Taking into account input mis-specification in direct UQ problems

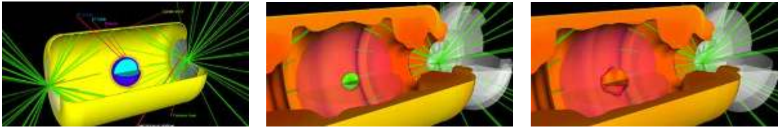
G. Perrin[1]

ETICS summer school | October 2020

[1] CEA/DAM/DIF, F-91297, Arpajon, France

Introduction

- Simulation plays a key role in the conception and the certification of new systems.
- In presence of uncertainties (model error, manufacturing tolerances,...), these analyses are based on a stochastic representation of the system.



(a) t=0 (b) t=t_1 (c) t=t_2

FIGURE: Implosion of a microballoon in laser experiment using indirect attack.

General framework

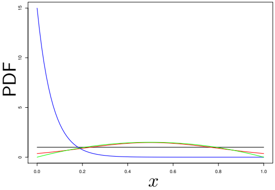
Notations

- $\mathcal{S} \leftrightarrow$ system of interest,
- $\boldsymbol{x} \in \mathbb{X} \subset \mathbb{R}^D \leftrightarrow$ system characteristics (dimensions, boundary conditions, material properties...),
- $\boldsymbol{x} \mapsto y(\boldsymbol{x}) \in \mathbb{R} \leftrightarrow$ quantity of interest for the monitoring of $\mathcal{S}$,

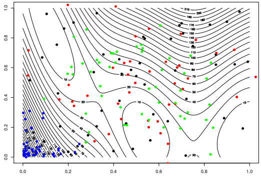
Asumptions

- $\boldsymbol{x}$ is not perfectly known $\Rightarrow$ it is modelled by a random vector with PDF $f_{\boldsymbol{x}}$.
- For $i \neq j$, x_i is supposed to be **independent** of $x_j \Rightarrow f_{\boldsymbol{x}}(\boldsymbol{x}) = \prod_{i=1}^D f_i(x_i)$.
- Each evaluation of y is **time-demanding**.

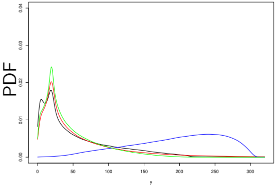
The key role of input specification



(a) Input PDFs



(b) iid realizations



(c) Output PDFs

The choice of $f_{\boldsymbol{x}}$ is central for

- a correct estimation of the distribution of y,
- carrying out sensitivity analyses,
- performing reliability analyses...

Imprecise specification of the input distribution

In many industrial applications, it is however not easy to associate distributions to the model inputs :

- not always reliable expert judgments...
- only few experimental measurements...

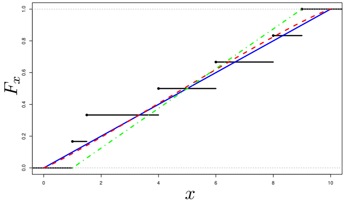


FIGURE: Black $\leftrightarrow$ empirical CDF, Blue $\leftrightarrow \mathcal{U}(0,10)$, Red dashed $\leftrightarrow$ Beta MLE, Green dotdashed $\leftrightarrow \mathcal{U}(1,9)$.

$\Rightarrow$ $f_{\boldsymbol{x}}$ is in many cases not perfectly known...

Problematics

Rather than hiding this potential misspecification of $f_{\boldsymbol{x}}$, it seems to me that it is necessary

1. to work to better **highlight** this source of uncertainty (problems of representations),
2. to work on a numerically **efficient management** of this source of uncertainty (problems of reusing samples),
3. to find the model inputs whose distribution has to be **particularly well-characterized** to make the results of the uncertainty analysis robust (problems of interpretation).

In the following, we therefore assume that $f_{\boldsymbol{x}}$ belongs to a set of PDFs defined over $\mathbb{R}^D$, noted $\mathcal{F}$.

Outline

Definition of the set of PDFs

- By construction, $\mathcal{F}$ is a subset of $\mathcal{P} = \left\{ f : \mathbb{R}^D \rightarrow [0, +\infty), \int_{\mathbb{R}^D} f = 1 \right\}$.
- However, even if $f_{\boldsymbol{x}}$ is affected by uncertainties, it is clear that we do not want to make $\mathcal{F}$ be equal to $\mathcal{P}$...
- In the following, we focus on three configurations :
  - → f_{x_i} belongs to a finite set of PDFs : $\mathcal{F} = \{f_1, f_2, \dots, f_M\}$ (example : presence of different expert judgments) ;
  - → $f_{\boldsymbol{x}}$ is in a parametric class of PDFs, with uncertainties on the parameters : $f_{\boldsymbol{x}} = f_{\boldsymbol{x}}(\cdot; \boldsymbol{\beta})$, with $\boldsymbol{\beta} \in \mathbb{B}$ a not perfectly known vector (example : the number of observations is too small to get a precise estimation of the PDF parameters) ;
  - → there only exists a reference PDF for each f_{x_i}. In that case, a normalized perturbation is added (more details are coming).

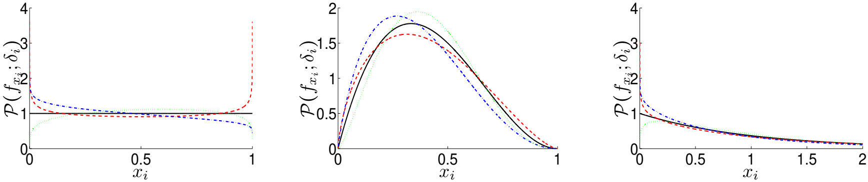
Normalization of the perturbations

- In the third case, to quantitatively compare the impact of each perturbation on $f_{\boldsymbol{x}}$, all the perturbations need to be normalized.
- To focus on interpretative perturbations, we propose to consider the following **parametric** perturbation on the mean and the standard deviation :

$$\mathcal{P}(f_{x_i}; \delta_i = (\mu_i, \varepsilon_i)) = \frac{d}{dx_i}\left(\Phi_{\mu_i, \varepsilon_i+1} \circ \Phi_{0,1}^{-1} \circ F_{x_i}\right)(x_i), \ x_i \in \mathbb{X}_i,$$

  - F_{x_i} is the **reference** CDF associated with f_{x_i},
  - $\Phi_{m,s}$, $\phi_{m,s}$ are the CDF/PDF of a Gaussian random variable $\mathcal{N}(m, s^2)$,
  - $\mu_i \in \mathbb{R}$ is the mean perturbation,
  - $\varepsilon > -1$ is the standard deviation perturbation.
- Due to the isoprobabilist transform, and to the inverse isoprobabilist transform, it is now possible to specify **normalized** perturbation levels that are characterized by the (2D)-dimensional vector $\boldsymbol{\delta} = (\mu_1, \varepsilon_1, \ldots, \mu_D, \varepsilon_D)$.

Examples of perturbations



(a) $x_i \sim U(0,1)$ (b) $x_i \sim Beta(2,3)$ (c) $x_i \sim Exp(1)$

FIGURE: Evolution of $\mathcal{P}(f_{x_i};\delta_i)$ for several values of μ_i, ε_i, and different f_{x_i}. Black solid line ↔ $(\mu_i, \varepsilon_i) = (0,0)$ (**NO PERTURBATION**). Red dashed line ↔ $(\mu_i, \varepsilon_i) = (0, 0.1)$. Blue dashed-dotted line ↔ $(\mu_i, \varepsilon_i) = (-0.2, 0)$. Green dotted line ↔ $(\mu_i, \varepsilon_i) = (0.05, -0.1)$.

Double source of uncertainty

Now, if we associate :

- → weights $p_1, \ldots, p_M$ such that $\sum_{m=1}^{M} p_m = 1$ to the elements of $\mathcal{F} = \{f_1, f_2, \ldots, f_M\}$,
- → a density $f_{\boldsymbol{\beta}}$ to the vector of parameters $\boldsymbol{\beta}$ characterizing $f_{\boldsymbol{x}} = f_{\boldsymbol{x}}(\cdot; \boldsymbol{\beta})$,
- → a density $f_{\boldsymbol{\delta}}$ to the vector of perturbations $\boldsymbol{\delta}$,

the PDF $f_{\boldsymbol{x}}$ of $\boldsymbol{x}$ becomes random, and it it easy to generate independent realization of it. In that case, we are confronted to a double source of input uncertainties.

Outline

Integrated uncertainty

A large part of the direct problems in UQ can be written in the form of an expectation computation :

- → robust conception : compute $\mathbb{E}\left[y(\boldsymbol{x})\right]$ and $\mathbb{E}\left[y(\boldsymbol{x})^2\right]$,
- → reliability analysis : compute $\mathbb{E}\left[1_{y(\boldsymbol{x})<0}\right]$,
- → sensitivity analysis : compute $\mathbb{E}\left[\mathbb{E}\left[y(\boldsymbol{x})|x_i\right]^2\right]$ and $\mathbb{E}\left[\mathbb{E}\left[y(\boldsymbol{x})|\boldsymbol{x}_{-i}\right]^2\right]$.

In that case, as :

$$\mathbb{E}\left[g(\boldsymbol{x})\right] = \mathbb{E}_{f_{\boldsymbol{x}}}\left[\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}\left[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}\right]\right]$$

taking into account this double source of uncertainty has a small impact in terms of computational cost.

Separated uncertainty

However, if the variability over $f_{\boldsymbol{x}}$ is due to **epistemic** uncertainties, in the sense that there exists a true (but unknown) PDF $f_{\boldsymbol{x}}^{\star}$ in $\mathcal{F}$ such that the true quantify of interest is equal to $\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}^{\star}}[g(\boldsymbol{x})]$, it seems necessary to

→ (at least) look for the mean and variance of $\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}]$,

→ (better) look for quantiles $q_{-,\alpha}$ and $q_{+,\alpha}$ such that :

$$\mathbb{P}(\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}] \in [q_{-,\alpha}, q_{+,\alpha}]) \geq 1-\alpha,$$

→ (if it is possible) analyze the whole distribution of random quantity $\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}]$.

Separated uncertainty

Using an Importance Sampling based approach, it is interesting to notice that if there exists an element $\widehat{f}_{\boldsymbol{x}}$ in $\mathcal{F}$ such that for all $\boldsymbol{x} \in \mathbb{X}$ and all $f_{\boldsymbol{x}} \in \mathcal{F}$, $\widehat{f}_{\boldsymbol{x}}(\boldsymbol{x}) = 0 \;\Rightarrow f_{\boldsymbol{x}}(\boldsymbol{x}) = 0$,

$$\begin{aligned}
\mathbb{E}\left[g(\boldsymbol{x})\right] &= \mathbb{E}_{f_{\boldsymbol{x}}}\left[\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}\left[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}\right]\right] \\
&= \mathbb{E}_{f_{\boldsymbol{x}}}\left[\mathbb{E}_{\boldsymbol{x}\sim \widehat{f}_{\boldsymbol{x}}}\left[\frac{g(\boldsymbol{x})f_{\boldsymbol{x}}(\boldsymbol{x})}{\widehat{f}_{\boldsymbol{x}}(\boldsymbol{x})} \mid f_{\boldsymbol{x}}\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}\sim \widehat{f}_{\boldsymbol{x}}, f_{\boldsymbol{x}}}\left[\frac{g(\boldsymbol{x})f_{\boldsymbol{x}}(\boldsymbol{x})}{\widehat{f}_{\boldsymbol{x}}(\boldsymbol{x})}\right].
\end{aligned}$$

Hence, starting from an iid sample of realizations of $\boldsymbol{x}$ under $\widehat{f}_{\boldsymbol{x}}$, in which we call the function, the computational costs associated with the evaluations of a (potentially high-dimensional) set of realizations of $\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}\left[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}\right]$ is almost similar to the one associated with $\mathbb{E}\left[g(\boldsymbol{x})\right]$.

Sensitivity analysis

- This possibility to easily compute realizations of $\mathbb{E}_{\boldsymbol{x}\sim f_{\boldsymbol{x}}}\left[g(\boldsymbol{x}) \mid f_{\boldsymbol{x}}\right]$ also invites us to look for the inputs whose PDF variability has to be reduced in priority to reduce this interval $q_{+,\alpha} - q_{-,\alpha}$ as much as possible.
- To this end, variance-based indices could be proposed as :

$$s_i = 1 - \frac{\mathbb{E}\left[\mathrm{Var}(g(\boldsymbol{x})|f_{x_i})\right]}{\mathrm{Var}(g(\boldsymbol{x}))}, \quad t_i = \frac{\mathbb{E}\left[\mathrm{Var}(g(\boldsymbol{x})|f_{\boldsymbol{x}_{-i}})\right]}{\mathrm{Var}(g(\boldsymbol{x}))}.$$

- By construction, the higher s_i or t_i, the more interest we have to reduce the uncertainty on f_{x_i}.
- Defining $\Delta q_\alpha(\boldsymbol{x})$ as the difference between quantiles $q_{+,\alpha}$ and $q_{-,\alpha}$ while considering uncertainties on the PDFs of each component of $\boldsymbol{x}$, sensitivity indices could also be proposed under the form :

$$\tau_i = 1 - \frac{\mathbb{E}\left[\Delta q_\alpha(\boldsymbol{x}|f_{x_i})\right]}{\Delta q_\alpha(\boldsymbol{x})}.$$

Outline

Application in a realibility context

Numerical illustration with R.

Outline

Conclusion

- In many applications, it is difficult to associate precise distributions to the model inputs.
- In that case, it is necessary to find methods to highlight and deal with this new source of uncertainty.
- In particular, if we want to model this uncertainty on the distributions with the probability theory, we need to associate distributions to these PDFs (for the isoprobabilist approach, how to choose δ ?).
- Once these distributions have been constructed, many computational aspects related to the uncertainty propagation can be solved when considering problems that can be written under the form of an expectation computation on a fixed support.
- However, there are still many things to be done if interested in more general problems.

Thank you for your attention.