An adaptive kriging method for characterizing uncertainty in inverse problems

Shuai Fu

Journée du GdR MASCOT-NUM - 23 mars 2011

advised by : Gilles Celeux (INRIA/SELECT)

EDF side : Mathieu Couplet, Nicolas Bousquet

Inverse problem

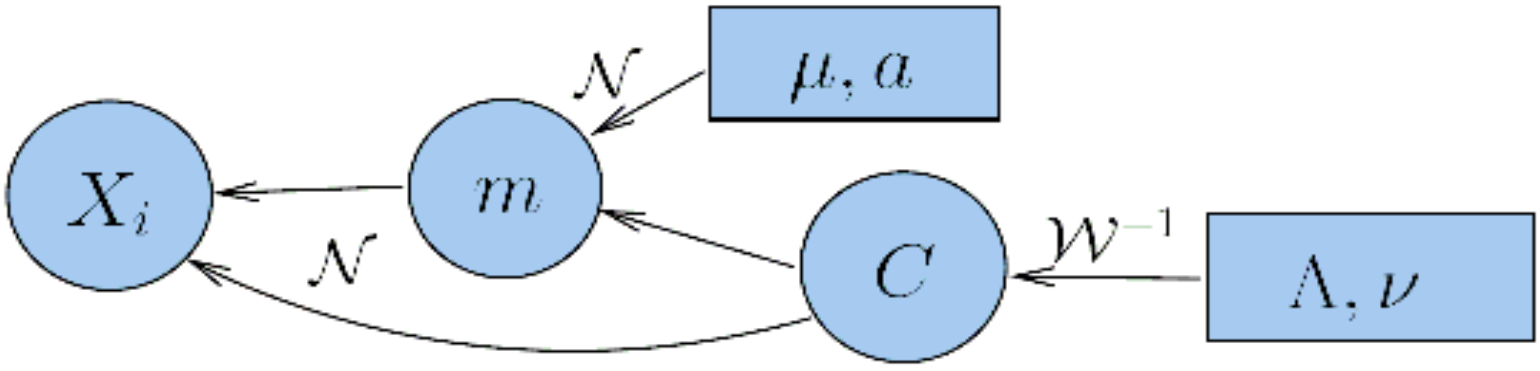
$$Y_i = H(X_i, d_i) + U_i,\ 1 \leq i \leq n$$

with

- Y_i : measured/observed vector data, in $\mathbb{R}^p$;
- H : known function "black box", expensive, describing the physical model, from $E \subset \mathbb{R}^{q+q_2}$ to $\mathbb{R}^p$;
- X_i : non observed random data, in $\mathbb{R}^q$, assumed i.i.d. ;
- d_i : observed variables related to experimental conditions, in $\mathbb{R}^{q_2}$;
- U_i : measurement-model errors, assumed i.i.d. ;
- Variables X_i and U_i are assumed to be independent.

Objective : to estimate the distribution of X_i from the data Y_i.

Bayesian approach



- *Prior* assumptions :

$$\begin{aligned} X_i|m, C &\sim \mathcal{N}_q(m, C) \\ m|C &\sim \mathcal{N}_q(\mu, C/a) \\ C &\sim \mathcal{IW}_q(\Lambda, \nu) \text{ (Inverse-Wishart)} \\ U_i &\sim \mathcal{N}_p(0, R), \ R \text{ assumed known} \end{aligned}$$

- estimate the *posterior* distribution : $\pi(\theta|Y_1, \ldots, Y_n)$, with $\theta = (m, C)$

Advantage of a Bayesian framework

- add the available expert knowledge
- with few data Y_i available
- reduce some potential identifiability problems

Comparison of methods

The SEM algorithm (frequentist) :

- E : conditional density of $X^{(k)} \sim p(\cdot|Y; m^{(k)}, C^{(k)})$
- S : simulation of $X^{(k)} \Rightarrow$ completed sample $Z^{(k)} = (Y, X^{(k)})$
- M : computation of $(m^{(k+1)}, C^{(k+1)})$ (EMV) on the basis of $Z^{(k)}$

The Gibbs algorithm (bayesian) :

- $C^{(k+1)}|m^{(k)}, Y, X^{(k)} \sim \mathcal{IW}\left(\Lambda + n\widetilde{C}^{(k)}_{\text{Exp}} + C^{(k)}_{\text{Bay}}, \nu + n + 1\right)$;
- $m^{(k+1)}|C^{(k+1)}, Y, X^{(k)} \sim \mathcal{N}\left(\frac{a}{n+a}\mu + \frac{n}{a+n}\overline{X^{(k)}}, \frac{C^{(k+1)}}{n+a}\right)$;
- S : $X^{(k+1)}|m^{(k+1)}, C^{(k+1)}, Y \propto$ complicated and unknown form

  $\Rightarrow$ numerical method (Metropolis-Hastings)

  $\Rightarrow$ identical to that of SEM

Kriging model

We note $z = (x, d) \in E \subset \mathbb{R}^{q+q_2}$, the function H is seen as the realization of a Gaussian process $\mathcal{H}$:

$$\mathcal{H}(z) \quad = \quad \sum_{i=1}^{k} \beta_i f_i(z) + G(z) \; = \; F(z)^T \beta + G(z),$$

with

- f_i known regression functions,
- β_i unknown parameters to estimate,
- G centered Gaussian process characterized by its covariance function

$$\mathrm{Cov}(G(z), G(z')) \quad = \quad \sigma^2 K_\epsilon(z, z')$$

where K_ϵ is a symmetric positive kernel with $K_\epsilon(z, z) = 1, \forall z$.

Kriging predictor

Choosing a design $D = \{z_1, \ldots, z_{N_{\max}}\}$: LHS-*maximin*

$$\mathcal{H} \,|\, \mathcal{H}_D = H_D = \{H(z_1), \ldots, H(z_{N_{\max}})\}$$

is still Gaussian, with explicit mean and covariance term.

Kriging approximation of H :
For all $z \in E$, H is approximated by

$$\hat{H}(z) \quad = \quad \mathbb{E}(\mathcal{H}(z) | \mathcal{H}_D = H_D),$$

which allows us to consider the covariance $\mathrm{Cov}(\mathcal{H}(z), \mathcal{H}(z') | \mathcal{H}_D = H_D)$ and its variance denoted by MSE :

$$\mathrm{MSE}(z) \quad = \quad \mathrm{Var}(\mathcal{H}(z) | \mathcal{H}_D = H_D) = \mathbb{E}((\mathcal{H}(z) - \hat{H}(z))^2 | \mathcal{H}_D = H_D)$$

This MSE can be seen as a measure of the prediction accuracy.

New kriging version of the model

H a realization of a Gaussian process $(\mathcal{H}(x,d))_{(x,d)\in E}$
$\Rightarrow$ Y_i a realization of another Gaussian process $\mathcal{Y}_i$

$$\begin{aligned}\mathcal{Y}_i &= \mathcal{H}(X_i,d_i)+U_i \\ &= \widehat{H}(X_i,d_i)+(\mathcal{H}-\widehat{H})(X_i,d_i)+U_i \\ &= \widehat{H}(X_i,d_i)+V_i(X_i,d_i)\,, 1\le i\le n\end{aligned}$$

where $\widehat{H}$ is a kriging approximation of $\mathcal{H}$:

$$(\mathcal{H}-\widehat{H})(x,d)\,|\,\mathcal{H}_D=H_D \quad\sim\quad \mathcal{N}(\mathbf{0},\,\mathrm{MSE}(x,d))$$

and so

$$V_i(x,d)\,|\,\mathcal{H}_D=H_D \quad\sim\quad \mathcal{N}(\mathbf{0},R+\mathrm{MSE}(x,d))$$

New kriging version of the model

More precisely, if Y_i is of size 2, our new model is written as :

$$\boldsymbol{\mathcal{Y}} = \begin{pmatrix} \mathcal{Y}_1^1 \\ \vdots \\ \mathcal{Y}_n^1 \\ \mathcal{Y}_1^2 \\ \vdots \\ \mathcal{Y}_n^2 \end{pmatrix} = \begin{pmatrix} \widehat{H}_1^1(Z_1) \\ \vdots \\ \widehat{H}_n^1(Z_n) \\ \widehat{H}_1^2(Z_1) \\ \vdots \\ \widehat{H}_n^2(Z_n) \end{pmatrix} + \begin{pmatrix} V_1^1(Z_1) \\ \vdots \\ V_n^1(Z_n) \\ V_1^2(Z_1) \\ \vdots \\ V_n^2(Z_n) \end{pmatrix} = \widehat{H}(\mathbf{Z}) + V(\mathbf{Z}),$$

with $Z_i = (X_i, d_i)$ et $\mathbf{Z} = (Z_1, \ldots, Z_n)$.

$$\Rightarrow \boldsymbol{\mathcal{Y}} \,|\, \mathbf{Z}, \mathcal{H}_D = H_D \quad \sim \quad \mathcal{N}_{2\times n}\left(\widehat{H}(\mathbf{Z}), \mathbf{R} + \mathrm{MSE}(\mathbf{Z})\right),$$

New kriging version of the model

$$\mathbf{R} = \left(\begin{array}{c|c} \begin{array}{ccc} R^{11} & & \\ & \ddots & \\ & & R^{11} \end{array} & \mathbf{0} \\ \hline \mathbf{0} & \begin{array}{ccc} R^{22} & & \\ & \ddots & \\ & & R^{22} \end{array} \end{array}\right), \quad \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \begin{array}{l} n \text{ times} \\ \\ n \text{ times} \end{array}$$

$$\mathrm{MSE}(\mathbf{Z}) = \left(\begin{array}{c|c} \mathrm{MSE}^1(\mathbf{Z}) & \mathbf{0} \\ \hline \mathbf{0} & \mathrm{MSE}^2(\mathbf{Z}) \end{array}\right) \quad \left.\begin{array}{l} \\ \\ \\ \end{array}\right\} \begin{array}{l} n \text{ times} \\ \\ n \text{ times} \end{array}$$

with R^{ii} the i-th diagonal component of R and $\mathrm{MSE}^i(\mathbf{Z})$ the variance-covariance matrix of $\mathcal{H}^i(\mathbf{Z})$, $i = 1, 2$.

Substitution (adaptive kriging)

1. Fix $N_{\max}$ as calculation budget and a proportion θ of points.
2. Choose a hypercubic domain E where proxy is valid.
3. Build a design LHS-*maximin* D with $\theta \times N_{\max}$ points in E.
4. Decide whether the design is satisfactory (quality criterion). If it is, we call the kriging predictor $\hat{H}$ and we run the Gibbs algorithm (Metropolis-Hastings).
5. If it isn't, we add the other $(1-\theta) \times N_{\max}$ points sequentially according to an adaptive procedure to improve the quality of design.

Problem associated to kriging method

Problem : MSE (10^{-3}) too large / $\mathbf{R}$ (10^{-5})

$\Rightarrow$ too much uncertainty

$\Rightarrow$ How to improve this kriging method ?

Three aspects of the problem :

1. choice of the experimental field E;
2. quality criterion of a design : Cross-validation *Leave-One-Out*;
3. choice of design points $D = \{Z_1, \ldots, Z_{N_{\max}}\}$ (adaptive kriging method).

Aspect 1 : choice of the experimental field E

- large area $\Rightarrow$ high MSE
- small area : fix E according to the *priori* predictive distribution of X_i

  (for example : Confidence interval with a certain confidence level)

$$
\begin{aligned}
X_i | m, C &\sim \mathcal{N}_q(m, C) \\
m | C &\sim \mathcal{N}_q(\mu, C/a) \\
C &\sim \mathcal{IW}_q(\Lambda, \nu) \\
\Rightarrow X_i &\sim \mathrm{St}_q\left(\mu, \frac{(1+\frac{1}{a})\cdot\Lambda}{\nu+1-q}, \nu+1-q\right)
\end{aligned}
$$

Aspect 2 : quality criterion of a design $Q(D)$

Two propositions :

$$Q_2(D) \quad = \quad 1 - \frac{\sum_{i=1}^{N_{\max}} \left\| H(z_i) - \widehat{H}_{-i}(z_i) \right\|^2}{\sum_{i=1}^{N_{\max}} \left\| H(z_i) - \overline{H}_D \right\|^2}$$

– Scheidt (2006), Cross-Validation

$$Q_{MD}(D) \quad = \quad \left(H(D^*) - \widehat{H}(D^*) \right)' \left(\mathrm{MSE}(D^*) \right)^{-1} \left(H(D^*) - \widehat{H}(D^*) \right)$$

– (Mahalanobis distance), with D^* a validation sample.

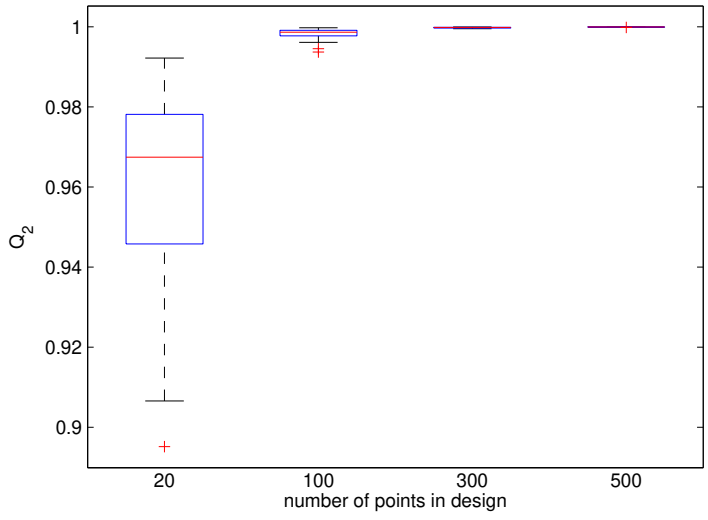


Figure: Q_2 depending on $N_{\max}$

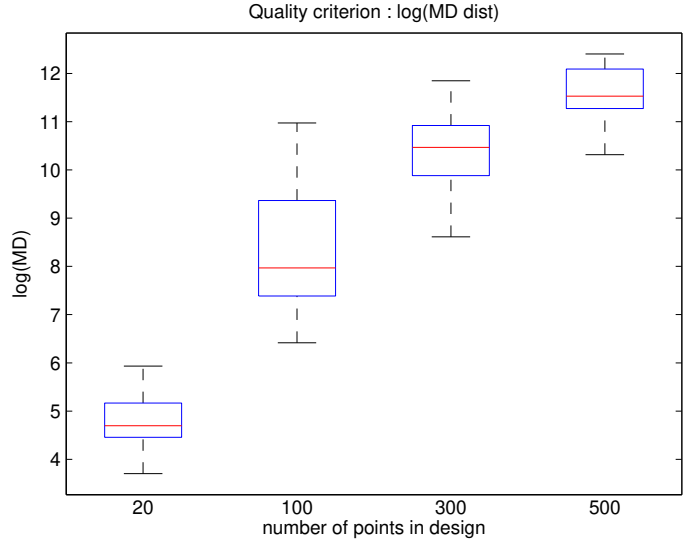


Figure: Q_{MD} depending on $N_{\max}$

Small experience : number of points N_{max} of design D

We consider 3 cases : $100 \rightarrow 300 \rightarrow 500$

- quality of approximation : bad $\rightarrow$ good
- computation time (budget) : light $\rightarrow$ heavy

Comparing the *posterior* distributions :

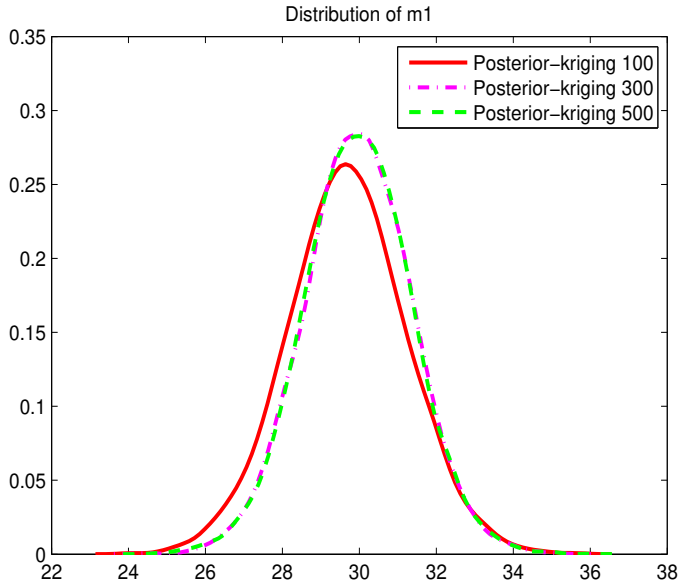


Figure: m_1

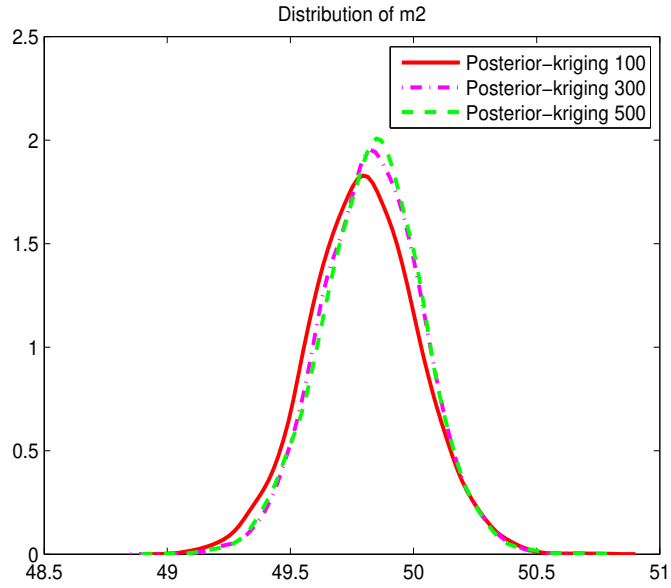


Figure: m_2

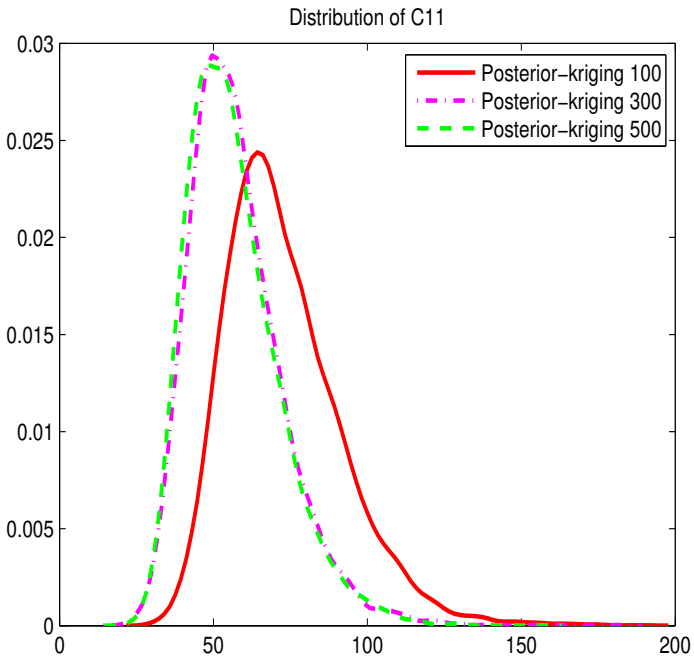


Figure: C_{11}

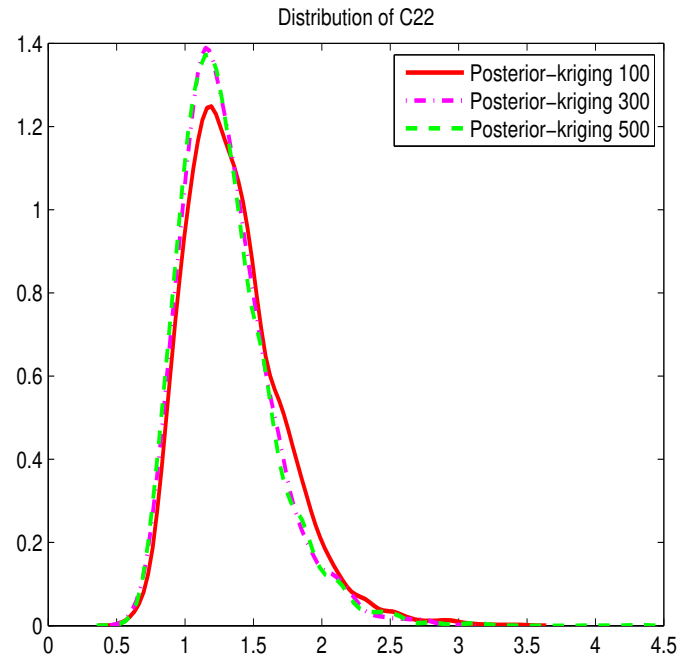


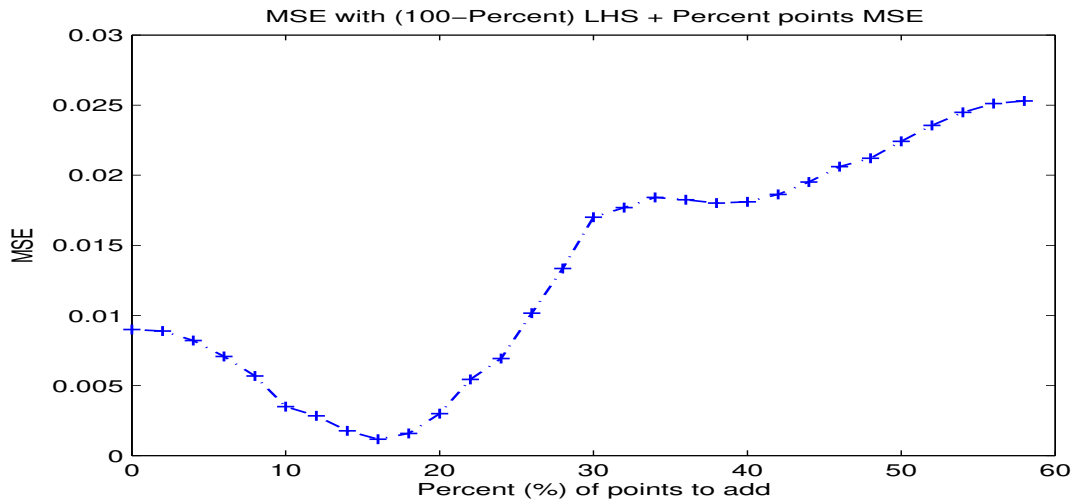
Figure: C_{22}

Aspect 3 : choice of design points $D = (Z_1, \ldots, Z_{N_{\max}})$

$\theta \times N_{\max}$ points according to LHS - maximin

$+$ $(1-\theta) \times N_{\max}$ added points according to a certain criterion :

1. $\max_{(x,d)\in E} \mathrm{MSE}(x,d)$
2. weighted criterion : $\max_{(x,d)\in E} \mathrm{MSE}(x,d)^{\alpha} \cdot \pi(x)^{1-\alpha}$
3. weighted criterion taking into account the y_i :
   $\max_{(x,d)\in E} \mathrm{MSE}(x,d)^{\alpha} \cdot \sum_i \pi(x|y_i, \theta^{[k]})^{1-\alpha}$ (at iteration k)



Aspect 3 : design D with added points : $\max_{(x,d)\in E} \mathrm{MSE}(x,d)^{\alpha} \cdot \pi(x)^{1-\alpha}$

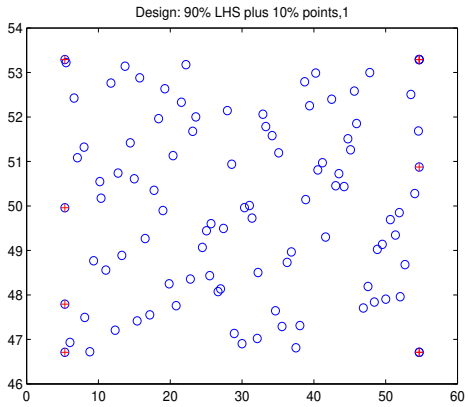


Figure: $\alpha = 1$

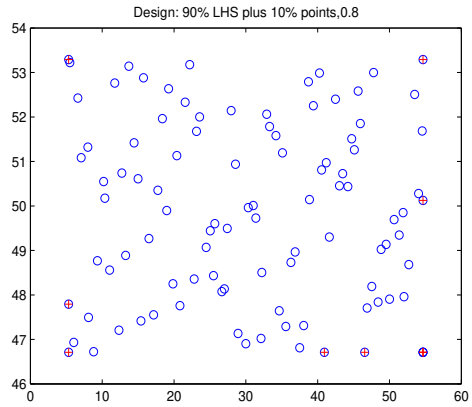


Figure: $\alpha = 0.8$

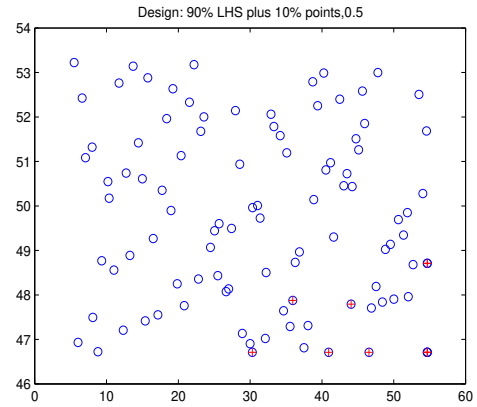


Figure: $\alpha = 0.5$

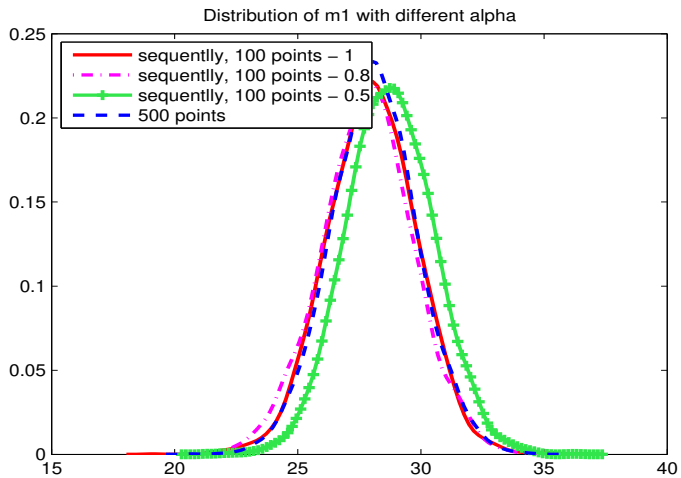


Figure: m_1

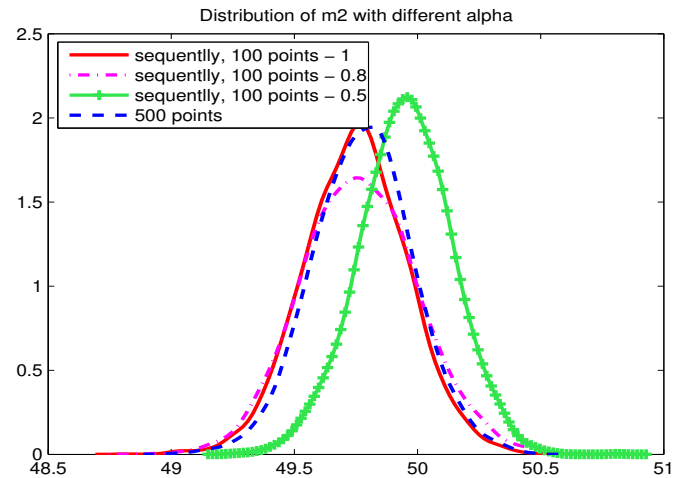


Figure: m_2

Aspect 3 : Comparing the *posterior* distributions (100 points Vs. 500 points)

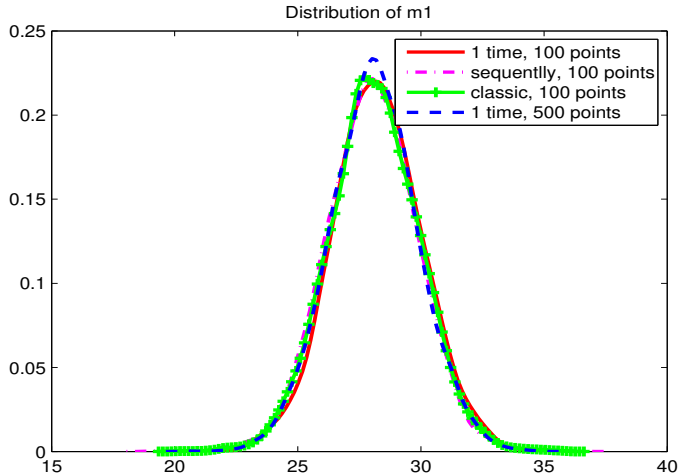


Figure: m_1

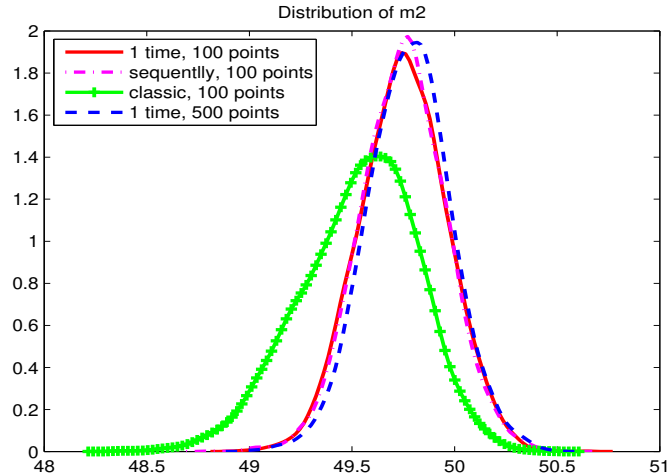


Figure: m_2

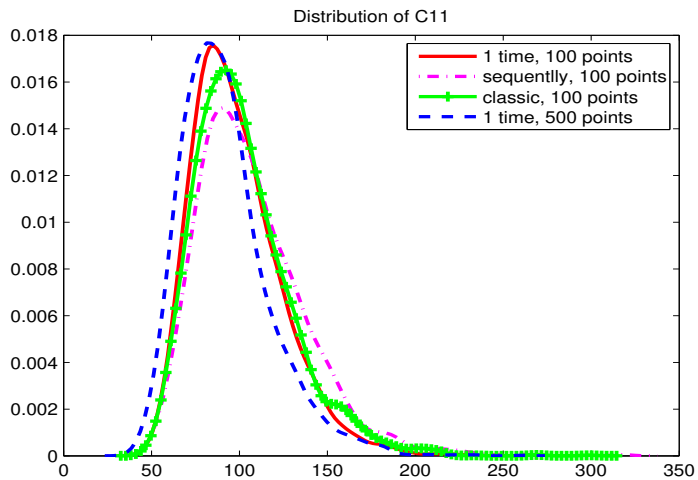


Figure: C_{11}

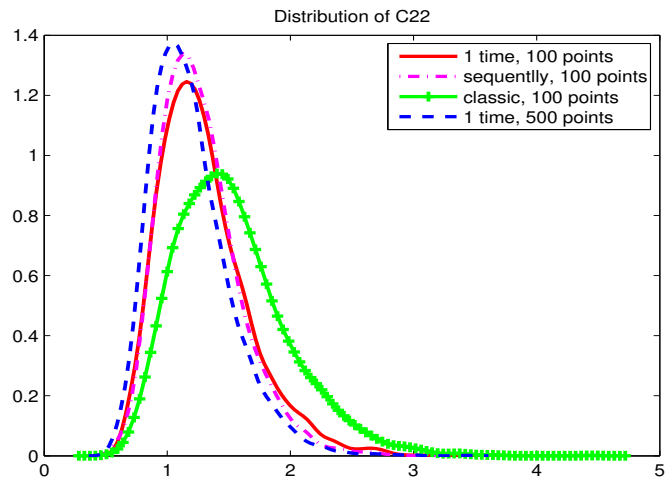


Figure: C_{22}

Small experience : elimination of a chain

Elimination procedure :

1. Step 1 : Run the MCMC algorithm and wait for the chosen burn-in period.
2. Step 2 : Calculate the BG (Brooks-Gelman) statistics based on all the parallel chains.
   - If $BG < 1.2$, nothing to do;
   - else, calculate the BG by removing a chain (leave-one-out), noted BG_{-i}. If $\exists i$, such that $BG_{-i} < 1.2$, candidate $i^* = i$.
3. Step 3 : Repeat the calculations of BG every 10 iterations (for example) for an extended period. If the candidate i is always the same, then the $i - th$ chain is to be eliminated.

Small experience : elimination of a chain

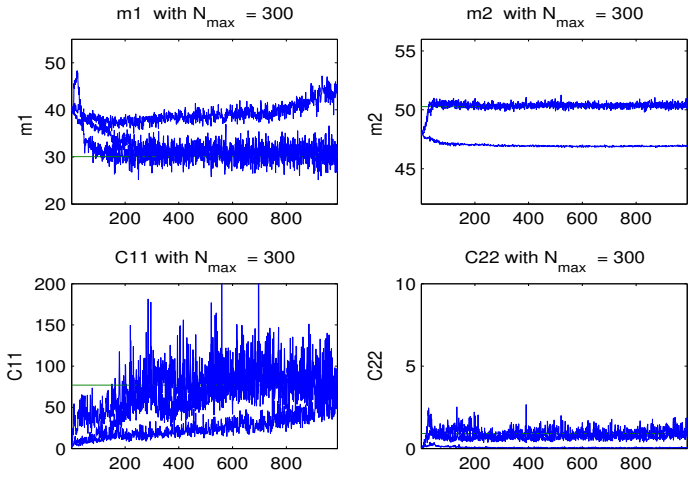


Figure: *before removing a chain*

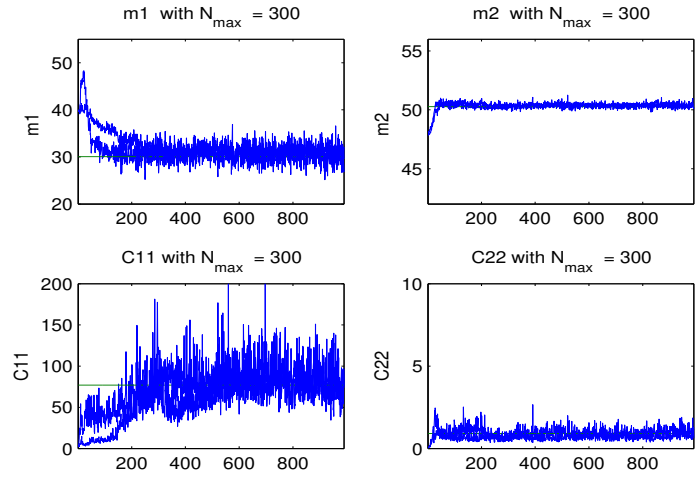


Figure: *after removing a chain*

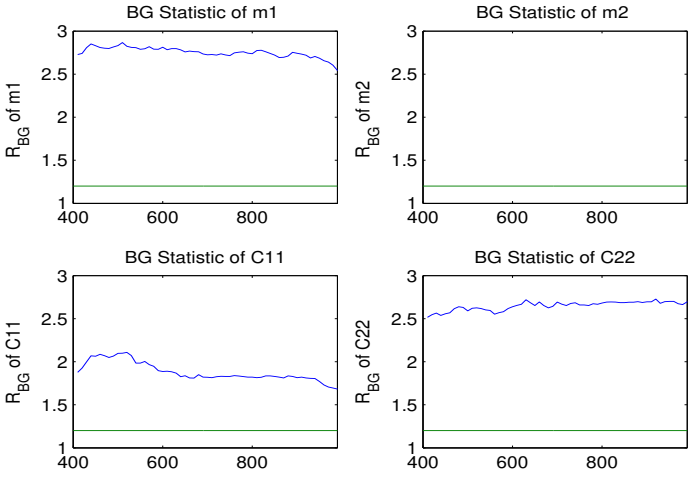


Figure: *BG before removing*

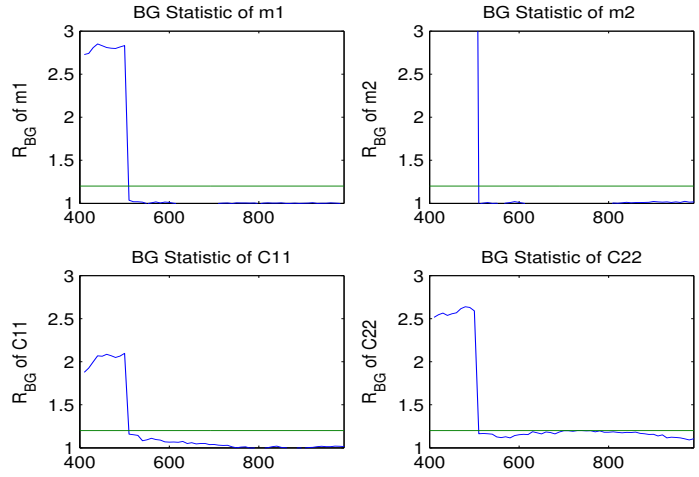


Figure: *BG after removing*

Conclusion and perspective

Conclusion :

- improvement of the performance of MCMC thanks to an "adaptive kriging" method
- proposition of two quality criteria of a design
- elimination of a bad chain to accelerate the convergence speed

Possible further investigations :

- *posterior* information to consider for a sequential kriging method
- Kullback Leibler Distance of two *posterior* distributions to calculate to measure the kriging quality
- calibration of the *prior* distribution in order to choose smartly the hyperparameters

Reference

- Leonardo S. Bastos and Anthony O'Hagan (2009) Diagnostics for Gaussian Process Emulators. University of Sheffield
- Dubourg V., Deheeger F. (2010) Une alternative à la substitution pour les méta-modèles en analyse de fiabilité. *Journées Nationales de la Fiabilité, Toulouse.*
- Picheny V., Ginsbourger D., Roustant O., Haftka R.T., Kim N-H. Adaptive Designs of Experiments for Accurate Approximation of a Target Region.
- Scheidt C. (2006) Analyse statistique d'expériences simulées: Modélisation adaptive de réponses non-régulières par krigeage et plans d'expériences. *Thesis, Louis Pasteur University, Strasbourg.*
- Bettinger R. (2009) Inversion d'un système par krigeage. *Thesis, Nice-Sophia Antipolis University.*
- Barbillon P., Celeux G., Grimaud A., Lefebvre Y. and De Rocquigny E. (2009) Non linear methods for inverse statistical problems. *Rapport de research INRIA.*
- Williams B., Santner T. and Notz W. (2000) Sequential design of computer experiments to minimize integrated response functions. *Statistica Sinica.*

Prior distribution :

Recall of the *prior* modeling :

$$\begin{aligned} X_i|m, C &\sim \mathcal{N}_q(m, C) \\ m|C &\sim \mathcal{N}_q(\mu, C/a) \\ C &\sim \mathcal{IW}_q(\Lambda, \nu) \\ U_i &\sim \mathcal{N}_q(0, R) \end{aligned}$$

Our choices :

- $a = 1$, $\Lambda = t \cdot \widetilde{C}_{\text{Exp}}$, $t = a + 1 = 2$, $\nu = t + 3 = 5$
- $\mu = (40, 47)' \neq m = (30, 50)$
- $R = \begin{pmatrix} 10^{-5} & 0 \\ 0 & 10^{-5} \end{pmatrix}$, $\widetilde{C}_{\text{Exp}} = C = \begin{pmatrix} 7.5^2 & 0 \\ 0 & 1 \end{pmatrix}$
- sample size $n = 30$