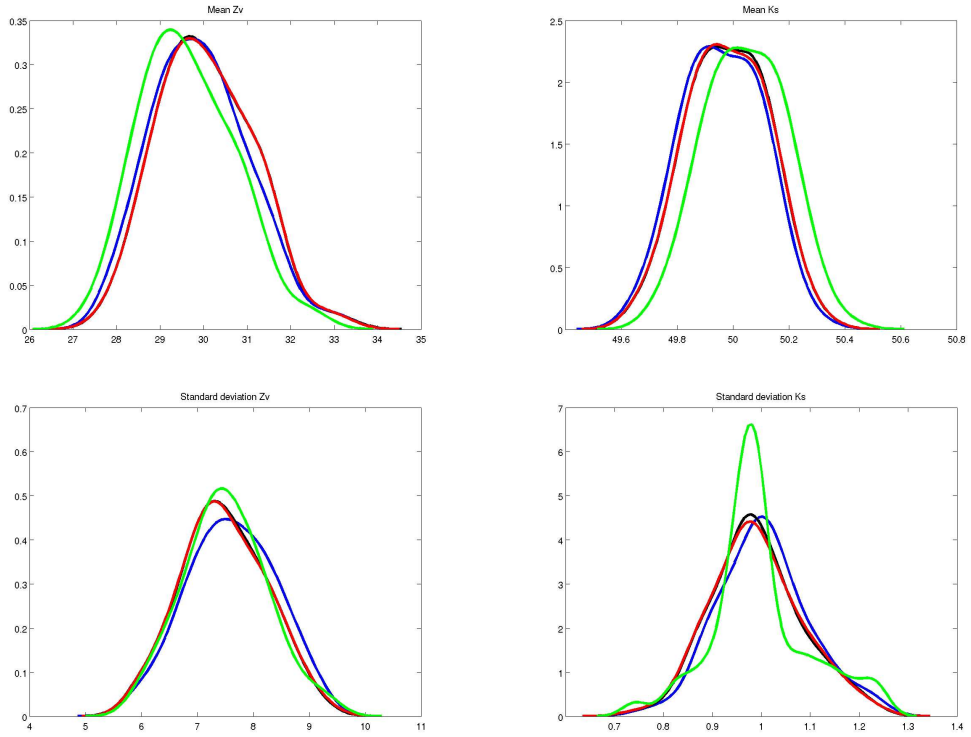

Figure 1: Smooth histograms for the four parameters to be estimated in the flooding example. The red line stands for the maximum likelihood estimates from the complete data, the black line for full SEM, the blue line for Kriging SEM and the green line for ECME.

Parameters	m_{K_s}	m_{Z_v}	σ_{K_s}	σ_{Z_v}	Numbers of calls to H
M.L. from complete data					N/A
Mean estimate	30.06	49.98	7.48	0.99	
Standard error	1.07	0.14	0.74	0.09	
ECME					between 750 and 1500
Mean estimate	29.63	50.04	7.50	1.01	
Standard error	1.06	0.14	0.74	0.12	
Full SEM					500 000
Mean estimate	30.06	49.98	7.48	0.99	
Standard error	1.07	0.14	0.74	0.09	
Kriging SEM					100
Mean estimate	29.92	49.96	7.61	1.00	
Standard error	1.09	0.14	0.76	0.09	

Table 1: Mean and standard error of the 100 computed estimates for the flooding example.

algorithm where H is to be evaluated for all the $n = 50$ points of the sample. Hence, $50 \cdot 200 \cdot 50$ calls to the H function were required with the full SEM algorithm.

4.2 A non linear example

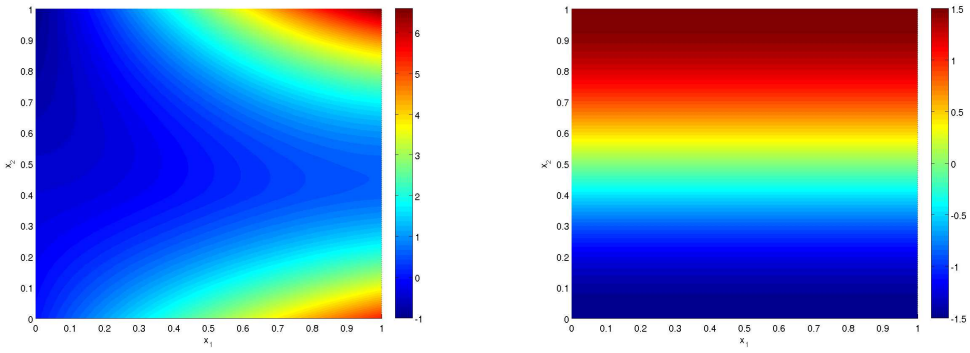

Figure 2: Colormaps corresponding to y_1 (on the left handside) and y_2 (on the right handside), where d is set to 0.5.

We have built an example to illustrate a problem which can occur when the function H cannot be locally linearly approximated. The model function is taken to be $H : [0,1]^3 \to \mathbb{R}^2$,

$$\begin{pmatrix} y_1 \\ y_2 \end{pmatrix} = H(x_1, x_2, d) = \begin{pmatrix} 5x_1(2x_2-1)^2 + x_2\cos(\pi(1-x_1)) + x_1|x_2 - 0.4| \\ (d+1)\sin(\pi(x_2 - 0.5)) \end{pmatrix}.$$

The (y_1, y_2) are observed values depending on d which is observed and follows an uniform distribution on $[0,1]$ and on non observed values: x_1 following a normal distribution with mean 0.4 and standard deviation $\sqrt{2}/10$ and x_2 following a normal distribution with mean 0.5 and standard deviation $\sqrt{2}/10$. As in the previous example, 100 samples of size $n = 50$ have been drawn to assess the estimation performances of each method: maximum likelihood estimator from the complete data, ECME with iterative linearisations of the function H, full SEM, Kriging SEM. $N_{\max} = 100$ evaluations of the function H have been used to obtain the Kriging approximation. Between six and ten linearisations were considered for the ECME algorithm. The domain where the Kriging approximation is done is $\Omega = [0,\ 1]^3$. In the Kriging predictor, the regression functions are set to be polynomials with degree equal or less than 2 and the kernel to be exponential i.e. $K_\theta(z, z') = \exp(-\theta\|z - z'\|_1)$. The initial values have been chosen as follows: for x_1, mean 0.2 and standard deviation 0.2; for x_2, mean 0.2 and standard deviation $2\sqrt{2}/10$. The histogram plots of all the methods are displayed in Figure 3.

Four typically spurious estimates (over 100) given by ECME method were not taken into account into the plots and in Table 2 which summarised the

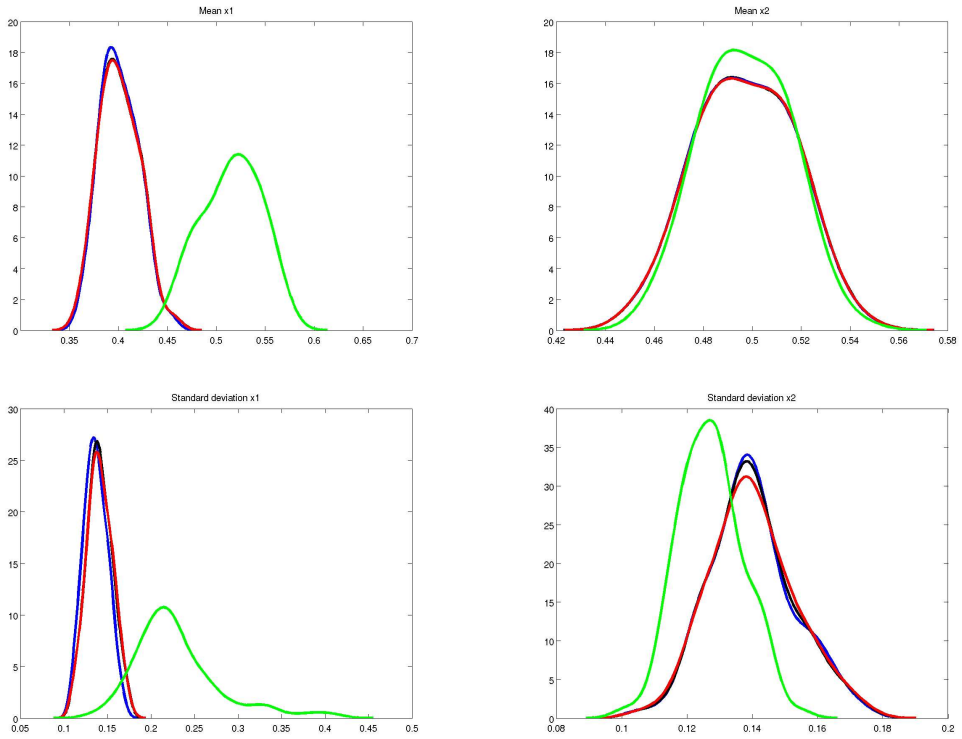

Figure 3: Smooth histograms for the four parameters to be estimated in the highly non linear example. The red line stands for the maximum likelihood estimates from the complete data, the black line for full SEM, the blue line for Kriging SEM and the green line for ECME.

Parameters	m_{x_1}	m_{x_2}	σ_{x_1}	σ_{x_2}	Numbers of calls to H
M.L. from complete data					N/A
Mean estimate	0.40	0.50	0.14	0.14	
Standard error	0.020	0.020	0.014	0.013	
ECME					between 750 and 1500
Mean estimate	0.52	0.50	0.23	0.13	
Standard error	0.030	0.020	0.050	0.010	
Full SEM					500 000
Mean estimate	0.40	0.50	0.14	0.14	
Standard error	0.020	0.020	0.014	0.013	
Kriging SEM					100
Mean estimate	0.40	0.50	0.14	0.14	
Standard error	0.019	0.020	0.013	0.013	

Table 2: Mean and standard error of the 100 computed estimates for the highly non linear example.

results. As it is apparent in Figure 3, the ECME algorithm cannot estimate reasonably well the parameters of the unobserved variable x_1. A linear approximation of H in a neighbourhood of the mean of (x_1, x_2) (i.e. $(0.4, 0.5)$) performs poorly and ECME is misleading. The Kriging approximation is much more flexible than a linear approximation. Thus, a design of $N_{\max} = 100$ points is enough to get an approximation of H on $[0, 1]^3$ leading to reasonable estimates with Kriging SEM. When the Jacobian matrices are computed at the different linearisation points, noticing a change in a sign of one of the coefficients can be a hint to think that the linear approximation would be misleading in ECME algorithm.

5 Discussion

A non linear method has been presented as an alternative to a linear method described in Celeux *et al.* (2009) to solve an inverse problem occurring often in an industrial context. The function H governing the model is supposed to be highly non linear and only known for a limited numbers of points because it is the output of an expensive black-box. To identify such a model, a non linear method based on a Stochastic EM (SEM) algorithm has been proposed. But, since the model function H cannot been made available for a large number of points, it is approximated by Kriging in order to simulate the non observed variables conditionally to the observed variables resulting in an approximated SEM algorithm, the so-called Kriging SEM algorithm. In this paper, examples have been studied to assess the error made when H is replaced by a Kriging approximation. No matter which method is used with the flooding model, where function H can be reasonably linearised, the estimators behave almost like the ideal maximum likelihood estimator based on the complete data. But, it can be noticed that Kriging SEM needs less exact values of H to be computed (namely, a design of 100 points to approximate the model function gives good results with Kriging SEM) than ECME algorithm (at least 750 exact values of H are

needed in the case where five iterated linearisations are enough). Furthermore, in ECME algorithm, the number of linearisations until the stopping criterion is reached is unknown a priori. Hence at the beginning of ECME algorithm, the number of needed calls to the model function H is not determined and this situation is somewhat uncomfortable. The second considered example where H was highly non linear illustrates that the linearisations at work in ECME can be misleading while SEM algorithm with a Kriging approximation continue to provide reasonable estimates. Linearisations are actually harmful if H has locally highly non linear behaviours.

An important and difficult issue is assessing the results: Has the algorithm converged? Are the estimates satisfactory? Unfortunately, there is no well ground criteria to answer those questions. Only experts can say if the estimates seem realistic. In particular, expert knowledge is required to decide which method is safer. Moreover, in the case where Kriging SEM could be recommended, experts are supposed to determine the domain where Kriging approximation is to be made and to propose a reasonable number of calls to the model function. Furthermore, the motivation for identifying the inputs probability distribution has to be kept in mind. This distribution is generally required for a further risk analysis: it will be propagated through a (possibly different) physical model to control the risk level of a key decision variable. Therefore, the sensitivity of this final variable as a function of this probability distribution would have to be taken into account in industrial applications in order to assess fairly the differences between the inversion algorithms investigated in the paper.

Contents

References

Auffray, Y. and Barbillon, P. (2009). Conditionally positive definite kernels : theoretical contribution, application to interpolation and approximation. Rapport de recherche INRIA.

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.

Celeux, G. and Diebolt, J. (1985). The SEM Algorithm: a Probabilistic Teacher Algorithm Derived from the EM Algorithm for the Mixture Problem. *Computational Statistics Quaterly*, **2**, 73-82.

Celeux, G. and Diebolt, J. (1987). A Probabilistic Teacher Algorithm for Iterative Maximum Likelihood Estimation. *Classification and related methods of Data Analysis*, 617-623, Amsterdam, North Holland.

Currin, C., Mitchell, T., Morris, M., and Ylvisaker, D. (1991). Bayesian prediction of deterministic functions, cith applications to the design and analysis of computer experiments. *Journal of the American Statistical Association*, **86**, 953-963.

De Crecy, A. (1996). Determination of the uncertainties of the constitutive relationships in the Cathare 2 Code. *Proceedings of the 1996 4th ASME/JSME International Conference on Nuclear Engineering*.

Delyon, B., Lavielle, M. and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, **27**, 94-128.

Dempster, E. J., Laird, N.M. and Rubin, D.B. (1977). Maximum likelihood from incomplete data via EM algorithm. *Annals of the Royal Statistical Society, Series B*, **39**, 1-38.

Fang, K.-T., Li, R. and Sudjianto, A. (2005). Design and Modeling for Computer Experiments (Computer Science & Data Analysis). *Chapman & Hall/CRC*.

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

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with a MCMC procedure. *ESAIM P&S*, **8**, 115-131.

Liu, C. and Rubin, D.B. (1994). The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, **81**, 633-648.

Lophaven, N.S., Nielsen H.B. and Sondergaard, J. (2002). *DACE, a Matlab Kriging Toolbox*. Technical Report IMM-TR-2002-12.
Available to : http://www2.imm.dtu.dk/ hbn/dace/dace.pdf

Marrel, A., Iooss, B.,Van Dorpe, F. and Volkova, E. (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Computational Statistics & Data Analysis*, **52**, 4731-4744.

McKay, M., Beckman, R. and Conover, W. (1979). A comparaison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, **21(2)**, 239-245

Morris, M.D. and Mitchell, T.J. (1995). Exploratory designs for computational experiments. *Journal of Statistical Planning and Inference*, **43**, 381-402.

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

Pasanisi A., Rocquigny (de) E. Bousquet N. Parent E. (2009). Some useful features of the Bayesian setting while dealing with uncertainties in industrial practice. *Proceedings of the ESREL 2009 Conference*, **3**, 1795-1802.

Rocquigny (de) E., Devictor N., Tarantola S. (2008). Uncertainty in industrial practice, a guide to quantitative uncertainty management. Wiley.

Rocquigny (de) E. (2009). Structural reliability under monotony: Properties of FORM, simulation or response surface methods and a new class of Monotonous Reliability Methods (MRM). *Structural Safety*, **31(5)**, 363-374.

Schaback, R. (2007). Kernel-based meshless methods. Technical report, Institute for Numerical and Applied Mathematics, Georg-August-University Goettingen.

Centre de recherche INRIA Saclay – Île-de-France
Parc Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 Orsay Cedex (France)

Centre de recherche INRIA Bordeaux – Sud Ouest : Domaine Universitaire - 351, cours de la Libération - 33405 Talence Cedex
Centre de recherche INRIA Grenoble – Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier
Centre de recherche INRIA Lille – Nord Europe : Parc Scientifique de la Haute Borne - 40, avenue Halley - 59650 Villeneuve d'Ascq
Centre de recherche INRIA Nancy – Grand Est : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex
Centre de recherche INRIA Paris – Rocquencourt : Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex
Centre de recherche INRIA Rennes – Bretagne Atlantique : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex
Centre de recherche INRIA Sophia Antipolis – Méditerranée : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399