Meta-model for sum of indicator random variables

RESIM 2021

Clément Rey
Joint work with Florian Bourgey and Emmanuel Gobet

CMAP (Ecole Polytechnique)

clement.rey@polytechnique.edu.fr

Outline

Introduction

Subject of study

- We study $\mathcal{L}_K$, a random variable taking values in $\mathbb{R}$ such that

$$\mathcal{L}_K = \sum_{k=1}^{K} \mathbf{1}_{c_k \leqslant X}$$

with $K \in \mathbb{N}^*$, X random variable with law ν and $(c_k)_{k \in \mathbb{N}^*}$ a sequence of independent random variables and independent from X.

Motivations

- Simulation methods (Monte Carlo) for the computation of : $\mathbb{E}[f(\mathcal{L}_K)]$, quantiles, etc...
- If K is large, the simulation cost of $\mathcal{L}_K$ is high.

Introduction

Objectives

- Propose a polynomial chaos expansion (truncated to order $N \in \mathbb{N}$) for the indicator and so for $\mathcal{L}_K$:

$$\mathbf{1}_{c_k \leqslant X} = \sum_{n=0}^{+\infty} \gamma_n(c_k) p_n(X) \approx \sum_{n=0}^{N} \gamma_n(c_k) p_n(X), \qquad \mathcal{L}_{K,N} = \sum_{k=1}^{K} \sum_{n=0}^{N} \gamma_n(c_k) p_n(X)$$

where $(p_n)_{n \in \mathbb{N}}$ is an Orthogonal Polynomial Set (OPS) for ν :

$$\mathbb{E}[p_n(X) p_m(X)] = \int_{\mathbb{R}} p_n(x) p_m(x) \nu(dx) = h_n \mathbf{1}_{n=m}.$$

Interest of the chaos decompositions

- Gaussian Approximation $\gamma^{\mathcal{G}}_{K,n}$ of $\sum_{k=1}^{K} \gamma_n(c_k)$ which satisfies a **CLT** when c_k are independent r.v. and $K \to \infty$:

$$\mathcal{L}^{\mathcal{G}}_{K,N} = \sum_{n=0}^{N} \gamma^{\mathcal{G}}_{K} p_n(X)$$

with $(\gamma^{\mathcal{G}}_{K,n})_{n=0,\dots,N}$ a **Gaussian** vector and $N << K$. Parameters computed only **one time** and not for each simulation.

Introduction

Organisation of the talk

- Chaos decomposition of the indicator function $1_{c \leqslant X}$.
- Chaos decomposition (truncated) $\mathcal{L}_{K,N} = \sum_{k=1}^{K} \sum_{n=0}^{N} \gamma_n(c_k) p_n(X)$ of $\mathcal{L}_K = \sum_{k=1}^{K} 1_{c_k \leqslant X}$.
- Meta-model $\mathcal{L}_{K,N}^{\mathcal{G}} = \sum_{n=0}^{N} \gamma_K^{\mathcal{G}} p_n(X)$ for $\mathcal{L}_K$ in the Gaussian Copula model

Outline

Indicator function

Chaos decomposition

- Consider a probability measure ν.
- Condition of existence of the OPS $(p_n)_{n\in\mathbb{N}}$ for ν : There exists $\kappa > 0$ such that

$$\int_{\mathbb{R}} e^{\kappa|x|}\nu(dx) < +\infty,$$

  Under this condition
  - $(p_n)_{n\in\mathbb{N}}$ is orthogonal for ν
  - $h_n := \|p_n\|^2_{L^2(\mathbb{R},\nu)} = \int_{\mathbb{R}} p_n(x)p_n(x)\nu(dx) < +\infty,$
  - $(p_n)_{n\in\mathbb{N}}$ is dense in $L^2(\mathbb{R},\nu)$
- For every $f \in L^2(\nu)$, the following Polynomial Chaos Expansion holds

$$f \overset{L^2(\nu)}{=} \sum_{n=0}^{+\infty} \gamma_n^f p_n, \qquad \gamma_n^f = \frac{\langle f, p_n\rangle_{L^2(\nu)}}{\|p_n\|^2_{L^2(\nu)}}.$$

Indicator function

Preliminaries

- Framework : c deterministic in $\mathbb{R}$ and X random variable in $\mathbb{R}$ with law ν such that $\nu(dx) = w(x)\mathbf{1}_{I_w}(x)dx$. **Classical Orthogonal Polynomial Set** :

COPS	Hermite	Laguerre	Jacobi	Legendre
Param.		$\alpha > -1$	$\alpha, \beta > -1$	$\alpha = \beta = 0$
ν	$\mathcal{N}(0,1)$	Gamma$(\alpha+1,1)$	$1 - 2$Beta$(\alpha+1, \beta+1)$	$\mathcal{U}(-1,1)$
$p_n(x)$	$\mathrm{He}_n(x)$	$\mathrm{L}_n(x)$	$\mathrm{J}_n(x)$	$\mathrm{Le}_n(x)$
I_w	$(-\infty, +\infty)$	$(0, +\infty)$	$(-1, 1)$	$(-1, 1)$
$w(x)$	$\frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}$	$\frac{x^\alpha e^{-x}}{\Gamma_{\alpha+1}}$	$\frac{(1-x)^\alpha (1+x)^\beta}{2^{\alpha+\beta+1}\mathrm{B}_{\alpha+1,\beta+1}}$	$\frac{1}{2}$

- Objective : For $c \in I_w$, study the PCE

$$\mathbf{1}_{c \leqslant X} = \sum_{n=0}^{\infty} \gamma_n(c) p_n(X).$$

Indicator function

Choice of polynomial basis

Let X with law ν and denote $\Psi_\nu(c) = \mathbb{P}(X \leqslant c)$. Let ν_1 be a COPS measure, we denote $T_{\nu,\nu_1} = \Psi_{\nu_1}^{-1} \circ \Psi_\nu$, and then $X_{\nu_1} := T_{\nu,\nu_1}(X) \sim \nu_1$

$$\mathbf{1}_{c \leq X} = \mathbf{1}_{T_{\nu,\nu_1}(c) \leq X_{\nu_1}}$$

Denoting $q = \Psi_\nu(c)$,

$$\mathbf{1}_{\Psi_\nu^{-1}(q) \leq X} = \mathbf{1}_{\Psi_{\nu_1}^{-1}(q) \leq X_{\nu_1}}$$

Indicator function

Proposition - PCE for the indicator function of a COPS

Let $(p_n)_{n\in\mathbb{N}}$ be a COPS w.r.t. $\nu(dx) = w(x)\mathbf{1}_{I_w}(x)dx$, and let $c \in I_w$. Then,

$$\forall x \in I_w \setminus \{c\}, \quad \mathbf{1}_{c\leq x} = \sum_{n=0}^{+\infty} \gamma_n(c)p_n(x),$$

with $\gamma_0(c) = \nu((c,+\infty))$ for every $n \geq 1$,

$$\gamma_n(c) = \frac{e^{-\frac{c^2}{2}}\mathrm{He}_{n-1}(c)}{n!\sqrt{2\pi}} \qquad \text{if } \nu = \mathcal{N}(0,1)$$

$$\gamma_n(c) = -\frac{(n-1)!}{\Gamma_{n+\alpha+1}}c^{\alpha+1}e^{-c}\mathrm{L}_{n-1}^{(\alpha+1)}(c) \qquad \text{if } \nu = \mathrm{Gamma}(\alpha+1,1)$$

$$\gamma_n(c) = C_n(\alpha,\beta)(1-c)^{\alpha+1}(1+c)^{\beta+1}\mathrm{J}_{n-1}^{(\alpha+1,\beta+1)}(c) \qquad \text{if } \nu = 1-2\mathrm{Beta}(\alpha+1,\beta+1)$$

with $C_n(\alpha,\beta) = \frac{(2n+\alpha+\beta+1)\Gamma_{n+\alpha+\beta+1}(n-1)!}{2^{\alpha+\beta+2}\Gamma_{n+\alpha+1}\Gamma_{n+\beta+1}}$.

Idea of the proof : Use the Rodrigue's formula $p_n(x) = \frac{1}{\kappa_n w(x)}\frac{\mathrm{d}^n}{\mathrm{d}x^n}[F(x)^n w(x)]$ where F (pol. of degree ≤ 2) and κ_n are known explicitly for COPS.

Indicator function

For OPS, $p_{n+1}(x) = (A_n x + B_n)\, p_n(x) - C_n p_{n-1}(x), \quad n \in \mathbb{N}^*$ where A_n, B_n, C_n are constants with $A_n > 0$, $C_n > 0$.

Proposition - Reccurrence relation for the Chaos coefficient

Let $(p_n)_{n\in\mathbb{N}}$ be a COPS w.r.t. $\nu\,(dx) = w(x)1_{I_w}(x)dx$ and $c \in I_w = (a, b)$, then

$$\gamma_{n+2}(c) = (B_{n+1} + A_{n+1}c)\,\frac{h_{n+1}}{h_{n+2}}\gamma_{n+1}(c) - C_{n+1}\,\frac{h_n}{h_{n+2}}\gamma_n(c) + \frac{h_{n+1}}{h_{n+2}}A_{n+1}\int_c^b \gamma_{n+1}(x)\mathrm{d}x.$$

Moreover, there exist $D_{\nu,n}, E_{\nu,n}, F_{\nu,n}$ ($D_{\nu,n}, F_{\nu,n} \geq 0$) such that

$$\gamma_{n+2}(c) = (D_{\nu,n} + cE_{\nu,n})\gamma_{n+1}(c) - F_{\nu,n}\gamma_n(c),$$

Indicator function - L_2 truncation error

Let us define

$$\mathcal{E}_N(c) := \mathbb{E}\left[\left|\mathbf{1}_{c\leq X} - \sum_{n=0}^{N}\gamma_n(c)p_n(X)\right|^2\right]^{\frac{1}{2}} = \left|\sum_{n=N+1}^{+\infty} h_n\gamma_n(c)^2\right|^{\frac{1}{2}}.$$

If ν is either $\mathcal{N}(0,1)$ (Hermite), Gamma$(\alpha+1,1)$ (Laguerre) or $1-2$Beta$(\alpha+1,\beta+1)$ (Jacobi)

$$\mathrm{E}_{\nu,N}(c) := e^{-\frac{c^2}{4}}\begin{cases} N^{-\frac{1}{4}}, & |c|\in[0,(4N)^{\frac{1}{2}}],\\ \dfrac{N^{1/24}}{|c|^{\frac{1}{4}}}, & |c|\in[(2N)^{\frac{1}{2}}, N^{\frac{39}{14}}],\\ \dfrac{|c|^{1/3}}{N^{19/12}}, & |c|\in[N^{\frac{39}{14}},+\infty). \end{cases}$$

$$\mathrm{E}_{\nu,N}(c) := e^{-\frac{c}{2}}\begin{cases} \dfrac{c^{\frac{\alpha}{2}+\frac{1}{4}}}{N^{1/4}}, & c\in(0,(4-\eta)N],\\ \dfrac{c^{\frac{\alpha}{2}+\frac{5}{12}}}{N^{1/4}}, & c\in[2N,+\infty). \end{cases}$$

$$\mathrm{E}_{\nu,N}(c) := \left(\frac{2+\sqrt{(\alpha+1)^2+(\beta+1)^2}}{2^{\alpha+\beta}\mathrm{B}_{\alpha+1,\beta+1}}\right)^{\frac{1}{2}}(1-c)^{\frac{\alpha}{2}+\frac{1}{4}}(1+c)^{\frac{\beta}{2}+\frac{1}{4}}N^{-\frac{1}{2}}, \quad c\in[-1,1].$$

Indicator function - L_2 truncation error

Proposition - L^2 approximation

Let $(p_n)_{n\in\mathbb{N}}$ be a COPS w.r.t. $\nu(dx) = w(x)\mathbf{1}_{I_w}(x)dx$, and $c \in I_w$, then there exists $A \geq 0$ such that

$$\mathcal{E}_N(c) = \mathbb{E}\left[\left|\mathbf{1}_{c\leq X} - \sum_{n=0}^{N} \gamma_n(c)p_n(X)\right|^2\right]^{\frac{1}{2}} \leq A\mathrm{E}_{\nu,N}(c).$$

Comparison of polynomial basis

For $q = \Psi_\nu(c)$, and ν_1 a COPS measure, and $X_{\nu_1} = \Psi_{\nu_1}^{-1} \circ \Psi_\nu(X)$,

$$\mathbf{1}_{\Psi_\nu^{-1}(q)\leq X} = \mathbf{1}_{\Psi_{\nu_1}^{-1}(q)\leq X_{\nu_1}}$$

To compare two COPS, we compare their L_2 error, $\mathcal{E}_{N,\nu}(\Psi_\nu^{-1}(q))$ and $\mathcal{E}_{N,\nu_1}(\Psi_{\nu_1}^{-1}(q))$ for $q \in (0,1)$.

Indicator function - L_2 truncation error

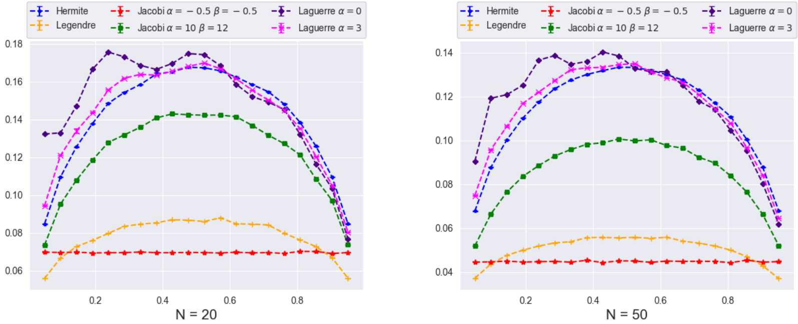

FIGURE – L^2 error $\mathcal{E}_N(c)$ for $c = \Psi_\nu^{-1}(q)$ for the COPS w.r.t. $q \in (0,1)$ and for $N \in \{20, 50\}$.

Indicator function - L_2 truncation error

Proposition - Extreme quantile

Let $N \in \mathbb{N}$ fixed. Take $c(q) = \Psi_\nu^{-1}(q)$. We have, for the Hermite polynomials

$$\mathrm{E}_{\nu,N}(c(q)) \sim \begin{cases} C\,|\ln q|^{\frac{5}{12}}\, q^{\frac{1}{2}}, & (q \to 0^+), \\ C\,|\ln(1-q)|^{\frac{5}{12}}\,(1-q)^{\frac{1}{2}}, & (q \to 1^-). \end{cases}$$

For the Laguerre polynomials,

$$\mathrm{E}_{\nu,N}(c(q)) \sim \begin{cases} Cq^{\frac{2\alpha+1}{4(\alpha+1)}}, & (q \to 0^+), \\ C\,|\ln(1-q)|^{\frac{5}{12}}\,(1-q)^{\frac{1}{2}}, & (q \to 1^-). \end{cases}$$

For the Jacobi polynomials,

$$\mathrm{E}_{\nu,N}(c(q)) \sim \begin{cases} Cq^{\frac{2\beta+1}{4(\alpha+1)}}, & (q \to 0^+), \\ C(1-q)^{\frac{2\alpha+1}{4(\beta+1)}}, & (q \to 1^-). \end{cases}$$

Indicator function - Extreme quantile

Best polynomials when $q \to 0^+$

- Jacobi when $\frac{2\beta_J+1}{4(\alpha_J+1)} > \frac{1}{2} \Leftrightarrow \alpha_J < \beta_J - \frac{1}{2}$,
- Hermite polynomials when $\alpha_J > \beta_J - \frac{1}{2}$,
- both Hermite and Jacobi when $\alpha_J = \beta_J - \frac{1}{2}$.

Worst L^2 estimates for Jacobi when $\frac{2\beta_J+1}{\alpha_J+1} < \frac{2\alpha_{La}+1}{\alpha_{La}+1}$.

Best polynomials when $q \to 1^-$

- Jacobi when $\frac{2\alpha_J+1}{4(\beta_J+1)} > \frac{1}{2} \Leftrightarrow \beta_J < \alpha_J - \frac{1}{2}$,
- Hermite and Laguerre when $\beta_J > \alpha_J - \frac{1}{2}$,
- All polynomials when $\beta_J = \alpha_J - \frac{1}{2}$ (up to some logarithmic error terms).

Worst L^2 estimates for Jacobi when $\alpha_J < \beta_J + \frac{1}{2}$.

Indicator function - L_2 truncation error

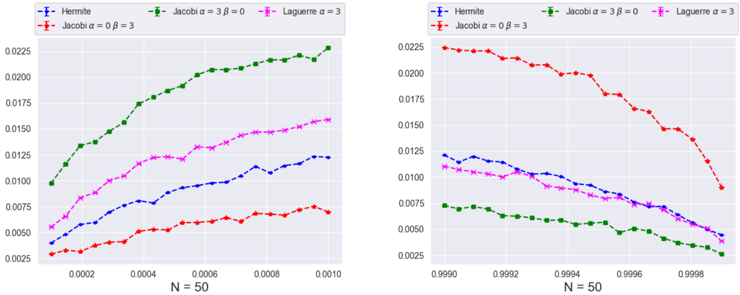

FIGURE – L^2 error for the COPS w.r.t. small quantiles $q \in \left(10^{-4}, 10^{-3}\right)$ (left) and large quantiles $q \in \left(1 - 10^{-3}, 1 - 10^{-4}\right)$ (right) for a fixed $N = 50$.

Outline

Sum of Indicator functions - L_2 truncation error

We consider (introduction of deterministic weights l_k)

$$\mathcal{L}_K = \sum_{k=1}^{K} l_k 1_{c_k \leqslant X}, \qquad \mathcal{L}_{K,N} = \sum_{k=1}^{K} l_k \sum_{n=0}^{N} \gamma_n(c_k) p_n(X)$$

and $\mathcal{E}_N(c) = \mathbb{E}\left[\left|\mathbf{1}_{c \leq X} - \sum_{n=0}^{N} \gamma_n(c) p_n(X)\right|^2\right]^{\frac{1}{2}}$.

Proposition - L_2 truncation error

Assume $(c_k)_{k=1,\dots,K}$ is a sequence of independent random varibales taking values in I_w. Then

$$\mathbb{E}\left[\left(\mathcal{L}_K - \mathcal{L}_{K,N}\right)^2\right]^{\frac{1}{2}} \leq \sum_{k=1}^{K} l_k^2 \mathbb{E}\left[\mathcal{E}_N(c_k)\right],$$

Outline

Meta-model - CLT

Recall $\mathcal{L}_{K,N} = \sum_{n=0}^{N} \sum_{k=1}^{K} l_k \gamma_n(c_k) p_n(X)$.

Consider the random vector $\epsilon_K := (\epsilon_{K,n})_{n=0,\dots,N}$ for N fixed, such that

$$\epsilon_{K,n} = \sum_{k=1}^{K} l_k \gamma_n(c_k), \qquad \text{so} \qquad \mathcal{L}_{K,N} = \sum_{n=0}^{N} \epsilon_{K,n} p_n(X).$$

and

$$\mu_{K,n} = \mathbb{E}\left[\epsilon_{K,n}\right] = \sum_{k=1}^{K} l_k \mathbb{E}\left[\gamma_n(c_k)\right],$$

$$\Sigma_{K,n,m} = \text{Cov}\left(\epsilon_{K,n}, \epsilon_{K,m}\right) = \sum_{k=1}^{K} l_k^2 \text{Cov}\left(\gamma_n(c_k), \gamma_m(c_k)\right).$$

Meta-model - CLT

Proposition - CLT

Assume that

$$\|\Sigma_K^{-1}\| \sup_{1\leq k\leq K} l_k^2 \xrightarrow[K\to+\infty]{} 0 \quad \text{with} \quad \|\Sigma_K^{-1}\| := \sup_{x\in\mathbb{R}^{N+1}, x\neq 0} \frac{|\Sigma_K^{-1}x|}{|x|}.$$

Then, the vector ϵ_K satisfies the following Central Limit Theorem

$$\Sigma_K^{-1/2}(\epsilon_K - \mu_K) \xrightarrow[K\to\infty]{d} \mathcal{N}(0, \mathrm{Id}_{N+1}).$$

We introduce the Gaussian approximation for ϵ_K :

$$\epsilon_K^{\mathrm{G}} = \left(\epsilon_{K,n}^{\mathrm{G}} : n = 0, \dots, N\right) \sim \mathcal{N}\left(\mu_K, \Sigma_K\right),$$

leading to the meta-model

$$\mathcal{L}_{K,N}^{\mathrm{G}} = \sum_{n=0}^{N} \epsilon_{K,n}^{\mathrm{G}} p_n\left(X\right).$$

Application to a Gaussien Copula model

Consider

- $K = 5 \times 10^5$.
- Let $\rho = 0.1, p = 0.01$, and

$$c_k = \frac{\sqrt{1-\rho^2}}{\rho}\varepsilon_k - \frac{1}{\rho}\Psi^{-1}_{\mathcal{N}(0,1)}(p).$$

- $X, \varepsilon_1, \ldots, \varepsilon_K$, i.i.d. $\mathcal{N}(0,1)$.
- $l_k = 1/\sqrt{k}$.

The Lindeberg condition holds as $\sup_{1\leq k\leq K} l_k^2 = 1$ and $\sum_{k=1}^{+\infty} \frac{1}{k} = +\infty$.

Simulation cost : About 17 times faster (when $N = 3$) than a naive Monte Carlo.

Application to a Gaussien Copula model

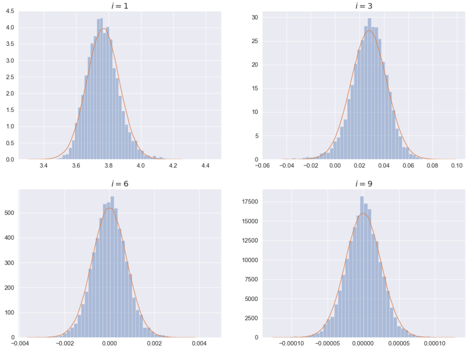

FIGURE – Histograms of $\epsilon_{K,i}$ with the p.d.f of $\mathcal{N}\left(\mathbb{E}\left[\epsilon_{K,i}\right], \text{Var}\left(\epsilon_{K,i}\right)\right)$ for $N \in \{1, 3, 6, 9\}$

Application to a Gaussien Copula model

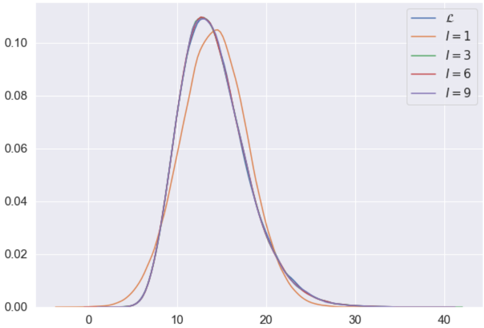

FIGURE – Distributions of $\mathcal{L}^{\mathrm{G}}_{K,N}$ for $N = I \in \{1, 3, 6, 9\}$ and $\mathcal{L}_K = \mathcal{L}$

Application to a Gaussien Copula model

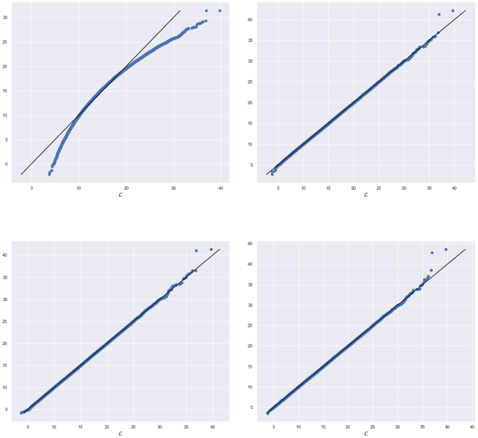

FIGURE – Q-Q plots of $\mathcal{L}^{\mathrm{G}}_{K,N}$ w.r.t. $\mathcal{L}_K$ for $N = 1, 3, 6, 9$ and 10^5 points.

Thank you for your attention !

- With Florian Bourgey and Emmanuel Gobet. *Meta-model of a large credit risk portfolio in the Gaussian copula model*, SIAM Journal on Financial Mathematics, 2020.
- With Florian Bourgey and Emmanuel Gobet. *A comparison study of polynomial-type chaos expansions for indicator functions*, 2021, Hal.