Sensitivity to statistical estimation uncertainties and probabilistic model identification

ETICS 2022

Charles Surget\textsuperscript{1,2}  
charles.surget@onera.fr

Sylvain Dubreuil\textsuperscript{1}, Jérôme Morio\textsuperscript{1}, Cécile Mattrand\textsuperscript{2}, Jean-Marc Bourinet\textsuperscript{2}, Nicolas Gayton\textsuperscript{2}.

7th of October 2022

\textsuperscript{1}ONERA/DTIS, F-31055 Toulouse, France  \textsuperscript{2}SIGMA Clermont, F-63000 Clermont-Ferrand, France
Table of contents

1 Context

2 Proposed approach

3 Sensitivity analysis

4 Illustration

5 Conclusion
Context - Uncertainty Quantification

Simulation code of a mechanical structure:

\[ d \text{ parameters} \quad (E, \nu, L, \ldots) \quad \rightarrow \quad \text{An output quantity} \quad (\sigma, u, \ldots) \]
Context - Uncertainty Quantification

Simulation code of a mechanical structure:

\[ d \text{ parameters} \quad (E, \nu, L, \ldots) \rightarrow \text{An output quantity} \quad (\sigma, u, \ldots) \]

In an uncertainty quantification context, those parameters are considered as an input continuous random vector:

\[ X = (X_1, \ldots, X_d)^t \] with values on the domain \( \mathcal{X} \subseteq \mathbb{R}^d \) and defined by a given Probability Density Function (PDF) \( f_X \).
One could be interested in assessing the following expectation of a particular function $\tau$ of $Y = \phi(X)$ (e.g. a mean or a probability of failure):

$$E_{f_X}[\tau(\phi(X))] = \int_{\mathcal{X}} \tau(\phi(x)) f_X(x) \, dx.$$  

(1.1)
Context - 1st uncertainty source

One could be interested in assessing the following expectation of a particular function \( \tau \) of \( Y = \phi(X) \) (e.g. a mean or a probability of failure):

\[
\mathbb{E}_{f_X}[\tau(\phi(X))] = \int_{\mathcal{X}} \tau(\phi(x)) f_X(x) \, dx.
\] (1.1)

Assuming \( \tau = \text{Id} \), the Monte Carlo (MC) estimator of this integral is given by:

\[
\hat{\mu}^{MC} = \frac{1}{N_X} \sum_{j=1}^{N_X} \phi(X^{(j)}),
\] (1.2)

with \( X^{(j)} \overset{i.i.d.}{\sim} f_X \) and \( N_X \) the size of the MC sample.
One could be interested in assessing the following expectation of a particular function \( \tau \) of \( Y = \phi(X) \) (e.g. a mean or a probability of failure):

\[
\mathbb{E}_{f_X}[\tau(\phi(X))] = \int_{\chi} \tau(\phi(x)) f_X(x) \, dx.
\] (1.1)

Assuming \( \tau = \text{Id} \), the Monte Carlo (MC) estimator of this integral is given by:

\[
\hat{\mu}^{MC} = \frac{1}{N_X} \sum_{j=1}^{N_X} \phi(X^{(j)}),
\] (1.2)

with \( X^{(j)} \sim f_X \) and \( N_X \) the size of the MC sample. A first uncertainty source is related to this sample, defined as \( \tilde{X} \) in the following process:

\[ \tilde{X} = \{X^{(j)}, j = 1, \ldots, N_X\} \]
In a realistic context, the PDF $f_X$ may be unknown [1]. Thus, the probabilistic model must be inferred from experimental tests:

$$\widetilde{D} = \{D^{(i)}, i = 1, \ldots, N_D\}$$

with $D^{(i)} \overset{i.i.d.}{\sim} f_X$ with $N_D$ the size of the database $\widetilde{D}$. The estimation $\hat{f}_X|\widetilde{D}$ [2, 3] of the PDF $f_X$ induces a second uncertainty source related to $\widetilde{D}$.


Context - Small-Data

\[ \mathcal{D} = \{D^{(i)}, i = 1, \ldots, N_D\} \]

with \(D^{(i)} \overset{i.i.d.}{\sim} f_X\)

\[ \hat{f}_{X|\mathcal{D}} \]

\[ \mathcal{X} = \{X^{(j)}, j = 1, \ldots, N_X\} \]

with \(X^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{X|\mathcal{D}}\)

\[ \Sigma \]

\[ \hat{\mu}^{MC} \]
The database is of limited size $N_D$: the small-data context is imposed by costly experimental tests.
The database is of limited size $N_D$: the small-data context is imposed by costly experimental tests.

The MC sample is of limited size $N_X$: the small-data context is imposed by the simulation time induced by the model complexity.
Context - Small-Data

- The **database** is of limited size $N_D$: the small-data context is imposed by costly experimental tests.

- The **MC sample** is of limited size $N_X$: the small-data context is imposed by the simulation time induced by the model complexity.

The Test-Simulation trade-off
The expectation (1.1) and its estimator are thus written as following for a given database $\tilde{D} = \tilde{d}$:

$$\mathbb{E}_{\tilde{D}=\tilde{d}} [\phi(X)] = \int_{\chi} \phi(x) \hat{f}_{X|\tilde{D}=\tilde{d}}(x) \, dx$$

$$\approx \frac{1}{N_X} \sum_{j=1}^{N_X} \phi(X^{(j)})$$

(1.3)

(1.4)

with $X^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{X|\tilde{D}=\tilde{d}}$. The estimator (1.4) is now subject to the first uncertainty source conditioned on the database $\tilde{D} = \tilde{d}$. 

The expectation (1.1) and its estimator are thus written as following for a given database $\tilde{D} = \tilde{d}$:
The expectation (1.1) and its estimator are thus written as following for a given database $\tilde{D} = \tilde{d}$:

$$
\mathbb{E}_{\hat{f}_{X|\tilde{D}=\tilde{d}}} [\phi(X)] = \int_{\mathcal{X}} \phi(x) \hat{f}_{X|\tilde{D}=\tilde{d}}(x) \, dx
$$

(1.3)

$$
\approx \frac{1}{N_X} \sum_{j=1}^{N_X} \phi(X^{(j)}) ,
$$

(1.4)

with $X^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{X|\tilde{D}=\tilde{d}}$. The estimator (1.4) is now subject to the first uncertainty source conditioned on the database $\tilde{D} = \tilde{d}$.

However, the uncertainty related to the database is not considered in the following variance :

$$
\nabla \hat{f}_{X|\tilde{D}=\tilde{d}} \left[ \hat{\mu}^{MC} \right] = \frac{1}{N_X} \nabla \hat{f}_{X|\tilde{D}=\tilde{d}} \left[ \phi(X) \right].
$$

(1.5)
Context - Problems

\[ \tilde{D} = \{D^{(i)}, i = 1, \ldots, N_D\} \]
with \( D^{(i)} \text{i.i.d.} \sim f_X \)

\[ \hat{f}_{X|\tilde{D}} \]

\[ \bar{X} = \{X^{(j)}, j = 1, \ldots, N_X\} \]
with \( X^{(j)} \text{i.i.d.} \sim \hat{f}_{X|\tilde{D}} \)

\[ \Sigma \]

\[ \hat{\mu}^{MC} \]
Context - Problems

Problem A
How to take into account the uncertainty of the database in the variance of the estimator?

\[ \tilde{D} = \{D^{(i)}, i = 1, \ldots, N_D\} \]
with \( D^{(i)} \text{i.i.d.} \sim f_X \)

\[ \hat{f}_{X|\tilde{D}} \]

\[ \bar{X} = \{X^{(j)}, j = 1, \ldots, N_X\} \]
with \( X^{(j)} \text{i.i.d.} \sim \hat{f}_{X|\tilde{D}} \)

\[ \Sigma \]

\[ \hat{\mu}^{MC} \]
Context - Problems

\[ \tilde{D} = \{ D^{(i)}, i = 1, \ldots, N_D \} \text{ with } D^{(i)} \overset{i.i.d.}{\sim} f_X \]

\[ \hat{f}_{X|\tilde{D}} \]

\[ \tilde{X} = \{ X^{(j)}, j = 1, \ldots, N_X \} \text{ with } X^{(j)} \overset{i.i.d.}{\sim} \hat{f}_{X|\tilde{D}} \]

\[ \Sigma \]

\[ \hat{\mu}^{MC} \]

**Problem A**
How to take into account the uncertainty of the database in the variance of the estimator?

**Problem B**
In order to improve efficiently the accuracy of the estimator, should the investment of data be made in the database or the MC sample?
# Table of contents

1. Context
2. Proposed approach
3. Sensitivity analysis
4. Illustration
5. Conclusion
Proposed approach - Double integral expectation

**Problem A**

How to take into account the uncertainty of the database in the variance of the estimator?
Proposed approach - Double integral expectation

**Problem A**

How to take into account the uncertainty of the database in the variance of the estimator?

The following expectation takes into account the variation of the database:

$$\mathbb{E}_{f(x,\hat{D})} [\phi(X)] = \int_{x^\mathcal{D}} \int_{x} \phi(x) f(x,\hat{D})(x,\tilde{d}) \, dx \, d\tilde{d}. \quad \text{(2.1)}$$
Proposed approach - Double integral expectation

**Problem A**

How to take into account the uncertainty of the database in the variance of the estimator?

The following expectation takes into account the variation of the database:

\[
\mathbb{E}_{f(x,D)} [\phi (X)] = \int_{\mathcal{X}^{ND}} \int_{\mathcal{X}} \phi (x) f(x,D) (x, \tilde{d}) dx \, d\tilde{d}.
\] (2.1)

An estimator of this integral [4] is the following:

\[
\hat{\mu}_{NRA-MC} = \frac{1}{N} \sum_{k=1}^{N} \frac{1}{N_X} \sum_{j=1}^{N_X} \phi \left( X_{(j)}^{(k)} \right) = \frac{1}{N} \sum_{k=1}^{N} \hat{\mu}_{MC}^{k},
\] (2.2)

with \( X_{(j)}^{(k)} \) i.i.d. \( \sim f_X|D_k \) and \( N \) the number of databases of size \( N_D \).

**Proposed approach - Empirical variance**

The process is repeated in order to estimate the variance with respect to the database.

Empirical variance:

\[
\hat{\mu}_{(X,D)}^{MC} = \frac{1}{N-1} \sum_{k=1}^{N} \left( \hat{\mu}_{MC}^{k} - \hat{\mu}_{NRA-MC} \right)^2
\]  

(2.3)
Proposed approach - Small-data context

In a small-data context, only a database \( \tilde{D} \) of limited size \( N_D \) is available.

**Resampling method**

Allows to generate \( N \) databases from the initial one. [5, 6] (e.g. Bootstrap)

**Solution A**

The nested estimator (2.2) is now conditioned on the initial database but the method allows to take into account the uncertainty related to it.

\[
\hat{\mu}^{NRA-MC} = \hat{\mu}^{MC} - \hat{\mu}^{MC-NRA}
\]

\[ (2.2) \]

\[
\hat{\mu}^{MC} = \sum \hat{f}_{X|\tilde{D}_i} \tilde{X}_i
\]

\[ (2.3) \]


# Table of contents

1. Context
2. Proposed approach
3. Sensitivity analysis
4. Illustration
5. Conclusion
Problem B

In order to improve efficiently the accuracy of the estimator, should the investment of data be made in the database or the MC sample?

\[ S_f = \frac{\text{VAR}(\hat{\mu}_{\text{MC}}|e)}{\text{VAR}(\hat{\mu}_{\text{MC}})} \]

\[ S_f = \frac{\text{VAR}(\hat{\mu}_{\text{MC}}|f)}{\text{VAR}(\hat{\mu}_{\text{MC}})} \]

Interpretation of Sobol' indices:

- \( S_f \rightarrow \) proportion of variance due to the database
- \( S_f \rightarrow \) proportion of variance due to the MC sample


Problem B

In order to improve efficiently the accuracy of the estimator, should the investment of data be made in the database or the MC sample?

An ANalysis Of VAriance [7, 8] is performed:

\[ S_D = \frac{\nabla \left[ E \left[ \hat{\mu}^{MC} | \tilde{D} \right] \right]}{\nabla \left[ \hat{\mu}^{MC} \right]} \]

\[ S_X = \frac{\nabla \left[ E \left[ \hat{\mu}^{MC} | \tilde{X} \right] \right]}{\nabla \left[ \hat{\mu}^{MC} \right]} \]  \hspace{1cm} \text{(3.1)}

Interpretation of Sobol' indices:

- \( S_D \to \) proportion of variance due to the database,
- \( S_X \to \) proportion of variance due to the MC sample.

Figure 3.1: Interpretation of Sobol' indices associated to the database and the Monte Carlo sample.
Sensitivity analysis - Interpretation

Figure 3.1: Interpretation of Sobol' indices associated to the database and the Monte Carlo sample.
Figure 3.1: Interpretation of Sobol' indices associated to the database and the Monte Carlo sample.
Figure 3.1: Interpretation of Sobol’ indices associated to the database and the Monte Carlo sample.
Sensitivity analysis - Interpretation

Figure 3.1: Interpretation of Sobol' indices associated to the database and the Monte Carlo sample.
Figure 3.1: Interpretation of Sobol' indices associated to the database and the Monte Carlo sample.
Sensitivity analysis - Independance of inputs

However, the dependance of $\tilde{X}$ to $\tilde{D}$ is an issue for the sensitivity analysis.

\[ T_D: [0, 1]^d \rightarrow X_U \rightarrow X, \quad (3.2) \]

Sensitivity analysis - Independence of inputs

However, the dependance of \( \tilde{X} \) to \( \tilde{D} \) is an issue for the sensitivity analysis.

**Isoprobabilistic transformation**

The transformation \( \mathcal{T}_D [9, 10, 11] \) is performed here to work with an independent sample \( \tilde{U} = \{ U^{(j)}, j = 1, \ldots, N_X \} \):

\[
\mathcal{T}_D : \begin{array}{ccc}
[0, 1]^d & \longrightarrow & \mathcal{X} \\
U & \mapsto & X
\end{array},
\]

(3.2)

with \( U^{(j)} \overset{i.i.d.}\sim \mathcal{U} [0, 1]^d \).


Solution B

The investment of data is guided by the highest index:

- $S_D > S_U \rightarrow$ Investment in the database (experimental tests),
- $S_D < S_U \rightarrow$ Investment in the MC sample (simulation).
Table of contents

1. Context
2. Proposed approach
3. Sensitivity analysis
4. Illustration
5. Conclusion
Mean deflection of the free end of a cantilever beam:

$$\phi (F, L, E, b, h) = \frac{4FL^3}{Ebh^3}$$

**Figure 4.1**: Representation of a cantilever beam where $F$ is the transverse load applied on the free end of the beam of length $L$, Young’s modulus $E$ and cross-section $bh$.

**Table 4.1**: Probabilistic models associated to independent input variables for a cantilever beam toy-case. [12]

<table>
<thead>
<tr>
<th>Input variable</th>
<th>Distribution</th>
<th>Mean</th>
<th>Coefficient of variation</th>
</tr>
</thead>
<tbody>
<tr>
<td>$F$</td>
<td>LogNormal</td>
<td>556.8 [N]</td>
<td>0.08</td>
</tr>
<tr>
<td>$L$</td>
<td>Normal</td>
<td>4290 [mm]</td>
<td>0.1</td>
</tr>
<tr>
<td>$E$</td>
<td>LogNormal</td>
<td>$2.10^5$ [MPa]</td>
<td>0.06</td>
</tr>
<tr>
<td>$b$</td>
<td>Normal</td>
<td>62 [mm]</td>
<td>0.1</td>
</tr>
<tr>
<td>$h$</td>
<td>Normal</td>
<td>98.7 [mm]</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Illustration - Results

**Figure 4.2:** Evolution of Sobol' indices for the cantilever beam toy-case at $N_D = [10, 100, 190]$ and (a) $N_X = 150$ (b) $N_X = 450$. Estimation of $n = 20$ indices for each combination.
# Table of contents

1. Context
2. Proposed approach
3. Sensitivity analysis
4. Illustration
5. Conclusion
Conclusion

Framework

- The probabilistic model is **unknown** and is inferred from experimental tests,
- A **small-data context** is imposed by costly experimental tests and a costly black box function.
Conclusion

Framework

- The probabilistic model is **unknown** and is inferred from **experimental tests**, 
- A **small-data context** is imposed by costly **experimental tests** and a costly **black box function**.

Current method

A) Takes into account the **database** uncertainty in the variance of the estimator, 
B) Answers the **test-simulation** trade-off by guiding the investment of data in the driving source of uncertainty.
Conclusion

Framework

- The probabilistic model is **unknown** and is inferred from experimental tests,
- A **small-data context** is imposed by costly experimental tests and a costly black box function.

Current method

A) Takes into account the **database** uncertainty in the variance of the estimator,
B) Answers the **test-simulation** trade-off by guiding the investment of data in the driving source of uncertainty.

Perspectives

- Reduction of the computational burden with importance sampling methods, [13]
- Quantification of the amount of data to invest while considering cost differences.


*Analyse de sensibilité fiabiliste avec prise en compte d’incertitudes sur le modèle probabiliste-Application aux systèmes aérospatiaux.*  

*Resampling methods: concepts, applications, and justification.*  

*The jackknife, the bootstrap and other resampling plans.*  
SIAM, 1982.


