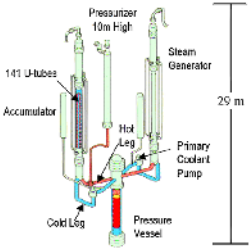
An information geometry approach for robustness analysis in uncertainty quantification of computer codes

Clément Gauchy

Supervisors: Jerôme STENGER, Roman SUEUR & Bertrand IOOSS

EDF R&D PRISME

CATHARE SIMULATION



Variables	Description
X1	Minimal film temperature
X2	Interfacial friction
X3	Interfacial friction
X4	Interfacial friction
X5	Interfacial friction
X6	Interfacial friction
X7	Critical flowrate

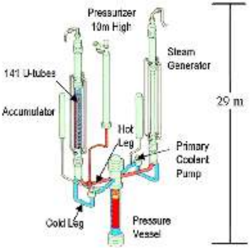
CATHARE code simulates a thermohydraulic transient during a specific accident.

Parameters values are tainted with uncertainties.

Parameters values are tainted with uncertainties. Input parameters are then modeled as random variables.

Parameters values are tainted with uncertainties. Input parameters are then modeled as random variables.

Hypothesis : Suppose X_i mutually independent.



Input variables	Probability distribution
X1	Uniform $\mathcal{U}([-44.9, 63.5])$
X2	Truncated Log Normal $\mathcal{LN}(0, 0.76)$ on $[0.1, 10]$
X3	Truncated Log Normal $\mathcal{LN}(0, 0.76)$ on $[0.1, 10]$
X4	Truncated Log Normal $\mathcal{LN}(0, 0.76)$ on $[0.1, 10]$
X5	Truncated Log Normal $\mathcal{LN}(0, 0.76)$ on $[0.1, 10]$
X6	Truncated Log Normal $\mathcal{LN}(-0.1, 0.45)$ on $[0.23, 3.45]$
X7	Truncated Normal $\mathcal{N}(6.4, 4.27)$ on $[0, 12.8]$

Experimental data and expert judgement help choosing probability distributions.

Internship objectives

- Input parameters probability distribution is a strong prior in risk assessment studies.

Internship objectives

- Input parameters probability distribution is a strong prior in risk assessment studies.
- The impact on the quantity of interest Y by a **probability density perturbation** has to be assessed

Internship objectives

- Input parameters probability distribution is a strong prior in risk assessment studies.

- The impact on the quantity of interest Y by a **probability density perturbation** has to be assessed

- The initial density f_i of variable X_i is **perturbed** into $f_{i\delta}$

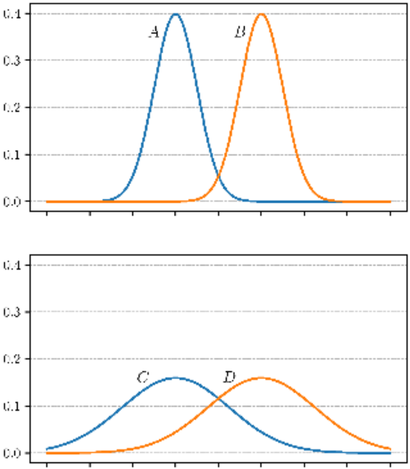
- Input parameters probability distribution is a strong prior in risk assessment studies.

- The impact on the quantity of interest Y by a **probability density perturbation** has to be assessed

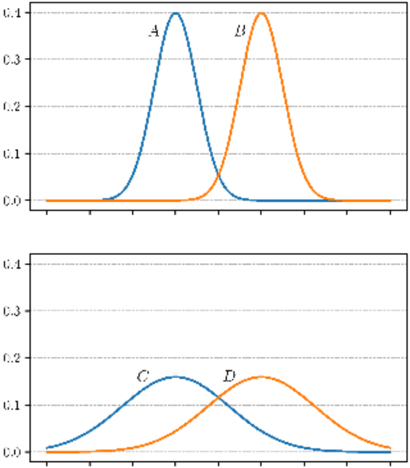
- The initial density f_i of variable X_i is **perturbed** into $f_{i\delta}$
- Main issue : How to define such a perturbation ?

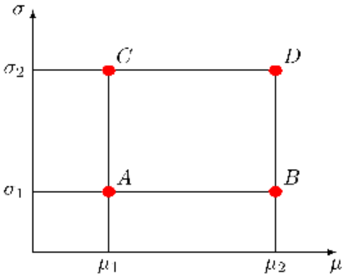
Perturbation example

Are A and B further apart than C and D?



Are A and B further apart than C and D?





For the 2-Wasserstein distance they are at equal distance. It computes $W_2(A, B) = W_2(C, D) = \sqrt{(\mu_2 - \mu_1)^2}$

VARIATIONAL APPROACH

- Pertubed density $f_{i\delta}$ is defined by minimizing the application $q \to KL(q||f_i)$ with moments constraints.
- Example : $\int x f_{i\delta}(x)dx = \delta_i, \ \int x^2 f_{i\delta}(x)dx = \delta_i$

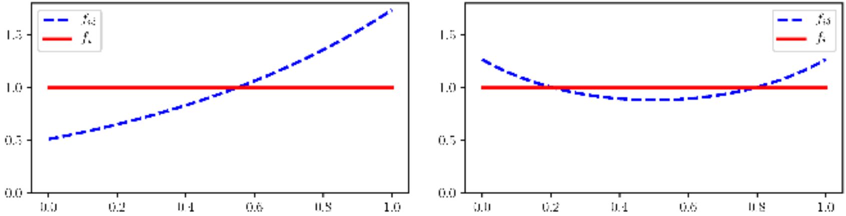


FIGURE 1 – Mean (left figure) and variance (right figure) perturbation of $\mathcal{U}(0,1)$ density.

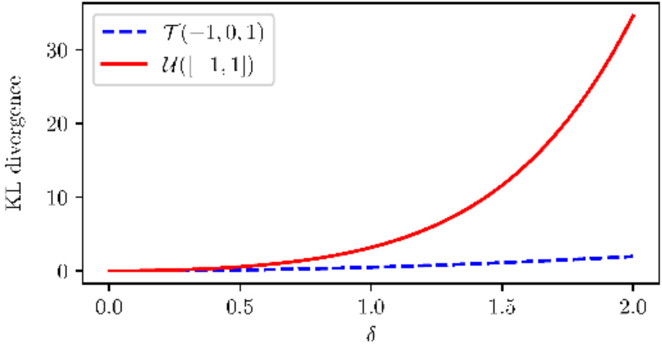
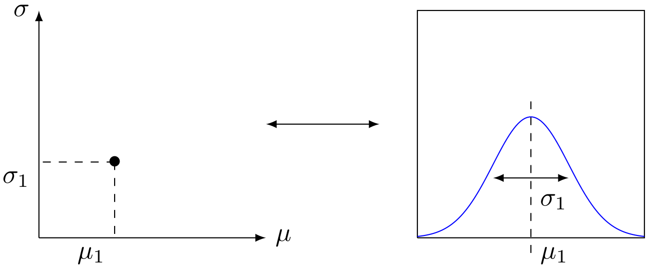
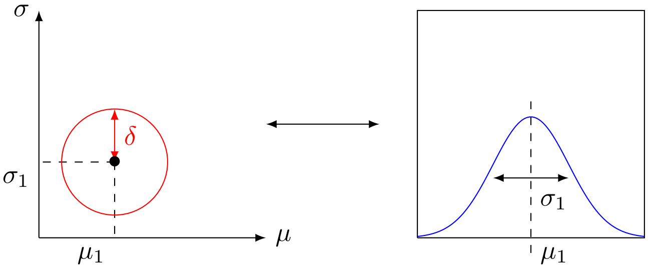


FIGURE 2 – KL divergence between initial density $\mathcal{T}(-1,0,1)$ and $\mathcal{U}(-1,1)$ and their associated perturbed density

- Only parametrical models are considered $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$
- Example : Gaussian distributions $\{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^{+*}\}$



- Only parametrical models are considered $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$
- Example : Gaussian distributions $\{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^{+*}\}$



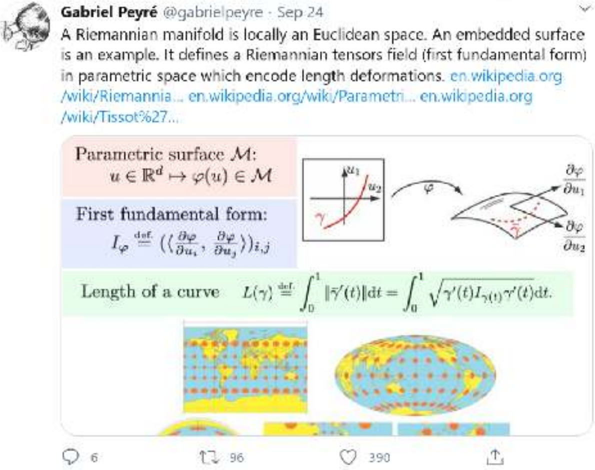
A NEW GEOMETRICAL APPROACH

- Fisher information endows statistical models with a remarkable geometric structure.

- Let $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ a parametrical statistical model

- Let $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ a parametrical statistical model
- A Riemannian manifold is defined on $\mathcal{S}$

- Let $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ a parametrical statistical model
- A Riemannian manifold is defined on $\mathcal{S}$
- To each point θ is associated a tangent space $T_\theta \mathcal{S} \simeq \mathbb{R}^d$



- Let $\mathcal{S} = \{f_\theta, \theta \in \Theta \subset \mathbb{R}^d\}$ a parametrical statistical model
- A Riemannian manifold is defined on $\mathcal{S}$
- To each point θ is associated a tangent space $T_\theta\mathcal{S} \simeq \mathbb{R}^d$
- The latter scalar product is defined in $T_\theta\mathcal{S}$:

$$\forall u, v \in T_\theta\mathcal{S},\ \langle u, v\rangle_\theta = u^T I(\theta) v\ ,$$

where $I(\theta)$ is the Fisher information matrix evaluated in θ.

$$I(\theta) = \mathbb{E}\Big[(\nabla_\theta \log f_\theta(X))(\nabla_\theta \log f_\theta(X))^T\Big]$$

Fisher information is a key feature in asymptotic statistics.

Borne de Cramer Rao :

Let $\hat{\theta}$ be an unbiaised estimator of θ, then

$$V(\hat{\theta}) \geq I(\theta)^{-1} \ , \tag{1}$$

where $V(\hat{\theta})$ is the covariance matrix of the estimator.

- The scalar product $\langle ., . \rangle_\theta$ could define an implicit distance

- The scalar product $\langle ., . \rangle_\theta$ could define an implicit distance
- This distance is called **Fisher distance**.

- The scalar product $\langle ., . \rangle_\theta$ could define an implicit distance
- This distance is called **Fisher distance**.
- Let $t \to q(t)$ be a $\mathcal{C}^1$ path in Θ, its length is defined by :

$$l(q) := \int_0^1 \sqrt{\langle \dot{q}(t), \dot{q}(t) \rangle_{q(t)}} dt \ ,$$

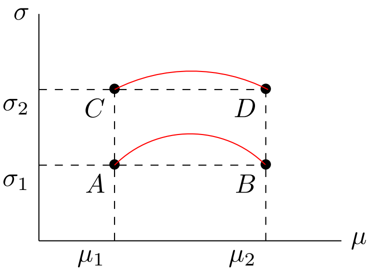
- the Fisher distance f_{θ_1} and f_{θ_2} is defined by :

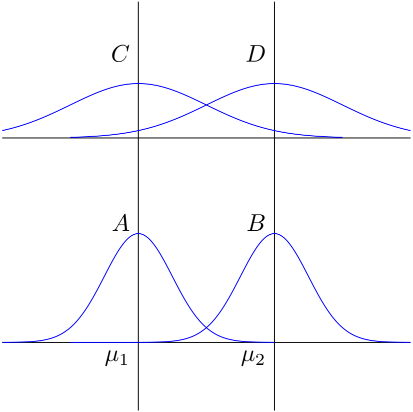
$$d_F(f_{\theta_1}, f_{\theta_2}) = \inf_{q \in \mathcal{C}(\theta_1, \theta_2)} l(q) \ ,$$

where $\mathcal{C}(\theta_1, \theta_2)$ is the set of $\mathcal{C}^1$ path between θ_1 and θ_2.

Interpretation (1/3)

Consider the space $\{\mathcal{N}(\mu, \sigma^2), (\mu, \sigma) \in \mathbb{R} \times \mathbb{R}^{+*}\}$





- Let $X_1, ..., X_n$ a n sized sample f_θ distributed.
- We denote by $\widehat{\theta}_n$ the maximum likelihood estimator

Central limit theorem :

$$\sqrt{n}(\widehat{\theta}_n - \theta) \xrightarrow{\mathcal{L}} \mathcal{N}(0, I(\theta)^{-1}) \ , \tag{2}$$

- The probability density of $\widehat{\theta}_n$ is :

$$p(\widehat{\theta}_n, \theta) \propto e^{-\frac{n}{2}\delta\theta^T I(\theta)\delta\theta}$$

The Fisher distance between two distributions f_θ and $f_{\theta'}$ represents the separability of the two distributions by a finite sample of independent observations sampled for the f_θ distribution.

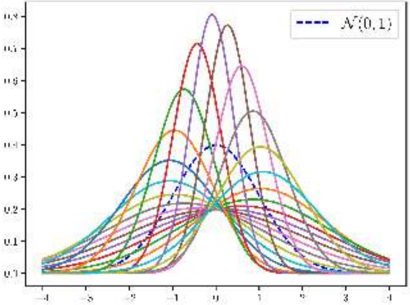


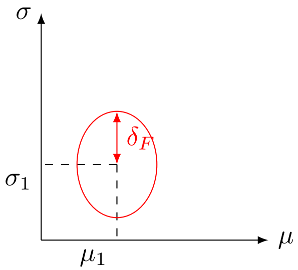
Figure 3 – All solid line distributions are at the same Fisher distance from the dashed one. They are located on the Fisher sphere of radius 1 centered in $\mathcal{N}(0,1)$

- All distributions on the Fisher sphere are equivalent perturbed densities from f_{θ_0}.

Geodesics computation

- All distributions on the Fisher sphere are equivalent perturbed densities from f_{θ_0}.
- We need to compute all geodesics such that $q(0) = \theta_0$ and $d_F(q(0), q(1)) = \delta$ for $\delta \in \mathbb{R}^+$ fixed.

- All distributions on the Fisher sphere are equivalent perturbed densities from f_{θ_0}.
- We need to compute all geodesics such that $q(0) = \theta_0$ and $d_F(q(0), q(1)) = \delta$ for $\delta \in \mathbb{R}^+$ fixed.



Let $t \to q(t)$ a path, with $p = I(q)\dot{q}$, the hamiltonian is written :

$$H(p,q) = \frac{1}{2} p^T I^{-1}(q) p \ .$$

If $t \to q(t)$ is a geodesic, then the function $t \to H(p(t), q(t))$ is constant.

A geodesic statisfies the following system of ordinary differential equations :

$$\begin{cases} \dot{q} = \frac{\partial H}{\partial p} \\ \dot{p} = -\frac{\partial H}{\partial q} \end{cases} \tag{3}$$

- The conservation of hamiltonian gives us the initial condition in "speed" $p(0)$ knowing that $d_F(q(0), q(1)) = \delta$

- The conservation of hamiltonian gives us the initial condition in "speed" $p(0)$ knowing that $d_F(q(0), q(1)) = \delta$
- With $(q(0), p(0))$ defined, the ODE system (3) has an unique solution thanks to Cauchy's theorem

- The conservation of hamiltonian gives us the initial condition in "speed" $p(0)$ knowing that $d_F(q(0), q(1)) = \delta$
- With $(q(0), p(0))$ defined, the ODE system (3) has an unique solution thanks to Cauchy's theorem
- Geodesics are computed using numerical methods.

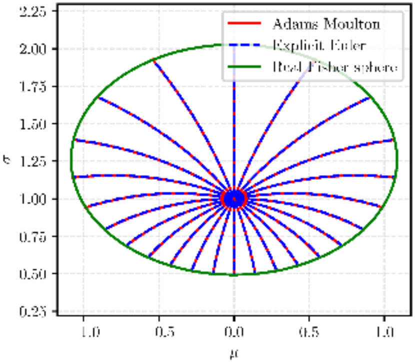


Figure 4 – Fisher sphere $\delta = 1$ - Coordinate space

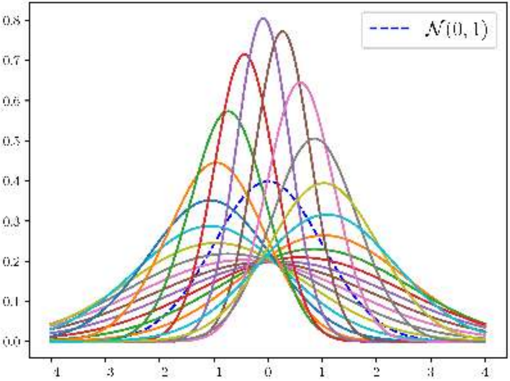


FIGURE 5 – Fisher sphere $\delta = 1$ - densities space

- We aim to measure the impact of density perturbation of input X_i to Y

- We aim to measure the impact of density perturbation of input X_i to Y
- We define the quantile-PLI (*Perturbated Law Index*) $S_{i\delta}$ by :

$$S_{i\delta} = \frac{q^{\alpha}_{i\delta} - q^{\alpha}}{q^{\alpha}}$$

- q^{α} and $q^{\alpha}_{i\delta}$ are respectively the quantiles of level α of Y with X_i distributed respectively according to f_i and $f_{i\delta}$

- We aim to measure the impact of density perturbation of input X_i to Y
- We define the quantile-PLI (*Perturbated Law Index*) $S_{i\delta}$ by :

$$S_{i\delta} = \frac{q_{i\delta}^{\alpha} - q^{\alpha}}{q^{\alpha}}$$

- q^{α} and $q_{i\delta}^{\alpha}$ are respectively the quantiles of level α of Y with X_i distributed respectively according to f_i and $f_{i\delta}$
- We obtain the **minimum** and the **maximum** of $S_{i\delta}$ for $f_{i\delta}$ in the Fisher sphere of radius δ centered in f_i.

- We aim to measure the impact of density perturbation of input X_i to Y
- We define the quantile-PLI (*Perturbated Law Index*) $S_{i\delta}$ by :

$$S_{i\delta} = \frac{q_{i\delta}^{\alpha} - q^{\alpha}}{q^{\alpha}}$$

- q^{α} and $q_{i\delta}^{\alpha}$ are respectively the quantiles of level α of Y with X_i distributed respectively according to f_i and $f_{i\delta}$
- We obtain the **minimum** and the **maximum** of $S_{i\delta}$ for $f_{i\delta}$ in the Fisher sphere of radius δ centered in f_i.
- This new methodology is called OF-PLI (*Optimal Fisher based PLI*).

PLI ESTIMATION

- Industrial simulation code are often time-expensive.

PLI ESTIMATION

- Industrial simulation code are often time-expensive.
- We want to estimate the PLI without resampling X_i from the perturbed density.

PLI ESTIMATION

- Industrial simulation code are often time-expensive.
- We want to estimate the PLI without resampling X_i from the perturbed density.
- We consider a sample $(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(N)})$ with X_i sampled from f_i and a simulation code G :

$$\hat{F}_{i\delta}(t) = \frac{\sum_{n=1}^{N} \frac{f_{i\delta}(X_i^{(n)})}{f_i(X_i^{(n)})} \mathbb{1}_{(G(\mathbf{X}^{(n)})<t)}}{\sum_{n=1}^{N} \frac{f_{i\delta}(X_i^{(n)})}{f_i(X_i^{(n)})}}$$

This is the reverse importance sampling (RIS) estimator of the cdf of $G(\mathbf{X})$

PLI ESTIMATION

- Industrial simulation code are often time-expensive.
- We want to estimate the PLI without resampling X_i from the perturbed density.
- We consider a sample $(\mathbf{X}^{(1)}, ..., \mathbf{X}^{(N)})$ with X_i sampled from f_i and a simulation code G :

$$\hat{F}_{i\delta}(t) = \frac{\sum_{n=1}^{N} \frac{f_{i\delta}(X_i^{(n)})}{f_i(X_i^{(n)})} \mathbb{1}_{(G(\mathbf{X}^{(n)})<t)}}{\sum_{n=1}^{N} \frac{f_{i\delta}(X_i^{(n)})}{f_i(X_i^{(n)})}}$$

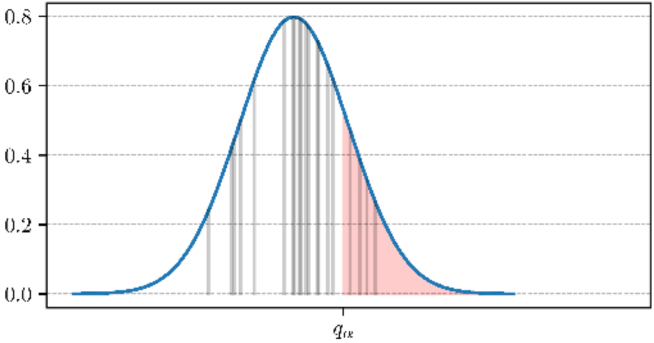
  This is the reverse importance sampling (RIS) estimator of the cdf of $G(\mathbf{X})$
- the perturbed quantile $q_{i\delta}^{\alpha}$ is estimated with the empirical quantile of $\hat{F}_{i\delta}$.

Theoretical results - PLI estimation

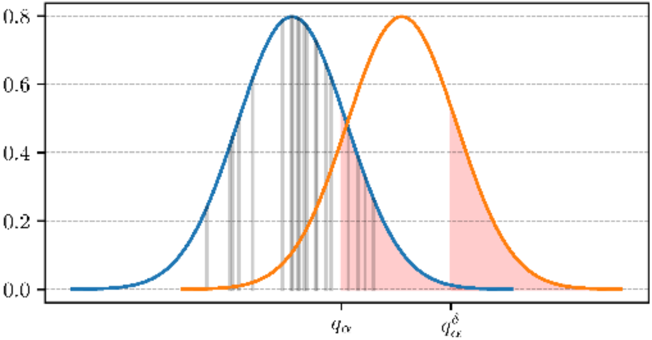
- Self normalized cdf estimator $\hat{F}_{i\delta}(t)$ is used because it is bounded. Moreover, it possess better asymptotic properties.
- The estimator $\hat{S}_{i\delta} = \frac{\hat{q}^{\alpha}_{i\delta} - \hat{q}^{\alpha}}{\hat{q}^{\alpha}}$ built verify a CLT.

- Self normalized cdf estimator $\hat{F}_{i\delta}(t)$ is used because it is bounded. Moreover, it possess better asymptotic properties.
- The estimator $\hat{S}_{i\delta} = \frac{\hat{q}^{\alpha}_{i\delta} - \hat{q}^{\alpha}}{\hat{q}^{\alpha}}$ built verify a CLT.
- Main hypothesis for the CLT : $\mathbb{E}\left[\left(\frac{f_{i\delta}(X)}{f_i(X)}\right)^2\right] < +\infty$

- Empirical criterion for choice of δ_{max} : Minimal number of $G(\mathbf{X}^{(i)})$'s values greater or lesser than the perturbed quantile.
- Due to the RIS estimator, we have to take care of the likelihood ratio value



- Empirical criterion for choice of δ_{max} : Minimal number of $G(\mathbf{X}^{(i)})$'s values greater or lesser than the perturbed quantile.
- Due to the RIS estimator, we have to take care of the likelihood ratio value

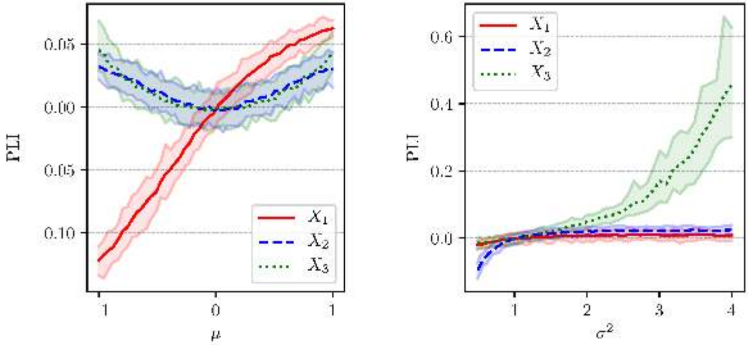


A toy case : Ishigami function

- We take 3 independent random variables (X_1, X_2, X_3) with a standard Gaussian distribution $\mathcal{N}(0,1)$.
- The output variable is the analytical function

$$G(x_1, x_2, x_3) = \sin(x_1) + 7\sin(x_2)^2 + 0.1x_3^4\sin(x_1)\ . \tag{4}$$

Numerical results : PLI with Kullback-Leibler minimization



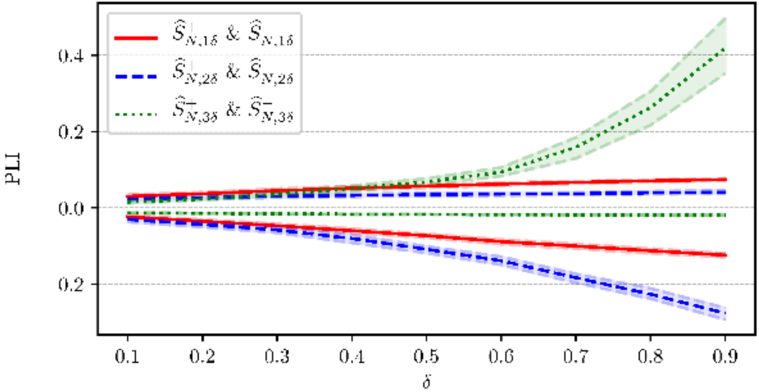


FIGURE 7 – OF-PLI for the Ishigami function with a 100 points grid on the Fisher sphere.

- OF-PLI computation for the flood model, quantifying the flooding risk of industrial facilities.

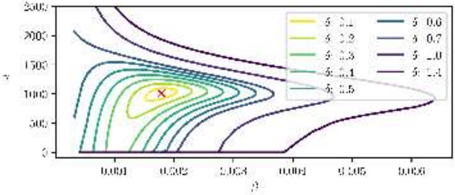
Flood model

Variable n°	Name	Description	Probability distribution	Admissible values
1	Q	Maximal annual flowrate	Gumbel $\mathcal{G}(1013, 558)$	$[500, 3000]$
2	K_s	Strickler coefficient	Normal $\mathcal{N}(30, 7.5)$	$[15, +\infty]$
3	Z_v	Upstream level of the river	Triangular $\mathcal{T}(50)$	$[49, 51]$
4	Z_m	Downstream level of the river	Triangular $\mathcal{T}(55)$	$[54, 56]$

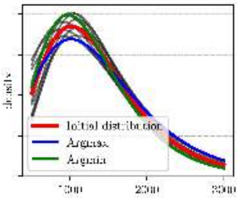
Input parameters of the flood model with their associated probability distribution

- We denote H the maximal annual water level.

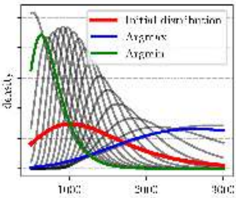
$$H = \left(\frac{Q}{300 K_s \sqrt{2.10^{-4}(Z_m - Z_v)}} \right)^{0.6}.$$



(a) Fisher sphere for an increasing δ.



(b) Densities on the Fisher sphere ($\delta = 0.1$).



(c) Densities on the Fisher sphere ($\delta = 1.4$).

FIGURE 8 – Analysis of the density perturbation of the variable Q.

Numerical results for the flood model

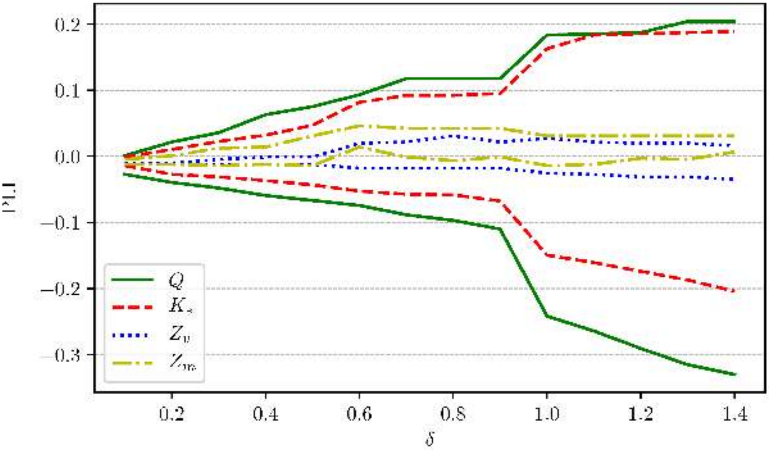


Figure 9 – OF-PLI for the flood model on 100 points on the Fisher sphere.

CODE CATHARE RESULTS

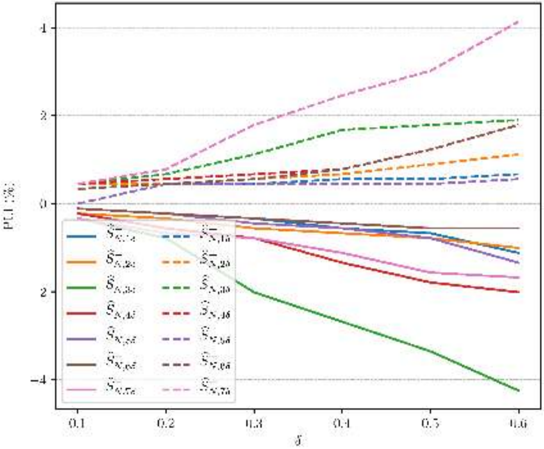


FIGURE 10 – OF-PLI for CATHARE code

Conclusion

- Definition of a new framework of density perturbation, development of a numerical solver in Python (OpenTurns inside).

Conclusion

- Definition of a new framework of density perturbation, development of a numerical solver in Python (OpenTurns inside).
- Theoretical results

Conclusion

- Definition of a new framework of density perturbation, development of a numerical solver in Python (OpenTurns inside).
- Theoretical results
- Writing of a scientific article (Arxiv link : `https://arxiv.org/pdf/2008.03060.pdf`)
- Perspectives : simultaneous pertubation of several density of input parameters, dependent input parameters.

QUESTIONS ?

Supposons que F_i soit différentiable en q^α avec $F_i'(q^\alpha) > 0$ et $F_{i\delta}$ soit différentiable en $q_{i\delta}^\alpha$ avec $F_{i\delta}'(q_{i\delta}^\alpha) > 0$. On note $\Sigma = \begin{pmatrix} \sigma_i^2 & \tilde{\theta}_i \\ \tilde{\theta}_i & \tilde{\sigma}_{i\delta}^2 \end{pmatrix}$ tel que :

$$\sigma_i^2 = \frac{\alpha(1-\alpha)}{f_i(q^\alpha)^2} \,.$$

$$\tilde{\sigma}_{i\delta}^2 = \frac{\mathbb{E}\left[\left(\frac{f_{i\delta}(X_i)}{f_i(X_i)}\right)^2 \left(\mathbb{1}_{(G(\mathbf{X}) \leq q_{i\delta}^\alpha)} - \alpha\right)^2\right]}{f_{i\delta}(q_{i\delta}^\alpha)^2} \,.$$

$$\tilde{\theta}_i = \frac{\mathbb{E}\left[\frac{f_{i\delta}(X_i)}{f_i(X_i)} \mathbb{1}_{(G(\mathbf{X}) \leq q^\alpha)} \mathbb{1}_{(G(\mathbf{X}) \leq q_{i\delta}^\alpha)}\right] - \alpha \mathbb{E}[\mathbb{1}_{(G(\mathbf{X}) \leq q_{i\delta}^\alpha)}]}{f_i(q^\alpha) f_{i\delta}(q_{i\delta}^\alpha)} \,.$$

Alors en supposant Σ inversible et $\mathbb{E}\left[\left(\frac{f_{i\delta}(X_i)}{f_i(X_i)}\right)^2\right] < +\infty$. On obtient :

$$\sqrt{N}\left(\hat{\theta}_N - \begin{pmatrix} q^\alpha \\ q^\alpha_{i\delta} \end{pmatrix}\right) \xrightarrow{\mathcal{L}} \mathcal{N}(0, \Sigma) \ .$$

La densité perturbée f_δ est défini par :

$$f_\delta = \underset{\pi \in \mathcal{P},\ s.t\ \mathbb{E}_\pi[X] = \mathbb{E}_f[X] + \delta}{\arg\min} KL(\pi || f) \ ,$$

où $KL(.||.)$ est la divergence de Kullback-Leibler.

Soit $X \sim f$ la transformation de Rosenblatt est défini par :

$$U = \Phi^{-1}(F(X)) \ ,$$

où Φ est la fonction de répartition de la loi $\mathcal{N}(0,1)$ et F la fonction de répartition de X. Ainsi, $U \sim \mathcal{N}(0,1)$