

Sequential design of experiments for estimating percentiles of black-box functions

T. Labopin-Richard V. Picheny

17 mai 2016

Contexte et objectif

- CONTEXTE : Modèle de **code numérique coûteux**

$$Y = g(X), \quad g : \mathbb{X} \subset \mathbb{R}^d \rightarrow \mathbb{R}.$$

- OBJECTIF : estimer le **quantile** de la sortie Y

$$q_\alpha(Y) := F_Y^{-1}(\alpha)$$

pour un niveau $\alpha \in]0, 1[$ fixé en évaluant g en le moins de points possibles.

Hypothèse gaussienne

Modèle de **krigeage universel** : g est la réalisation d'un processus gaussien G centré et de fonction de covariance $c(.,.)$ fixée.

Proposition

Pour $\mathcal{A}_n = \{(x_1, g_1 := g(x_1)), \dots, (x_n, g(x_n) := g_n)\}$, on sait que

$$\mathcal{L}(G|\mathcal{A}_n) = GP(m_n(\cdot), k_n(\cdot, \cdot)),$$

où $\forall \mathbf{x} \in \mathbb{X}$,

$$m_n(\mathbf{x}) = \mathbb{E}(G(\mathbf{x})|\mathcal{A}_n) = c_n(\mathbf{x})^T C_n^{-1} \mathbf{g}_n,$$

$$k_n(\mathbf{x}, \mathbf{x}') = \text{Cov}(G(\mathbf{x}), G(\mathbf{x}')|\mathcal{A}_n) = c(\mathbf{x}, \mathbf{x}') - c_n(\mathbf{x})^T C_n^{-1} c_n(\mathbf{x}'),$$

avec $c_n(\mathbf{x}) = [c(\mathbf{x}_1, \mathbf{x}), \dots, c(\mathbf{x}_n, \mathbf{x})]^T$, $C_n = [c(\mathbf{x}_i, \mathbf{x}_j)]_{1 \leq i, j \leq n}$ et $\mathbf{g}_n = [g_1, \dots, g_n]$.

Stratégie séquentielle et bayésienne de planification d'expériences

- Pour un budget initial N_0 , on construit un échantillon d'initialisation $(\mathbf{x}_0^i, g(\mathbf{x}_0^i))_{i=1\dots N_0}$, et on calcule l'estimateur initial du quantile q_{N_0} .
- A chaque étape $n + 1 \geq N_0 + 1$ et jusqu'à ce que le budget d'évaluations autorisées N soit atteint : connaissant les précédentes évaluations \mathcal{A}_n et l'estimateur q_n , on choisit le futur point à évaluer \mathbf{x}_{n+1}^* grâce à un certain critère. On évalue $g(\mathbf{x}_{n+1}^*) := g_{n+1}$ et on met à jour \mathcal{A}_{n+1} et q_{n+1} .
- q_N est l'estimateur du quantile à retourner.

Stratégie séquentielle et bayésienne de planification d'expériences

- Pour un budget initial N_0 , on construit un échantillon d'initialisation $(\mathbf{x}_0^i, g(\mathbf{x}_0^i))_{i=1\dots N_0}$, et on calcule l'estimateur initial du quantile q_{N_0} .
- A chaque étape $n + 1 \geq N_0 + 1$ et jusqu'à ce que le budget d'évaluations autorisées N soit atteint : connaissant les précédentes évaluations \mathcal{A}_n et l'estimateur q_n , on choisit le futur point à évaluer \mathbf{x}_{n+1}^* grâce à un certain critère. On évalue $g(\mathbf{x}_{n+1}^*) := g_{n+1}$ et on met à jour \mathcal{A}_{n+1} et q_{n+1} .
- q_N est l'estimateur du quantile à retourner.

BESOIN : un estimateur séquentiel du quantile et un critère de sélection du nouveau point à évaluer.

Deux estimateurs du quantile

- Conditionnellement à \mathcal{A}_n , la meilleure approximation de $G(\mathbf{x})$ est $m_n(\mathbf{x})$, un estimateur intuitif est donc le quantile de la moyenne du processus gaussien.

$$q_n := q(m_n(X)) = q(\mathbb{E}[G(X)|\mathcal{A}_n]). \quad (1)$$

- L'estimateur minimisant l'erreur en moyenne quadratique $\mathbb{E}((q - q_n)^2)$ parmi tous les estimateurs \mathcal{A}_n -mesurables est

$$q_n = \mathbb{E}(q(G(X))|\mathcal{A}_n). \quad (2)$$

Critère de sélection et méthode SUR

- **Stepwise Uncertainty Reduction** : Geman et al. (1996).
- On recherche des critères de la forme

$$\mathbf{x}_{n+1}^* = \underset{\mathbf{x}_{n+1} \in \mathbb{X}}{\operatorname{argmin}} J_n(\mathbf{x}_{n+1}), \quad (3)$$

- Difficulté : **Evaluer l'impact potentiel du point candidat x_{n+1} sans avoir accès à son évaluation g_{n+1} .**

Application au quantile : Arnaud et al. (2010), Jala et al. (2012)

⇒ Minimisation de la variance conditionnelle de l'estimateur (2).

⇒ Méthode performante en petite dimension mais **trop coûteuse**.

Une expression *explicite* de l'estimateur (1)

- L'estimateur (1) est $q_n = q(m_n(X))$.
- Grâce à un nouvel échantillon $\mathbf{X}_{MC} = (\mathbf{x}_{MC}^1, \dots, \mathbf{x}_{MC}^l)$ de X , on a accès à une **version empirique**

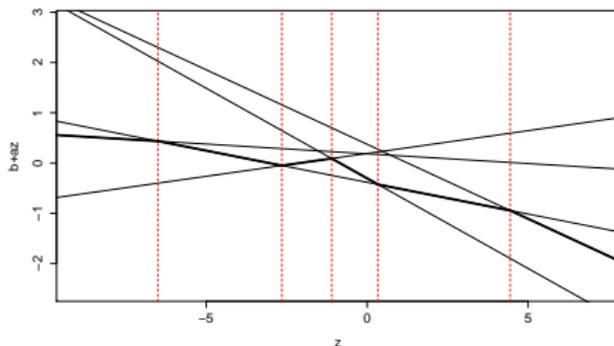
$$q_n = m_n(\mathbf{X}_{MC})_{(\lfloor l\alpha \rfloor + 1)} := m_n(x_n^q).$$

- **BUT** : exprimer $q_{n+1} = m_{n+1}(x_{n+1}^q)$ en fonction du modèle à l'étape n et du nouveau couple de données (x_{n+1}, g_{n+1}) .

Formule de mise à jour à un pas : Chevalier et al. (2014)

$$m_{n+1}(\mathbf{X}_{MC}) = m_n(\mathbf{X}_{MC}) + \frac{k_n(\mathbf{X}_{MC}, \mathbf{x}_{n+1})}{s_n(\mathbf{x}_{n+1})^2} (g_{n+1} - m_n(\mathbf{x}_{n+1})). \quad (4)$$

Une expression *explicite* de l'estimateur (1)



Proposition

Il existe une suite d'intervalles (B_i) connus telle que

$$q_{n+1}(\mathbf{x}_{n+1}, g_{n+1}) = \sum_{i=0}^L m_{n+1}(\mathbf{x}_{n+1}^q(B_i)) \mathbf{1}_{g_{n+1} \in B_i}$$

Le critère *proba*

- Par définition du quantile, on a

$$\mathbb{P}(G(X) \geq q(G(X))) = 1 - \alpha.$$

- On propose donc le critère suivant :

$$J_n^{\text{prob}} = \left| \int_{\mathbb{X}} \mathbb{P}(G(\mathbf{x}) \geq q_n | \mathcal{A}_n) d\mathbf{x} - (1 - \alpha) \right| = |\Gamma_n - (1 - \alpha)|,$$

avec $\Gamma_n = \int_{\mathbb{X}} \mathbb{P}(G(\mathbf{x}) \geq q_n | \mathcal{A}_n) d\mathbf{x}$.

- **PROBLEME** : J_{n+1}^{prob} dépend de g_{n+1} .

- **SOLUTION** : Sous l'hypothèse gaussienne et à l'étape n , g_{n+1} est la réalisation d'une variable aléatoire $G_{n+1} \sim \mathcal{N}(m_n(x_{n+1}), s_n^2(x_{n+1})) \Rightarrow$ on moyennise.
- Le critère devient alors

$$J_n^{\text{prob}}(\mathbf{x}_{n+1}) = |\mathbb{E}_{G_{n+1}}(\Gamma_{n+1}(x_{n+1})) - (1 - \alpha)|$$

avec

$$\Gamma_{n+1}(x_{n+1}) = \int_{\mathbb{X}} \mathbb{P}(G(\mathbf{x}) \geq q_{n+1} | A_{n+1}) d\mathbf{x},$$

et $A_{n+1} = \mathcal{A}_{n+1} \cup (\mathbf{x}_{n+1}, G_{n+1})$.

Proposition

$$J_n^{\text{prob}}(\mathbf{x}_{n+1}) = \left| \int_{\mathbb{X}} \sum_{i=1}^{L-1} \left[\Phi_{r_i^n} \left(e_n^i(\mathbf{x}_{n+1}; \mathbf{x}), f_n^i(\mathbf{x}_{n+1}, l_{i+1}) \right) \right. \right. \\ - \Phi_{r_i^n(\mathbf{x}_{n+1}, \mathbf{x})} \left(e_n^i(\mathbf{x}_{n+1}; \mathbf{x}), f_n(\mathbf{x}_{n+1}, l_i) \right) \\ + \Phi_{r_i^n} \left((e_n^i(\mathbf{x}_{n+1}; \mathbf{x}), f_n^i(\mathbf{x}_{n+1}, l_1)) \right) \\ \left. \left. + \Phi_{-r_i^n} \left(e_n^i(\mathbf{x}_{n+1}; \mathbf{x}), -f_n^i((\mathbf{x}_{n+1}, l_L)) \right) \right] d\mathbf{x} - (1 - \alpha) \right|$$

Le critère *variance*

- q_n doit converger vers le quantile \Rightarrow on veut un estimateur **de plus en plus *stable***.
- La **variance de $q_{n+1} | \mathcal{A}_n \cup (\mathbf{x}_{n+1}, G_{n+1})$** est un bon indicateur de *stabilité*.
- Choix du point qui maximise cette variance pour obtenir le point dont l'évaluation a un large impact sur la valeur de l'estimateur \Rightarrow **Réduction de l'instabilité sur l'estimateur**.

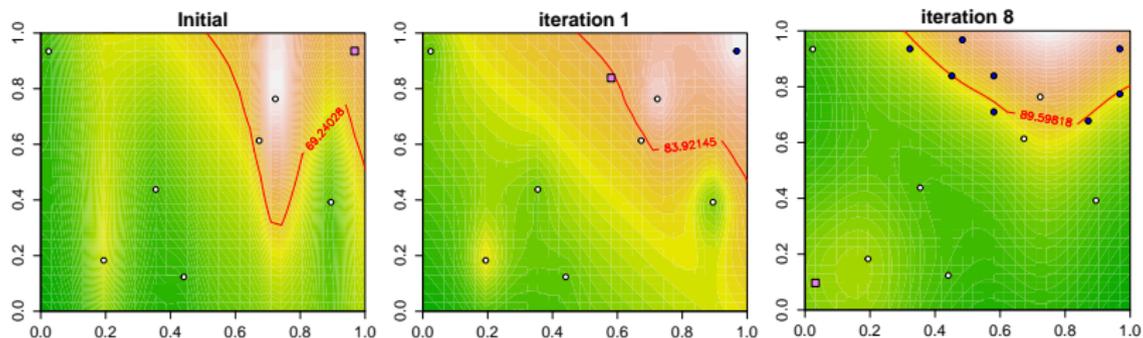
$$J_n^{\text{Var}}(\mathbf{x}_{n+1}) = \text{Var}_{G_{n+1}}(q_{n+1} | \mathcal{A}_{n+1})$$

avec $\mathcal{A}_{n+1} = \mathcal{A}_n \cup (\mathbf{x}_{n+1}, G_{n+1})$.

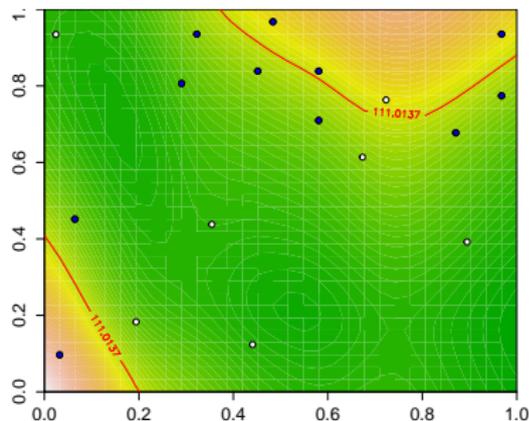
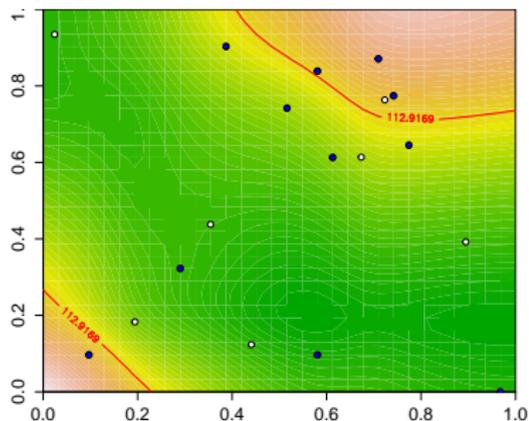
Proposition

$$\begin{aligned}
 J_n^{\text{Var}}(\mathbf{x}_{n+1}) = & \\
 & \sum_{i=1}^L [k_n(\mathbf{x}_{n+1}^q(B_i), \mathbf{x}_{n+1})]^2 V(s_n(\mathbf{x}_{n+1}), l_{i+1}, l_i) P_i \\
 & + \sum_{i=1}^L [m_n(\mathbf{x}_{n+1}^q(B_i) - k_n(\mathbf{x}_{n+1}^q(B_i), \mathbf{x}_{n+1}) E(s_n(\mathbf{x}_{n+1}), l_{i+1}, l_i))]^2 (1 - P_i) P_i \\
 & - 2 \sum_{i=2}^L \sum_{j=1}^{i-1} [m_n(\mathbf{x}_{n+1}^q(B_i) - k_n(\mathbf{x}_{n+1}^q(B_i), \mathbf{x}_{n+1}) E(s_n(\mathbf{x}_{n+1}), l_{i+1}, l_i))] P_i \\
 & [m_n(\mathbf{x}_{n+1}^q(B_j) - k_n(\mathbf{x}_{n+1}^q(B_j), \mathbf{x}_{n+1}) E(s_n(\mathbf{x}_{n+1}), l_{j+1}, l_j))] P_j,
 \end{aligned}$$

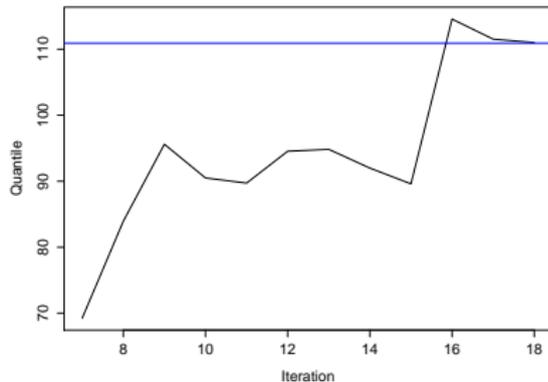
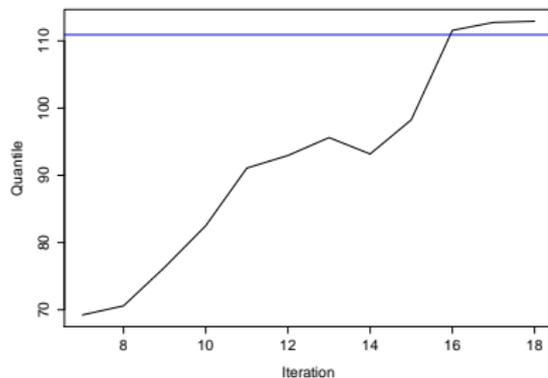
Critère *variance* en dimension 2



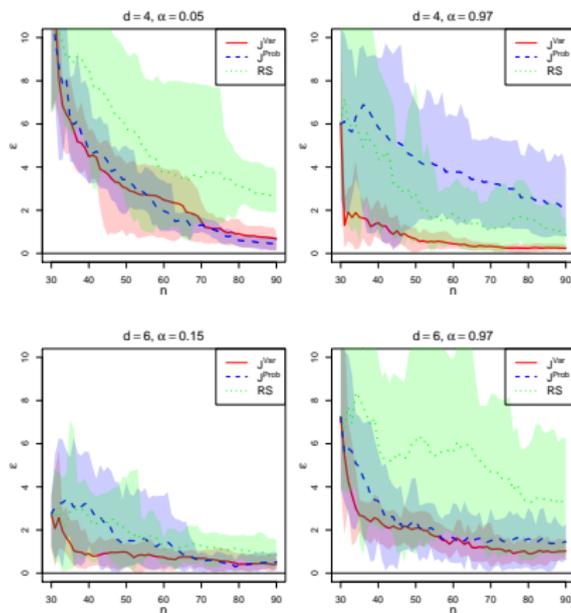
Comparaison des deux critères en dimension 2



Convergence des deux estimateurs en dimension 2



En dimension supérieure





Aurélié Arnaud, Julien Bect, Mathieu Couplet, Alberto Pasanisi, and Emmanuel Vazquez.

Évaluation d'un risque d'inondation fluviale par planification séquentielle d'expériences.

In *42èmes Journées de Statistique*, Marseille, France, France, 2010.



Julien Bect, David Ginsbourger, Ling Li, Victor Picheny, and Emmanuel Vazquez.

Sequential design of computer experiments for the estimation of a probability of failure.

Statistics and Computing, 22(3) :773–793, 2012.



Clément Chevalier, David Ginsbourger, and Xavier Emery.

Corrected kriging update formulae for batch-sequential data assimilation.

In *Mathematics of Planet Earth*, pages 119–122. Springer, 2014.



Donald Geman and Bruno Jedynak.

An active testing model for tracking roads in satellite images.
Pattern Analysis and Machine Intelligence, IEEE Transactions on, 18(1) :1–14, 1996.



Marjorie Jala, Céline Lévy-Leduc, Eric Moulines, Emmanuelle Conil, and Joe Wiart.

Sequential design of computer experiments for parameter estimation with application to numerical dosimetry.
In *Signal Processing Conference (EUSIPCO), 2012 Proceedings of the 20th European*, pages 909–913. IEEE, 2012.