Estimation d'une probabilité de dépassement de seuil via l'utilisation d'un ordre stochastique convexe.

Lucie Bernard [1] & Arnaud Guyader [2] & Florent Malrieu [3] & Philippe Leduc [4]

[1] *LSTA, Université Pierre et Marie Curie, Paris, lucie.bernard@live.fr*
[2] *LSTA, Université Pierre et Marie Curie, Paris, arnaud.guyader@upmc.fr*
[3] *LMPT, Université François Rabelais, Tours, florent.malrieu@lmpt.univ-tours.fr*
[4] *STMicroelectronics, Tours, philippe.leduc@st.com*

Résumé. On s'intéresse à l'estimation d'une probabilité de dépassement de seuil d'une variable aléatoire $g(\mathbf{X})$, où la loi de $\mathbf{X}$ est connue mais la fonction déterministe $g : \mathbb{X} \to \mathbb{R}$ est une boîte-noire coûteuse en temps de calcul. Le peu d'observations de la fonction g dont on dispose n'est pas suffisant pour pouvoir utiliser directement des méthodes de simulations Monte Carlo. On propose alors de modéliser la fonction g par un processus aléatoire gaussien et de considérer la probabilité de dépassement de seuil comme étant la réalisation d'une variable aléatoire. La principale contribution consiste ici à construire une variable aléatoire alternative, dont les bonnes propriétés, en termes de simulation, permettent d'améliorer l'estimation de la probabilité de dépassement de seuil que l'on peut mener en étudiant la distribution la première variable. On montre notamment qu'il existe un ordre convexe entre ces deux variables.

Mots-clés. Méthodes bayésiennes, Processus, Modèles non-paramétriques, Statistique mathématique.

Abstract. We are interested in estimating the threshold-exceeding probability of a random variable $g(\mathbf{X})$, where the distribution of $\mathbf{X}$ is known but the deterministic function $g : \mathbb{X} \to \mathbb{R}$ is an expensive to evaluate black-box function. We do not have enough observations of the function g to use classical Monte Carlo simulation methods. We propose to model the function g by a Gaussian process and to consider the threshold-exceeding probability as being the realization of a random variable. The main contribution consists in building an alternative random variable, whose good properties, in terms of simulation, improve the estimate of the threshold-exceeding probability that can be made by considering the first variable. In particular, we show that there exists a convex order between these two variables.

Keywords. Bayesian methods, Process, Non-parametric models, Mathematical statistics.

1 Contexte

Durant la mise œuvre d'une production industrielle, le processus de fabrication ne peut être entièrement et parfaitement contrôlé (on peut penser, par exemple, aux machineries lourdes et complexes qui se succèdent, ou aux environnements de travail dont la température ou l'humidité sont difficiles à maintenir constantes et peuvent varier au cours du temps), et peut engendrer des produits défaillants qui ne répondent pas aux exigences du cahier des charges. On s'intéresse ici à l'estimation d'une probabilité de dépassement de seuil pour évaluer la rentablité de la production.

Le cadre théorique est le suivant: on considère que les produits étudiés sont caractérisés par d paramètres physiques, généralement appelés *facteurs*. Les valeurs numériques associées à ces facteurs sont fluctuantes, compte tenu de la variabilité inhérente au processus de fabrication. On introduit alors une variable aléatoire $\mathbf{X}$ définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathbb{P})$ et prenant ses valeurs dans l'espace mesurable $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, où $\mathbb{X} \subseteq \mathbb{R}^d$ est un ensemble connu, qui contient l'ensemble des valeurs pouvant être prises par les facteurs. La loi jointe $\mathbb{P}_{\mathbf{X}}$ de la variable aléatoire $\mathbf{X}$ est également supposée connue. On note $\mathbf{D}^n = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ l' ensemble constitué de n realisations $\mathbf{x}_i \in \mathbb{X}$ de $\mathbf{X}$, généralement appelé gensemble d'apprentissage (ou *design of experiments* en anglais).

La performance d'un produit est mesurée au travers de la sortie d'une simulation numérique, qui consiste en l'évaluation d'une fonction g mesurable, déterministe et à valeurs réelles. Plus précisement, on considère que le produit, dont les facteurs prennent les valeurs numériques $\mathbf{x}$, est défaillant, i.e ne satisfait pas les exigences du cahier des charges, si $g(\mathbf{x})$ dépasse un seuil $T \in \mathbb{R}$. Ainsi, la rentabilité de la production industrielle est évaluée à travers le calcul d'une probabilité de dépassement de seuil p définie par

$$p = \mathbb{P}_{\mathbf{X}}(\{\mathbf{x} \in \mathbb{X} : g(\mathbf{x}) \geq T\}) = \mathbb{P}(g(\mathbf{X}) \geq T) = \int_{\mathbb{X}} \mathbb{1}_{g(\mathbf{x}) \geq T} \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}), \tag{1}$$

et appelée généralement *probabilité de défaillance*.

La méthode de référence pour l'estimation d'une telle probabilité est la méthode de Monte Carlo naïve, qui consiste à considérer l'estimateur

$$\widehat{p}_{g,N}^{MC} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{g(X_i) \geq T},$$

où $X_1, .., X_N \overset{i.i.d.}{\sim} \mathbb{P}_{\mathbf{X}}$. Il est exclu de l'utiliser ici car les simulations numériques étant requises pour l'étude de systèmes dont les modèles mathématiques ne sont pas disponibles ou excessivement complexes, on est typiquement dans le cas ici où g est une fonction boîte noire coûteuse en temps de calcul. La seule information dont on dispose sur g consiste en ses évaluations aux points de l'ensemble d'apprentissage, c'est-à-dire que l'on connaît les valeurs de $g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)$.

2 Régression par processus gaussien

Dans ce contexte, une idée naturelle consiste à adopter un point de vue bayésien et à considérer que les observations $g(\mathbf{x}_1), \ldots, g(\mathbf{x}_n)$ représentent de l'information incomplète sur la réalisation d'un processus aléatoire ξ indexé par $\mathbb{X}$. Dans la suite, ce processus est défini sur un espace probabilisé $(\Omega_0, \mathcal{F}_0, \mathrm{P}_0)$. Aussi, en choisissant une distribution a priori gaussienne pour ξ, on peut construire conditionnellement aux évaluations de g aux points de l'échantillon d'apprentissage $\mathbf{D}^n$, un processus gaussien ξ_n dont chaque réalisation, i.e. chaque fonction $\mathbf{x} \longmapsto \xi_n(\mathbf{x}, \cdot)$, passe par les points de coordonnées $(\mathbf{x}_i, g(\mathbf{x}_i))_{1 \leq i \leq n}$ (voir figures 1 et 2 pour un exemple en dimension $d = 1$.) Cette approche fait référence à la méhode de régression par processus gaussiens, appelée aussi Krigeage (voir [1] et [2] pour plus d'informations sur l'intérêt de cette méthode dans le contexte des simulations numériques). Pour tout $\mathbf{x} \in \mathbb{X}$, la moyenne et la variance de la variable aléatoire (gaussienne) $\xi_n(\mathbf{x})$ sont respectivement notées $m_n(\mathbf{x})$ et $\sigma_n^2(\mathbf{x})$ dans la suite.

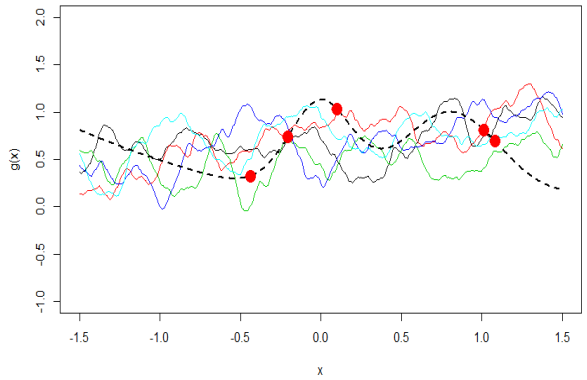

Figure 1: *Exemple en dimension $d = 1$: Les points rouges sont les points de coordonnées $(\mathbf{x}_i, g(\mathbf{x}_i))_{1 \leq i \leq n}$. La ligne en pointillés noirs est la vraie fonction g. Des réalisations du processus gaussien ξ sont représentées.*

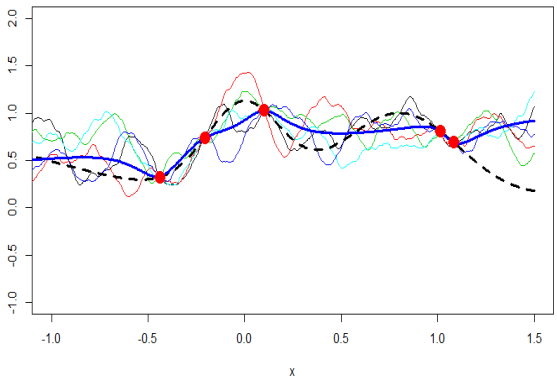

Figure 2: *Exemple en dimension $d = 1$: Des réalisations du processus gaussien conditionnel ξ_n sont représentées. Ligne pleine bleue est sa fonction moyenne, i.e $\mathbf{x} \longmapsto m_n(\mathbf{x})$.*

3 Estimateurs de la probabilité de défaillance

Aussi, comme proposé dans [3], on peut envisager de poursuivre les principes du raisonnement bayésien et de considérer la probabilité de défaillance p comme étant, conditionnellement aux observations, la réalisation d'une variable aléatoire P_n définie sur l'espace probabilisé $(\Omega_0, \mathcal{F}_0, \mathrm{P}_0)$ et à valeurs dans $[0, 1]$, telle que pour tout $\omega_0 \in \Omega_0$,

$$P_n(\omega_0) = \mathbb{P}_{\mathbf{X}}(\{\mathbf{x} \in \mathbb{X} : \xi_n(\mathbf{x}, \omega_0) \geq T\}) = \int_{\mathbb{X}} \mathbb{1}_{\xi_n(\mathbf{x},\omega_0)\geq T} \, \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}).$$

Les réalisations du processus ξ_n étant des fonctions qui interpolent les points de coordonnées $(\mathbf{x}_1, g(\mathbf{x}_1)), \ldots, (\mathbf{x}_n, g(\mathbf{x}_n))$, un estimateur "naturel" de p est la valeur moyenne de la variable aléatoire P_n, c'est-à-dire $\mathrm{E}_0[P_n]$. Cette moyenne est facilement calculable puisque l'on a,

$$\begin{aligned} \mathrm{E}_0[P_n] &= \int_{\mathbb{X}} \mathrm{P}_0\left(\xi_n(\mathbf{x}) \geq T\right) \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}) \\ &= \mathbb{E}\left[\mathrm{P}_0\left(\xi_n(\mathbf{x}) \geq T\right)\right] \\ &= \mathbb{E}\left[s_n(\mathbf{x})\right], \end{aligned} \tag{2}$$

où $\mathbb{E}$ désigne l'espérance par rapport à la loi $\mathbb{P}_{\mathbf{X}}$ et

$$s_n(\cdot) = \mathrm{P}_0\left(\xi_n(\cdot) \geq T\right) = \Phi\left(\frac{m_n(\cdot) - T}{\sigma_n(\cdot)}\right),$$

avec $\Phi : \mathbb{R} \to [0,1]$ la fonction de répartition d'une loi normale centrée réduite. En pratique, une estimation de p est alors donné par $1/N \sum_{i=1}^{N} s_n(\tilde{\mathbf{X}}_i)$, où $\tilde{\mathbf{X}}_1, \ldots, \tilde{\mathbf{X}}_N$ est un N-échantillon de variables i.i.d. de loi $\mathbb{P}_{\mathbf{X}}$.

Cependant, il est très difficile de simuler la variables P_n car cela nécessite de simuler le processus ξ_n et d'effectuer une integration par rapport à la loi $\mathbb{P}_{\mathbf{X}}$. De plus, la loi de P_n depend de la loi finie-dimensionnelle du processus ξ_n, ce qui rend ses moments, ou encore sa variance, difficiles à calculer. Par exemple, on peut facilement montrer que pour tout $m \geq 1$, le moment d'ordre m de la variable P_n vérifie

$$\mathrm{E}_0[P_n^m] = \int_{\mathbb{X}^m} \mathrm{P}_0(\xi_n(\mathbf{x}_1) \geq T, \ldots \xi_n(\mathbf{x}_m) \geq T) \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}_1) \ldots \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}_m), \tag{3}$$

et donc que toute estimation par simulation de Monte Carlo d'un moment de P_n nécessite un temps de calcul assez lourd .

Ainsi, on va chercher à fournir des solutions pour approcher les moments, la variance ou encore les quantiles de P_n, i.e. pour apprendre sur la distribution de P_n. En particulier, on propose d'introduire une variable aléatoire R, définie sur un espace probabilisé $(\Omega, \mathcal{F}, \mathsf{P})$ et définie pour tout $\boldsymbol{\omega} \in \Omega$ par

$$R(\boldsymbol{\omega}) = \mathbb{P}_{\mathbf{X}}\left(\{\mathbf{x} \in \mathbb{X} : s_n(\mathbf{x}) > U(\boldsymbol{\omega})\}\right) = \int_{\mathbb{X}} \mathbb{1}_{s_n(\mathbf{x}) > U(\boldsymbol{\omega})} \, \mathrm{d}\mathbb{P}_{\mathbf{X}}(\mathbf{x}),$$

où $U : \Omega \to [0,1]$ est une variable aléatoire continue. La moyenne U est notée $\mathsf{E}[\cdot]$.

La distribution de R dépend mesurablement de la distribution de U. Elle a donc l'avantage d'être unidimensionelle, de seulement tenir compte des lois marginales du processus ξ_n, et d'être facilement simulable. Aussi, on peut montrer que les espérances de P_n et R sont égales si et seulement si U suit une loi uniforme sur $[0,1]$. Autrement dit, si U suit une loi uniforme sur $[0,1]$, alors un estimateur de p peut aussi être donné par l'espérance de R. On peut également montrer que pour toute fonction convexe φ, on a

$$E_0[\varphi(P_n)] \leq \mathsf{E}[\varphi(R)] \tag{4}$$

si et seulement si U suit une loi uniforme sur $[0,1]$. On dit que P_n *is smaller than* R *in the convex order*. On invite à consulter [6] pour une présentation des différents ordres stochastiques pouvant exister entre des variables aléatoires.

De part l'inégalité (4), il est alors facile de fournir des majorants des moments d'ordre m et de la variance de P_n. On peut également en déduire un estimateur des quantiles de P_n, qui est plus précis que celui issu de l'inégalité de Markov proposé dans [3]. On propose des résultats de simulations numériques qui comparent les capacités de R à estimer p à celles de d'autres méthodes proposées dans la littérature (voir [3], [4] en particulier ou encore [5] pour d'autres exemples de méthodologies mêlant régression par processus gaussien et simulations Monte Carlo).

Bibliographie

[1] C. E. Rasmussen and C. K. I. Williams. (2006), Gaussian processes for machine learning. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.
[2] J. Sacks, B. S. Schiller, and W.J. Welch. (1989), Designs for computer experiments. Technometrics, 31(1).
[3] Y. Aufray, P. Barbillon, and J-M. Marin. (2014), Bounding rare event probabilities in computer experiments. Comput. Statist. Data Anal., 80: 153-166.
[4] J. Bect, D. Ginsbourger, L. Li, V. Picheny, E. Vazquez. (2012), Sequential design of computer experiments for the estimation of a probability of failure. Stat. Comput, 22(3):773-793.
[5] V. Dubourg, F. Deheeger, B. Sudret. (2013), Metamodel-based importance sampling for structural reliability analysis. Probabilistic Engineering Mechanics.
[6] M. Shaked, J. Shanthikumar. (2017), Stochastic Orders. Springer Series in Statistics. Springer, New-York.