

Estimation of a threshold exceeding probability by using the convex order

MMR 2017

Lucie Bernard

Philippe Leduc - Florent Malrieu

LSTA - STMicroelectronics - LMPT

July 6 2017

Framework

- Let $\mathbf{X} \in \mathbb{R}^d$ be a **random variable** and $\mathbf{X} \sim P_{\mathbf{X}}$.
- Let $g : \mathbb{R}^d \rightarrow \mathbb{R}$ be an **expensive to evaluate black-box function**.
- **Goal**: Estimate the **probability** p of $g(\mathbf{X})$ exceeding a threshold T ,

$$p = \mathbb{P}(g(\mathbf{X}) \geq T) = \int_{\mathbb{R}^d} \mathbb{1}_{g(\mathbf{x}) \geq T} P_{\mathbf{X}}(d\mathbf{x}),$$

from a **restricted number of observations** $\{g(\mathbf{x}_1), \dots, g(\mathbf{x}_n)\}$ of the random variable $g(\mathbf{X})$.

Gaussian process regression

Basic ideas

- The function g is supposed to be a **realization of a Gaussian process** $\xi = \{\xi(\mathbf{x})\}_{\mathbf{x} \in \mathbb{R}^d}$.
- Given the observations $\{g(\mathbf{x}_1), \dots, g(\mathbf{x}_n)\}$, the **posterior probability distribution** of $\xi(\mathbf{x})$ is still **Gaussian**, i.e.

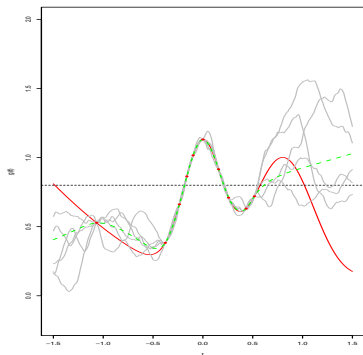
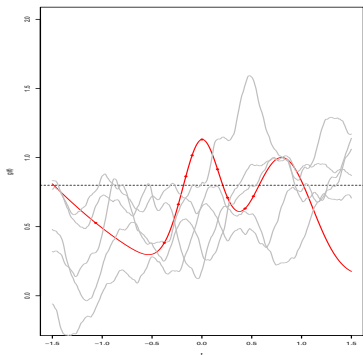
$$\forall \mathbf{x} \in \mathbb{R}^d, [\xi(\mathbf{x}) \mid \xi(\mathbf{x}_1) = g(\mathbf{x}_1), \dots, \xi(\mathbf{x}_n) = g(\mathbf{x}_n)] \sim \mathcal{N}(m_n(\mathbf{x}), \sigma_n^2(\mathbf{x})),$$

where the posterior mean and variance $m_n(\mathbf{x})$ and $\sigma_n^2(\mathbf{x})$ can be easily computed.

- **Each realization** of the conditioned process is an **interpolation model** for g .

Example in dimension $d = 1$

- $T = 0.8$, $n = 10$, $P_X = \mathcal{N}(0, 0.4^2)$.
- The true function is $g(\mathbf{x}) = (0.4\mathbf{x} - 0.3)^2 + e^{-11.534|\mathbf{x}|^{1.95}} + e^{-5(\mathbf{x}-0.8)^2}$.



Left: Simulations of realizations of ξ with given unconditional mean and variance functions.

Right: Conditional simulations and mean function m_n .

A first estimation

As a result, the probability

$$\rho = \int_{\mathbb{R}^d} \mathbb{1}_{g(\mathbf{x}) \geq T} P_{\mathbf{X}}(d\mathbf{x})$$

is a **realization of the random variable** $P_n \in [0, 1]$ defined by

$$P_n = \mathbb{P}(\xi_n(\mathbf{X}) \geq T \mid \xi_n) = \int_{\mathbb{X}} \mathbb{1}_{\xi_n(\mathbf{x}) \geq T} P_{\mathbf{X}}(d\mathbf{x}),$$

where ξ_n is the process conditioned on the n available observations.

A first estimation

- A natural estimator \hat{p}_n of p is the mean value of P_n :

$$\hat{p}_n = \mathbb{E}[P_n] = \int_{\mathbf{X}} \mathbb{E}[\mathbb{1}_{\xi_n(\mathbf{x}) \geq T}] P_{\mathbf{X}}(d\mathbf{x}) = \mathbb{E}[s_n(\mathbf{X})],$$

where the function s_n is easy to evaluate because

$$s_n(\mathbf{x}) = \mathbb{E}[\mathbb{1}_{\xi_n(\mathbf{x}) \geq T}] = \mathbb{P}(\xi_n(\mathbf{x}) \geq T) = \Phi\left(\frac{m_n(\mathbf{x}) - T}{\sigma_n(\mathbf{x})}\right).$$

- Finally, the estimator is

$$\hat{p}_{N, P_n}^{MC} = \frac{1}{N} \sum_{i=1}^N s_n(\mathbf{X}_i),$$

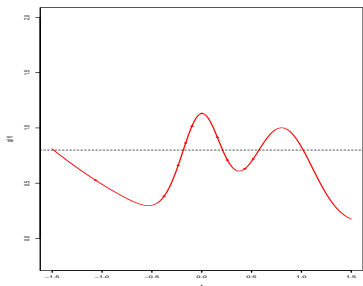
where $\mathbf{X}_1, \dots, \mathbf{X}_N \sim P_{\mathbf{X}}$.

Example in dimension $d = 1$

- We want to estimate the probability

$$p = \int_{\mathbb{R}} \mathbb{1}_{g(\mathbf{x}) \geq T} P_{\mathbf{X}}(d\mathbf{x}),$$

where $P_{\mathbf{X}} = \mathcal{N}(0, 0.4^2)$.



- We use a Monte Carlo sample of size $N = 10^4$,

$$\mathbf{X}_1, \dots, \mathbf{X}_N \sim P_{\mathbf{X}}.$$

Estimator	Estimated value of p
$\hat{p}_N^{MC} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{g(\mathbf{X}_i) \geq T}$	0,45
$\hat{p}_{N, P_n}^{MC} = \frac{1}{N} \sum_{i=1}^N s_n(\mathbf{X}_i)$	0.42

Limits

- The distribution of P_n is unknown.
- **Simulating** P_n require realizations of a Gaussian process: it is **time consuming** and could lead to **numerical issues**.
- For example, the m -th moment of P_n is not easy to compute:

$$\mathbb{E}[P_n^m] = \int_{\mathbb{R}^d} \dots \int_{\mathbb{R}^d} \mathbb{P}\left(\bigcap_{i=1}^m \xi_n(\mathbf{x}_i) \geq T\right) P_{\mathbf{X}}(d\mathbf{x}_1) \dots P_{\mathbf{X}}(d\mathbf{x}_m).$$

↔ How to learn about the distribution of P_n ?

↔ For all $\alpha \in [0, 1]$, what about the α -quantile $F_{P_n}^{-1}(\alpha)$ of P_n ?

An alternative estimation

- Let $U \in [0, 1]$ be a real random variable.
- We introduce an **alternative random variable** $R_n = R_n(U)$ defined by

$$R_n = \mathbb{P}(s_n(\mathbf{X}) > U \mid U) = \int_{\mathbf{X}} \mathbb{1}_{s_n(\mathbf{x}) > U} dP_{\mathbf{X}}(\mathbf{x}).$$

Proposition

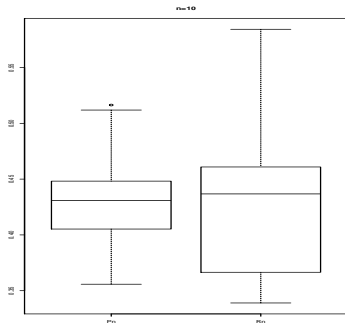
$$U \sim \mathcal{U}[0, 1] \Leftrightarrow \mathbb{E}[P_n] = \mathbb{E}[R_n]$$

- The **mean value** of R_n is an **estimator** of p .

Example in dimension $d = 1$

- Necessary time to generate 10^3 realizations of P_n : 9.95 minutes.
 - Necessary time to generate 10^3 realizations of R_n : 0.14 seconds.
- ▶ $M = 10^3$, $U_1, \dots, U_M \sim \mathcal{U}[0, 1]$.
- ▶ $N = 10^4$, $\mathbf{X}_1, \dots, \mathbf{X}_N \sim P_{\mathbf{X}}$.
- ▶ For all $j = 1, \dots, M$, $R_{n,j} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{s_n(\mathbf{X}_i) > U_j}$.

Estimator	Estimated value of p
$\hat{p}_N^{MC} = \frac{1}{N} \sum_{i=1}^N \mathbb{1}_{g(\mathbf{X}_i) \geq T}$	0,45
$\hat{p}_{N,P_n}^{MC} = \frac{1}{N} \sum_{i=1}^N s_n(\mathbf{X}_i)$	0.42
$\hat{p}_{N,R_n}^{MC} = \frac{1}{M} \sum_{j=1}^M R_{n,j}$	0.42



An alternative estimation

Proposition

$$U \sim \mathcal{U}[0, 1] \Leftrightarrow P_n \leq_{\text{cx}} R_n.$$

- The inequality $P_n \leq_{\text{cx}} R_n$ means that for all **convex function** φ ,
- $$\mathbb{E}[\varphi(P_n)] \leq \mathbb{E}[\varphi(R_n)].$$

As a result,

- $\text{Var}[P_n] \leq \text{Var}[R_n]$.
- $\mathbb{E}[P_n^m] \leq \mathbb{E}[R_n^m]$, $m \geq 1$.
- $\frac{1}{1-\alpha} \int_{\alpha}^1 F_{P_n}^{-1}(t) dt \leq \frac{1}{1-\alpha} \int_{\alpha}^1 F_{R_n}^{-1}(t) dt$, $\alpha \in (0, 1)$.

An alternative estimation

- Finally, by the **convex order** and **Markov inequality**, the α -quantile $F_{P_n}^{-1}(\alpha)$ of P_n satisfies

$$\max(0, \mathbb{E}[R_n] + \alpha - 1) \leq \frac{1}{\alpha} \int_0^\alpha F_{R_n}^{-1}(t) dt \leq F_{P_n}^{-1}(\alpha) \leq \frac{1}{1-\alpha} \int_\alpha^1 F_{R_n}^{-1}(t) dt \leq \min\left(1, \frac{\mathbb{E}[R_n]}{1-\alpha}\right).$$

- Explicit formulas for the **lower and upper bounds**

Proposition

For all $\alpha \in (0, 1)$,

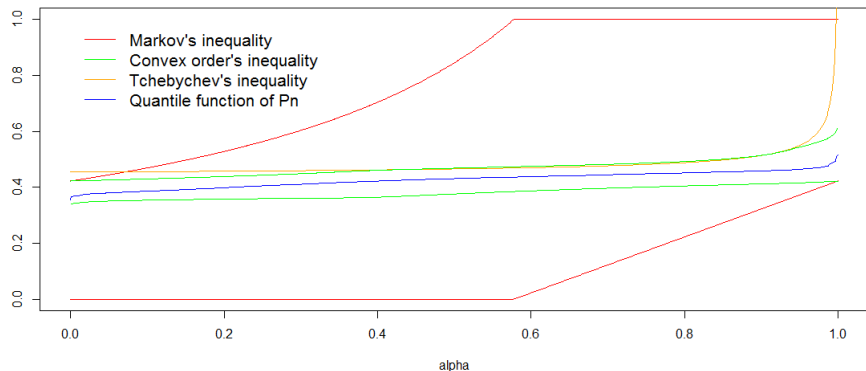
- $\frac{1}{\alpha} \int_0^\alpha F_{R_n}^{-1}(t) dt = 1 - \int_{\mathbb{R}^d} \min\left(1, \frac{1 - s_n(\mathbf{x})}{\alpha}\right) P_{\mathbf{X}}(d\mathbf{x}).$
- $\frac{1}{1-\alpha} \int_\alpha^1 F_{R_n}^{-1}(t) dt = \int_{\mathbb{R}^d} \min\left(1, \frac{s_n(\mathbf{x})}{1-\alpha}\right) P_{\mathbf{X}}(d\mathbf{x}).$

Example in dimension $d = 1$, $n = 10$

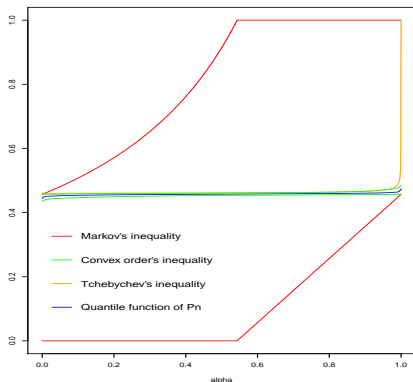
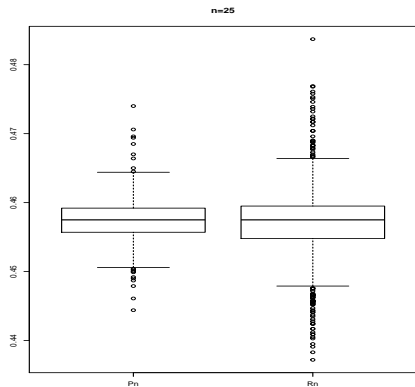
- ▶ The boundaries given by the convex order tell us that we have with probability $\geq 0,95$,

$$0,35 \leq P_n \leq 0,54.$$

- ▶ Tchebychev's inequality: $F_{P_n}^{-1}(\alpha) \leq \mathbb{E}[P_n] + \sqrt{\frac{\text{Var}[P_n]}{1-\alpha}}$.



Exemple in dimension $d = 1$, $n = 25$



- ▶ The boundaries given by the convex order tell us that we have with probability $\geq 0,95$,

$$0,44 \leq P_n \leq 0,47.$$

- ▶ A sample Monte Carlo $\mathbf{X}_1, \dots, \mathbf{X}_N \sim P_X$ of size $N = 10^6$ gives the estimate $\hat{p}_N^{MC} = 0,448$.

Conclusion & Perspectives

- The distribution of the random variable P_n is unknown. We use an alternative random variable, which is easy to simulate and has the same mean value.
- Can we find other results about the convex order to **refine** our credible interval?

References



Y. Auffray, P. Barbillon, and J-M. Marin.

Bounding rare event probabilities in computer experiments.

Comput. Statist. Data Anal., 80 :153–166, 2014.



J. Oger, P. Leduc, and E Lesigne.

A random field model and decision support in industrial production.

J. SFdS, 156(3) :1–26, 2015.



Nicole Bäuerle and Alfred Müller.

Stochastic orders and risk measures : consistency and bounds.

Insurance Math. Econom., 38(1) :132–148, 2006.



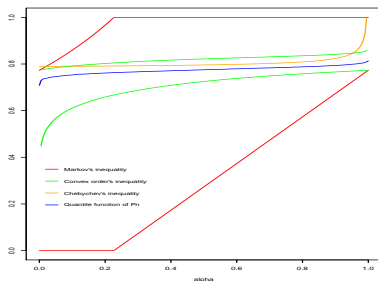
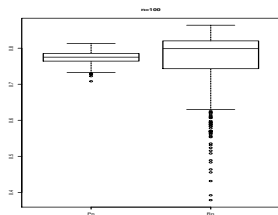
M. Shaked and J.G. Shanthikumar.

Stochastic Orders.

Springer Series in Statistics. Springer, New York, 2007.

Annexe : Example in dimension $d = 3$.

- $T = 1$, $n = 100$, $P_X = \mathcal{U}[-1, 1]$.
- $g(x_1, x_2, x_3) = 4.5 \sin(x_1) e^{-4x_2} - 0.6x_2^2 + 2.3(5 - x_3^2)$.
- A Monte Carlo sample of size $N = 10^4$ gives the estimate $\hat{\rho}_N^{MC} = 0,8$.
- Necessary time to generate 10^3 realizations of $P_n \approx 20$ minutes.
- Necessary time to generate 10^3 realizations of $R_n \approx 0.32$ seconds.
- The mean value of the random variable P_n gives the estimates $\hat{\rho}_{N, P_n}^{MC} = 0,77$.
- The boundaries given by the convex order tell us that we have with probability $\geq 0,95$, $0,57 \leq P_n \leq 0,86$.



Annexe : Example in dimension $d = 3$.

- $T = 1$, $n = 200$, $P_{\mathbf{X}} = \mathcal{U}[-1, 1]$.
- $g(x_1, x_2, x_3) = 4.5 \sin(x_1) e^{-4x_2} - 0.6x_2^2 + 2.3(5 - x_3^2)$.
- A Monte Carlo sample of size $N = 10^4$ gives the estimate $\hat{\rho}_N^{MC} = 0,8$.
- Necessary time to generate 10^3 realizations of $P_n \approx 20$ minutes.
- Necessary time to generate 10^3 realizations of $R_n \approx 0.32$ seconds.
- The mean value of the random variable P_n gives the estimates $\hat{\rho}_{N, P_n}^{MC} = 0,79$.
- The boundaries given by the convex order tell us that we have with probability $\geq 0,95$, $0,68 \leq P_n \leq 0,82$.

