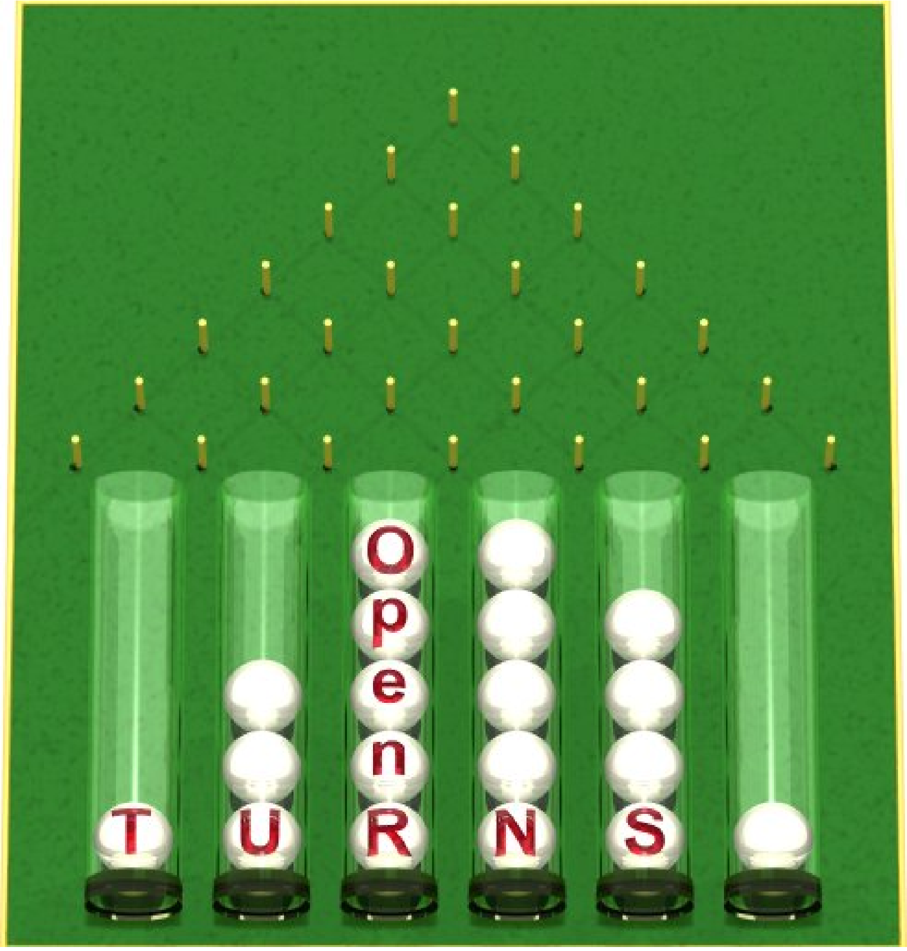
Reference Guide

Open TURNS version 0.12.1

November 8, 2008



Contents

1 Introduction

This document is part of Open TURNS' documentation. Its aim is to introduce the global methodology for the quantification of uncertainties by a model-based approach, and the methods proposed by Open TURNS to carry out the different steps of such a study. Indeed, even if each industrial study exhibits some particularities, a common framework composed of four steps is proposed.

- **Step A**: *specification of the case-study: uncertainty sources, model and criteria*
  The objective is to identify the sources of uncertainty, and the major characteristics of interest of the system studied that are influenced by these sources.
- **Step B**: *quantifying the sources of uncertainty*
  The characteristics of the uncertainty sources are determined in a deterministic or probabilistic framework.
- **Step C**: *uncertainty propagation*
  A numerical method is carried out to compute the uncertainty of the system's characteristics of interest.
- **Step C'**: *ranking of the sources of uncertainty / sensitivity analysis*
  Chosen indicators are used to rank the uncertainty sources with respect to their impact on the uncertainty of the system's characteristics of interest.

The first part of this document briefly presents each step, and illustrates the key points through a realistic – even though simplified – example.
In the second part of the document, the focus is placed on the methods that an analyst may use in Open TURNS to carry out steps B, C and C'. For each method, a synthetic form is given to highlight:

- the mathematical principles,
- the position of the method in the global methodology,
- some basic recommendations on when and how to use (and not to use) the method,
- some key bibliographic references.

For the practical side i.e. the use of these functions in Open TURNS Textual User Interface, the user is referred to the [TUI user manual] and [TUI use-cases guide].

Presentation of the flood example
Suppose that the industrial study concerns a dyke built along a river to protect an industrial facility from floods. For the industry that runs this facility, a risk exists: even if the dyke would have been high enough to contain the major floods of the last century, one does not know if the protection will be sufficient to face the next flood. For instance, meteorological events vary from year to year and are thus considered random, at least given our current knowledge in this scientific field. Therefore, an uncertainty study becomes valuable to ensure risk control (the dyke should be high enough to limit the risk of inundation) and economical optimization (the construction and maintenance cost increase with the dyke height, which should consequently not be either under or over-dimensioned).

2 Global methodology of an uncertainty study

2.1 Step A: specification of the case-study

The first step of an uncertainty study can be roughly described as “the definition of the problem". This may seem obvious, but starting an uncertainty study requires an analysis of some key issues – the foundations that will ensure that the industrial goals have been correctly translated in mathematical terms.

2.1.1 Variables of interest, model and input variables

In our framework, a *variable of interest* denotes a scalar variable on which the uncertainty is to be quantified. A *model* denotes a mathematical function that enables the computation of a set variable of interest, being given several *input variables* on which the user may have data and/or expert/engineering judgement. The basis of the uncertainty study is the following mathematical equation:

$$\underline{y} = h\left(\underline{x}, \underline{d}\right)$$

where:

- $\underline{y} = \left(y^1, \ldots, y^{n_y}\right) \in \mathbb{R}^{n_y}$ is a vector that regroups the variables of interest,
- h denotes the model,
- $\underline{x} = \left(x^1, \ldots, x^{n_x}\right) \in \mathbb{R}^{n_x}$ denotes the vector of input variables of the model on which uncertainties are to be studied,
- $\underline{d} = \left(d^1, \ldots, d^{n_d}\right) \in \mathbb{R}^{n_d}$ denotes the vector of input variables of the model treated as certain (uncertainties are negligible/neglected, or a penalized value is used).

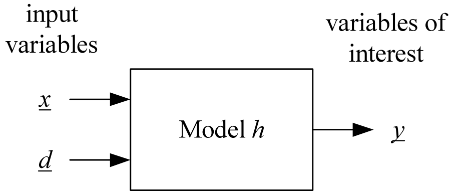


Illustration on the flood example
A key variable to be studied is the annual maximum water level; in addition, one may also want to consider the annual cost including damage caused by possible floods and maintenance of the dyke. Therefore, two *variables of interest* $\underline{y} = \left(y^1, y^2\right)$ can be studied: y^1 denotes the annual maximum water level, and y^2 denotes the overall annual cost. y^1 can be evaluated via more or less complex hydrological models, the main input factors being the river flow and some characteristics of the river bed (such as Strickler's coefficient to represent the friction i.e. the bed roughness). y^2 requires in addition an economical model to assess the costs (systematic maintenance and damages repair).

Some of the models input variables are uncertain: the river flow and bed's characteristics are naturally variable from year to year, and damage cost may not be well known. They are therefore part of $\underline{x}$, even if some of them may be put in $\underline{d}$ by using penalized value (e.g. a maximal damage cost or a “worst possible" Strickler's coefficient). This last approach could be chosen if too scarce information is available on these sources of uncertainty.

Note that every model is a simplified view of reality, which introduces another source of uncertainty in the analysis. Thus, one has to keep in mind the importance of a compromise between model uncertainty (complex models usually offer a more accurate evaluation of the variable of interest) and input variables uncertainty (complex models may involve much more uncertain factors on which information has to be available).

2.1.2 Criteria of the uncertainty study

Now that the general context has been staged, one major question is still to be addressed before moving to the core of the uncertainty study. The variable(s) of interest for the user are known to be uncertain, and this uncertainty is to be quantified; but what exactly could we or should we use to measure uncertainty? Open TURNS' methodology proposes deterministic and probabilistic criteria that meet many industrial cases requirements.

Deterministic criteria
In a deterministic context, one may want to assess the range of possible values of $\underline{y}$, that is to say a subset $D_y \subset \mathbb{R}^{n_y}$ in which we are *sure* to find $\underline{y}$. In the following, we will refer to this type of uncertainty measurement as a *deterministic criterion*; Open TURNS proposes methods that can be used to estimate the minimum and the maximum of a variable of interest.
This approach is the easiest to understand from a conceptual point of view, easier anyway than the probabilistic approach that we will now address. But we will see in step C that it is not always the less demanding approach in terms of CPU time.

Probabilistic criteria: probability of exceeding a threshold / failure probability, and quantile
Most of the methods proposed in Open TURNS use a probabilistic framework. In such a context, the vector $\underline{y}$ of variables of interest is seen as a mathematical object called *random vector*, usually noted in capital letters $\underline{Y}$. Roughly speaking, this means that one associates a probability to each interval (and more generally to each subset of values). Note that in such an approach, the range of possible values of $\underline{Y}$ may be infinite e.g. the water level in our flood problem may be somewhere between 0 and $+\infty$, even if very large values will be associated to probabilities that are extremely close to zero.
The most complete measure of uncertainty when dealing with a random vector is the *probability distribution*. One way to characterize a probability distribution is the following function F_Y, called *cumulative distribution function*:

$$F_Y\left(y^1, \ldots, y^{n_y}\right) = \mathbb{P}\left(Y^1 \leq y^1, \ldots, Y^{n_y} \leq y^{n_y}\right)$$

In an uncertainty study, one may want to assess the value of the cumulative distribution function at least in certain points. More precisely, focus may be placed on the following quantities.

- *Probability of exceeding a threshold*: the aim is to assess the probability of the event $\mathcal{D}$ = "the variable of interest Y^i exceeds a threshold important for the industrial goals at stakes (e.g. safety)":

  $$\mathbb{P}\left(Y^i > \text{threshold}\right) = 1 - F_{Y^i}\left(\text{threshold}\right)$$

  In industrial applications concerning structural reliability, one often talks of "failure probability", term that will also be used in Open TURNS' documentation. By convention (also derived from the field of structural reliability), the event "threshold exceeded" is often re-written as:

  $$\mathcal{D}_f = \left\{\underline{x} \in \mathbb{R}^{n_x} \left| g\left(\underline{x}, \underline{d}\right) < 0\right.\right\}$$

- *Quantiles*: the aim is to assess the threshold that a variable of interest may exceed with a probability equal to a given value. For $\alpha \in]0, 1[$, the quantile of level α of a scalar variable of interest Y^i is defined as follows:

  $$q_{Y^i}(\alpha) \text{ is the scalar such that } \mathbb{P}\left(Y^i \leq q_{Y^i}(\alpha)\right) = F_{Y^i}\left(q_{Y^i}(\alpha)\right) = \alpha$$

These criteria are very rich in terms of industrial meanings. But their assessment may be sometimes quite demanding in terms of CPU time (step C) and/or knowledge on the sources of uncertainty (step B). This is why in some applications, practitioners may be interested in more simple probabilistic criteria.

Probabilistic criteria: central dispersion
The *expectation/average value* μ_i and *variance* σ_i^2 of a variable of interest Y^i are defined as follows:

$$\mu_i = \mathbb{E}\left(Y^i\right),\ \sigma_i^2 = \mathbb{E}\left[\left(Y^i - \mu_i\right)^2\right]$$

Exception made of very particular cases, these two quantities are not sufficient to compute the probability of exceeding a threshold, or a quantiles. But they provide an “order of magnitude" of uncertainty: the standard deviation σ_i (square root of the variance) – normalized by the average value μ_i in order to remove scale effects – is an indicator of the *dispersion* of the variable of interest Y^i. Values distant from μ_i are more likely if σ_i is large.

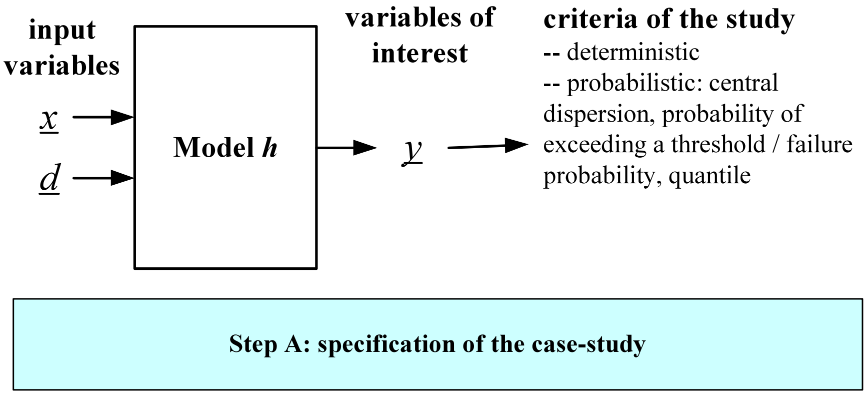

input variables
x
d
Model h
variables of interest
y
criteria of the study
-- deterministic
-- probabilistic: central dispersion, probability of exceeding a threshold / failure probability, quantile
Step A: specification of the case-study

Illustration on the flood example
In our flood example, practitioners may be interested is the probability of a flood over a year. Since Y^1 denotes the annual maximum water level:

$$\mathbb{P}\left(Y^1 > \text{dyke height}\right) = 1 - F_{Y^1}\left(\text{dyke height}\right)$$

Another probabilistic quantity of interest would be the 99%-quantile of the variable of interest Y^1, that is to say the level of water that is exceeded only 1 time per century on average (probability of exceeding the threshold equal to 1%). Note that here, one has in mind very low probabilities. But if the description of the methods proposed in Open TURNS often place the focus on low probabilities assessment – which yields specific difficulties – it is obviously possible to use these methods in order to adress “non-rare" events.
The value of these indicators (probability of flood and quantiles) is relevant *only* if one is able to provide an accurate probabilistic model of the uncertainty sources (e.g. the river flow and the bed's characteristics), problem that will be addressed in step B. If information on the uncertainty sources is scarce or difficult to collect, a first uncertainty study could focus on the expectation and standard deviation of the variable Y^1, which will bring some first useful – even though limited – informations on uncertainty.

2.2 Step B: quantifying uncertainty sources

Once step A has been carried out, the next step is to define a model to represent the uncertainties on the vector $\underline{x}$. The methods to be used depend mainly on the type of criteria chosen (deterministic or probabilistic) and on the information available (statistical datasets and/or expert/engineering judgement).

Deterministic criteria
In a deterministic framework, the range of possible values has to be determined for each component of the uncertainty sources $\underline{x}$.

Probabilistic criteria
In a probabilistic framework, the vector $\underline{x}$ of uncertainty sources is seen as a random vector denoted by $\underline{X}$. The uncertainty study then requires to assess the probability distribution of $\underline{X}$.
The first question that has to be investigated concerns the possible dependencies between uncertain variables. Common physical phenomenon may link several components of vector $\underline{X}$; then obtaining an information on X^i would change our knowledge of X^j. If such dependencies are suspected, a multi-dimensional analysis is required in order not to bias the results of the uncertainty study. In case of independence, a uni-dimensional analysis for each X^i is sufficient.
In this version, Open TURNS proposes a way of building a multi-dimensional probability distribution of $\underline{X}$ in two sub-steps.

- First, a uni-dimensional analysis has to be carried out for each uncertainty source X^i. The methods proposed by Open TURNS are described below.
- Second, some measures of the dependencies between the sources of uncertainty are to be determined through expert/engineering judgement or statistical tools provided by Open TURNS. The measures used by Open TURNS are correlation coefficients; the underlying mathematical tools are so-called “copulas".

In the uni-dimensional case, the way to build a probability distribution depends on the available data.

- Sometimes, the only available information is an expert/engineering judgement based on an analysis of the underlying physics, feedback of experience from other studies, dedicated literature, etc. Then, Open TURNS proposes a list of *parametric models* that describe various types of uncertainty thanks to a small number of parameters; these parameters can be chosen according to expert/engineering judgement.
- Suppose now that datasets are available: several measurements of the variable X^i have been carried out previously. Then, one may use again a *parametric model*, but this time with the help of statistical tools provided by Open TURNS in order to choose the most relevant model, estimate its parameters and validate the resulting model. Anyway, there still exists a risk of choosing a non-relevant parametric model, which may result in an inaccurate uncertainty study. The user may avoid this risk by choosing a *non-parametric model* proposed by Open TURNS: the result is only “data-driven" – which ensures robustness – but the number of data required is much larger than for a parametric model, especially if the uncertainty study focus on rare events.

  Note that whatever the method used to build a probability distribution (parametric or non-parametric), two phases can be distinguished: the construction of the model, and its critical analysis regarding the objectives of the study (based on data or expert/engineering judgement). This second phase should focus on the “important" parts of the probability distribution: for instance, if the criterion of the study is a rare quantile, a special attention has often to be paid to extreme values of the uncertain variables. If the criterion deals with central dispersion, the requirement on extreme values are less important.

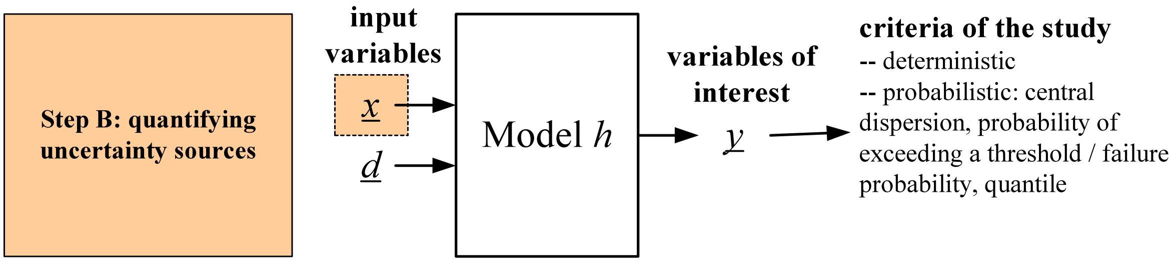

Step B: quantifying uncertainty sources
input variables
x
d
Model h
variables of interest
y
criteria of the study
-- deterministic
-- probabilistic: central dispersion, probability of exceeding a threshold / failure probability, quantile

Illustration on the flood example
In a deterministic framework, note that the upper limit for the river flow is always relative: whatever “realistic" value is proposed, one has to be aware that there is still a residual risk of exceeding this limit.
If a probabilistic framework is considered, some uncertainty sources can be reasonably assumed independent: there is no physical reason that may justify a dependancy between the river flow and Strickler's friction coefficient (knowing the flow of arriving water does not give any information on the state of the river bed). But if several uncertain variables characterize the river bed (e.g. Strickler's coefficient and some indicators of topography), the question of dependency should be investigated in order not to false the results of the study, even if it is an additional source of complexity.
Finally, note that some relationships between the variable of interests and some uncertain variables are monotonic. For instance, the maximum value of the water level will be reached for the highest possible value considered for the river flow, since a non-decreasing relation intuitively exists between these variables. Therefore, studying a high quantile of the water level requires a good confidence in the probabilistic model of extreme river flow values.

2.3 Step C: uncertainty propagation

Now that the analysis on the uncertainty sources has been carried out, the next goal is to translate the model chosen in step B in terms of uncertainty on the variables of interest via the relation:

$$\underline{y} = h\left(\underline{x}, \underline{d}\right)$$

The method to be used depends on the criteria of the study, and on some characteristics of the model h.

Deterministic criteria
In this situation, range of values have been determined for $\underline{x}$. Finding the minimum and maximum values of $\underline{y}$ is quite easy if the model h is monotonous with respect to $\underline{x}$ (one only has to consider the boundary values of $\underline{x}$). But in a more general context, this is a potentially complex optimization problem. Open TURNS proposes a simplified approach based on design of experiments to estimate extreme values of $\underline{y}$.

Probabilistic criteria
Step B has provided the probability distribution of $\underline{X}$. The objective is then to assess some characteristics of interest of the distribution of $\underline{Y} = h\left(\underline{X}, \underline{d}\right)$: probability of exceeding a threshold, quantile, or expectation and variance. Open TURNS proposes a set of relevant methods for each of these quantities.

- For the assessment of expectation/variance or threshold exceeding probability, Open TURNS proposes both approximation methods (numerically efficient whatever the CPU cost of a run of h, but only valid if the analyst can justify some properties of h e.g. regular, close to linear, etc.) and robust sampling methods (no assumption is made on h, but CPU-cost becomes a more critical issue).
- For the assessment of a quantile, Open TURNS proposes a sampling method.

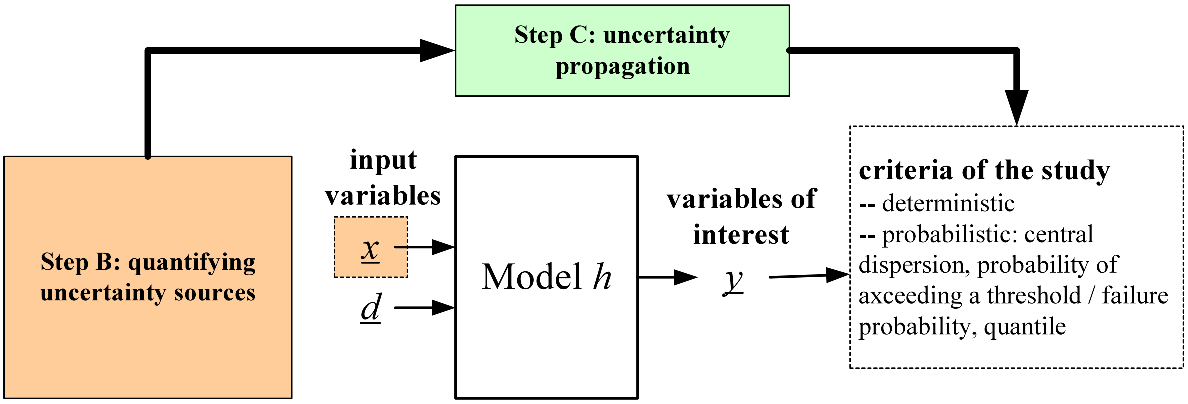

Step C: uncertainty propagation
Step B: quantifying uncertainty sources
input variables
$\underline{x}$
$\underline{d}$
Model h
variables of interest
$\underline{y}$
criteria of the study
-- deterministic
-- probabilistic: central dispersion, probability of axceeding a threshold / failure probability, quantile

Illustration on the flood example

In a deterministic framework, the computation of extremum values is facilitated by the fact that some relationships between the variable of interests and some uncertain variables are monotonic, as mentioned above: the maximum value of the water level will be reached for the highest possible value considered for the river flow.

In a probabilistic framework, the complexity of the hydrological model h plays an important role in the propagation method to be chosen. If a simple model with a low CPU cost is used, robust sampling methods are the most natural candidates. Otherwise, approximation methods and/or accelerated sampling methods may be attractive. Note that one does not have to choose a *unique* method: cross-validating the results by using several propagation methods may be fruitful!

2.4 Step C': Ranking uncertainty sources / Sensitivity analysis

In a probabilistic framework, a better understanding of uncertainties can be achieved by analysing the contribution of the different uncertainty sources to the uncertainty of the variables of interest. For each couple "criteria of the study / propagation method used in step C", post-treatment procedures are proposed by Open TURNS in order to rank the uncertainty sources.

It is important to note that an uncertainty study rarely stops after a first processing of steps A, B, C and C', and the last step then plays a crucial role. Indeed, the ranking results highlight the variables that truly determine the relevancy of the final results of the study. If the uncertainty model of some of these variables has been chosen a bit roughly in step B e.g. because of time constraints or any practical difficulties, collecting further informations on these meaningful sources would be a relevant move to refine the analysis.

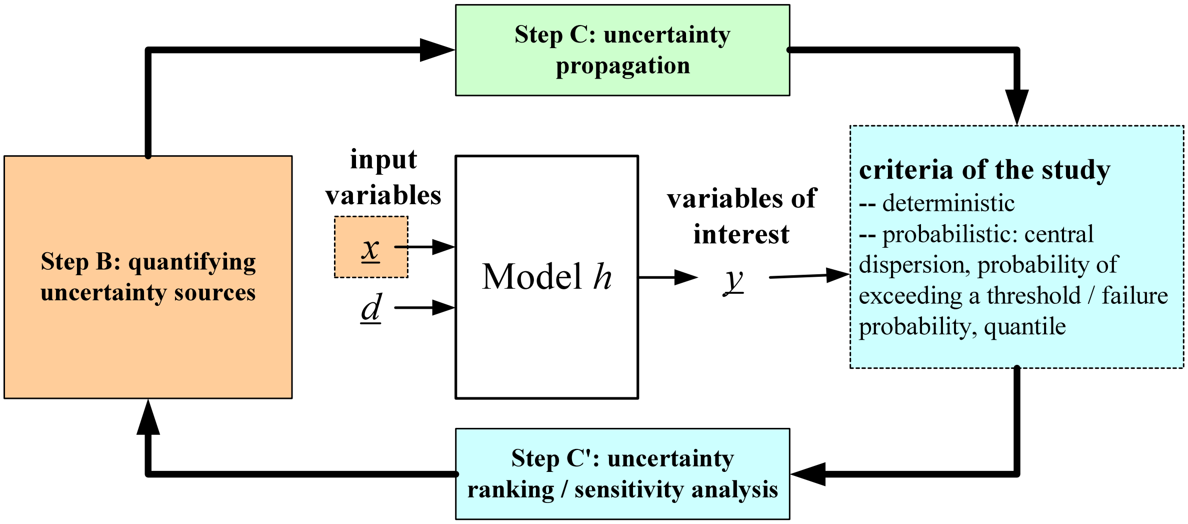

Step C: uncertainty propagation
Step B: quantifying uncertainty sources
input variables
$\underline{x}$
$\underline{d}$
Model h
variables of interest
$\underline{y}$
criteria of the study
-- deterministic
-- probabilistic: central dispersion, probability of exceeding a threshold / failure probability, quantile
Step C': uncertainty ranking / sensitivity analysis

Illustration on the flood example

It is important to note that the result of the uncertainty ranking is strongly linked to the type of criterion considered. For instance, suppose that the central dispersion of the annual maximum water level is studied. Suppose also that the river flow is pointed out by uncertainty ranking as the most important uncertain variable, the other ones having almost a negligible impact. However, it would be dangerous to say without further investigation that this would be the same if the focus is shifted towards extreme values of the variable of interest (high quantile or rare probability). it is quite possible that the role of bed's roughness uncertainty will be increased since extreme values of the water level may come only from the conjunction of a high flow *and* a high roughness.

3 Open TURNS' methods for Step B: quantifying uncertainty sources

This section is organized in three parts. The first one gives the list of probabilistic uncertainty models proposed by Open TURNS. The second part gives an overview of the content of the statistical toolbox that may be used to build these uncertainty models if data are available. The last part is dedicated to the mathematical description of each method.

3.1 Probabilistic models proposed in Open TURNS

Open TURNS proposes two different types of probabilistic models: non-parametric and parametric ones.

3.1.1 Non-parametric models

- [Empirical Cumulative Distribution Function] – see page 14
- [Kernel smoothing] – see page 16

3.1.2 Parametric models

- [Usual uni- and multi-dimensional probability distribution functions] – see page 24
- [Copulas: a mathematical tool for multi-dimensional distributions] – see page 35

3.2 Classical statistical tools for uncertainty quantification

Building a dataset may require to aggregate several data sources; Open TURNS offers some techniques to check beforehand if these data sources are indeed related to the same probability distribution.
Moreover, when a parametric model is used, Open TURNS provide statistical tools to estimate the parameters, validate the resulting model and address the important issue of dependencies among uncertainty sources.

3.2.1 Aggregation of two samples

- Qualitative analysis
  - [Graphical analysis using QQ-plot] – see page 39
- Quantitative analysis
  - [Smirnov test] – see page 42

3.2.2 Estimation of a parametric models

- [Maximum Likelihood method] – see page 44

3.2.3 Analysis of the goodness of fit of a parametric model

- Qualitative goodness-of-fit analysis
  - [Graphical analysis] – see page 47
- Quantitative goodness-of-fit analysis
  - [Chi-square test] – see page 51
  - [Kolmogorov-Smirnov test] – see page 53
  - [Cramer-Von-Mises test] – see page 56
  - [Anderson-Darling test] – see page 58
  - [Bayesian Information Criterion (BIC)] – see page 60

3.2.4 Detection and quantification of dependencies among uncertainty sources

- Linear correlations
  - [Pearson correlation coefficient] – see page 62
  - [Pearson independence test] – see page 65
- Monotonous correlations
  - [Spearman correlation coefficient] – see page 67
  - [Spearman independance test] – see page 70
- Model-free dependency analysis
  - [Chi-square independence test] – see page 72
- Regression methods
  - [Linear regression] – see page 74

3.3 Methods description

3.3.1 Step B – Empirical cumulative distribution function

Mathematical description

Goal

The empirical cumulative distribution function provides a graphical representation of the probability distribution of a random vector without implying any prior assumption concerning the form of this distribution. It concerns a non-parametric approach which enables the description of complex behaviour not necessarily detected with parametric approaches.
Therefore, using general notation, this means that we are looking for an estimator $\widehat{F}_N$ for the cumulative distribution function F_X of the random variable $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$:

$$\widehat{F}_N \leftrightarrow F_X$$

Principle of the method for $n_X = 1$

Let us first consider the uni-dimensional case, and let us denote $\underline{X} = X^1 = X$. The empirical probability distribution is the distribution created from a sample of observed values $\{x_1, x_2, \ldots, x_N\}$. It corresponds to a discrete uniform distribution on $\{x_1, x_2, \ldots, x_N\}$: where X' follows this distribution,

$$\forall\, i \in \{1, \ldots, N\}\,,\ \Pr\left(X' = x_i\right) = \frac{1}{N}$$

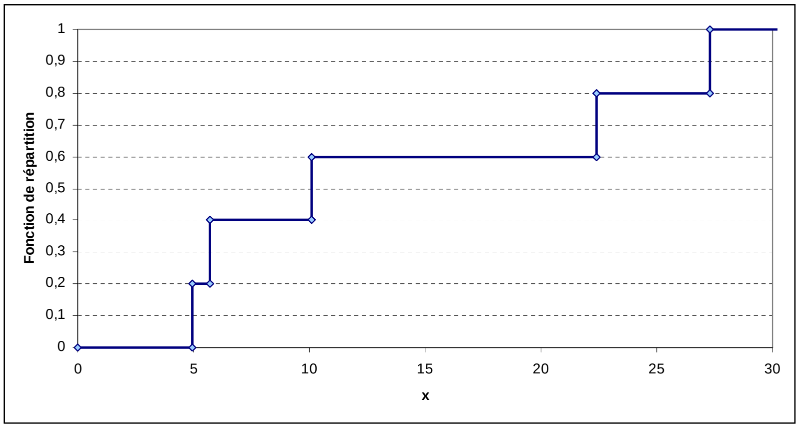
The empirical cumulative distribution function $\widehat{F}_N$ with this distribution is constructed as follows:

$$F_N(x) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\{x_i \leq x\}}$$

The empirical cumulative distribution function $F_N(x)$ is defined as the proportion of observations that are less than (or equal to) x and is thus an approximation of the cumulative distribution function $F_X(x)$ which is the probability that an observation is less than (or equal to) x.

$$F_X(x) = \Pr\left(X \leq x\right)$$

The diagram below provides an illustration of an ordered sample $\{5, 6, 10, 22, 27\}$.



Principle of the method for $n_X > 1$

The method is similar for the case $n_X > 1$. The empirical probability distribution is a distribution created from a sample $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$. It corresponds to a discrete uniform distribution on $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$: where $\underline{X}'$ follows this distribution,

$$\forall\, i \in \{1, \ldots, N\}\,,\ \Pr\left(\underline{X}' = \underline{x}_i\right) = \frac{1}{N}$$

Thus we have:

$$F_N(\underline{x}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}_{\left\{x_i^1 \leq x^1, \ldots, x_N^{n_X} \leq x^{n_X}\right\}}$$

in comparison with the theoretical probability density function F_X:

$$F_X(x) = \mathbb{P}\left(X^1 \leq x^1, \ldots, X^{n_X} \leq x^{n_X}\right)$$

Other notations

This method is also referred to in the literature as the empirical distribution function.

Link with OpenTURNS methodology

This method is used in step B “Quantifying Sources of Uncertainty". It enables us to obtain a representation of the distribution of the vector $\underline{X}$ of uncertain variables defined in step A “Specifying Criteria and the Case Study", without applying any a priori modelling hypotheses.

References and theoretical basics

This method has the advantage of depending only on the observed values, without any other modelling assumptions (as in the [kernel smoothing method]). Nevertheless, in the case where little data is available, the estimation of the criteria defined in step A can be less precise with this non-parametric method than with a parametric approach (e.g. the models described in [standard parametric models]).
The following bibliographical references provide main starting points for further study of this method:

- Saporta G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon W.J. & Massey F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill

3.3.2 Step B – Density by Kernel Smoothing

Mathematical description

Goal

Kernel smoothing methods enable to estimate probability density functions without any reference to a statistical model. It can uncover structural features in the data which a parametric approach might not reveal.
So, following the general notations, it means we are looking for an estimator $\hat{f}_N^h(x)$ of the probability density function of the input random vector $\underline{X}$:

$$\hat{f}_N^h(x) \leftrightarrow f_X(x)$$

These methods require a set of initial data to build this estimator.

Principles of the method for $n_X = 1$

The principles of the method are first presented in dimension $n_X = 1$. The principles are similar in dimension $n_X > 1$ and the formulas will be presented in the following paragraph of this file.
Thus, let us note $\underline{x} = x^1 = x$ and $(x_1, ..., x_N)$ a sample of the input random variable X. This random variable X is supposed to follow a continuous density f_X, which is of course unknown. This method enables to build an estimate of the probability density function driven by the initial data set. This 'estimated' function, called the *kernel density estimator* $\hat{f}_N^h$, will replace the 'true' density function f_X in the rest of the modelling steps. This function is built to converge in a certain sense to this density f_X. The *kernel density estimator* is a random variable defined by:

$$\hat{f}_N^h(x) = \frac{1}{Nh}\sum_{i=1}^{N} K\left(\frac{x - X_i}{h}\right)$$

where:

- K is called the Kernel. Its properties are the followings:
  - It can be a proper pdf, usually chosen to be unimodal and symmetric about zero.
  - The center of the kernel is placed right over each data point.
  - The influence of each data point is spread about its neighborhood.
  - The contribution from each point is summed to overall estimate.
- h is the bandwidth. Its properties are the followings:
  - It represents a scaling factor,
  - It controls how wide the probability mass is spread around each point.
  - It controls the smoothness or roughness of a density estimate.
  - The bandwidth selection bears danger of under- or over-smoothing.
- $X_1, \ldots, X_N$ are N random variables following the same probability density function f_X. Its realization is the sample of points $x_1, \ldots, x_N$.

- x is a value of the input variable.

We recap hereafter different questions linked to the use of these methods:

- How to choose the optimal bandwidth h ($= h_{opt}$)?
- Which Kernels K are to be used?

How to choose the optimal bandwidth h_{opt}?

Therefore, $\hat{f}_N^h(x)$ is the weighted mean of the probability density functions K centered on the variables X_k. The influence of each variable X_k is controlled by the bandwidth h. The optimal bandwidth is defined towards the following $AMISE$ criterion (Asymptotic Mean Integrated Squared Error). The optimization of the $AMISE$ criterion traduces the trade-off to be found between the convergence of the expectation and the convergence of the variance of the estimator of the pdf.

$$AMISE(h) \quad = \quad \int_{\mathbb{R}} [AB^2(x,h) + AV(x,h)]\, dx$$

where:

$$\begin{aligned} AB(x,h) &\approx \mathbb{E}[\hat{f}_N^h(x) - f_X(x)] \\ AV(x,h) &\approx \text{Var}\left[\hat{f}_N^h(x)\right] = \mathbb{E}[(\hat{f}_N^h(x) - f_X(x))^2] \end{aligned}$$

Finally, it can be demonstrated that $AMISE(h)$ is minimum for:

$$h_{opt} = \left(\frac{\|K\|_2^2}{\sigma_K^4 \left\|f''_X\right\|_2^2} \right)^{\frac{1}{5}} \times \frac{1}{N^{1/5}}$$

where:

- N is the size of the sample of data,
- $\|K\|_2^2$ is the L_2 norm of the Kernel,
- σ_K is the standard deviation of the Kernel,
- $\|f''_X\|_2^2$ is the L_2 norm of the 'true' probability density function f_X. It is usually unknown but one has to make the assumption that it is finite.

$AMISE(h)$ is thus equal to:

$$AMISE(h_{opt}) = \frac{5}{4}(\sigma_K \|K\|_2^2)^{4/5} \ \left\|f''\right\|_2^{2/5} \ \frac{1}{N^{4/5}}$$

As a remark, a pdf estimation by histogram is less 'efficient' (see references) as it decreases the AMISE evaluation by a power $2/3$ in the number of samples N.

Which Kernels K are used?

Different Kernels are available in the literature. Within Open TURNS, the only Kernel to be proposed is the Gaussian one $\mathcal{N}(0, \sigma_K)$, defined by:

$$K(x) = \frac{1}{\sqrt{2.\pi}.\sigma_K}.\exp^{-x^2/(2.\sigma_K^2)}$$

The *Plug-in method* within Open TURN'S

A general rule can be given to use the Kernel smoothing method:

1. Choose h_{opt} as a first guess as if the pdf f_X to be built were a Gaussian law $\mathcal{N}(\mu, \sigma)$,

$$h_{opt}^1 \cong 1.364 * \sigma * \left(\frac{\|K\|_2^2}{\sigma_K^4}\right)^{1/5} N^{-1/5}$$

2. Estimate $\|f''\|_2$ with the first estimation of the bandwidth h_{opt}^1,
3. Inject the new value of $\|f''\|_2$ to re adjust h_{opt}^2.

Application in dimension $n_X > 1$

The process is very similar in higher dimension. The Kernel estimator is the following:

$$\hat{f}_N^h(\underline{x}) = K_h(\underline{x}) = \frac{1}{(\det H)^{1/2}} * K\left(\frac{{}^t\underline{x}.H^{-1}.\underline{x}}{2}\right)$$

Within Open TURNS, only the Gaussian Kernel Φ is used:

$$K(\underline{x}) = \Phi_{n_X}(\underline{x}) = \frac{1}{(2.\pi)^{\frac{n_X}{2}}}.\exp(-\frac{{}^t\underline{x}.\underline{x}}{2})$$

For one given sample of size N, the optimal bandwidth h_{opt}^i is obtained following:

$$h_{opt}^i = \left(\frac{4}{n_X+2}\right)^{1/(n_X+4)} \frac{\hat{\sigma}_N^i}{N^{1/(n_X+4)}}$$

where:

- n_X is the dimension of the space.
- $(\hat{\sigma}_N^i)^2$ is an estimator of the standard deviation of the $i-th$ component obtained from N the realizations.
- $H = diag((h_{opt}^1)^2, \ldots, (h_{opt}^N)^2)$, it represents the covariance matrix.

Other notations

En français, reconstruction à noyaux, lissage à noyaux.

Link with OpenTURNS methodology

This method is part of the step B of the global methodology. It requires a sample of the input variables for the problem defined in the step A of the global methodology. It enables to build an expression of the probability density function of the input variables without any *a priori* reference to a parametric probability density function. The number of parameters to be determined is not known *a priori* and depends on the size of the sample.

References and theoretical basics

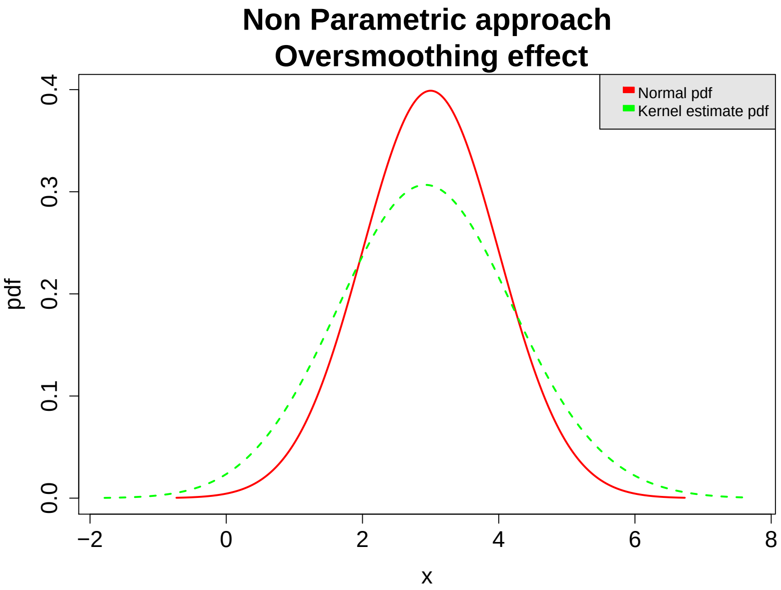
This method is very convenient in the sense that it is only attached to the real dataset. No more assumption is required on the feature of the probability density function. Nevertheless, for the practical cases where the data are scarce, the accuracy of the estimation of a criterion can be worse by a non parametric approach than the one obtained by a parametric approach for the same size of the dataset. A parametric approach is well adapted when the pdf can be justified either by expert judgement or by return of experience.
David W. Scott, 'Multivariate Density Estimation'.

[Parametric Analysis]

Examples

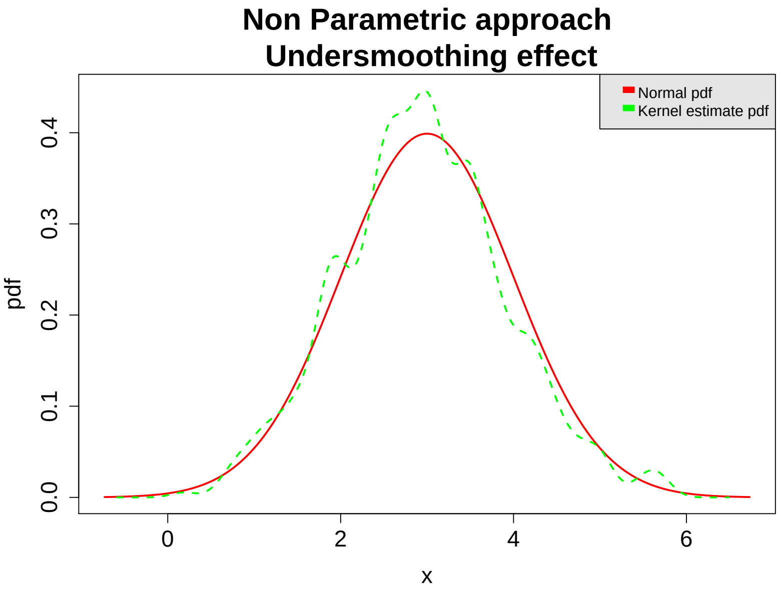
Example n°1: Choice of the bandwidth h

This example illustrates the effect of the choice of the bandwidth h on the estimation of the pdf compared to the optimal one (Figure n°xx_3). Depending on the choice of h, one could observe for the same size N of input values over-smoothing effects (figure N°xx_1) or under-smoothing (Figure n°xx_2) effects. In any case, the Kernel smoothing method is still converging when one adds samples but more slowly as if an optimal bandwidth had been chosen.
Oversmoothing effect
In this case, h is bigger than the optimal choice h_{opt}. The effect of the values is more widely spread as in the optimal case. The 'vicinity' (in the AMISE sense) with the 'true' pdf is deteriorated at a given number of sample.



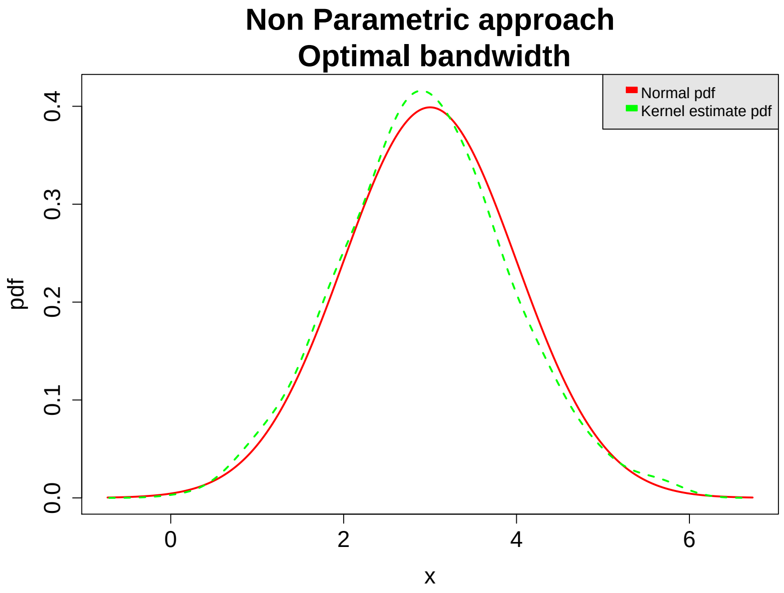
Undersmoothing effect

In this case, h is smaller than the optimal choice h_{opt}. The effect of the values is more locally focused on the values obtained in the data set than in the optimal case. The 'vicinity' (in the AMISE sense) with the 'true' pdf is deteriorated at a given number of sample N.



Optimal smoothing

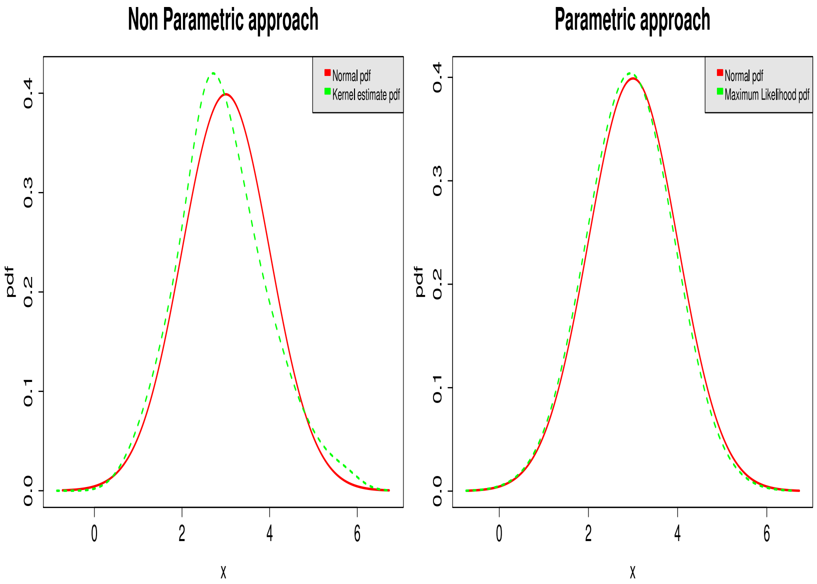
Following the previous rule, for a Gaussian law



Examples n°2 and 3: Difference between parametric and non parametric modelling

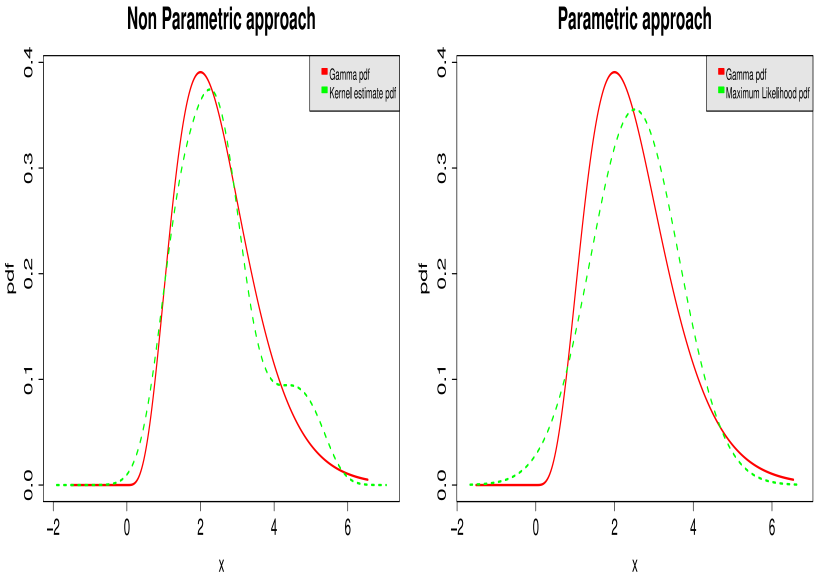
Within these two examples, we try to illustrate the differences (advantages and drawbacks) of a parametric approach and a non parametric approach when one tries to build the pdf of an unknown pdf.

1. **Example n°2: Symmetric law**
   The 'true' pdf f_X to be built is a Gaussian law $\mathcal{N}(3, 1)$. It has to be estimated by a sample of size 200.
   The parametric approach pre supposes the knowledge of this statistical model and estimates the parameters of the Gaussian law by a maximum likelihood method.
   The non parametric approach does pre suppose the statistical model and fits the estimated pdf to the data set.
   If one can justify the statistical model, it is preferable to choose a parametric approach.

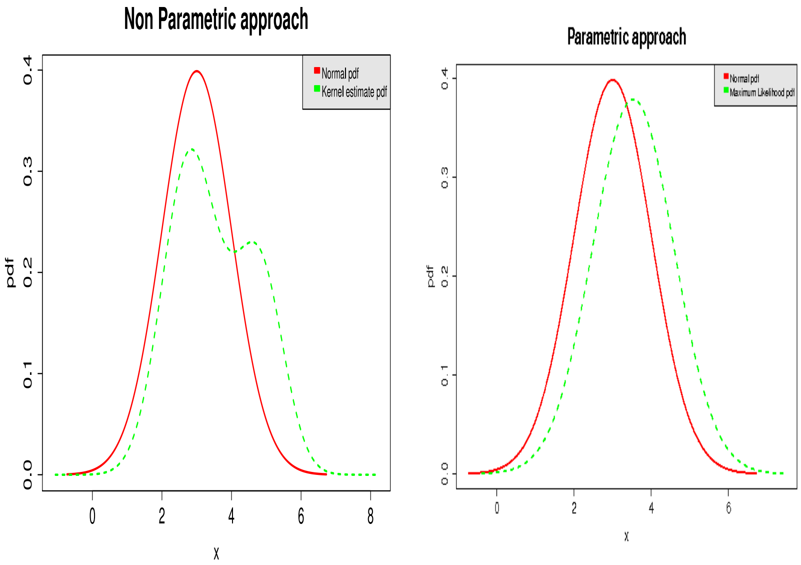


2. **Example n°3: Non Symmetric law**
   The 'true' pdf f_X to be built is a Gamma law $-(5, 2)$. It has to be estimated by a sample of size 200.

   The parametric approach Parametric estimation tries to estimate a Gaussian law by a maximum likelihood approach.
   The non parametric approach is still only sticked to the data set and does not require the knowledge of the statistical model.
   If the statistical model is not known, it is more robust to choose a non parametric approch.



Example n°4: Effect of the number of samples
When the statistical model is the 'true' one, a small number of data has a stronger impact on the non parametric approach than on the parametric approach.


Non Parametric approach
🟥 Normal pdf
🟩 Kernel estimate pdf
pdf
0.0
0.1
0.2
0.3
0.4
0
2
4
6
8
x
Parametric approach
🟥 Normal pdf
🟩 Maximum Likelihood pdf
pdf
0.0
0.1
0.2
0.3
0.4
0
2
4
6
x

3.3.3 Step B – Standard parametric models

Mathematical description

Objective

Parametric models aim to describe probability distributions of a random variable with the aid of a limited number of parameters $\underline{\theta}$. Therefore, in the case of continuous variables (i.e. where all possible values are continuous), this means that the probability density of $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ can be expressed as $f_X(\underline{x};\underline{\theta})$. In the case of discrete variables (i.e. those which take only discrete values), their probabilities can be described in the form $\mathbb{P}\left(\underline{X} = \underline{x};\underline{\theta}\right)$.

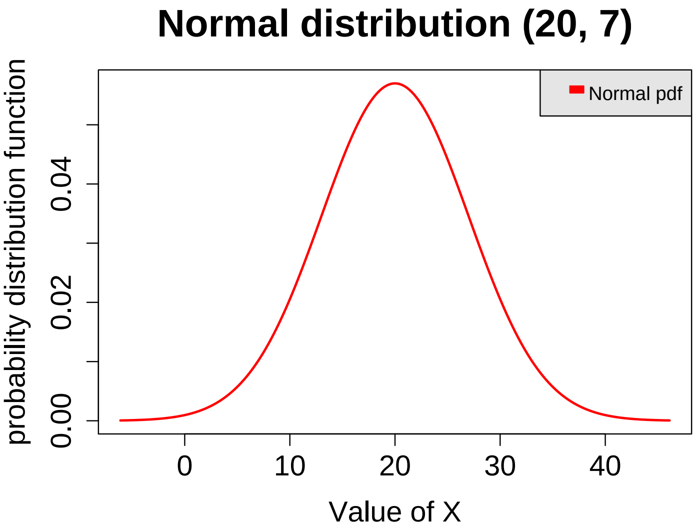
Available distributions for $n_X = 1$

Let us first consider 1 dimension and let $\underline{X} = X^1 = X$. The standard distributions available in Open TURNS are listed in this section. We start with continuous distributions.

- **Normal distribution (or Gaussian distribution)** : $\underline{\theta} = (\mu, \sigma)$, with the constraint $\sigma > 0$. The probability density is given as:

$$f_X(x;\underline{\theta}) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

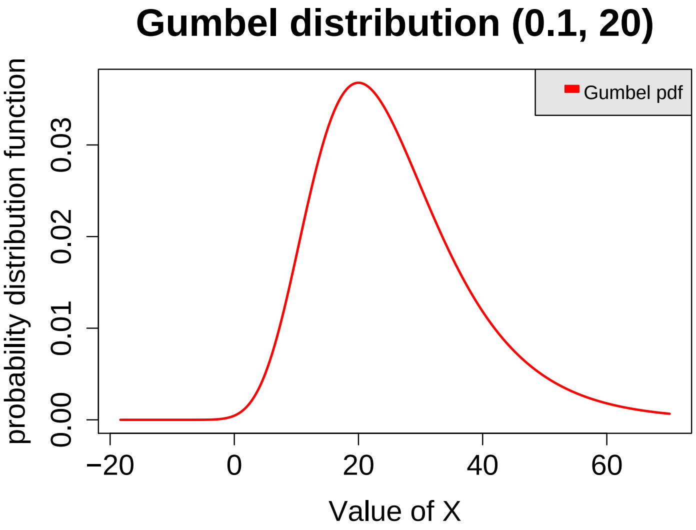
We note that a random variable which follows a Normal distribution as defined here takes real values from $\mathbb{R}$. μ provides the most likely value (for which the probability density function is at its highest), and the density function is symmetric around this value (the values $\mu - a$ and $\mu + a$ are equally likely); μ is also the expected value (mean) of this distribution. Whilst σ provides a measure of dispersion: the larger it is, the flatter the probability density function is (i.e. values far away from μ are still likely, or in other words possible values are more spread out).



- **Gumbel distribution** : $\underline{\theta} = (\alpha, \beta)$, with the constraint $\alpha > 0$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \alpha \exp\left(-\alpha(x-\beta) - e^{-\alpha(x-\beta)}\right)$$

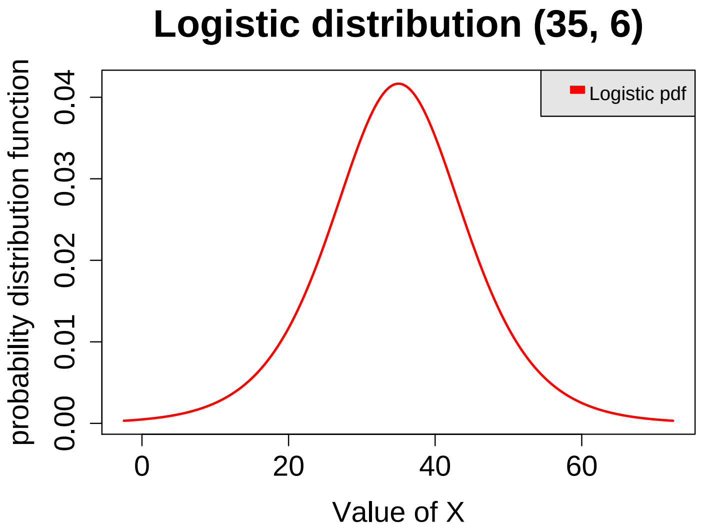
We note that a random variable which follows a Gumbel distribution as defined here takes real values from $\mathbb{R}$. β describes the most likely value, but this is less than the expected value of the distribution because the distribution is asymmetric (right skewed): the probability values in the distribution's right tail (i.e. values greater than β) decrease more gradually than those in the left tail (i.e. values less than β). a provides a measure of dispersion: the probability density function flattens as α decreases.



- **Logistic distribution** : $\underline{\theta} = (\alpha, \beta)$, with the constraint $\beta \geq 0$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{\exp\left(\frac{x-\alpha}{\beta}\right)}{\beta\left[1 + \exp\left(\frac{x-\alpha}{\beta}\right)\right]^2}$$

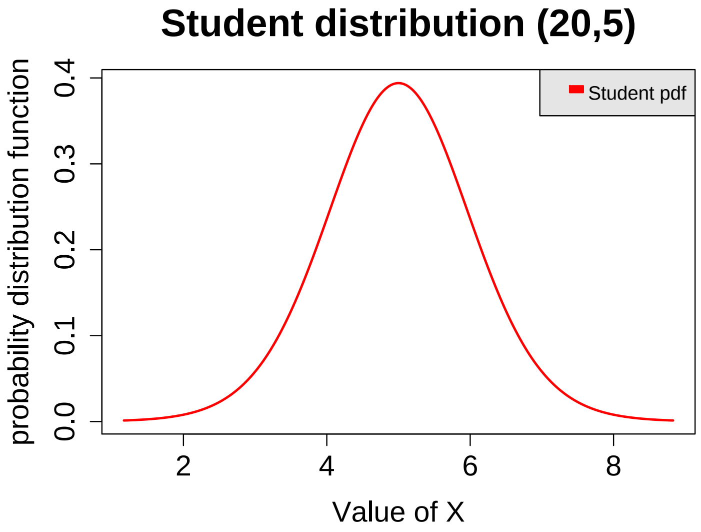
We note that a random variable which follows a logistic distribution as defined here takes real values from $\mathbb{R}$. α describes the most likely value. β provides a measure of dispersion: the probability density function flattens as β decreases.



- **Student's t-distribution** : $\underline{\theta} = (\nu, \mu)$, with the constraint $\nu \geq 2$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{1}{\sqrt{\nu} B\left(\frac{1}{2}, \frac{\nu}{2}\right)} \left(1 + \frac{(x-\mu)^2}{\nu}\right)^{-\frac{1}{2}(\nu+1)}$$

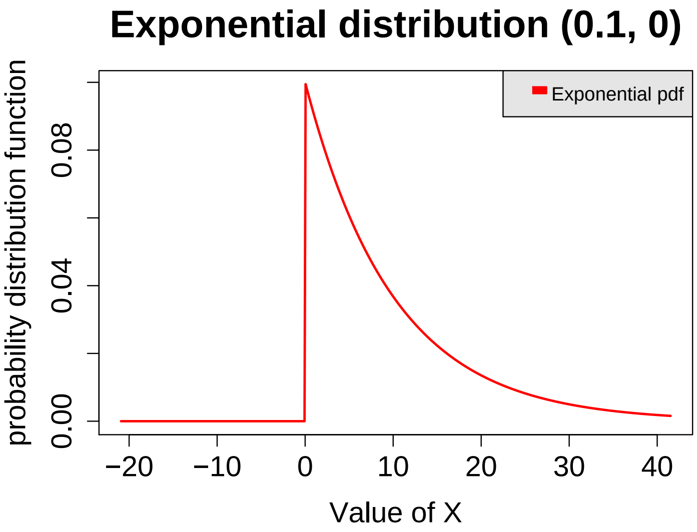
  where B describes the function beta. We note that a random variable which follows a Student's t-distribution as defined here takes real values from $\mathbb{R}$. μ describes the most likely value. ν is a measure of dispersion: the probability density function flattens as ν decreases.



- **Exponential distribution** : $\underline{\theta} = (\lambda, \gamma)$, with the constraint $\lambda > 0$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \lambda \exp\left(-\lambda(x-\gamma)\right) \mathbf{1}_{\gamma \leq x}$$

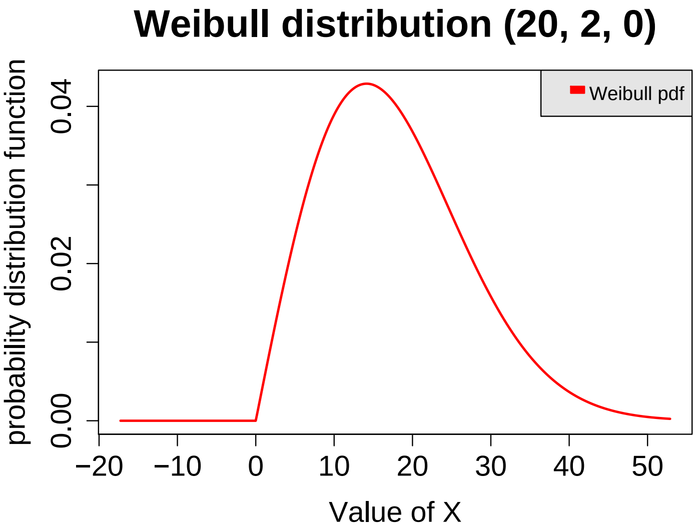
We note that a random variable which follows an Exponential distribution as defined here takes values in the range $[\gamma, +\infty)$, and is right skewed. Both a and ß influence the dispersion. The expected value of the distribution is $\gamma + 1/\lambda$. The coefficient of variation (standard deviation / mean) is constant and equal to 1 whatever the value of λ.



- **Weibull distribution** : $\underline{\theta} = (\alpha, \beta, \gamma)$, with the constraints $\alpha > 0$, $\beta > 0$. probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{\beta}{\alpha}\left(\frac{x-\gamma}{\alpha}\right)^{\beta-1} \exp\left(-\left(\frac{x-\gamma}{\alpha}\right)^{\beta}\right) \mathbf{1}_{\gamma \leq x}$$

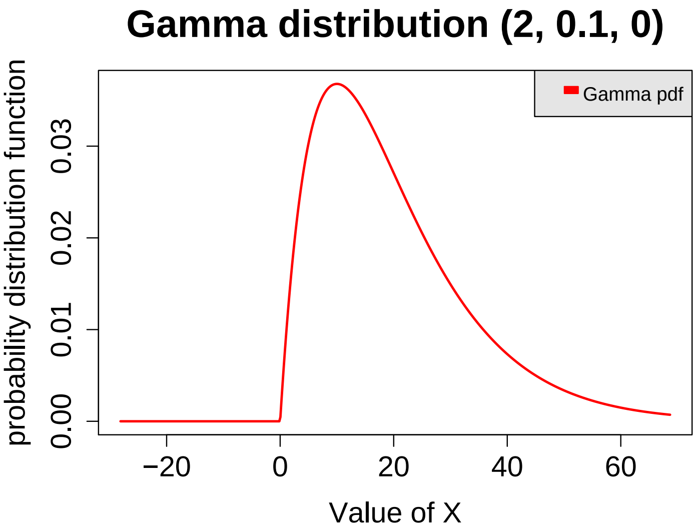
We note that a random variable which follows a Weibull distribution as defined here takes values in the range $[\gamma, +\infty)$, and is right skewed. Both α and β influence the dispersion. We note that the distribution becomes more skewed as β decreases. In the case where $\beta = 1$ this is corresponds to the Exponential distribution.



- **Gamma distribution** : $\underline{\theta} = (\lambda, k, \gamma)$, with the constraints $\lambda > 0$, $k > 0$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{\lambda}{\Gamma(k)} \left(\lambda(x-\gamma)\right)^{k-1} \exp\left(-\lambda(x-\gamma)\right) \mathbf{1}_{\gamma \leq x}$$

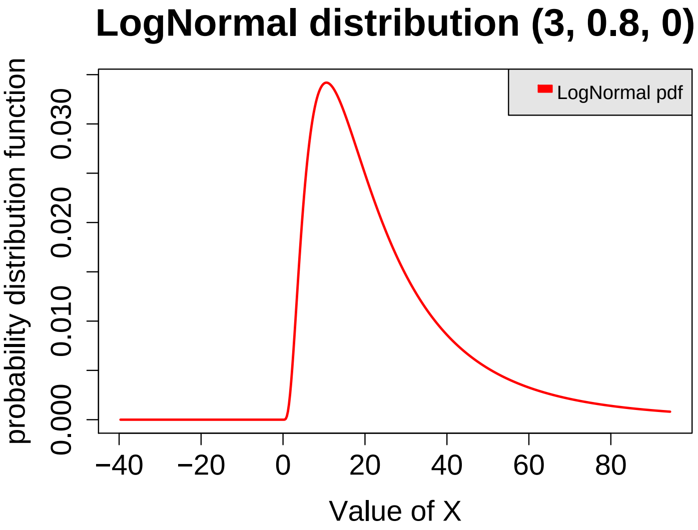
  where Γ is the gamma function. We note that a random variable which follows a gamma Distribution as defined takes values in the range $[\gamma, +\infty)$, and is right skewed.



- **Lognormal distribution** : $\underline{\theta} = (\mu_l, \sigma_l, \gamma)$, with the constraint $\sigma_l > 0$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{1}{\sigma_l (x-\gamma)\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{\ln(x-\gamma)-\mu_l}{\sigma_l}\right)^2\right) \mathbf{1}_{\gamma \leq x}$$

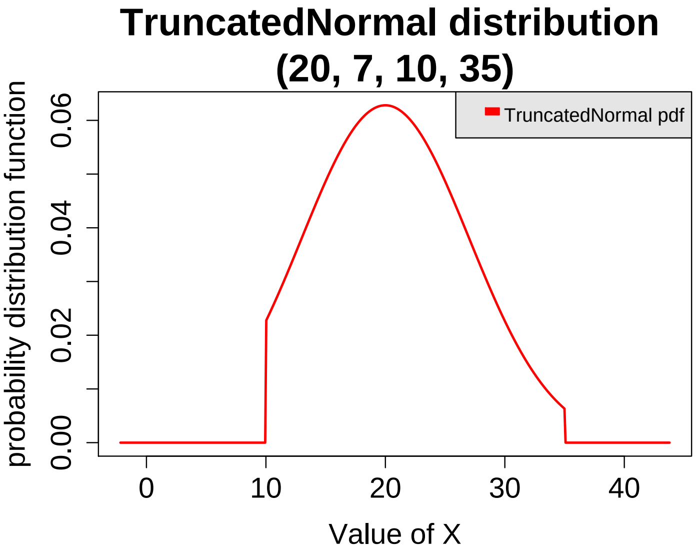
We note that a random variable which follows a Log-normal distribution as defined here takes values in the range $[\gamma, +\infty)$, and is right skewed.



- **Truncated Normal Distribution** : $\underline{\theta} = (\mu_n, \sigma_n, a, b)$, with the constraints $\sigma_n > 0$, $b > a$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{\varphi(\frac{x-\mu_n}{\sigma_n})/\sigma_n}{\Phi(\frac{b-\mu_n}{\sigma_n}) - \Phi(\frac{a-\mu_n}{\sigma_n})} \mathbf{1}_{a \leq x \leq b}$$

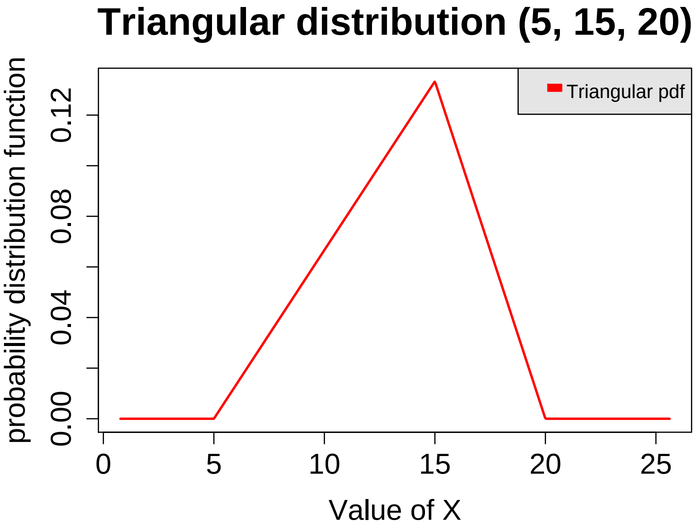
where φ and Φ represent the probability density and the cumulative distribution function respectively of the reduced centred Normal distribution (i.e. the mean μ zero and standard deviation σ equal to 1). We note that a random variable that follows a Truncated Normal Distribution takes values in the interval $[a, b]$. μ describes the most likely value. Whilst σ provides a measure of dispersion: the probability density function flattens as s increases (the probability density becomes zero for values outside the interval $[a, b]$).



- **Triangular distribution** : $\underline{\theta} = (a, b, m)$, with the constraints $a \leq m$, $m \leq b$, $b > a$. The probability density function is expressed as:

$$f_X(x;\underline{\theta}) = \begin{cases} 2\frac{x-a}{(m-a)(b-a)} & \text{if } a \leq x \leq m \\ 2\frac{b-x}{(b-m)(b-a)} & \text{if } m \leq x \leq b \\ 0 & \text{otherwise} \end{cases}$$

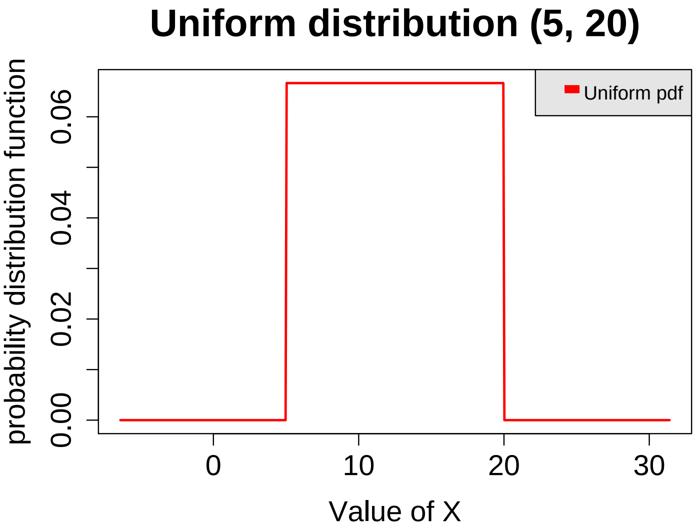
  We note that a random variable that follows a triangular distribution as defined here takes in the interval $[a, b]$. m describes the most likely value.



- **Uniform distribution** : $\underline{\theta} = (a, b)$, with the constraint $a < b$. The probability density function is expressed as:

$$f_X(x;\underline{\theta}) = \frac{1}{b-a}\mathbf{1}_{a \leq x \leq b}$$

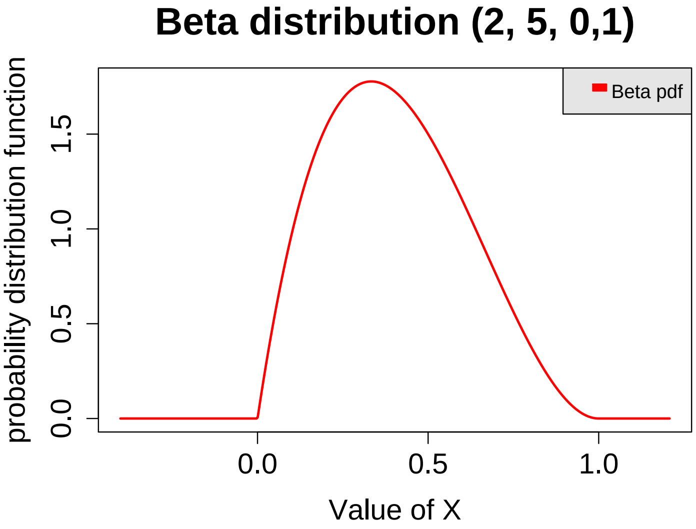
We note that a random variable that follows a uniform distribution as defined here takes values in the interval $[a, b]$. All values in this interval are equally-likely.



- **Beta distribution** : $\underline{\theta} = (r, t, a, b)$, with the constraints $r > 0$, $t > r$, $b > a$. The probability density function is expressed as:

$$f_X(x; \underline{\theta}) = \frac{(x-a)^{r-1}(b-x)^{t-r-1}}{(b-a)^{t-1}B(r, t-r)} \mathbf{1}_{a \leq x \leq b}$$

where B denotes the Beta function. We note that a random variable that follows a Beta distribution as described here takes values in the interval $[a, b]$.

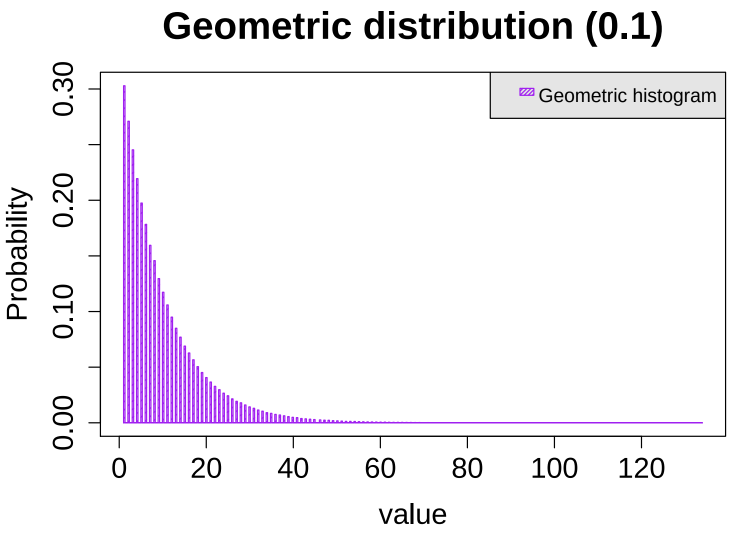


Still in 1 dimension, Open TURNS also offers Discrete Distributions.

- **Geometric distribution** : $\underline{\theta} = p$, with the constraint $0 < p < 1$. all natural numbers $x \in \mathbb{N}^*$,

$$\mathbb{P}\left(X = x; \underline{\theta}\right) = p\left(1-p\right)^{x-1}$$

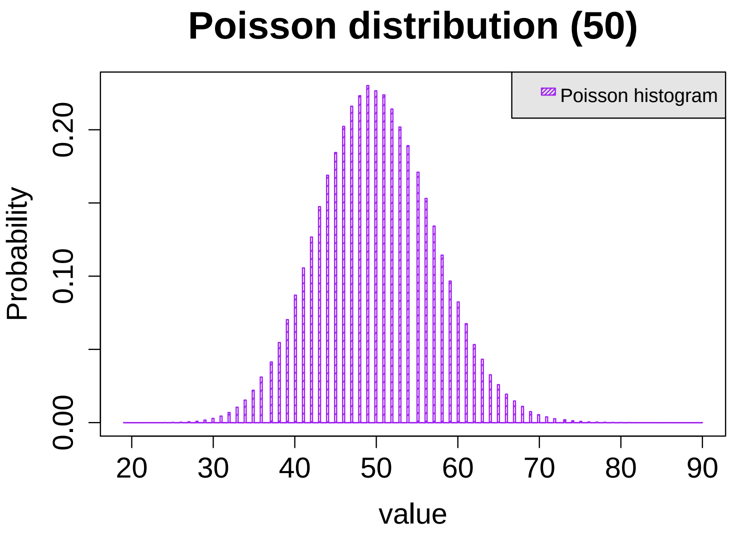
We note that a random variable that follows a Geometric Distribution as defined here takes values from $\mathbb{N}^*$.



- **Poisson distribution** : $\underline{\theta} = \lambda$, with the constraint $\lambda > 0$. For all $x \in \mathbb{N}$,

$$\mathbb{P}\left(X = x; \underline{\theta}\right) = \frac{\lambda^x}{x!}\exp\left(-\lambda\right)$$

We note that a random variable that follows a Poisson distribution as defined here takes values from $\mathbb{N}$.



Available distributions for $n_X > 1$

Let us now consider $n_X > 1$. Currently only one continuous distribution is available in Open TURNS.

- **Multi-Normal Distribution (or Multivariate Normal Distribution)** : $\underline{\theta} = (\mu, \mathbf{C})$, where μ is a vector of size n_X and $\mathbf{C}$ is a n_X by n_X positive definite symmetric matrix. The probability density function is expressed as:

$$f_X(\underline{x};\underline{\theta}) = \frac{1}{\sqrt{\det \mathbf{C}}(2\pi)^{n_X/2}} \exp\left(-\frac{1}{2}(x-\mu)^t\mathbf{C}^{-1}(x-\mu)\right)$$

  where $(x-\mu)^t$ the transpose vector. We note that a random variable that follows a Normal distribution as described here takes values from $\mathbb{R}^{n_X}$.

Other notations

Link with OpenTURNS methodology

These probability distributions can be used in step B “Quantifying Sources of Uncertainty". Choosing a probability distribution is equivalent to implicitly making a hypothesis on the type of uncertainty of one of the variables $\underline{X}$ defined in step A “Specifying Criteria and the Case Study".

References and theoretical basics

This parametric approach has the advantage of characterizing the uncertainty using a reduced number of parameters. This is particularly useful when there is little data available for the unknown variables (situation in which a non-parametric approach would be limited – see [empirical distribution function] and [kernel smoothing]) and even when there is no data (the analysis can thus only rely on expert judgement, easier to interpret when there are few distribution parameters).

Moreover, a parametric approach is often preferable when the uncertainty study criterion defined in step A is concerned with a rare event, obtaining a precise evaluation of the necessary criteria generally necessitates the extrapolation of X values from the observed data. Beware however! An unwise modelling assumption (bad choice of distribution) can lead to an erroneous extrapolation and thus the results of the study may be false!

The correct choice of probability distribution is thus crucial. Statistical tools are available to validate or invalidate the choice of distribution given a set of data (see for example [Graphical analysis] [Kolmogorov-Smirnov test]). But consideration of the underlying context is also recommended. For example:

- the Normal distribution is relevant in metrology to represent certain measures of uncertainty.
- the Exponential distribution is useful for modelling uncertainty when considering the life duration of material that is not subject to ageing,

- the Gumbel distribution is defined to describe extreme phenomenon (e.g. maximal annual flow of a river or of wind speed)

Certain distributions are often used to express expert judgement in simple terms:

- the Uniform distribution expresses knowledge concerning the absolute limits of variables (i.e. the probability to exceed these limits is strictly zero) without any other prior assumption about the distribution (such as, for example the mean value or the most likely value),
- the Triangular distribution expresses knowledge concerning the absolute limits of variables and the most likely value.

Finally, an important point concerning the multi-dimensional case where $n_X > 1$. Choosing the type of distribution implies an assumption about the uncertainty of each of the variables X^i, but also on the potential inter-dependencies between variables. These inter-dependencies between unknown variables can consequently have an impact on the results of the uncertainty study.
Readers wishing to consider the dependencies in their study more deeply are referred to, for example, [copula method], [linear correlation], [rank correlation].

The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill
- Bhattacharyya, G.K., and R.A. Johnson, (1997). “Statistical Concepts and Methods", John Wiley and Sons, New York.

3.3.4 Step B – Independent Copula - Normal Copula

Mathematical description

Goal

To define the joined probability density function of the random input vector $\underline{X}$ by composition, one needs:

- the specification of the copula of interest C with its parameters,
- the specification of the n_X marginal laws of interest F_{X_i} of the n_X input variables X_i.

The joined cumulative density function is therefore defined by :

$$\mathbb{P}\left(X^1 \leq x^1, X^2 \leq x^2, \cdots, X^{n_X} \leq x^{n_X}\right) = C\left(F_{X^1}(x^1), F_{X^2}(x^2), \cdots, F_{X^{n_X}}(x^{n_X})\right)$$

Within this part, we define the concept of copula and its use within Open TURNS.

Principles

The copulas enable to represent the part of the joined cumulative density function which is not described by the marginal laws. It enables to represent the dependence structure of the input variables. A copula is a special cumulative density function defined on $[0,1]^{n_X}$ whose marginal laws are uniform on $[0,1]$. The choice of the dependence structure is disconnected from the choice of the marginal laws.

Basic properties of copulas

Roughly speaking, a copula is a n_U-dimensional cumulative density function with uniform marginals.

- $C(\underline{u}) \geq 0$, $\forall \underline{u} \in [0,1]^{n_U}$
- $C(\underline{u}) = u_i$, $\forall \underline{u} = (1, \ldots, 1, u_i, 1, \ldots, 1)$
- For all N-box $\mathcal{B} = [a_1, b_1] \times \cdots \times [a_{n_U}, b_{n_U}] \in [0,1]^{n_U}$, we have $\mathcal{V}_C(\mathcal{B}) \geq 0$, where:
  - $\mathcal{V}_C(\mathcal{B}) = \sum_{i=1,\cdots,2^{n_U}} sign(\underline{v}_i) \times C(\underline{v}_i)$, the summation being made over the 2^{n_U} vertices $\underline{v_i}$ of $\mathcal{B}$.
  - $sign(\underline{v}_i) = +1$ if $v_i^k = a_k$ *for an even number of* $k's$, $sign(\underline{v}_i) = -1$ *otherwise.*

Copulas available within Open TURNS

Different copulas are available within Open TURNS:

- *Independent Copula*: It means that all the input variables are independent the ones from the others. The independent copula is defined by:

$$C^{Indep}(u_1, u_2, \cdots, u_{n_U}) = \prod_{i=1}^{n_U} u_i$$

- *Gaussian Copula*: The Gaussian copula is parametrized by a correlation matrix $\mathbf{R}$. The Gaussian copula is thus defined by:

  $$C_{\mathbf{R}}^{Gauss} = \Phi_{\mathbf{R}}^{n_U}\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2), \cdots, \Phi^{-1}(u_{n_U})\right)$$

  where:

  – $\Phi_{\mathbf{R}}^{n_X}$ is the multinormal cumulative density function in dimension n_X:

  $$\Phi_{\mathbf{R}}^{n_X}(\underline{x}) = \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_{n_X}} \frac{1}{(2\pi.\det\mathbf{R})^{\frac{n_X}{2}}} . e^{-\frac{{}^t\underline{u}.\mathbf{R}.\underline{u}}{2}} du_1 \dots du_{n_X}$$

  – Φ is the cumulative distribution function of the normal law in dimension 1:

  $$\Phi(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}} dt$$

  – $\mathbf{R}$ is the correlation matrix. This matrix is defined by its algebric properties: symmetric, definite and positive.

The correlation matrix $\mathbf{R}$ can be obtained by different means:

– If one knows the Spearmann correlation Matrix, that is to say,

$$\rho_{ij}^S = \rho^S(X_i, X_j) = \rho^P(F_{X_i}(X_i), F_{X_j}(X_j))$$

the correlation matrix $\mathbf{R}$ is deduced by the following formula:

$$\mathbf{R}_{ij} = 2\sin(\frac{\pi}{6}\rho_{ij}^S)$$

– If one knows the Kendall measure of correlation, that is to say,

$$\tau_{ij} = \tau(X_i, X_j) = \mathbb{P}\left((X_{i_1} - X_{i_2}).(X_{j_1} - X_{j_2}) > 0\right) - P\left((X_{i_1} - X_{i_2}).(X_{j_1} - X_{j_2}) < 0\right)$$

where (X_{i_1}, X_{j_1}) and (X_{i_2}, X_{j_2}) follow the law of (X_i, X_j), the correlation matrix $\mathbf{R}$ is deduced by the following formula:

$$\mathbf{R}_{ij} = \sin(\frac{\pi}{2}.\tau_{ij})$$

– If one knows the Pearson correlation Matrix $\mathbf{R}^P$, there are two possibilities:

1. If and only if all the marginal laws are Gaussian,

   $$\mathbf{R} \equiv \mathbf{R}^P$$

2. In the other cases, one has to build the correlation matrix $\mathbf{R}$ by inversion of the following formula from the Pearson Correlation Matrix $\mathbf{R}^P$:

   $$\mathbf{R}_{ij}^P = \int\int_{\mathbb{R}^2} (x^i - \mathbb{E}[X^i])(x^j - \mathbb{E}[X^j])\Phi_{ij}(x^i, x^j, \mathbf{R}_{ij}) dx^i dx^j$$

Other notations

Link with OpenTURNS methodology

This method of modelling the dependencies between the input variables is part of the step B of the global methodology ("quantify sources of uncertainty"). It enables to build an expression of the probability density function of the input variables $\underline{X}$ defined in step A ("specification of the model and criteria") by composition with the marginal distributions of each X^i. This method requires the knowledge of the Spearman correlation matrix or the Kendall correlation measure. It can also be used if one knows the Pearson correlation matrix, but only with the assumption of Gaussian marginal laws for all the input variables.

References and theoretical basics

One has to pay attention that the composition of the marginal distributions and the copulas available in Open TURNS is not sufficient to represent all types of dependencies (see examples in the next section). Previous statistical and/or justifications should be done to justify this choice of modeling dependencies. Besides, as previously discussed, the use of Copula is totally decoupled from the knowledge of the marginal laws of the input variables.
The following references give a first entry point to the Copulas:

- Nelsen, 'Introduction to Copulas'
- Embrechts P., Lindskog F., Mc Neil A., 'Modelling dependence with copulas and application to Risk Management', ETZH 2001.

Examples

First, let us present two examples of copulas
Second, we are going to illustrate our way to build pdf for different combinations of copulas and marginal laws. The following examples present the building of the joined pdf for of the couple (X_1, X_2).

1. **Independent Copula** $C^{Indep}(u_1, u_2) = u_1.u_2$
   (a) $X_1 \hookrightarrow \mathcal{U}(-0.5, 0.5)$, $X_2 \hookrightarrow \mathcal{U}(-0.5, 0.5)$
   (b) $X_1 \hookrightarrow \mathcal{U}(-0.5, 0.5)$, $X_2 \hookrightarrow \mathcal{G}(0, 1)$
   (c) $X_1 \hookrightarrow \mathcal{G}(0, 1)$, $X_2 \hookrightarrow \mathcal{G}(0, 1)$
2. **Gaussian Copula** $C^G_{\mathbf{R}}(u_1, u_2) = \Phi^2_{\mathbf{R}}(\Phi^{-1}(u_1), \Phi^{-1}(u_2))$
   (a) $\mathbf{R}$ is a Spearman correlation Matrix, $\mathbf{R} = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}$
      i. $X_1 \hookrightarrow \mathcal{U}(-0.5, 0.5)$, $X_2 \hookrightarrow \mathcal{U}(-0.5, 0.5)$
      ii. $X_1 \hookrightarrow \mathcal{U}(-0.5, 0.5)$, $X_2 \hookrightarrow \mathcal{G}(0, 1)$

iii. $X_1 \hookrightarrow \mathcal{G}(0,1)$, $X_2 \hookrightarrow \mathcal{G}(0,1)$

(b) $\mathbf{R}$ is a Spearman correlation Matrix, $\mathbf{R} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$

i. $X_1 \hookrightarrow \mathcal{U}(-0.5,0.5)$, $X_2 \hookrightarrow \mathcal{U}(-0.5,0.5)$

ii. $X_1 \hookrightarrow \mathcal{U}(-0.5,0.5)$, $X_2 \hookrightarrow \mathcal{G}(0,1)$

iii. $X_1 \hookrightarrow \mathcal{G}(0,1)$, $X_2 \hookrightarrow \mathcal{G}(0,1)$

(c) $\mathbf{R}$ is a Pearson correlation Matrix, $\mathbf{R} = \begin{pmatrix} 1 & -0.8 \\ -0.8 & 1 \end{pmatrix}$

i. $X_1 \hookrightarrow \mathcal{G}(0,1)$, $X_2 \hookrightarrow \mathcal{G}(0,1)$

3.3.5 Step B – Using QQ-plot to compare two samples

Mathematical description

Goal

Let X be a scalar uncertain variable modelled as a random variable. This method is concerned with the construction of a dataset prior to the choice of a probability distribution for X. A QQ-plot (where "QQ" stands for "quantile-quantile") is a tool that may be used to compare two samples $\{x_1, \ldots, x_N\}$ and $\{x'_1, \ldots, x'_M\}$; the goal is to determine graphically whether these two samples come from the same probability distribution or not. If this is the case, the two samples should be aggregated in order to increase the robustness of further statistical analyses.

Principle of the method

A QQ-plot is based on the notion of quantile. The α-quantile $q_X(\alpha)$ of X, where $\alpha \in (0, 1)$, is defined as follows:

$$\mathbb{P}\left(X \leq q_X(\alpha)\right) = \alpha$$

If a sample $\{x_1, \ldots, x_N\}$ of X is available, the quantile can be estimated empirically:

1. the sample $\{x_1, \ldots, x_N\}$ is first placed in ascending order, which gives the sample $\left\{x_{(1)}, \ldots, x_{(N)}\right\}$;
2. then, an estimate of the α-quantile is:

$$\widehat{q}_X(\alpha) = x_{([N\alpha]+1)}$$

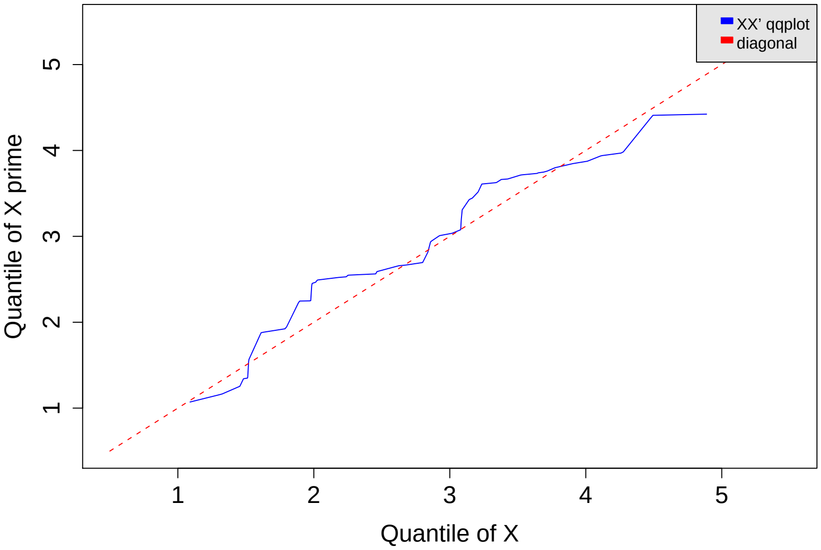
   where $[N\alpha]$ denotes the integral part of $N\alpha$.

Thus, the j^{th} smallest value of the sample $x_{(j)}$ is an estimate $\widehat{q}_X(\alpha)$ of the α-quantile where $\alpha = (j-1)/N$ ($1 < j \leq N$). Let us then consider our second sample $\{x'_1, \ldots, x'_M\}$; this one also provides an estimate $\widehat{q}'_X(\alpha)$ of this same quantile:

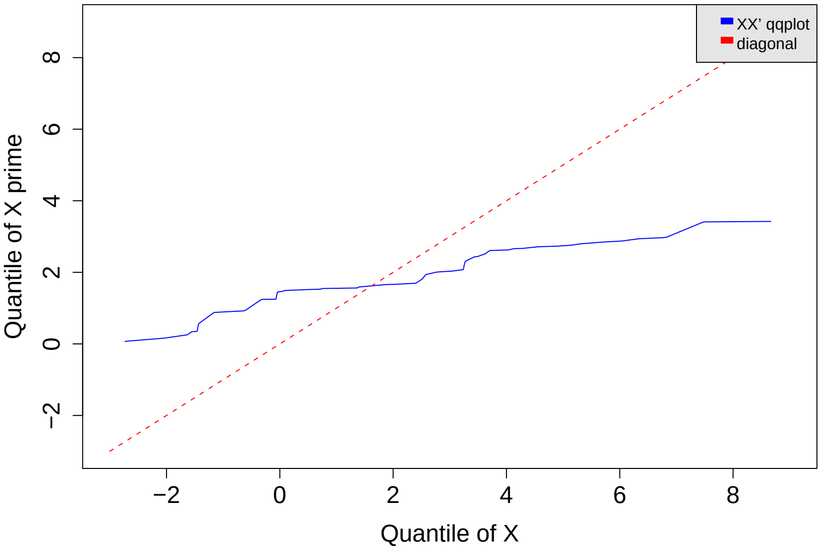
$$\widehat{q}'_X(\alpha) = x'_{([M\times(j-1)/N]+1)}$$

If the the two samples correspond to the same probability distribution, then $\widehat{q}_X(\alpha)$ and $\widehat{q}'_X(\alpha)$ should be close. Thus, graphically, the points $\{(\widehat{q}_X(\alpha), \widehat{q}'_X(\alpha)),\ \alpha = (j-1)/N,\ 1 < j \leq N\}$ should be close to the diagonal.

The following figure illustrates the principle of a QQ-plot with two samples of size $M = 50$ and $N = 50$. Note that the unit of the two axis is that of the variable X studied. In this example, the points remain close to the diagonal and the hypothesis "the two samples come frome the same distribution" does not seem irrelevant, even if a more quantitative analysis (see [Smirnov test]) should be carried out to confirm this.



In this second example, the two samples clearly arise from two different distributions.



Other notations

Link with OpenTURNS methodology

This method is used in step B “Quantifying Sources of Uncertainty". It is a tool for the construction of a dataset that can be used afterwards to choose a probability distribution for some uncertain variables defined in step A “Specifying Criteria and the Case Study".

References and theoretical basics

A QQ-plot is a graphical analysis, the conclusion of which remains obviously subjective. The reader is referred to [Smirnov test] for a more quantitative analysis. The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill
- D'Agostino, R.B. and Stephens, M.A. (1986). “Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Bhattacharyya, G.K., and R.A. Johnson, (1997). “Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). “Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.6 Step B – Comparison of two samples using Smirnov's test

Mathematical description

Goal

Let X be a scalar uncertain variable modelled as a random variable. This method is concerned with the construction of a dataset prior to the choice of a probability distribution for X. Smirnov's test is a tool that may be used to compare two samples $\{x_1,\dots,x_N\}$ and $\{x'_1,\dots,x'_M\}$; the goal is to determine whether these two samples come from the same probability distribution or not. If this is the case, the two samples should be aggregated in order to increase the robustness of further statistical analyses.

Principle of the method

Smirnov's test is a statistical test based on the maximum distance between the cumulative distribution function $\widehat{F}_N$ and $\widehat{F}'_M$ of the samples $\{x_1,\dots,x_N\}$ and $\{x'_1,\dots,x'_M\}$ (see [empirical cumulative distribution function]). This distance is expressed as follows:

$$\widehat{D}_{M,N} = \sup_x \left| \widehat{F}_N\left(x\right) - \widehat{F}'_M\left(x\right) \right|$$

The probability distribution of the distance $\widehat{D}_{M,N}$ is asymptotically known (i.e. as the size of the samples tends to infinity). If M and N are sufficiently large, this means that for a probability α, one can calculate the threshold / critical value d_α such that:

- if $\widehat{D}_{M,N} > d_\alpha$, we conclude that the two samples are not identically distributed, with a risk of error α,
- if $\widehat{D}_{M,N} \leq d_\alpha$, it is reasonable to say that both samples arise frome the same distribution.

An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the "identically-distributed" hypothesis is rejected. Thus, the two samples will be supposed identically distributed if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

Other notations

This test is also referred to as the Kolmogorov-Smirnov's test for two samples.

Link with OpenTURNS methodology

This method is used in step B "Quantifying Sources of Uncertainty". It is a tool for the construction of a dataset that can be used afterwards to choose a probability distribution for some uncertain variables defined in step A "Specifying Criteria and the Case Study".

References and theoretical basics

The test is concerned with the maximum deviation between the tw empirical distributions; it is by nature highly sensitive to presence of local deviations (two samples may be rejected even if they seem similar for almost the whole domain of variation).

We remind the reader that the underlying theoretical results of the test are asymptotic. There is no rule to determine the minimum number of data values one needs to use this test; but it is often considered a reasonable approximation when N is of an order of a few dozen.
The following bibliographical references provide main starting points for further study of this method:

- Saporta G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon W.J. & Massey F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill

3.3.7 Step B – Maximum Likelihood Method

Mathematical description

Goal

This method is concerned with the parametric modelling of a probability distribution for a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. The appropriate probability distribution is found by using a sample of data $\{\underline{x}_1, \ldots, \underline{x}_N\}$. Such an approach can be described in two steps as follows:

- Choose a probability distribution (e.g. the Normal distribution, or any other distribution available in OpenTURNS see [standard parametric models]),
- Find the parameter values $\underline{\theta}$ that characterize the probability distribution (e.g. the mean and standard deviation for the Normal distribution) which best describes the sample $\{\underline{x}_1, \ldots, \underline{x}_N\}$.

The maximum likelihood method is used for the second step.

Principle

In the current version of Open TURNS this method is restricted to the case where $n_X = 1$ and continuous probability distributions. Please note therefore that $\underline{X} = X^1 = X$ in the following text. The maximum likelihood estimate (MLE) of $\underline{\theta}$ is defined as the value of $\underline{\theta}$ which maximizes the likelihood function $L\left(X, \underline{\theta}\right)$:

$$\hat{\underline{\theta}} = \mathrm{argmax}\ L\left(X, \underline{\theta}\right)$$

Given that $\{x_1, \ldots, x_N\}$ is a sample of independent identically distributed (i.i.d) observations, $L\left(x_1, \ldots, x_N, \underline{\theta}\right)$ represents the probability of observing such a sample assuming that they are taken from a probability distribution with parameters $\underline{\theta}$. In concrete terms, the likelihood $L\left(x_1, \ldots, x_N, \underline{\theta}\right)$ is calculated as follows:

$$L\left(x_1, \ldots, x_N, \underline{\theta}\right) = \prod_{j=1}^{N} f_X\left(x_j; \underline{\theta}\right) \text{ if the distribution is continuous, with density } f_X\left(x; \underline{\theta}\right)$$

For example, if we suppose that X is a Gaussian distribution with parameters $\underline{\theta} = \{\mu, \sigma\}$ (i.e. the mean and standard deviation),

$$\begin{aligned} L\left(x_1, \ldots, x_N, \underline{\theta}\right) &= \prod_{j=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x_j - \mu}{\sigma}\right)^2\right] \\ &= \frac{1}{\sigma^N (2\pi)^{N/2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{j=1}^{N} \left(x_j - \mu\right)^2\right] \end{aligned}$$

The following figure graphically illustrates the maximum likelihood method, in the particular case of a Gaussian probability distribution.

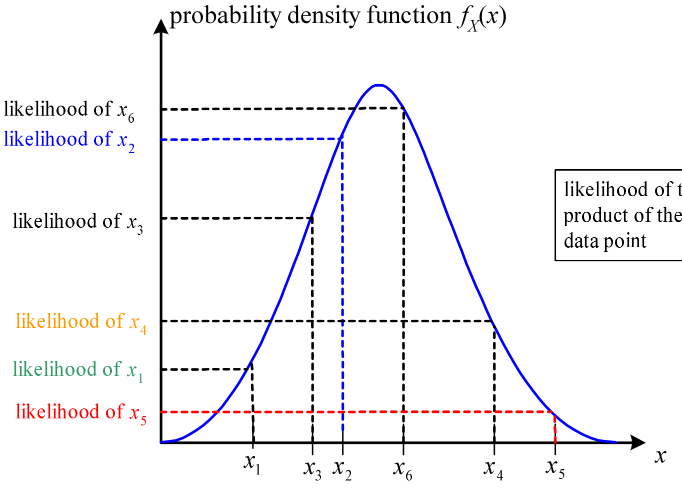


Illustration of likelihood calculation for N=6 and a normal distribution

In general, in order to maximize the likelihood function classical optimisation algorithms (e.g. gradient type) can be used. The Gaussian distribution case is an exception to this, as the maximum likelihood estimators are obtained analytically:

$$\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} x_i, \ \widehat{\sigma^2} = \frac{1}{N} \sum_{i=1}^{N} (x_i - \widehat{\mu})^2$$

Other notations

Link with OpenTURNS methodology

Having specified the variable of interest and having defined a criterion (step A “Specifying Criteria and the Case Study"), the uncertainty of the input variable X^i must be then quantified in step B. The superscript i is omitted, as only a single component is used here, that is a single unknown variable (or source of uncertainty).

Input:
$\{x_1, \ldots, x_N\}$: sample data
Distribution : : Distribution type chosen from the proposed continuous 1-dimensional distributions in [standard parametric models]

Output :
$\underline{\hat{\theta}}$: maximum likelihood estimate of $\underline{\theta}$

References and theoretical basics

The sample size used in the maximum likelihood method has an effect on the quality of results. In fact:

- as N tends to infinity, the asymptotic theory results assure, under certain assumptions concerning the regularity of the model, that the MLE is the best possible estimator (its bias tends towards 0 i.e. no tendency towards under- or over-estimation, the uncertainty of $\underline{\hat{\theta}}$ is lesser than in all other unbiased estimation methods); in practice, one often considers the asymptotic behaviour to be reached when $N \geq$ a few dozens, even if no theoretical rule can assure this with certitude.
- if N is smaller, the MLE is still useful but $\underline{\hat{\theta}}$ is less robust (uncertainty greater and bias possible).

A more advanced study of the goodness-of-fit of the selected probability distribution with the given sample data is described in [Graphical analysis] [Kolmogorov-Smirnov test] , [Cramer-Von Mises test] , [Anderson-Darling test] and [BIC criterion].
The following bibliographical references provide main starting points for further study of this method:

- Saporta G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon W.J. & Massey F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill

3.3.8 Step B – Graphical goodness-of-fit analysis

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to verify the compatibility between a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and a candidate probability distribution previous chosen. Open TURNS enables the use of graphical tools to answer this question in the one dimensional case $n_X = 1$, and with a continuous distribution.

Principle of the method

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. The first graphical tool provided by Open TURNS is a QQ-plot (where “QQ" stands for “quantile-quantile"). In the specific case of a Normal distribution (see [standard parametric models]), Henry’s line may also be used.

QQ-plot A QQ-Plot is based on the notion of quantile. The α-quantile $q_X(\alpha)$ of X, where $\alpha \in (0, 1)$, is defined as follows:

$$\mathbb{P}\left(X \leq q_X(\alpha)\right) = \alpha$$

If a sample $\{x_1, \ldots, x_N\}$ of X is available, the quantile can be estimated empirically:

1. the sample $\{x_1, \ldots, x_N\}$ is first placed in ascending order, which gives the sample $\left\{x_{(1)}, \ldots, x_{(N)}\right\}$;

2. then, an estimate of the α-quantile is:

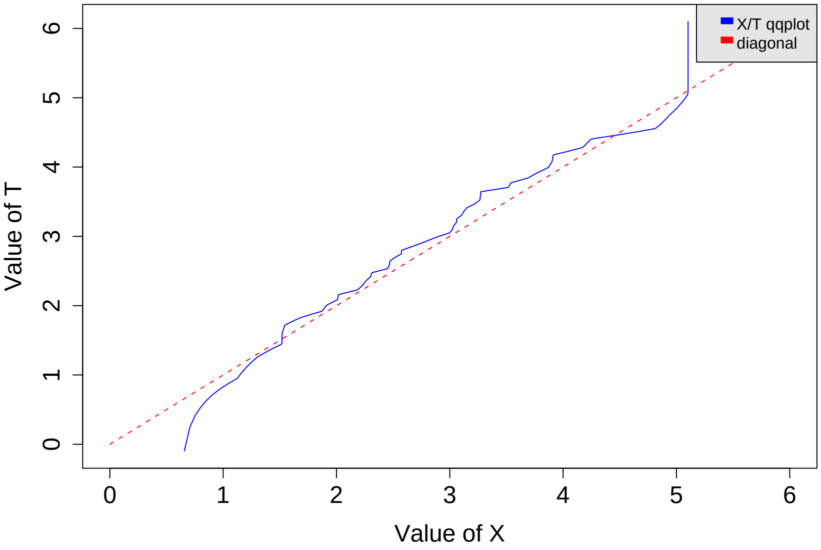
$$\widehat{q}_X(\alpha) = x_{([N\alpha]+1)}$$

   where $[N\alpha]$ denotes the integral part of $N\alpha$.

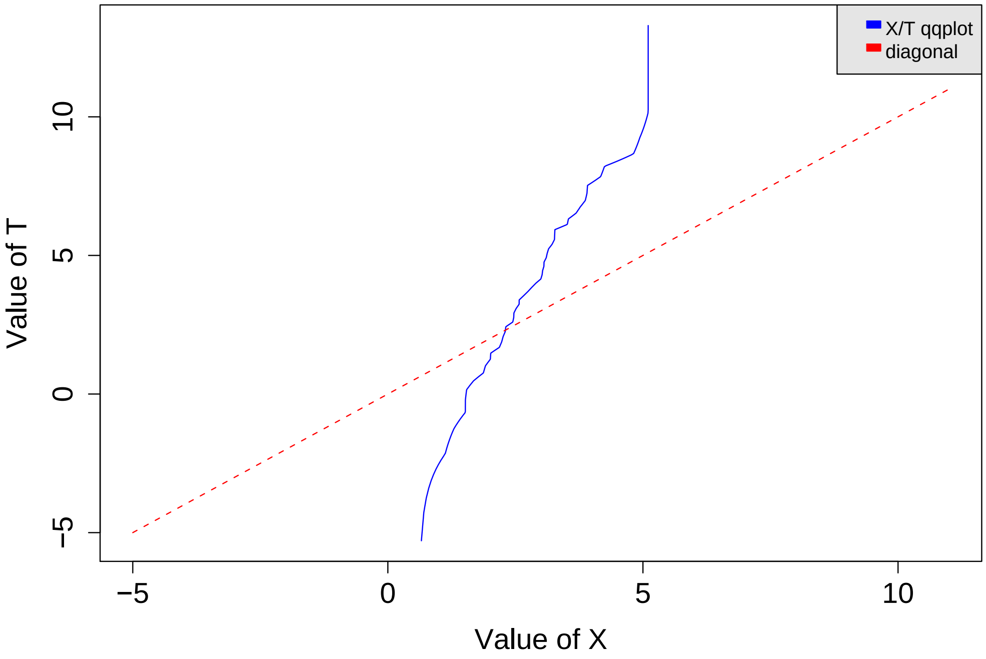
Thus, the j^{th} smallest value of the sample $x_{(j)}$ is an estimate $\widehat{q}_X(\alpha)$ of the α-quantile where $\alpha = (j-1)/N$ $(1 < j \leq N)$.
Let us then consider the candidate probability distribution being tested, and let us denote by F its cumulative distribution function. An estimate of the α-quantile can be also computed from F:

$$\widehat{q}'_X(\alpha) = F^{-1}\left((j-1)/N\right)$$

If F is really the cumulative distribution function of F, then $\widehat{q}_X(\alpha)$ and $\widehat{q}'_X(\alpha)$ should be close. Thus, graphically, the points $\{(\widehat{q}_X(\alpha), \widehat{q}'_X(\alpha)), \ \alpha = (j-1)/N, \ 1 < j \leq N\}$ should be close to the diagonal.
The following figure illustrates the principle of a QQ-plot with a sample of size $N = 50$. Note that the unit of the two axis is that of the variable X studied; the quantiles determined via F are called here “value of T". In this example, the points remain close to the diagonal and the hypothesis “F is the cumulative distribution function of X" does not seem irrelevant, even if a more quantitative analysis (see for instance [Kolmogorov-Smirnov goodness-of-fit test]) should be carried out to confirm this.



In this second example, the candidate distribution function is clearly irrelevant.



Henry's line This second graphical tool is only relevant if the candidate distribution function being tested is gaussian. It also uses the ordered sample $\left\{x_{(1)}, \ldots, x_{(N)}\right\}$ introduced for the QQ-plot, and the empirical cumulative distribution function $\widehat{F}_N$ presented in [empirical cumulative distribution function].

By definition,

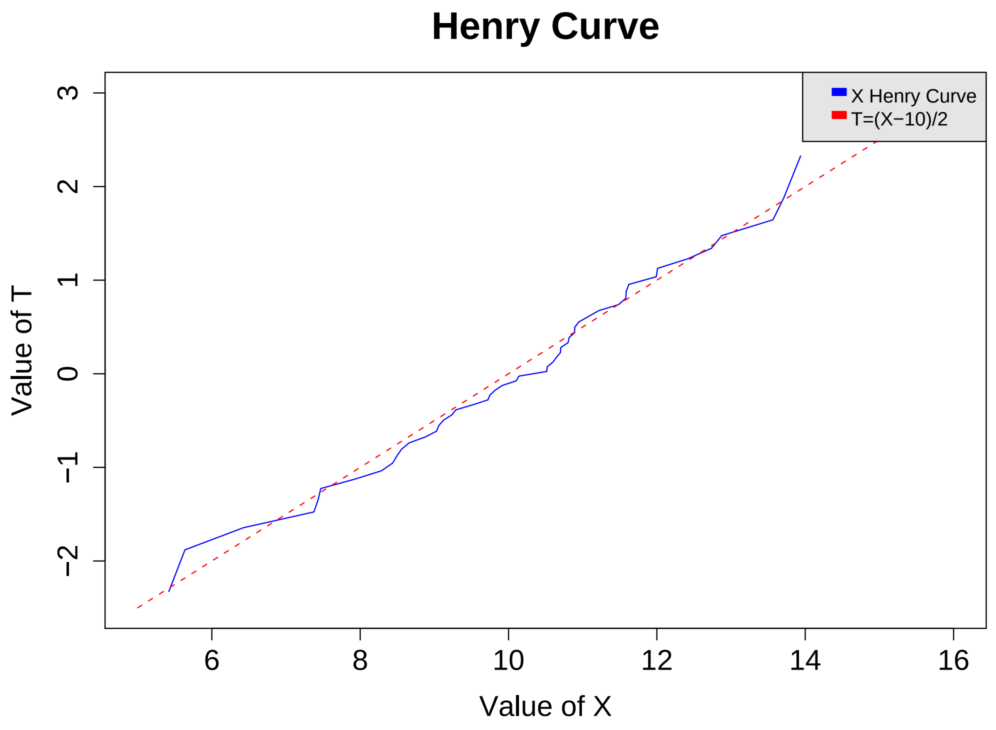
$$x_{(j)} = \widehat{F}_N^{-1}\left(\frac{j}{N}\right)$$

Then, let us denote by Φ the cumulative distribution function of a Normal distribution with mean 0 and standard deviation 1. The quantity $t_{(j)}$ is defined as follows:

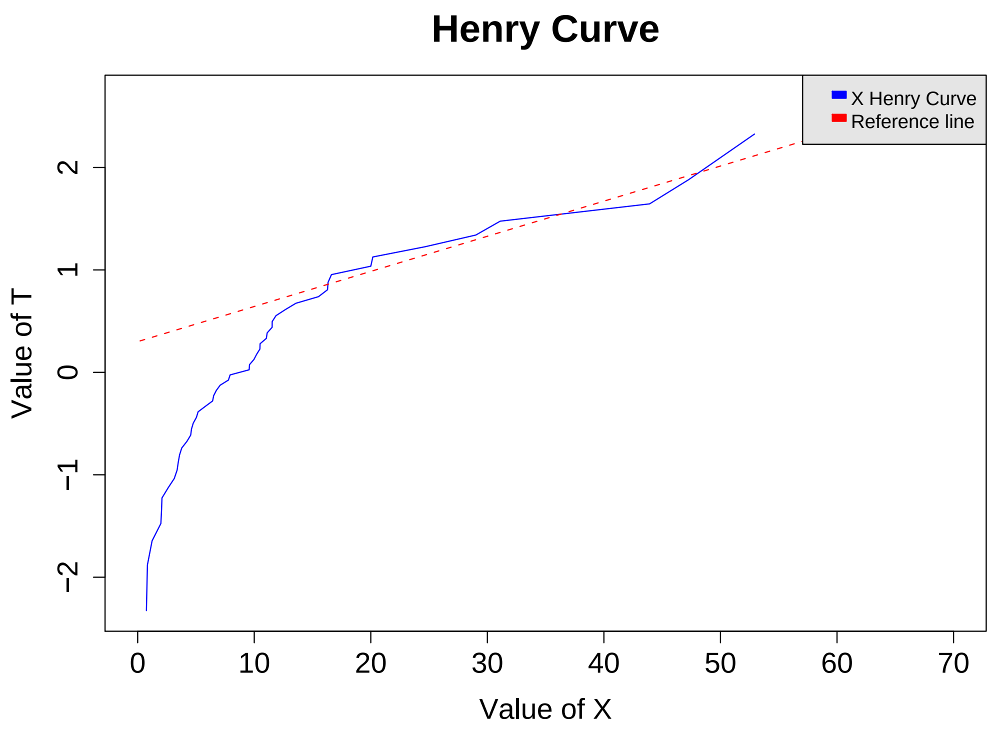
$$t_{(j)} = \Phi^{-1}\left(\frac{j}{N}\right)$$

If X is distributed according to a normal probability distribution with mean μ and standard-deviation σ, then the points $\left\{\left(x_{(j)}, t_{(j)}\right),\ 1 \leq j \leq N\right\}$ should be close to the line defined by $t = (x-\mu)/\sigma$. This comes from a property of a normal distribution: it the distribution of X is really $\mathcal{N}(\mu, \sigma)$, then the distribution of $(X-\mu)/\sigma$ is $\mathcal{N}(0,1)$.

The following figure illustrates the principle of Henry's graphical test with a sample of size $N = 50$. Note that only the unit of the horizontal axis is that of the variable X studied. In this example, the points remain close to a line and the hypothesis “the distribution function of X is a gaussian one" does not seem irrelevant, even if a more quantitative analysis (see for instance [Kolmogorov-Smirnov goodness-of-fit test]) should be carried out to confirm this.



In this second example, the hypothesis of a gaussian distribution seems far less relevant because of the behaviour for small values of X.



Other notations

Link with OpenTURNS methodology

This method is used in step B "Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A "Specifying Criteria and the Case Study".

References and theoretical basics

Since QQ-plot and Henry's line are graphical analysis, their conclusion remain obviously subjective. The reader is referred to [Komogorov-Smirnov test], [Cramer-Von-Mises test], [Anderson-Darling test] for a more quantitative analysis.
The following bibliographical references provide main starting points for further study of this method:

- Saporta G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon W.J. & Massey F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill

3.3.9 Step B – Chi-squared goodness of fit test

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to verify the compatibility between a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and a candidate probability distribution previous chosen. Open TURNS enables the use of the χ^2 Goodness-of-Fit test to answer this question in the one dimensional case $n_X = 1$, and with a discrete distribution.

Principle

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. We also note that as we are considering discrete distributions i.e. those for which the possible values of X belong to a discrete set $\mathcal{E}$, the candidate distribution is characterised by the probabilities $\{p(x;\underline{\theta})\}_{x\in\mathcal{E}}$.
The chi squared test is based on the fact that if the candidate distribution is appropriate, the number of values in the sample x1, x2, ..., xN that are equal to x should be on average equal to $Np(x;\underline{\theta})$. The idea is therefore to compare the “theoretical values" with the actual observed values. This comparison is performed with the aid of the following “distance".

$$\widehat{D}_N^2 = \sum_{x\in\mathcal{E}_N} \frac{\left(Np(x) - n(x)\right)^2}{n(x)}$$

where $\mathcal{E}_N$ denotes the elements of $\mathcal{E}$ which have been observed at least once in the data sample and where $n(x)$ denotes the number of data values in the sample that are equal to x.

The probability distribution of the distance $\widehat{D}_N^2$ is asymptotically known (i.e. as the size of the sample tends to infinity), and this asymptotic distribution does not depend on the candidate distribution being tested. If N is sufficiently large, this means that for a probability α, one can calculate the threshold / critical value) d_α such that:

- if $\widehat{D}_N > d_\alpha$, we reject the candidate distribution with a risk of error α,
- if $\widehat{D}_N \leq d_\alpha$, the candidate distribution is considered acceptable.

An important notion is the so-called “p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the candidate distribution is rejected. Thus, the candidate distribution will be accepted if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

Other notations

Link with OpenTURNS methodology

This method is used in step B “Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A "Specifying Criteria and the Case Study".

Input data:
$\{x_1, \ldots, x_N\}$: data sample
Distribution : probability distribution that we are testing for goodness-of-fit

Parameters:
α : Level of significance for the test

Outputs:
Result : Binary variable specifying whether the candidate distribution is rejected (0) or not (1)
α_{lim} : p-value of the test

References and theoretical basics

The test is suitable for discrete distributions. It cannot be used for continuous distributions except by means of an arbitrary discretisation of possible values of X, an important source of potential error. Readers interested in Goodness of Fit tests for continuous variables are referred to [Kolmogorov-Smirnov test] , [Cramer-Von Mises test], [Anderson-Darling test] in the reference documentation.

Even for discrete distributions, certain precautions must be taken when using this test. Firstly, the critical value d_α is only valid for a sufficiently large sample size. No rule exists to determine the minimum number of data values necessary in order to use this test; it is often thought, however, that the approximation is reasonable when N is of the order of a few dozen. But whatever the value of N, the distance – and similarly the p-value – remains a useful tool for comparing different probability distributions to a sample. The distribution which minimizes $\widehat{D}_N$ – or maximizes the p-value – will be of interest to the analyst.
On the other hand, the calculation of d_α and of the p-value should in theory be modified if we are testing the goodness of fit of a parametric model and if the parameters of the candidate distribution have been estimated from the same sample. The current version of Open TURNS, however, does not permit such a modification, and so the results must be used with care when the p-value α_{lim} and the desired error risk α are very close.

The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- D'Agostino, R.B. and Stephens, M.A. (1986). "Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). "Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.10 Step B – Kolmogorov-Smirnov goodness-of-fit test

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to verify the compatibility between a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and a candidate probability distribution previous chosen. Open TURNS enables the use of the Kolmogorov-Smirnov Goodness-of-Fit test to answer this question in the one dimensional case $n_X = 1$, and with a continuous distribution.

Principle

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. This goodness-of-fit test is based on the maximum distance between the cumulative distribution function $\widehat{F}_N$ of the sample $\{x_1, x_2, \ldots, x_N\}$ (see [empirical cumulative distribution function]) and that of the candidate distribution, denoted F. This distance may be expressed as follows:

$$D = \sup_x \left|\widehat{F}_N\left(x\right) - F\left(x\right)\right|$$

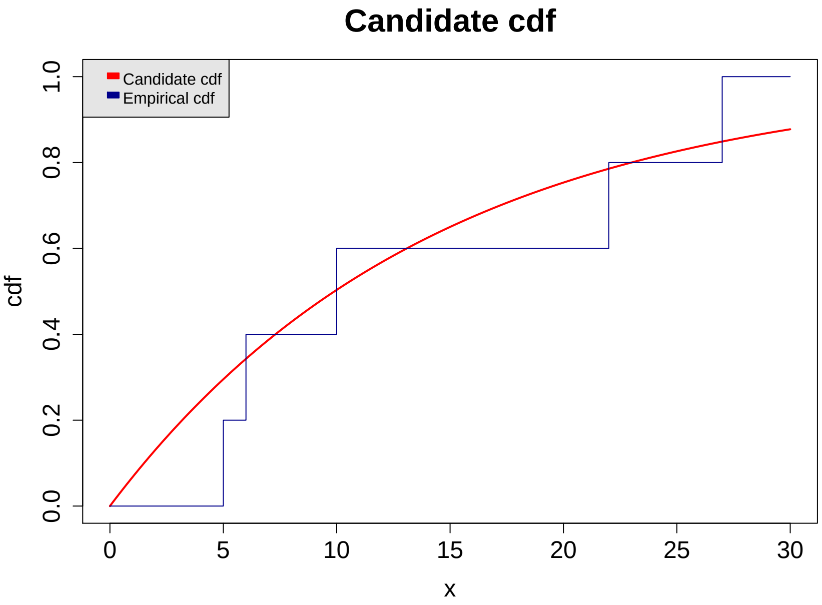
With a sample $\{x_1, x_2, \ldots, x_N\}$, the distance is estimated by:

$$\widehat{D}_N = \sup_{i=1\ldots N} \left|F\left(x_i\right) - \frac{i-1}{N}; \frac{i}{N} - F\left(x_i\right)\right|$$

The probability distribution of the distance $\widehat{D}_N$ is asymptotically known (i.e. as the size of the sample tends to infinity). If N is sufficiently large, this means that for a probability α and a candidate distribution type, one can calculate the threshold / critical value d_α such that:

- if $\widehat{D}_N > d_\alpha$, we reject the candidate distribution with a risk of error α,
- if $\widehat{D}_N \leq d_\alpha$, the candidate distribution is considered acceptable.

Note that d_α does not depend on the candidate distribution F being tested, and the test is therefore relevant for any continuous distribution.
An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the candidate distribution is rejected. Thus, the candidate distribution will be accepted if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.
The diagram below illustrates the principle of comparison with the empirical cumulative distribution function for an ordered sample $\{5, 6, 10, 22, 27\}$; the candidate distribution considered here is the Exponential distribution with parameters $\lambda = 0.07$, $\gamma = 0$ (see [standard parametric models]).



Other notations

This method is also referred to in the literature as Kolmogorov's Test.

Link with OpenTURNS methodology

This method is used in step B "Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A "Specifying Criteria and the Case Study".

Input data:
$\{x_1, \ldots, x_N\}$: data sample
Distribution : probability distribution that we are testing for goodness-of-fit

Parameters:
α : Level of significance for the test

Outputs:
Result : Binary variable specifying whether the candidate distribution is rejected (0) or not (1)
α_{lim} : p-value of the test

References and theoretical basics

The test is concerned with the maximum deviation between the empirical distribtuion and the candidate distribution, it is by nature highly sensitive to presence of local deviations (a candidate distribution may be rejected even if it correctly describes the sample for almost the whole domain of variation).

We remind the reader that the underlying theoretical results of the test are asymptotic. There is no rule to determine the minimum number of data values one needs to use this test; but it is often considered a reasonable approximation when N is of an order of a few dozen. But whatever the value of N, the distance – and similarly the p-value – remains a useful tool for comparing different probability distributions to a sample. The distribution which minimizes $\widehat{D}_N$ – or maximizes the p-value – will be of interest to the analyst. We also point out that the calculation of d_α should in theory be modified if on is testing the goodness-of-fit to a parametric model where the parameters have been estimated from the same sample. The current version of Open TURNS does not allow this modification, and the results should be therefore used with caution when the p-value $\alpha_{\textrm{lim}}$ and the desired error risk α are very close.
Readers interested in Goodness of Fit tests for continuous distributions are referred to[Cramer-Von Mises test] and [Anderson-Darling test] in the reference documentation.
The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill
- NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/
- D'Agostino, R.B. and Stephens, M.A. (1986). “Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Bhattacharyya, G.K., and R.A. Johnson, (1997). “Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). “Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.11 Step B – Cramer-Von Mises goodness-of-fit test

Mathematical description

Objective

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to verify the compatibility between a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and a candidate probability distribution previous chosen. Open TURNS enables the use of the Cramer-von-Mises Goodness-of-Fit test to answer this question in the one dimensional case $n_X = 1$, and with a continuous distribution. The current version is limited to the case of the Normal distribution.

Principle

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. This goodness-of-fit test is based on the distance between the cumulative distribution function $\widehat{F}_N$ of the sample $\{x_1, x_2, \ldots, x_N\}$ (see [empirical cumulative distribution function]) and that of the candidate distribution, denoted F. This distance is no longer the maximum deviation as in the [Kolmogorov-Smirnov test] but the distance squared and integrated over the entire variation domain of the distribution:

$$D = \int_{-\infty}^{\infty} \left[F\left(x\right) - \widehat{F}_N\left(x\right)\right]^2 dF$$

With a sample $\{x_1, x_2, \ldots, x_N\}$, the distance is estimated by:

$$\widehat{D}_N = \frac{1}{12N} + \sum_{i=1}^{N} \left[\frac{2i-1}{2N} - F\left(x_i\right)\right]^2$$

The probability distribution of the distance $\widehat{D}_N$ is asymptotically known (i.e. as the size of the sample tends to infinity). If N is sufficiently large, this means that for a probability α and a candidate distribution type, one can calculate the threshold / critical value d_α such that:

- if $\widehat{D}_N > d_\alpha$, we reject the candidate distribution with a risk of error α,
- if $\widehat{D}_N \leq d_\alpha$, the candidate distribution is considered acceptable.

Note that d_α depends on the candidate distribution F being tested; the current version of Open TURNS is limited to the case of the Normal distribution.
An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the candidate distribution is rejected. Thus, the candidate distribution will be accepted if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

Other notations

-

Link with OpenTURNS methodology

This method is used in step B "Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A "Specifying Criteria and the Case Study".

Input data:
$\{x_1, \ldots, x_N\}$: data sample
Distribution : normal probability distribution that we are testing for goodness-of-fit

Parameters:
α : Level of significance for the test

Outputs:
Result : Binary variable specifying whether the candidate distribution is rejected (0) or not (1)
α_{lim} : p-value of the test

References and theoretical basics

The test concerns the deviation squared and integrated over the entire variation domain, it often appears to be more robust than the Kolmogorov-Smirnov test.

We remind the reader that the underlying theoretical results of the test are asymptotic. There is no rule to determine the minimum number of data values one needs to use this test; but it is often considered a reasonable approximation when N is of an order of a few dozen. But whatever the value of N, the distance – and similarly the p-value – remains a useful tool for comparing different probability distributions to a sample. The distribution which minimizes $\widehat{D}_N$ – or maximizes the p-value – will be of interest to the analyst. We also point out that the calculation of d_α should in theory be modified if on is testing the goodness-of-fit to a parametric model where the parameters have been estimated from the same sample. The current version of Open TURNS does not allow this modification, and the results should be therefore used with caution the p-value α_{lim} and the desired error risk α are very close.
Readers interested in Goodness of Fit tests for continuous distributions are referred to [Kolmogorov-Smirnov test] and [Anderson-Darling test] in the reference documentation.

The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- D'Agostino, R.B. and Stephens, M.A. (1986). "Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). "Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.12 Step B – Anderson-Darling goodness-of-fit test

Mathematical description

Objective

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to verify the compatibility between a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$ and a candidate probability distribution previous chosen. Open TURNS enables the use of the Anderson-Darling Goodness-of-Fit test to answer this question in the one dimensional case $n_X = 1$, and with a continuous distribution. The current version is limited to the case of the Normal distribution.

Principle

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. This goodness-of-fit test is based on the distance between the cumulative distribution function $\widehat{F}_N$ of the sample $\{x_1, x_2, \ldots, x_N\}$ (see [empirical cumulative distribution function]) and that of the candidate distribution, denoted F. This distance is a quadratic type, as in the [Cramer-Von Mises test], but gives more weight to deviations of extreme values:

$$D = \int_{-\infty}^{\infty} \frac{\left[F\left(x\right) - \widehat{F}_N\left(x\right)\right]^2}{F(x)\left(1 - F(x)\right)}\, dF(x)$$

With a sample $\{x_1, x_2, \ldots, x_N\}$, the distance is estimated by:

$$\widehat{D}_N = -N - \sum_{i=1}^{N} \frac{2i-1}{N} \left[\ln F(x_{(i)}) - \ln\left(1 - F(x_{(N+1-i)})\right)\right]$$

where $\left\{x_{(1)}, \ldots, x_{(N)}\right\}$ describes the sample placed in ascending order.
The probability distribution of the distance $\widehat{D}_N$ is asymptotically known (i.e. as the size of the sample tends to infinity). If N is sufficiently large, this means that for a probability α and a candidate distribution type, one can calculate the threshold / critical value d_α such that:

- if $\widehat{D}_N > d_\alpha$, we reject the candidate distribution with a risk of error α,
- if $\widehat{D}_N \leq d_\alpha$, the candidate distribution is considered acceptable.

Note that d_α depends on the candidate distribution F being tested; the current version of Open TURNS is limited to the case of the Normal distribution.
An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the candidate distribution is rejected. Thus, the candidate distribution will be accepted if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

Other notations

-

Link with OpenTURNS methodology

This method is used in step B “Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A “Specifying Criteria and the Case Study".

Input data:
$\{x_1, \ldots, x_N\}$: data sample
Distribution : normal probability distribution that we are testing for goodness-of-fit

Parameters:
α : Level of significance for the test

Outputs:
$\widehat{D}_N$: Distance between theoretical and empirical values
d_α : Threshold / Critical value which if exceeded the tested probability is rejected
Result : Binary variable specifying whether the candidate distribution is rejected or not

References and theoretical basics

The Anderson-Darling test is theoretically designed to be more sensitive to the quality of fit in the tails of the distribution. A user interested in the extreme values of the source of uncertainty being studied will find this particularly interesting but we stress that both tails of the distribution, upper and lower, will influence the test results.

We remind the reader that the underlying theoretical results of the test are asymptotic. There is no rule to determine the minimum number of data values one needs to use this test; but it is often considered a reasonable approximation when N is of an order of a few dozen. But whatever the value of N, the distance – and similarly the p-value – remains a useful tool for comparing different probability distributions to a sample. The distribution which minimizes $\widehat{D}_N$ – or maximizes the p-value – will be of interest to the analyst. We also point out that the calculation of d_α should in theory be modified if on is testing the goodness-of-fit to a parametric model where the parameters have been estimated from the same sample. The current version of Open TURNS does not allow this modification, and the results should be therefore used with caution the p-value $\alpha_{\lim}$ and the desired error risk α are very close.
Readers interested in Goodness of Fit tests for continuous distributions are referred to [Kolmogorov-Smirnov test] and [Cramer-von-Mises test] in the reference documentation.

The following bibliographical references provide main starting points for further study of this method:

- NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/
- D'Agostino, R.B. and Stephens, M.A. (1986). “Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Sprent, P., and Smeeton, N.C. (2001). “Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.13 Step B – Bayesian Information Criterion (BIC)

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It seeks to rank variable candidate distributions by using a sample of data $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_N\}$. Open TURNS enables the use of the Bayesian Information Criterion (BIC) to answer this question in the one dimensional case $n_X = 1$.

Principle

Let us limit the case to $n_X = 1$. Thus we denote $\underline{X} = X^1 = X$. Moreover, let us denote by $\mathcal{M}_1,\ldots, \mathcal{M}_K$ the parametric models envisaged by the user among the [standard parametric models]. We suppose here that the parameters of these models have been estimated previously by the [maximum likelihood method] on the basis of the sample $\{\underline{x}_1, \underline{x}_2, \ldots, \underline{x}_n\}$. We denote by L_i the maximized likelihood for the model $\mathcal{M}_i$.
By definition of the likelihood, the higher L_i, the better the model describes the sample. However, using the likelihood as a criterion to rank the candidate probability distributions would involve a risk: one would almost always favour complex models involving many parameters. If such models provide indeed a large numbers of degrees-of-freedom that can be used to fit the sample, one has to keep in mind that complex models may be less robust that simpler models with less parameters. Actually, the limited available information (N data points) does not allow to estimate robustly too many parameters.
The BIC criterion can be used to avoid this problem. The principle is to rank $\mathcal{M}_1,\ldots, \mathcal{M}_K$ according to the following quantity:

$$\mathrm{BIC}_i = \log\left(L_i\right) - \frac{p_i}{2}\log(n)$$

where p_i denotes the number of parameters being adjusted for the model $\mathcal{M}_i$. The larger BIC_i, the better the model. Note that the idea is to introduce a penalization term that increases with the numbers of parameters to be estimated. A complex model will then have a good score only if the gain in terms of likelihood is high enough to justify the number of parameters used.
The term “Bayesian Information Criterion" comes the interpretation of the quantity BIC_i. In a bayesian context, the unknow “true" model may be seen as a random variable. Suppose now that the user does not have any informative prior information on which model is more relevant among $\mathcal{M}_1,\ldots, \mathcal{M}_K$; all the models are thus equally likely from the point of view of the user. Then, one can show that BIC_i is an approximation of the posterior distribution's logarithm for the model $\mathcal{M}_i$.

Other notations

Link with OpenTURNS methodology

This method is used in step B “Quantifying Sources of Uncertainty", to verify if the probability distribution is appropriate to describe the uncertainty of a component X^i of the vector of unknown variables defined in step A “Specifying Criteria and the Case Study".

References and theoretical basics

Compared to other criteria proposed in literature for model selection and based on the same idea of penalization (such as the AIC criterion), the BIC criterion tends to favour models with a small number of parameters. Moreover, note that the undelying hypothesis is that the user does not have any significant prior information on which model is more relevant; if such prior information is available (for instance via literature or expert judgement), the BIC criterion becomes less relevant.
Readers interested in other ways to rank candidate models referred to [Kolmogorov-Smirnov test] , [Cramer-Von Mises test] and [Anderson-Darling test] in the reference documentation.
The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- D'Agostino, R.B. and Stephens, M.A. (1986). "Goodness-of-Fit Techniques", Marcel Dekker, Inc., New York.
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.
- Burnham, K.P., and Anderson, D.R (2002). "Model Selection and Multimodel Inference: A Practical Information Theoretic Approach", Springer

3.3.14 Step B – Pearson Correlation Coefficient

Mathematical description

Goal

This method is concerned with the parametric modelling of a probability distribution for a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It aims to measure a type of dependence (here a linear correlation) which may exist between two components X^i and X^j.

Principle

The Pearson's correlation coefficient $\rho_{U,V}$ aims to measure the strength of a linear relationship between two random variables U and V. It is defined as follows:

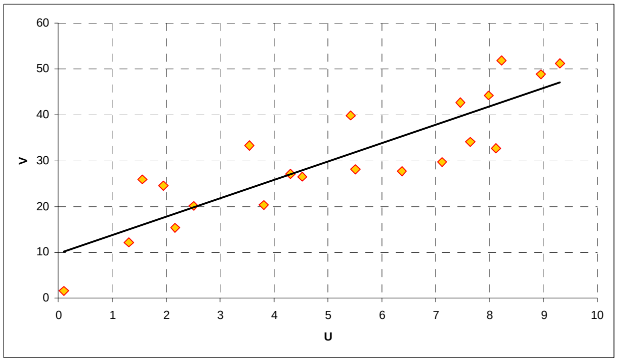
$$\rho_{U,V} = \frac{\mathrm{Cov}\left[U,V\right]}{\sigma_U \sigma_V}$$

where $\mathrm{Cov}\left[U,V\right] = \mathbb{E}\left[\left(U - m_U\right)\left(V - m_V\right)\right]$, $m_U = \mathbb{E}\left[U\right]$, $m_V = \mathbb{E}\left[V\right]$, $\sigma_U = \sqrt{\mathrm{Var}\left[U\right]}$ and $\sigma_V = \sqrt{\mathrm{Var}\left[V\right]}$. If we have a sample made up of a set of N pairs $\{(u_1,v_1),(u_2,v_2),\ldots,(u_N,v_N)\}$, Pearson's correlation coefficient can be estimated using the formula:

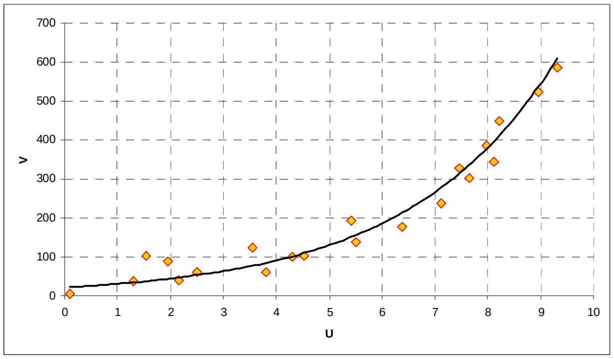
$$\widehat{\rho}_{U,V} = \frac{\displaystyle\sum_{i=1}^{N} \left(u_i - \overline{u}\right)\left(v_i - \overline{v}\right)}{\sqrt{\displaystyle\sum_{i=1}^{N} \left(u_i - \overline{u}\right)^2 \left(v_i - \overline{v}\right)^2}}$$

where $\overline{u}$ and $\overline{v}$ represent the empirical means of the samples $(u_1,\ldots,u_N)$ and $(v_1,\ldots,v_N)$.
Pearson's correlation coefficient takes values between -1 and 1. The closer its absolute value is to 1, the stronger the indication is that a linear relationship exists between variables U and V. The sign of Pearson's coefficient indicates if the two variables increase or decrease in the same direction (positive coefficient) or in opposite directions (negative coefficient). We note that a correlation coefficient equal to 0 does not necessarily imply the independence of variables U and V: this property is in fact theoretically guaranteed only if U and V both follow a Normal distribution. In all other cases, there are two possible situations in the event of a zero Pearson's correlation coefficient:

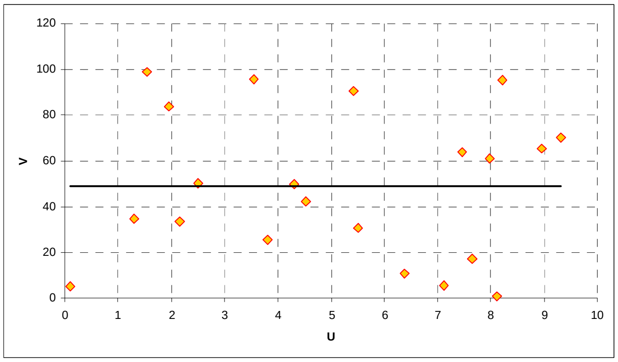
- the variables U and V are in fact independent,
- or a non-linear relationship exists between U and V.



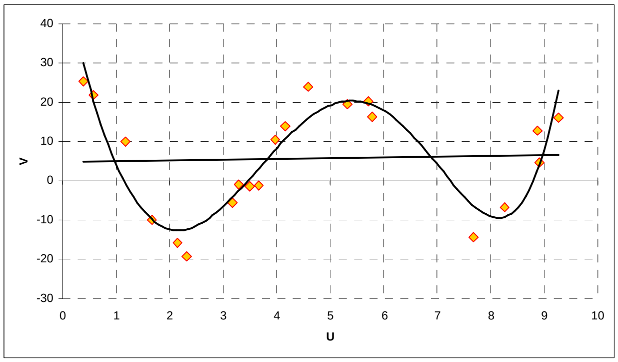
A linear relationship exists between U and V : Pearson's coefficient is a relevant measure of dependency in this case



There is a strong relationship between U and V, but it is non-linear : Pearson's coefficent is not a relevant measure of dependency in his case



The estimate of Pearson's coefficient is quite close to 0 in this case because U and V are independant



The estimate of Pearson's coefficient is quite close to 0 in this case, even though U and V are far from being independent (this is because the non-monotonic trend is to complex to be handled by Pearson's coefficent)

Other notations

The estimate $\widehat{\rho}$ of Pearson's correlation coefficient is sometimes denoted by r.

Link with OpenTURNS methodology

Pearson's correlation coefficient can be used in step B “Quantifying Sources of Uncertainty". Having defined the vector $\underline{X}$ of input variables in step A “Specifying Criteria and the Case Study", [Pearson's Independence Test] shows how to test for the existence of a linear type of dependency between two components X^i and X^j. Such a relationship should in fact be taken in to account so as not to falsify the results of step C “Propagation of Uncertainty".
Pearson's correlation coefficient is also used in step C' “Sensitivity Analysis and Ranking of Sources of Uncertainty". If a propagation of uncertainty with Monte-Carlo simulation (step C, [Mean and Variance Estimation using Standard Monte Carlo]) has been carried out, [Pearson's Ranking] shows the user how to class the components of the input vector $\underline{X}$ according to their impact on the uncer-

tainty of a final variable / output variable defined in step A.

References and theoretical basics

Regardless of the method used in step B or step C', we recall that the Pearson's coefficient is only useful in measuring a linear relationship between two variables. Readers are referred to the following references:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.

3.3.15 Step B – Pearson's correlation test

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{nx}\right)$. It seeks to find a type of dependency (here a linear correlation) which may exist between two components X^i and X^j.

Principle

The Pearson's correlation coefficient $\rho_{U,V}$, defined in [Pearson's Coefficient], measures the strength of a linear relationship between two random variables U and V. If we have a sample made up of N pairs $\{(u_1, v_1), (u_2, v_2), (u_N, v_N)\}$, we denote $\widehat{\rho}_{U,V}$ to be the estimated coefficient.
Even in the case where two variables U and V have a Pearson's coefficient $\rho_{U,V}$ equal to zero, the estimate $\widehat{\rho}_{U,V}$ obtained from the sample may be non-zero: the limited sample size does not provide the perfect image of the real correlation. Pearson's test nevertheless enables one to determine if the value obtained by $\widehat{\rho}_{U,V}$ is significantly different from zero. More precisely, the user first chooses a probability α. From this value the critical value d_α is calculated such that:

- if $|\widehat{\rho}_{U,V}| > d_\alpha$, one can conclude that the real Pearson's correlation coefficient $\rho_{U,V}$ is not zero; the risk of error in making this assertion is controlled and equal to α;
- if $|\widehat{\rho}_{U,V}| \leq d_\alpha$, there is insufficient evidence to reject the null hypothesis $\rho_{U,V} = 0$.

An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the null correlation hypothesis is rejected. Thus, Pearson's coefficient is supposed non zero if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

Other notations

-

Link with OpenTURNS methodology

The Pearson's test is used in step B "Quantifying Sources of Uncertainty". It enables us to verify if a linear type of dependency exists between the two components X^i and X^j of the input variable vector $\underline{X}$ defined in step A "Specifying Criteria and the Case Study". Such a relationship should in fact be taken into account to avoid distortion of results in step C "Propagation of Uncertainty".

Input data :

Two samples $\{x^i_1, \ldots, x^i_N\}$ and $\left\{x^j_1, \ldots, x^j_N\right\}$ of variables X^i and X^j, each pair $\left(x^i_k, x^j_k\right)$ corresponding to a simultaneous sampling of the two variables

Parameters :

a probability α taking values strictly between 0 and 1, defining the risk of permissible decision error (significance level)

Outputs :

Result : Binary variable specifying whether the hypothesis of a correlation coefficient equal to 0 is rejected (0) or not (1)

$\alpha_{\lim}$: p-value of the test

References and theoretical basics

Certain precautions should be taken when interpreting the Pearson's test results.

- The underlying theory of the Pearson test assumes in fact that the variables X^i and X^j are both normally distributed. In all other cases, the decision produced by the test is only valid if the sample size N is sufficiently large (in practice $N \geq$ a few dozen, even if there is no theoretical result that enables us to prove that asymptotic behaviour has been attained).

- Still considering the case of distributions other than the Normal distribution, whatever the value of N, we recall that $\rho_{X^i,X^j} = 0$ does not enable us to conclude that X^i and X^j are independent (see [Pearson's Correlation Coefficient]).

- More generally, the numerical value of Pearson's correlation coefficient can only be interpreted when the two variables studied X^i and X^j are related in a linear way; the scatter plot of points $\left\{(x^i_1, x^j_1), \ldots, (x^i_N, x^j_N)\right\}$ provides some indication concerning the validity of this hypothesis.

The following pages describe methods which enable us to test the hypothesis of the Normal distribution using the available sample $\{x^i_1, \ldots, x^i_N\}$ and $\left\{x^j_1, \ldots, x^j_N\right\}$: [Kolmogorov-Smirnov Goodness of Fit Test], [Cramer-von Mises Goodness of Fit Test], [Anderson-Darling Goodness of Fit Test].

Out of Pearson's test validity domain (i.e. linear relationship, Normal distributions), [Spearman's test] provides some answers.

The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip

- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill

- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.

3.3.16 Step B – Spearman correlation coefficient

Mathematical description

Goal

This method is concerned with the parametric modelling of a probability distribution for a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It aims to measure a type of dependence (here a monotonous correlation) which may exist between two components X^i and X^j.

Principle

The Spearman's correlation coefficient $\rho^S_{U,V}$ aims to measure the strength of a monotonic relationship between two random variables U and V. It is in fact equivalent to the Pearson's correlation coefficient after having transformed U and V to linearize any monotonic relationship (remember that Pearson's correlation coefficient may only be used to measure the strength of linear relationships, see[Pearson's Correlation Coefficient]):

$$\rho^S_{U,V} = \rho_{F_U(U),F_V(V)}$$

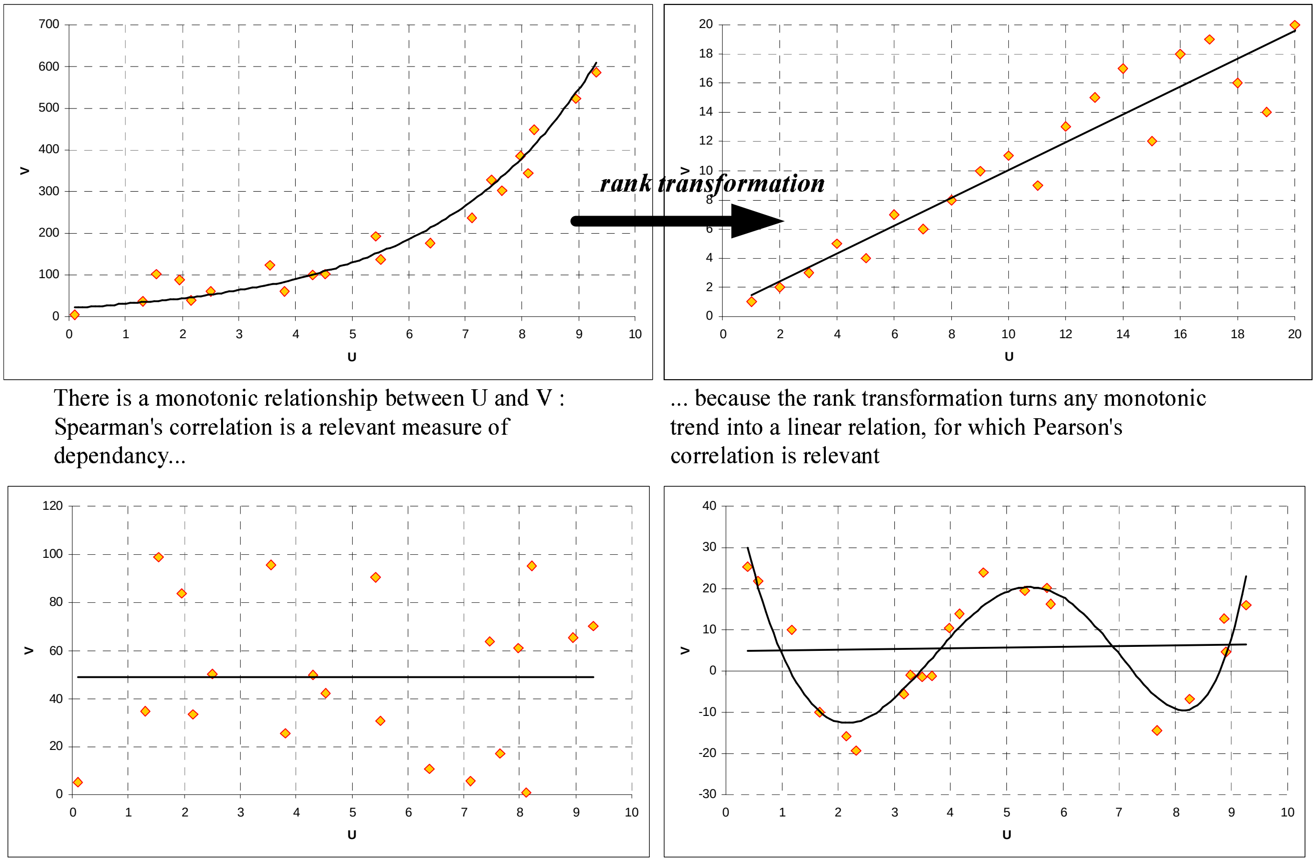
where F_U and F_V denote the cumulative distribution functions of U and V.
If we arrange a sample made up of N pairs $\{(u_1,v_1),(u_2,v_2),\ldots,(u_N,v_N)\}$, the estimation of Spearman's correlation coefficient first of all requires a ranking to produce two samples $(u_1,\ldots,u_N)$ and $(v_1,\ldots,v_N)$. The ranking $u_{[i]}$ of the observation u_i is defined as the position of u_i in the sample reordered in ascending order: if u_i is the smallest value in the sample $(u_1,\ldots,u_N)$, its ranking would equal 1; if u_i is the second smallest value in the sample, its ranking would equal 2, and so forth. The ranking transformation is a procedure that takes the sample $(u_1,\ldots,u_N))$ as input data and produces the sample $(u_{[1]},\ldots,u_{[N]})$ as an output result.
For example, let us consider the sample $(u_1,u_2,u_3,u_4) = (1.5, 0.7, 5.1, 4.3)$. We therefore have $(u_{[1]},u_{[2]}u_{[3]},u_{[4]}) = (2,1,4,3)$. $u_1 = 1.5$ is in fact the second smallest value in the original, $u_2 = 0.7$ the smallest, etc.
The estimation of Spearman's correlation coefficient is therefore equal to Pearson's coefficient estimated with the aid of the N pairs $(u_{[1]},v_{[1]})$, $(u_{[2]},v_{[2]})$, …, $(u_{[N]},v_{[N]})$:

$$\widehat{\rho}^S_{U,V} = \frac{\displaystyle\sum_{i=1}^{N}\left(u_{[i]}-\overline{u}_{[]}\right)\left(v_{[i]}-\overline{v}_{[]}\right)}{\sqrt{\displaystyle\sum_{i=1}^{N}\left(u_{[i]}-\overline{u}_{[]}\right)^2\left(v_{[i]}-\overline{v}_{[]}\right)^2}}$$

where $\overline{u}_{[]}$ and $\overline{v}_{[]}$ represent the empirical means of the samples $(u_{[1]},\ldots,u_{[N]})$ and $(v_{[1]},\ldots,v_{[N]})$.
The Spearman's correlation coefficient takes values between -1 and 1. The closer its absolute value is to 1, the stronger the indication is that a monotonic relationship exists between variables U and V. The sign of Spearman's coefficient indicates if the two variables increase or decrease in the same direction (positive coefficient) or in opposite directions (negative coefficient). We note that a correlation coefficient equal to 0 does not necessarily imply the independence of variables U and V. There are two possible situations in the event of a zero Spearman's correlation coefficient:

- the variables U and V are in fact independent,

- or a non-monotonic relationship exists between U and V.



There is a monotonic relationship between U and V : Spearman's correlation is a relevant measure of dependancy...

... because the rank transformation turns any monotonic trend into a linear relation, for which Pearson's correlation is relevant

In this case, the estimate of Spearman's correlation coefficient is close to 0 because U and V are independant

In this case, the estimate of Spearman's correlation coefficient is close to 0 even though U and V are not independant (this is because the non-monotonic trend cannot be handled by Spearman's coefficient)

Other notations

Spearman's coeeficient is often referred to as the rank correlation coefficient.

Link with OpenTURNS methodology

Spearman's correlation coefficient can be used in step B “Quantifying Sources of Uncertainty". Having defined the vector $\underline{X}$ of input variables in step A “Specifying Criteria and the Case Study", [Spearman's Independence Test] shows how to test for the existence of a monotonous type of dependency between two components X^i and X^j. Such a relationship should in fact be taken in to account so as not to falsify the results of step C “Propagation of Uncertainty".
Spearman's correlation coefficient is also used in step C' “Sensitivity Analysis and Ranking of Sources of Uncertainty". If a propagation of uncertainty with Monte-Carlo simulation (step C, [Mean and Variance Estimation using Standard Monte Carlo]) has been carried out, [Spearman's Ranking]

shows the user how to class the components of the input vector $\underline{X}$ according to their impact on the uncertainty of a final variable / output variable defined in step A.

References and theoretical basics

Regardless of the method used in step B or step C', we recall that the Spearman's coefficient is only useful in measuring a monotonous relationship between two variables. Readers are referred to the following references:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). "Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.17 Step B – Spearman correlation test

Mathematical description

Goal

This method is concerned with the modelling of a probability distribution of a random vector $\underline{X} = \left(X^1, \ldots, X^{nx}\right)$. It seeks to find a type of dependency (here a monotonous correlation) which may exist between two components X^i and X^j.

Principle

The Spearman's correlation coefficient $\rho^S_{U,V}$, defined in [Spearman's Coefficient], measures the strength of a monotonous relationship between two random variables U and V. If we have a sample made up of N pairs $\{(u_1, v_1), (u_2, v_2), (u_N, v_N)\}$, we denote $\widehat{\rho}^S_{U,V}$ to be the estimated coefficient.
Even in the case where two variables U and V have a Spearman's coefficient $\rho^S_{U,V}$ equal to zero, the estimate $\widehat{\rho}^S_{U,V}$ obtained from the sample may be non-zero: the limited sample size does not provide the perfect image of the real correlation. Pearson's test nevertheless enables one to determine if the value obtained by $\widehat{\rho}^S_{U,V}$ is significantly different from zero. More precisely, the user first chooses a probability α. From this value the critical value d_α is calculated automatically such that:

- if $\left|\widehat{\rho}^S_{U,V}\right| > d_\alpha$, one can conclude that the real Spearman's correlation coefficient $\rho^S_{U,V}$ is not zero; the risk of error in making this assertion is controlled and equal to α;
- if $\left|\widehat{\rho}^S_{U,V}\right| \leq d_\alpha$, there is insufficient evidence to reject the null hypothesis $\rho^S_{U,V} = 0$.

An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability $\alpha_{\textrm{lim}}$ under which the null correlation hypothesis is rejected. Thus, Spearman's's coefficient is supposed non zero if and only if $\alpha_{\textrm{lim}}$ is greater than the value α desired by the user. Note that the higher $\alpha_{\textrm{lim}} - \alpha$, the more robust the decision.

he

Other notations

-

Link with OpenTURNS methodology

The Spearman's test is used in step B "Quantifying Sources of Uncertainty". It enables us to verify if a monotonous type of dependency exists between the two components X^i and X^j of the input variable vector $\underline{X}$ defined in step A "Specifying Criteria and the Case Study". Such a relationship should in fact be taken into account to avoid distortion of results in step C "Propagation of Uncertainty".

Input data :

Two samples $\{x_1^i, \ldots, x_N^i\}$ and $\left\{x_1^j, \ldots, x_N^j\right\}$ of variables X^i and X^j, each pair $\left(x_k^i, x_k^j\right)$ corresponding to a simultaneous sampling of the two variables

Parameters :

a probability α taking values strictly between 0 and 1, defining the risk of permissible decision error (significance level)

Outputs :

Result : Binary variable specifying whether the hypothesis of a correlation coefficient equal to 0 is rejected (0) or not (1)

$\alpha_{\lim}$: p-value of the test

References and theoretical basics

Certain precautions should be taken when interpreting the Spearman's test results.

- Remember that $\rho_{X^i,X^j} = 0$ does not enable us to conclude that X^i and X^j are independent (see [Spearman's correlation coefficient]).
- More generally, the numerical value of Spearman's correlation coefficient can only be interpreted when the two variables studied X^i and X^j are related in a monotonous way; the scatter plot of points $\left\{(x_1^i, x_1^j), \ldots, (x_N^i, x_N^j)\right\}$ provides some indication concerning the validity of this hypothesis.

The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). "Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) "Introduction to statistical analysis (4th ed.)", McGraw-Hill
- Bhattacharyya, G.K., and R.A. Johnson, (1997). "Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). "Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.18 Step B – Chi-squared test for independence

Mathematical description

Goal

This method is concerned with the parametric modelling of a probability distribution for a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. We seek here to detect possible dependencies that may exist between two components X^i and X^j. In response to this, Open TURNS offers the use of the χ^2 test for Independence for discrete probability distributions.

Principle

As we are considering discrete distributions, the possible values for X^i and X^j respectively belong to the discrete sets $\mathcal{E}_i$ and $\mathcal{E}_j$. The χ^2 test of independence can be applied when we have a sample consisting of N pairs $\left\{(x_1^i, x_1^j), (x_2^i, x_2^j), (x_N^i, x_N^j)\right\}$. We denote:

- $n_{u,v}$ the number of pairs in the sample such that $x_k^i = u$ and $x_k^j = v$,
- n_u^i the number of pairs in the sample such that $x_k^i = u$,
- n_v^j the number of pairs in the sample such that $x_k^j = v$.

The test thus uses the quantity denoted $\widehat{D}_N^2$:

$$\widehat{D}_N^2 = \sum_{u \in \mathcal{E}_i} \sum_{v \in \mathcal{E}_2} \frac{\left(p_{u,v} - p_v^j p_u^i\right)^2}{p_u^i p_v^j}$$

where:

$$p_{u,v} = \frac{n_{u,v}}{N},\ p_u^i = \frac{n_u^i}{N},\ p_v^j = \frac{n_v^j}{N}$$

The probability distribution of the distance $\widehat{D}_N^2$ is asymptotically known (i.e. as the size of the sample tends to infinity). If N is sufficiently large, this means that for a probability α, one can calculate the threshold (critical value) d_α such that:

- if $\widehat{D}_N > d_\alpha$, we conclude, with the risk of error α, that a dependency exists between X^i and X^j,
- if $\widehat{D}_N \leq d_\alpha$, the independence hypothesis is considered acceptable.

An important notion is the so-called "p-value" of the test. This quantity is equal to the limit error probability α_{lim} under which the independence hypothesis is rejected. Thus, independence is assumed if and only if α_{lim} is greater than the value α desired by the user. Note that the higher $\alpha_{\text{lim}} - \alpha$, the more robust the decision.

Other notations

This method is also referred to in the literature as the χ^2 test of contingency.

Link with OpenTURNS methodology

The χ^2 independence test is used in step B “Quantifying Sources of Uncertainty". It enables the existence of a dependency between two components X^i and X^j of the input vector $\underline{X}$, defined in step A “Specifying Criteria and the Case Study", to be verified.

Input data :

Two samples $\left\{x_1^i, \ldots, x_N^i\right\}$ and $\left\{x_1^j, \ldots, x_N^j\right\}$ of variables X^i and X^j, each pair $\left(x_k^i, x_k^j\right)$ corresponding to a simultaneous sampling of the two variables

Parameters :

a probability α taking values strictly between 0 and 1, defining the risk of permissible decision error (significance level)

Outputs :

Result : Binary variable specifying whether the hypothesis of independence is rejected (0) or not (1)

α_{lim} : p-value of the test

References and theoretical basics

The χ^2 test of independence can be applied when the two variables of study are discrete. Its use for continuous distributions is only possible by means of an arbitrary discretisation of possible values of X, a high source of potential error.

On the other hand, no hypothesis is made in the form of the relationship between the two tested variables. Readers interested in the detection of dependencies between two continuous variables are referred to [Pearson’s Test] and [Spearman’s test] in the reference documentation.

The following bibliographical references provide main starting points to further study of this method:

- Saporta, G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill
- Bhattacharyya, G.K., and R.A. Johnson, (1997). “Statistical Concepts and Methods", John Wiley and Sons, New York.
- Sprent, P., and Smeeton, N.C. (2001). “Applied Nonparametric Statistical Methods – Third edition", Chapman & Hall

3.3.19 Step B – Linear regression

Mathematical description

Goal

This method is concerned with the parametric modelling of a probability distribution for a random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$. It aims to measure a type of dependence (here a linear relation) which may exist between a component X^i and other uncertain variables X^j.

Principle of the method

The principle of the multiple linear regression model is to find the function that links the variable X^i to other variables X^{j_1},…,X^{j_K} by means of a linear model:

$$X^i = a_0 + \sum_{j \in \{j_1,\ldots,j_K\}} a_j X^j + \varepsilon$$

where ε describes a random variable with zero mean and standard deviation σ independent of the input variables X^i. For given values of X^{j_1},…,X^{j_K}, the average forecast of X^i is denoted by $\widehat{X}^i$ and is defined as:

$$\widehat{X}^i = a_0 + \sum_{j \in \{j_1,\ldots,j_K\}} a_j X^j$$

The estimators for the regression coefficients $\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_K$, and the standard deviation σ are obtained from a sample of $(X^i, X^{j_1}, \ldots, X^{j_K})$, that is a set of N values $(x_1^i, x_1^{j_1}, \ldots, x_1^{j_K})$,…,$(x_n^i, x_n^{j_1}, \ldots, x_n^{j_K})$. They are determined via the least-squares method:

$$\{\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_K\} = \text{argmin} \sum_{k=1}^{n} \left[x_k^i - a_0 - \sum_{j \in \{j_1,\ldots,j_K\}} a_j x_k^j \right]^2$$

In other words, the principle is to minimize the total quadratic distance between the observations x_k^i and the linear forecast $\widehat{x}_k^i$.

Some estimated coefficient $\widehat{a}_\ell$ may be close to zero, which may indicate that the variable X^{j_ℓ} does not bring valuable information to forecast X^i. Open TURNS includes a classical statistical test to identify such situations: Fisher's test. For each estimated coefficient $\widehat{a}_\ell$, an important characteristic is the so-called "p-value" of Fisher's test. The coefficient is said to be "significant" if and only if $\alpha_{\ell \mathrm{lim}}$ is greater than a value α chosen by the user (typically 5% or 10%). The higher the p-value, the more significant the coefficient.

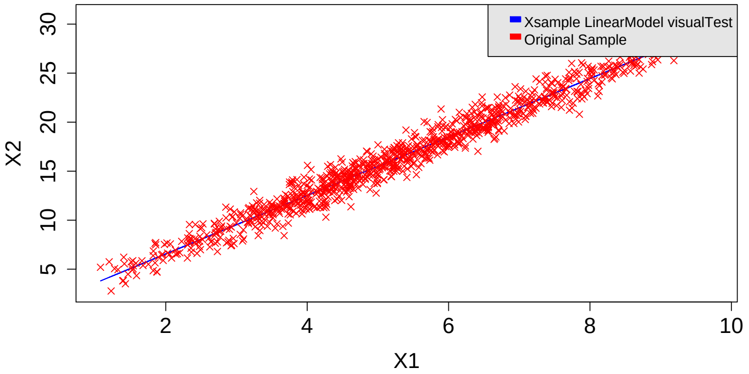
Another important characteristic of the adjusted linear model is the coefficient of determination R^2. This quantity indicates the part of the variance of X^i that is explained by the linear model:

$$R^2 = \frac{\displaystyle\sum_{k=1}^{n} \left(x_k^i - \overline{x}^i\right)^2 - \sum_{k=1}^{n} \left(x_k^i - \widehat{x}_k^i\right)^2}{\sum_{k=1}^{n} \left(x_k^i - \overline{x}^i\right)^2}$$

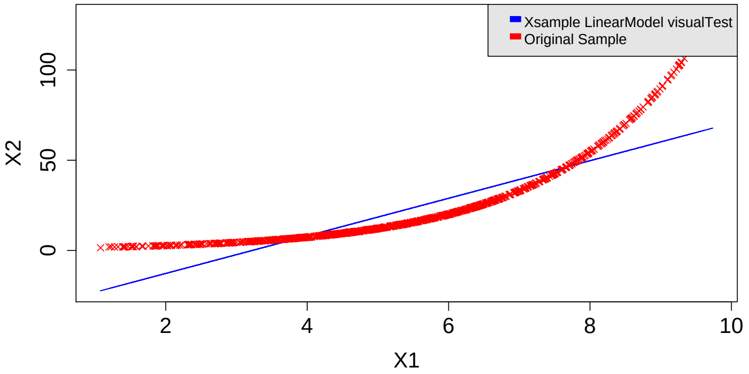
where $\overline{x}^i$ denotes the empirical mean of the sample $\left\{x_1^i, \ldots, x_n^i\right\}$.

Thus, $0 \leq R^2 \leq 1$. A value close to 1 indicates a good fit of the linear model, whereas a value close to 0 indicates that the linear model does not provide a relevant forecast. A statistical test allows to detect

significant values of R^2. Again, a p-value is provided: the higher the p-value, the more significant the coefficient of determination.
By definition, the multiple regression model is only relevant for linear relationships, as in the following simple example where $X^2 = a_0 + a_1 X^1$.



In this second example (still in dimension 1), the linear model is not relevant because of the exponential shape of the relation. But a linear approach would be useful on the transformed problem $X^2 = a_0 + a_1 \exp X^1$. In other words, what is important is that the relationships between X^i and the variables X^{j_1},...,X^{j_K} is linear with respect to the regression coefficients a_j.

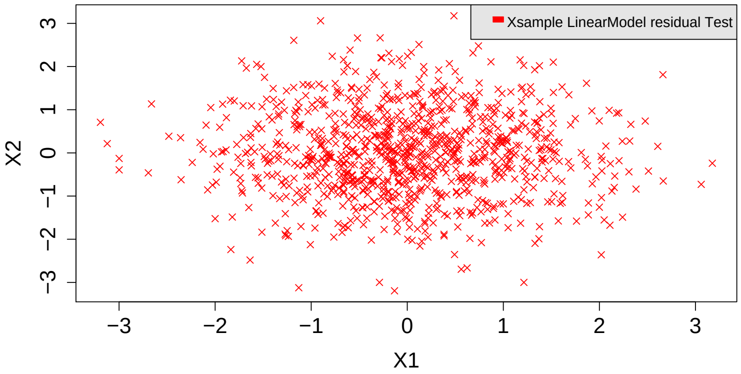


The value of R^2 is a good indication of the goodness-of fit of the linear model. However, several other verifications have to be carried out before concluding that the linear model is satisfactory. For instance, one has to pay attentions to the “residuals" $\{u_1, \ldots, u_N\}$ of the regression:

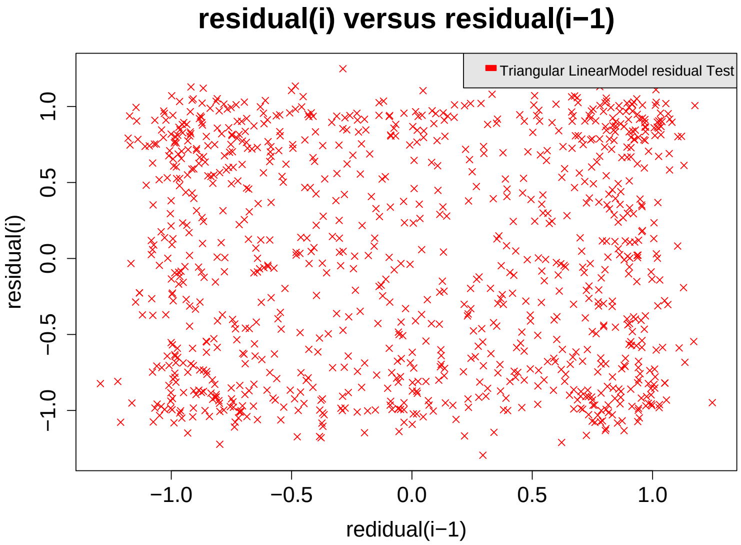
$$u_j = x^i - \widehat{x}^i$$

A residual is thus equal to the difference between the observed value of X^i and the average forecast provided by the linear model. A key-assumption for the robustness of the model is that the characteristics of the residuals do not depend on the value of X^i, X^{j_1},...,X^{j_K}: the mean value should be close to 0 and the standard deviation should be constant. Thus, plotting the residuals versus these variables can fruitful.

In the following example, the behaviour of the residuals is satisfactory: no particular trend can be detected neither in the mean nor in he standard deviation.



The next example illustrates a less favourable situation: the mean value of the residuals seems to be close to 0 but the standard deviation tends to increase with X. In such a situation, the linear model should be abandoned, or at least used very cautiously.



Other notations

Link with OpenTURNS methodology

Multiple linear regression can be used in step B “Quantifying Sources of Uncertainty". Having defined the vector $\underline{X}$ of input variables in step A “Specifying Criteria and the Case Study", linear regression allows to detect a linear type of dependency between uncertain variables. Such a relationship should in fact be taken in to account so as not to bias the results of step C “Propagation of Uncertainty".

References and theoretical basics

As we have seen in the mathematical description, there is a consequent list of verifications that have to be carried to validate the linear model. In particular, underlying assumptions on the residuals are important to ensure the robustness of the average forecast. Detecting a non-conform behaviour of the residuals can also provide leads on transformations that could be carried out before applying linear regression (such as considering the logarithm of a variable instead of the variable itself).
The following bibliographical references provide main starting points for further study of this method:

- Saporta, G. (1990). “Probabilités, Analyse de données et Statistique", Technip
- Dixon, W.J. & Massey, F.J. (1983) “Introduction to statistical analysis (4th ed.)", McGraw-Hill
- NIST/SEMATECH e-Handbook of Statistical Methods, http://www.itl.nist.gov/div898/handbook/
- Bhattacharyya, G.K., and R.A. Johnson, (1997). “Statistical Concepts and Methods", John Wiley and Sons, New York.

4 Open TURNS' methods for Step C: uncertainty propagation

This section is organized according to the different uncertainty criterion defined in step A: deterministic min-max criterion, probabilist criterion on central dispersion (expectation and variance), probability of exceeding a threshold / failure probability, and probabilistic criterion based on quantiles. Each method proposed for these criteria is described at the end of the section.

4.1 Deterministic min-max criterion

Only a simplified approach is available in the current version of Open TURNS.

- [Seeking extreme values on a discrete set of inputs chosen through desig of experiment] – see page 79

4.2 Probabilistic criteria

4.2.1 Central dispersion

Two categories of method are proposed: approximation methods and sampling methods.

- Approximation methods
  - [Quadratic combination / Perturbation method] – see page 81
- Sampling methods
  - [Standard Monte-Carlo simulation] – see page 84

4.2.2 Probability of exceeding a threshold / failure probability / probability of an event

Again, two categories of method are proposed: approximation methods and sampling methods.

- Approximation methods
  - FORM-SORM methods
    - [Preliminary iso-probabilistic transformation] – see page 87
    - [FORM algorithm] – see page 91
    - [SORM algorithm] – see page 95
    - [Reliability index] – see page 99
  - Validation of FORM-SORM underlying hypothesis
    - [Preliminary sphere sampling] – see page 102
    - [Strong-maximum test] – see page 104
- Sampling methods
  - [Standard sampling method] – see page 109
  - Accelerated simulation
    - [Importance sampling] – see page 112
    - [Directional simulation] – see page 114
    - [Latin Hypercube Sampling] – see page 118

4.2.3 Quantile of a variable of interest

Only one sampling approach is available in the current version of Open TURNS.

- [Standard sampling method and Wilk's formula] – see page 121

4.3 Methods description

4.3.1 Step C – Min-Max Approach using Design of Experiments

Mathematical description

Goal

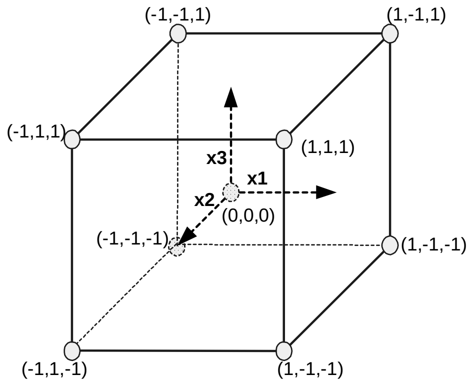
The method is used in the following context: $\underline{x} = \left(x^1, \ldots, x^{n_X}\right)$ is a vector of unknown variables, $\underline{d}$ a vector considered to be well known or where uncertainty is negligible, and $\underline{y} = h(\underline{x}, \underline{d}) = \left(y^1, \ldots, y^{n_Y}\right)$ describes the variables of interest. The objective here is to determine the extreme (minimum and maximum) values of the components of $\underline{y}$ for all possible values of $\underline{x}$.

Principle

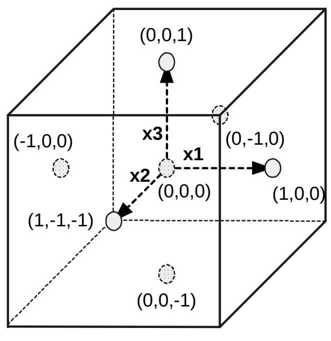
Determining the extreme (minimum and maximum) values of the variables $\underline{y}$ for the set of all possible values of $\underline{x}$ can prove to be a complex optimisation problem when this set of values are continuous. This complex problem is simplified here, the extreme values of $\underline{y}$ are sought for only a finite set of combinations $\{\underline{x}_1, \ldots, \underline{x}_N\}$ chosen using a design of experiments. This technique aims to explore in the most appropriate manner, the set of possible values of $\underline{x}$ for a fixed value of N.
The method is made up of three steps:

- choice of experiment design used to determine the combinations $\{\underline{x}_1, \ldots, \underline{x}_N\}$ of unknown variables (crossed, factorial or combined design of experiments),
- calculation of $\underline{y}_i = h(\underline{x}_i, \underline{d})$ for $i = 1, \ldots, N$,
- calculation of $\min_{1 \leq i \leq N} y_i^k$ and of $\max_{1 \leq i \leq N} y_i^k$, together with the combinations related to these extreme values: $\underline{x}_{k,\min} = \text{argmin}_{1 \leq i \leq N} y_i^k$ and $\underline{x}_{k,\max} = \text{argmax}_{1 \leq i \leq N} y_i^k$.

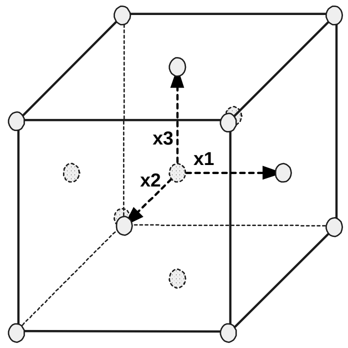
To construct a design of experiment in Open TURNS, the user provides a "central" point $\underline{x}_0$ for $\underline{x}$, as well as a set of levels $k_1, \ldots, k_m$. The diagrams below illustrate this principle showing different types of 3-dimensional designs ($n_X = 3$, $\underline{x} = \left(x^1, x^2, x^3\right)$). The central value is taken to be equal to 0, and only one single level $k_1 = 1$ is used.



factorial design



axial design



composite design

The factorial design contains a central point $\underline{x}_0$ as well as the points $\left\{\left(x_0^1 \pm k_j, \ldots, x_0^{n_X} \pm k_j\right)\right\}_{1 \leq j \leq m}$. In other words, the design includes $1 + m \times 2^{n_X}$ different points.

The axial design contains the central point $\underline{x}_0$ as well as the points $\left\{\left(x_0^1 \pm k_j, x_0^2, \ldots, x_0^{n_X}\right),\ \left(x_0^1, x_0^2 \pm k_j, x_0^3, \ldots, x_0^{n_X}\right),\ \left(x_0^1, x_0^2, \ldots, x_0^{n_X-1}, x_0^{n_X} \pm k_j\right)\right\}_{1 \leq j \leq m}$. In other words, the design includes $1 + m \times 2n_X$ different points.

The composite design defined in Open TURNS includes the set of points defined in a factorial design and in a crossed design.

Other notations

One can also refer to the terms “deterministic" study of uncertainty or the study of uncertainty “by intervals", when fixing a lower and upper value for each of the input components x^i, and by seeking the minimum and maximum values in a complete (factorial) design with the combinations of $\underline{x}$ also generated.

Link with OpenTURNS methodology

This method is used in step C “Propagation of uncertainty" to evaluate a deterministic minimum-maximum type of criterion for the output value defined in step A “Specifying the Criteria and the Case Study".
Input Data:

- $\underline{x}$: vector of unknown variables defined in step A,
- $\underline{d}$: vector of deterministic calculation parameters,
- $\underline{y} = h(\underline{x}, \underline{d})$: output variables / variables of interest specified in step A,

Method Parameters:

- type of design of experiment to be used (factorial, axial, composite),
- calculation parameters for the design of the experiment,

Output Data:

- $\{\underline{x}_1, \ldots, \underline{x}_N\}$: combinations of unknown variable determined by the design of the experiment,
- $\min_{1 \leq i \leq N} y_i^k$ and $\max_{1 \leq i \leq N} y_i^k$: extremes for the variable of interest,
- $\text{argmin}_{1 \leq i \leq N} y_i^k$ and $\underline{x}_{\max} = \text{argmax}_{1 \leq i \leq N} y_i^k$: combinations of uncertain variables associated with these extremes.

References and theoretical basics

This approach using the design of experiments does not require the function h to have any special property. The extremes thus determined, however, only give an approximate idea of the variation range for the variable of interest; in this simplified approach which does not make use of a real optimisation algorithm, it does in no way guarantee in general that one has approached or contained the function extremes, except in particular cases e.g. monotonic model h.

4.3.2 Step C – Quadratic Combination / Perturbation Method

Mathematical description

Goal

The quadratic combination approach is a probabilistic approach designed to propagate the uncertainties of the input variables $\underline{X}$ through the model h towards the output variables $\underline{Y}$. It enables to access the central dispersion (Expectation, Variance) of the output variables.

Principles

This method is based on a Taylor decomposition of the output variable $\underline{Y}$ towards the $\underline{X}$ random vectors around the mean point $\underline{\mu}_X$. Depending on the order of the Taylor decomposition (classically first order or second order), one can obtain different formulas. For easiness of the reading, we first present the formulas with $n_Y = 1$ before the ones obtained for $n_Y > 1$.

Case $n_Y = 1$

As $Y = h(\underline{X})$, the Taylor decomposition around $\underline{x} = \underline{\mu}_X$ at the second order yields to:

$$Y = h(\underline{\mu}_X) + < \underline{\nabla} h(\underline{\mu}_X), \underline{X} - \underline{\mu}_X > + \frac{1}{2} << \underline{\underline{\nabla}}^2 h(\underline{\mu}_X, \underline{\mu}_X), \underline{X} - \underline{\mu}_X >, \underline{X} - \underline{\mu}_X > + o(\mathrm{Cov}\left[\underline{X}\right])$$

where:

- $\underline{\mu}_X = \mathbb{E}\left[\underline{X}\right]$ is the vector of the input variables at the mean values of each component.
- $\mathrm{Cov}\left[\underline{X}\right]$ is the covariance matrix of the random vector $\underline{X}$. The elements are the followings : $(\mathrm{Cov}\left[\underline{X}\right])_{ij} = \mathbb{E}\left[\left(X^i - \mathbb{E}\left[X^i\right]\right) \times \left(X^j - \mathbb{E}\left[X^j\right]\right)\right]$
- $\underline{\nabla} h(\underline{\mu}_X) = {}^t\left(\frac{\partial y}{\partial x^j}\right)_{\underline{x} = \underline{\mu}_X} = {}^t\left(\frac{\partial h(\underline{x})}{\partial x^j}\right)_{\underline{x} = \underline{\mu}_X}$ is the gradient vector taken at the value $\underline{x} = \underline{\mu}_X$ and $j = 1, \ldots, n_X$.
- $\underline{\underline{\nabla}}^2 h(\underline{x}, \underline{x})$ is a matrix. It is composed by the second order derivative of the output variable towards the i^{th} and j^{th} components of $\underline{x}$ taken around $\underline{x} = \underline{\mu}_X$. It yields to: $\left(\nabla^2 h(\underline{\mu}_X, \underline{\mu}_X)\right)_{ij} = \left(\frac{\partial^2 h(\underline{x}, \underline{x})}{\partial x^i \partial x^j}\right)_{\underline{x} = \underline{\mu}_X}$
- $<, >$ is a scalar product between two vectors.

Approximation at the order 1 - *Case* $n_Y = 1$

Expectation:

$$\mathbb{E}\left[Y\right] = h(\underline{\mu}_X)$$

Variance:

$$\mathrm{Var}\left[Y\right] = \sum_{i,j=1}^{n_X} \frac{\partial h(\underline{\mu}_X)}{\partial X^i} . \frac{\partial h(\underline{\mu}_X)}{\partial X^j} . (\mathrm{Cov}\left[\underline{X}\right])_{ij}$$

Approximation at the order 2 - *Case* $n_Y = 1$

Expectation:

$$\mathbb{E}\left[Y\right] = h(\underline{\mu}_X) + \frac{1}{2} . \sum_{i,j=1}^{n_X} \frac{\partial^2 h(\underline{\mu}_X, \underline{\mu}_X)}{\partial x^i \partial x^j} . (\mathrm{Cov}\left[\underline{X}\right])_{ij}$$

Variance:
The decomposition of the variance at the order 2 is not implemented in the standard version of Open TURNS. It requires both the knowledge of higher order derivatives of the model and the knowledge of moments of order strictly greater than 2 of the pdf.

Case $n_Y > 1$

The quadratic combination approach can be developped at different orders from the Taylor decomposition of the random vector $\underline{Y}$. As $\underline{Y} = h(\underline{X})$, the Taylor decomposition around $\underline{x} = \underline{\mu}_X$ at the second order yields to:

$$\underline{Y} = h(\underline{\mu}_X) + < \underline{\underline{\nabla}} h(\underline{\mu}_X), \, \underline{X} - \underline{\mu}_X > + \frac{1}{2} << \underline{\underline{\underline{\nabla}}}^2 h(\underline{\mu}_X, \, \underline{\mu}_X), \, \underline{X} - \underline{\mu}_X >, \, \underline{X} - \underline{\mu}_X > + o(\mathrm{Cov}\left[\underline{X}\right])$$

where:

- $\underline{\mu}_X = \mathbb{E}\left[\underline{X}\right]$ is the vector of the input variables at the mean values of each component.
- $\mathrm{Cov}\left[\underline{X}\right]$ is the covariance matrix of the random vector $\underline{X}$. The elements are the followings : $(\mathrm{Cov}\left[\underline{X}\right])_{ij} = \mathbb{E}\left[\left(X^i - \mathbb{E}\left[X^i\right]\right)^2\right]$
- $\underline{\underline{\nabla}} h(\underline{\mu}_X) = {}^t\left(\frac{\partial y^i}{\partial x^j}\right)_{\underline{x} = \underline{\mu}_X} = {}^t\left(\frac{\partial h^i(\underline{x})}{\partial x^j}\right)_{\underline{x} = \underline{\mu}_X}$ is the transposed Jacobian matrix with $i = 1, \ldots, n_Y$ and $j = 1, \ldots, n_X$.
- $\underline{\underline{\underline{\nabla}}}^2 h(\underline{x}\,, \underline{x})$ is a tensor of order 3. It is composed by the second order derivative towards the i^{th} and j^{th} components of $\underline{x}$ of the k^{th} component of the output vector $h(\underline{x})$. It yields to: $\left(\nabla^2 h(\underline{x})\right)_{ijk} = \frac{\partial^2 (h^k(\underline{x}))}{\partial x^i \partial x^j}$
- $< \underline{\underline{\nabla}} h(\underline{\mu}_X), \, \underline{X} - \underline{\mu}_X >= \sum_{j=1}^{n_X} \left(\frac{\partial \underline{y}}{\partial x^j}\right)_{\underline{x} = \underline{\mu}_X} . \left(X^j - \underline{\mu}^j_X\right)$
- $<< \underline{\underline{\underline{\nabla}}}^2 h(\underline{\mu}_X, \, \underline{\mu}_X), \, \underline{X} - \underline{\mu}_X >, \, \underline{X} - \underline{\mu}_X >= \left({}^t(\underline{X}^i - \underline{\mu}^i_X) . \left(\frac{\partial^2 y^k}{\partial x^i \partial x^k}\right)_{\underline{x} = \underline{\mu}_X} . (\underline{X}^j - \underline{\mu}^j_X)\right)_{ijk}$

Approximation at the order 1 - *Case* $n_Y > 1$

Expectation:

$$\mathbb{E}\left[\underline{Y}\right] \approx \underline{h}(\underline{\mu}_X)$$

Pay attention that $\mathbb{E}\left[\underline{Y}\right]$ is a vector. The k^{th} component of this vector is equal to the k^{th} component of the output vector computed by the model h at the mean value. $\mathbb{E}\left[\underline{Y}\right]$ is thus the computation of the model at mean.

Variance:

$$\mathrm{Cov}\left[\underline{Y}\right] \approx {}^t\, \underline{\underline{\nabla}}\,\underline{h}(\underline{\mu}_X).\mathrm{Cov}\left[\underline{X}\right].\underline{\underline{\nabla}}\,\underline{h}(\underline{\mu}_X)$$

Approximation at the order 2 - *Case* $n_Y > 1$

Expectation:

$$\mathbb{E}\left[\underline{Y}\right] \approx \underline{h}(\underline{\mu}_X) + \frac{1}{2}.\underline{\underline{\underline{\nabla}}}^2\underline{h}(\underline{\mu}_X, \underline{\mu}_X) \odot \mathrm{Cov}\left[\underline{X}\right]$$

This last formulation is the reduced writing of the following expression:

$$(\mathbb{E}\left[\underline{Y}\right])_k \approx (\underline{h}(\underline{\mu}_X))_k + \left(\sum_{i=1}^{n_X} \frac{1}{2}(\mathrm{Cov}\left[\underline{X}\right])_{ii}.(\nabla^2\, h(\underline{X}))_{iik} + \sum_{i=1}^{n_X}\sum_{j=1}^{i-1}(\mathrm{Cov}\left[X\right])_{ij}.(\nabla^2\, h(\underline{X}))_{ijk} \right)_k$$

Variance:

The decomposition of the variance at the order 2 is not implemented in the standard version of Open TURNS. It requires both the knowledge of higher order derivatives of the model and the knowledge of moments of order strictly greater than 2 of the pdf.

Other notations

Perturbation methods

Link with OpenTURNS methodology

This method is part of the step C 'Propagation of Uncertainties' of the global methodology. It requires the definition of the input random vector $\underline{X}$, the definition of the model of interest h ((both should have been done in step specification of model and criteria).

References and theoretical basics

This method is well fitted when one wants to obtain the parameters of the central dispersion. Be careful, if the model is largely non linear or not monotonous, the Taylor approximation at the order 2 may not be accurate on the domain of the input variables and thus the assessment of the first and second order moments may be largely false. Besides, one has to pay attention that this method is generally not justified to compute low probabilities. Pay attention that the mean and variance obtained by quadratic decomposition should not be used tu deduce low probabilities. For instance, the 95 % quantile of Y^i is generally not equal to the $\mu_Y^i + 1,64.\sigma^i$ - except if one may prove that Y^i follows a gaussian distribution - and the error is potentially huge.

4.3.3 Step C – Estimating the mean and variance using the Monte Carlo Method

Mathematical description

Goal

Let us denote $\underline{Y} = h\left(\underline{X}, \underline{d}\right) = \left(Y^1, \ldots, Y^{n_Y}\right)$, where $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ is a random vector, and $\underline{d}$ a deterministic vector. We seek here to evaluate, the characteristics of the central part (central tendency and spread i.e. mean and variance) of the probability distribution of a variable Y^i, using the probability distribution of the random vector $\underline{X}$.

Principle

The Monte Carlo method is a numerical integration method using sampling, which can be used, for example, to determine the mean and standard deviation of a random variable Y^i (if these quantities exist, which is not the case for all probability distributions):

$$m_{Y^i} = \int u\, f_{Y^i}(u)\, du,\ \sigma_{Y^i} = \sqrt{\int \left(u - m_{Y^i}\right)^2\, f_{Y^i}(u)\, du}$$

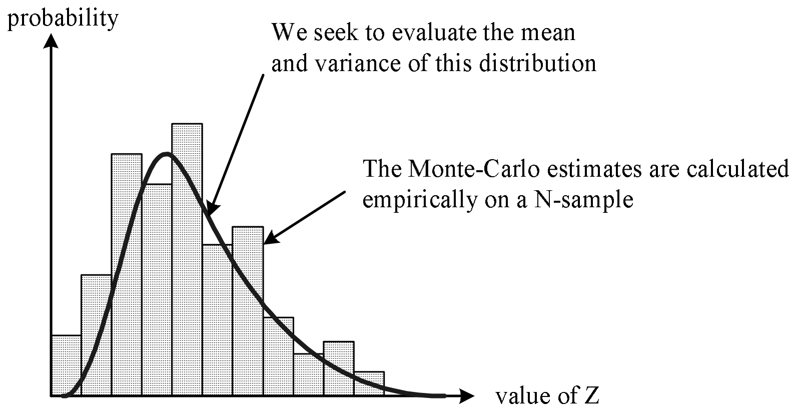
where f_{Y^i} represents the probability density function of Y^i.
Suppose now that we have the sample $\left\{y_1^i, \ldots, y_N^i\right\}$ of N values randomly and independently sampled from the probability distribution f_{Y^i}; this sample can be obtained by drawing a N sample $\{\underline{x}_1, \ldots, \underline{x}_N\}$ of the random vector $\underline{X}$ (the distribution of which is known) and by computing $\underline{y}_j = h\left(\underline{x}_j, \underline{d}\right)\ \forall 1 \leq j \leq N$. Then, the Monte-Carlo estimations for the mean and standard deviation are the empirical mean and standard deviations of the sample:

$$\widehat{m}_{Y^i} = \frac{1}{N}\sum_{j=1}^{N} y_j^i,\ \widehat{\sigma}_{Y^i} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}\left(y_j^i - \widehat{m}_{Y^i}\right)^2}$$

These are just estimations, but by the law of large numbers their convergence to the real values m_{Y^i} and σ_{Y^i} is assured as the sample size N tends to infinity. The Central Limit Theorem enables the difference between the estimated value and the sought value to be controlled by means of a confidence interval (especially if N is sufficiently large, typically $N >$ a few dozens even if there is now way to say for sure if the asymptotic behaviour is reached). For a probability α strictly between 0 and 1 chosen by the user, one can, for example, be sure with a confidence α, that the true value of m_{Y^i} is between $\widehat{m}_{i,\inf}$ and $\widehat{m}_{i,\sup}$ calculated analytically from simple formulae. To illustrate, for $\alpha = 0.95$:

$$\widehat{m}_{i,\inf} = \widehat{m}_{Y^i} - 1.96\frac{\widehat{\sigma}_{Y^i}}{\sqrt{N}},\ \widehat{m}_{i,\sup} = \widehat{m}_{Y^i} + 1.96\frac{\widehat{\sigma}_{Y^i}}{\sqrt{N}},\ \text{that is to say } \Pr\left(\widehat{m}_{i,\inf} \leq m_{Y^i} \leq \widehat{m}_{i,\sup}\right) = 0.95$$

The size of the confidence interval, which represents the uncertainty of this mean estimation, decreases as N increases but more gradually (the rate is proportional to $\sqrt{N}$: multiplying N by 100 reduces the length of the confidence interval $|\widehat{m}_{i,\inf} - \widehat{m}_{i,\sup}|$ by a factor 10).



Other notations

Direct sampling, crude Monte Carlo method, Classical Monte Carlo integration

Link with OpenTURNS methodology

In the overall process, the Monte Carlo simulation method for estimating the variance appears in step C “Propagation of Uncertainty" when the study of uncertainty is concerned with the dispersion of the variable of interest Y^i defined in step A “Specifying Criteria and the Case Study". To be more precise, this method requires that the following steps have previously been previously completed:

- step A: specification of input variables $\underline{X}$ and $\underline{d}$ and the output variable of interest $\underline{Y} = h(\underline{X}, \underline{d})$,
- step B: use of one of the proposed techniques for determining the probability distribution of the variable $\underline{X}$,

The method's parameters are the following:

- number N of simulations,
- probability α giving the required confidence level for the confidence intervals,

The method described here returns the following results:

- the Monte-Carlo estimates $\widehat{m}_{Y^i}$ and $\widehat{\sigma}_{Y^i}$ for the mean and standard deviations of the variable of interest Y^i,
- the confidence interval $[\widehat{m}_{i,\text{inf}}, \widehat{m}_{i,\text{sup}}]$ for the mean m_{Y^i}.

References and theoretical basics

The Monte-Carlo method does not require any assumptions on the form of the function h which relates $\underline{X}$

and $\underline{Y}$, except that the expected value and standard deviation of Y^i should exist (which is not the case, for example, if Y^i follows the Cauchy distribution).
Actually, the only limitation resides in N, the number of simulations, which if not sufficiently high (because of the CPU time required for an estimation of $\underline{Y} = h(\underline{X}, \underline{d})$), can result in too greater uncertainty for the estimations of $\widehat{m}_{Y^i}$ and $\widehat{\sigma}_{Y^i}$. It is fitting then to verify the convergence of the estimators, especially by plotting the graph of the coefficient de variation $\widehat{\sigma}_{Y^i}/\widehat{m}_{Y^i}$ as a function of N: if convergence is not visible, it is necessary to increase N or if needed to choose another propagation method to estimate the central uncertainty of $\underline{Y}$ (see [Quadratic combination / Perturbation method]).
The following references provide a bibliographic starting point for interested readers for further study of the method described here:

- Robert C.P., Casella G. (2004). “Monte Carlo Statistical Methods", Springer, ISBN 0-387-21239-6, 2nd ed.
- Rubinstein R.Y. (1981). “Simulation and The Monte Carlo methods", John Wiley & Sons
- “Guide to the expression of Uncertainty in Measurements (GUM)", ISO publication

4.3.4 Step C – Iso-probabilistic transformation preliminary to FORM-SORM methods

Mathematical description

Goal

The isoprobabilistic transformation is used under the following context: $\underline{X}$ is a probabilistic input vector, $f_{\underline{X}}(\underline{x})$ its joint probability density function, F_i the marginals if its components, $R_X = [r_{ij}]$ its linear correlation matrix whose generic term is $r_{ij} = \mathbb{E}\left[\left(\frac{X^i - m_i}{\sigma_i}\right)\left(\frac{X^j - m_j}{\sigma_j}\right)\right]$, with $m_i = \mathbb{E}\left[X^i\right]$ et $\sigma_i = \sqrt{\mathrm{Var}\left[X^i\right]}$.

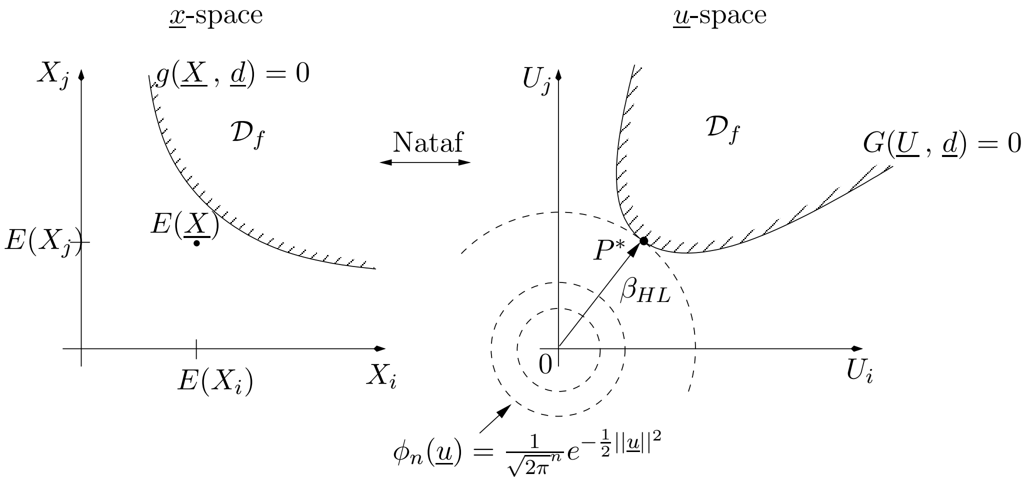
Let us denote by $\underline{d}$ a determinist vector, $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \leq 0\}$ the event considered here and $\mathrm{g}(\underline{\mathrm{X}}\,,\underline{\mathrm{d}}) = 0$ its boundary.

One way to evaluate the probability content of the event $\mathcal{D}_f$:

$$P_f = \int_{\mathcal{D}_f} f_{\underline{X}}(\underline{x})\, d\underline{x}, \qquad (1)$$

is to introduce the Nataf isoprobabilistic transformation wich maps the probabilistic model in terms of $\underline{X}$ onto an equivalent model in terms of n independant standard normal random $\underline{U}$.

Principle


$\underline{x}$-space
$\underline{u}$-space
X_j
$g(\underline{X}\,,\,\underline{d}) = 0$
$\mathcal{D}_f$
Nataf
U_j
$\mathcal{D}_f$
$G(\underline{U}\,,\,\underline{d}) = 0$
$E(X_j)$
$E(\underline{X})$
P^*
β_{HL}
0
$E(X_i)$
X_i
U_i
$\phi_n(\underline{u}) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}||\underline{u}||^2}$

Nataf isoprobabilistic transformation

The Nataf isoprobabilistic transformation wich maps the probabilistic model in terms of $\underline{X}$ onto an equivalent model in terms of n independent standard normal random variables $\underline{U}$ in the following two steps :

- Step 1 : T_1 : the input random vector $\underline{X}$ is mapped onto a random vector $\underline{Y}$, that is supposed to have standard normal components.
- Step 2 : T_2 : $\underline{Y}$ is mapped onto the random vector $\underline{U}$ whose components are standard normal and independant.

The first step is $T_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$\underline{Y} = T_1(\underline{X}) = \left(\begin{array}{c} \Phi^{-1}(F_1(X^1)) \\ \Phi^{-1}(F_2(X^2)) \\ \vdots \\ \Phi^{-1}(F_n(X^n)) \end{array} \right). \tag{2}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{u^2}{2})\, du.$

Let us denote by $R_Y = (\rho_{ij}]$ the linear correlation matrix of the gaussian random vector $\underline{Y}$. Due to the above assumptions, one gets:

$$\begin{aligned} r_{ij} &= \mathbb{E}\left[\left(\frac{X^i - m_i}{\sigma_i}\right)\left(\frac{X^j - m_j}{\sigma_j}\right)\right] \\ &= \mathbb{E}\left[\left(\frac{F_{X^i}^{-1}(\Phi(Y^i)) - m_i}{\sigma_i}\right)\left(\frac{F_{X^j}^{-1}(\Phi(Y^j)) - m_j}{\sigma_j}\right)\right] \\ &= \int\!\!\int \left(\frac{F_{X^i}^{-1}(\Phi(y_i)) - m_i}{\sigma_i}\right)\left(\frac{F_{X^j}^{-1}(\Phi(y_j)) - m_j}{\sigma_j}\right) \phi_2(y_i, y_j, \rho_{ij}) dy_i dy_j \end{aligned} \tag{3}$$

The probability density function of $\underline{Y}$ is the multinormal distribution (4):

$$\phi_n(\underline{y}, R_Y) = \frac{1}{\sqrt{(2\pi)^n det(R_Y)}} \exp(-\frac{1}{2}\underline{y}^t R_Y^{-1} \underline{y}) \tag{4}$$

The second step is $T_2 : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$\underline{Y} = T_2(\underline{Y}) = \Gamma_0 \underline{Y} \tag{5}$$

where $\Gamma_0 = \mathcal{B}^{-1}$ whith $\mathcal{B}$ is the lower triangular Cholesky factor of R_Y : $R_Y = \mathcal{B}.\mathcal{B}^t$.

The isoprobabilistic transform is used in the First and Second Order reliability Method to evaluate the probability content of the event $\mathcal{D}_f$ (refer to [FORM] and [SORM]).

Other notations

–

Link with OpenTURNS methodology

Within the global methodology, the isoprobabilistic transformation is used in the First and Second Order reliability Method to evaluate the probability content of the event $\mathcal{D}_f$.

References and theoretical basics

The following difficulties have been mentioned in the literature:

- the determination of such a gaussian random vector from (3) is not always possible, which happens in particularly when the coefficients r_{ij} are too close to 1 or -1,
- even in the case where the determination of the coefficients ρ_{ij} is possible, the matrix R_Y obtained this way might not be a correlation matrix (i.e. it might not be positive definite),
- the numerical resolution of equation (3) is computationally demanding.

This last point has generated some analytical approximations as :

$$\rho_{ij} = f(\text{density parameters }, r_{ij}) \times r_{ij}, \tag{6}$$

where f is a function of the marginal distributions of X^i and X^j, and r_{ij} (DerKiureghian).

Let's note some usefull references:

- O. Ditlevsen and H.O. Madsen, 2004, "Structural reliability methods," Department of mechanical engineering technical university of Denmark - Maritime engineering, internet publication.
- J. Goyet, 1998,"Sécurité probabiliste des structures - Fiabilité d'un élément de structure," Collège de Polytechnique.
- A. Der Kiureghian, P.L. Liu, 1986,"Structural Reliability Under Incomplete Probabilistic Information", Journal of Engineering Mechanics, vol 112, n°1, pp85-104.
- H.O. Madsen, Krenk, S., Lind, N. C., 1986, "Methods of Structural Safety," Prentice Hall.

Examples

Let's apply this method to the following analytical example which considers a cantilever beam, of Young's modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E, F, L, I).
The input random vector is $\underline{X} = (E, F, L, I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{ \begin{array}{lcl} E & = & Normal(50, 1) \\ F & = & Normal(1, 1) \\ L & = & Normal(10, 1) \\ I & = & Normal(5, 1) \end{array} \right.$$

The four random variables are independant.

The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E,F,L,I) \in \mathbb{R}^4 \,/\, y(E,F,L,I) \geq 3\}$.

In that case, the isoprobabilistic transformation maps the random vector (E,F,L,I) into the random vector $\underline{U}$ such as :

$$\left\{\begin{array}{lcl} U_1 & = & \frac{E-50}{1} \\ U_2 & = & \frac{F-1}{1} \\ U_3 & = & \frac{L-10}{1} \\ U_4 & = & \frac{I-5}{1} \end{array}\right.$$

The limit state function is :

$$\left\{\begin{array}{ll} \text{in the } \underline{x}\text{-space :} & g(E,F,L,I) = -y(E,F,L,I) + 3 \\ \text{in the } \underline{u}\text{-space :} & G(\underline{U}) = g(50+U_1, 1+U_2, 10+U_3, 5+U_4). \end{array}\right.$$

4.3.5 Step C – FORM

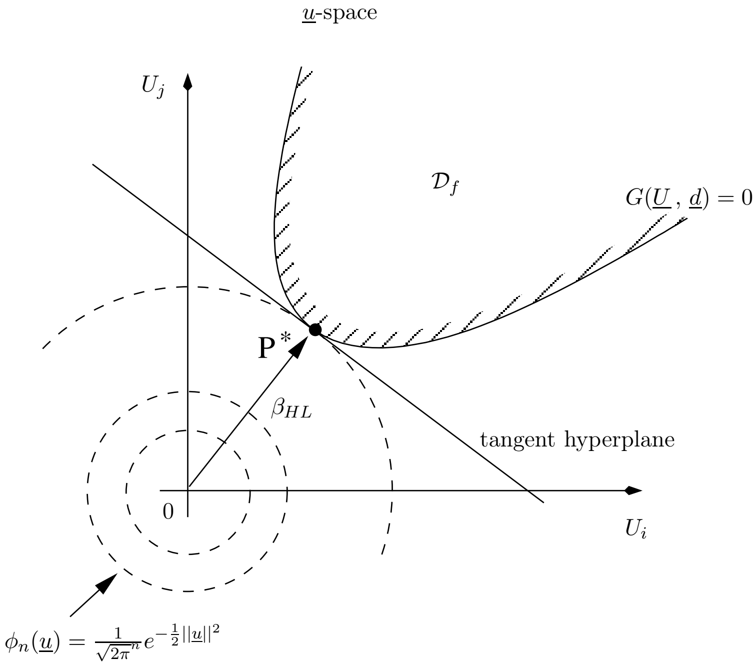
Mathematical description

Goal

The First Order Reliability Method is used in the following context: $\underline{X}$ denotes a random input vector, representing the sources of uncertainties, $f_{\underline{X}}(\underline{x})$ its joint density probability, $\underline{d}$ a deterministic vector, representing the fixed variables $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \leq 0\}$ the event considered here and $g(\underline{X}\,,\,\underline{d}) = 0$ its boundary (also called limit state surface).
The objective of FORM is to evaluate the probability content of the event $\mathcal{D}_f$:

$$P_f = \int_{g(\underline{X}\,,\,\underline{d}) \leq 0} f_{\underline{X}}(\underline{x})\, d\underline{x}. \tag{7}$$

Principle


$\underline{u}$-space
U_j
$\mathcal{D}_f$
$G(\underline{U}\,,\,\underline{d}) = 0$
$\mathbf{P}^*$
β_{HL}
tangent hyperplane
0
U_i
$\phi_n(\underline{u}) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}||\underline{u}||^2}$

FORM approximation

The principle is:

1. Map the probabilistic model in terms of $\underline{X}$ onto an equivalent model in terms of n independent standard normal random variables gathered in the vetor $\underline{U}$. Refer to [Iso Probabilistic Transformation] to obtain details on the mapping function denoted by T: $\underline{U} = T(\underline{X})$. The mapping of the limit state function is $G(\underline{U}\,,\,\underline{d}) = g(T^{-1}(\underline{U}\,,\,\underline{d}))$. Then, the event considered becomes : $\mathcal{D}_f = \{\underline{U} \in \mathbb{R}^n \,/\, G(\underline{U}\,,\,\underline{d}) \leq 0\}$ and eq.(7) becomes:

$$P_f = \int_{G(\underline{U}\,,\,\underline{d}) \leq 0} \phi_n(\underline{u})\, d\underline{u}). \tag{8}$$

In the $\underline{u}$-space, the joint probability density function is the standard multi-normal density, whose most interesting characteristics are its rotational symmetry and its rapid decay with increasing distance form the origin.

2. Approximate the limit state surface in the $\underline{u}$-space by a linear surface at the design point $\underline{P}^*$, where $\underline{P}^*$ is the point located on the limit state surface of maximum likelihood: the characteristics of the $\underline{u}$-space mas such that $\underline{P}^*$ is the point on the limit state surface closest to the origin. $\underline{P}^*$ is the result of a constrained optimisation problem.

3. In the $\underline{u}$-space, the probability content eq. (7) where the limit state surface has been approximated by a linear surface (hyperplane) by can be obtained exactly:

$$P_{f,FORM} = \left| \begin{array}{ll} \Phi(-\beta_{HL}) & \text{if the origin of the } \underline{u}\text{-lies in the domain } \mathcal{D}_f \\ \Phi(+\beta_{HL}) & \text{otherwise} \end{array} \right. \tag{9}$$

where β_{HL} is the Hasofer-Lind reliability index, which means the distance of the design point $\underline{P}^*$ to the origin of the $\underline{u}$-space, and Φ is the gaussian cumulative density probability.

Other notations

Here, the event considered is explicited directly from the limit state function $g(\underline{X}\,,\,\underline{d})$: this is the classical structural reliability formulation.
However, if the event is a threshold exceedance, it is useful to explicite the variable of interest $Z = \tilde{g}(\underline{X}\,,\,\underline{d})$, evaluated from the model $\tilde{g}(.)$. In that case, the event considered, associated to the threshold z_s has the formulation : $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, Z = \tilde{g}(\underline{X}\,,\,\underline{d}) > z_s\}$ and the limit state function is : $g(\underline{X}\,,\,\underline{d}) = z_s - Z = z_s - \tilde{g}(\underline{X}\,,\,\underline{d})$. P_f is the threshold exceedance probability, defined as : $P_f = P(Z \geq z_s) = \int_{g(\underline{X}\,,\,\underline{d}) \leq 0} f_{\underline{X}}(\underline{x})\, d\underline{x}$.

Link with OpenTURNS methodology

Within the global methodology, the First Order Reliability Method is used in the step C: "Uncertainty propagation" in the case of the evaluation of the probability of an event by an approximation method.
It requires to have fulfilled the following steps beforehand:

- step A: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest $Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$; identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C: select an appropriate optimisation algorithm among those proposed.

The First Order Reliability Method provides the following results:

- the FORM probability calculated by eq.9,

- the importance factors associated to the event (refer to [Importance Factors]),
- if asked by the user, the sensitivity factors associated to the event (refer to [Sensitivity Factors]).

References and theoretical basics

One is usually interested in the evaluation of a very small probability $\underline{P}^*$ where the evaluation of the limit state function of the model requires computationally expensive subroutines. The FORM method has been designed specifically for such cases for which simulation techniques (see for instance [standard sampling approach]) are computationally prohibitive.

The quality of the results obtained by the First Order Reliability Method depends on:

- the hypothesis of the mapping T of the $\underline{x}$-space in the $\underline{u}$-space: [IsoProbabiliticFunction] shows cases where the mapping is not feasible. In such cases, it may imply to modify the probabilistic modelisation of the problem if one wants to apply the Form method with the Nataf isoprobabilistic transformation.
- the quality of the optimisation algorithm used to find the design point: it is important that the optimisation converges towards the global minimum of the distance function
- the quality of the computation of the gradients of the limit state function. It is important to choose an optimisation algorithm adapted to the model considered
- the quality of the design point in the $\underline{u}$-space. It has several fields:
  - the shape of the limit state surface: the boundary is supposed to be well approximated by a plane near the design point,
  - the unicity of the design point in the $\underline{u}$-space: FORM is valid when there is only one point on the limit state surface at a distance minimal to the origin,
  - the strongness of the design point: FORM is valid under the hypothesis that most of the contribution to P_f is concentrated in the vicinity of the design point, which is the case both when around $\underline{P}^*$, the contribution decreases rapidly with the distance to $\underline{P}^*$ and when there is no local maximum with comparable density.

  The first hypothesis can be checked by testing other method to evaluate P_f : SORM (refer to [SORM]) that takes into account the curvatures of the surface, or importance sampling techniques (refer to [Importance sampling]) that makes no hypothesis on the shape of the surface.
  The unicity and the strongness of the design point can be checked thanks to the Strong Maximum Test (refer to [Strong Max Test]).
  Accelerated sampling techniques such as directional sampling (refer to [Directional sampling]) are also still valid if the unicity or strongness are doubtful.
  A limitation of FORM (or SORM) approximation is that it is generally impossible to quantify the approximation error. Although the method has been used satisfactorily in many circumstances, it is generally useful, if computationally possible, to validate Form/Som using at least one of the techniques above mentioned.

Let's note some usefull references:

- Breitung, 1984, “Asymptotic Approximation for multinormal Integrals,” Journal of Engineering Mechanics, ASCE, 110(3), 357-366.
- O. Ditlevsen and H.O. Madsen, 2004, “Structural reliability methods,” Department of mechanical engineering technical university of Denmark - Maritime engineering, internet publication.
- H. O. Madsen, Krenk, S., Lind, N. C., 1986, “Methods of Structural Safety,” Prentice Hall.

Examples

Let’s apply this method to the following analytical example which considers a cantilever beam, of Young’s modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E,F,L,I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E,F,L,I).
The input random vector is $\underline{X} = (E,F,L,I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{\begin{array}{lcl} E & = & Normal(50,1) \\ F & = & Normal(1,1) \\ L & = & Normal(10,1) \\ I & = & Normal(5,1) \end{array}\right.$$

The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E,F,L,I) \in \mathbb{R}^4 \,/\, y(E,F,L,I) \geq 3\}$ We obtain the following results :

- design point in the $\underline{x}$-space, $P^* = (E^* = 49.97, F^* = 1.842, l^* = 10.45, I^* = 4.668)$
- the generalized and Hasofer reliability index : $\beta_g = \beta_{HL} = 1.009$
- the FORM probability : $P_{f,FORM} = 1.564e^{-1}$

4.3.6 Step C – SORM

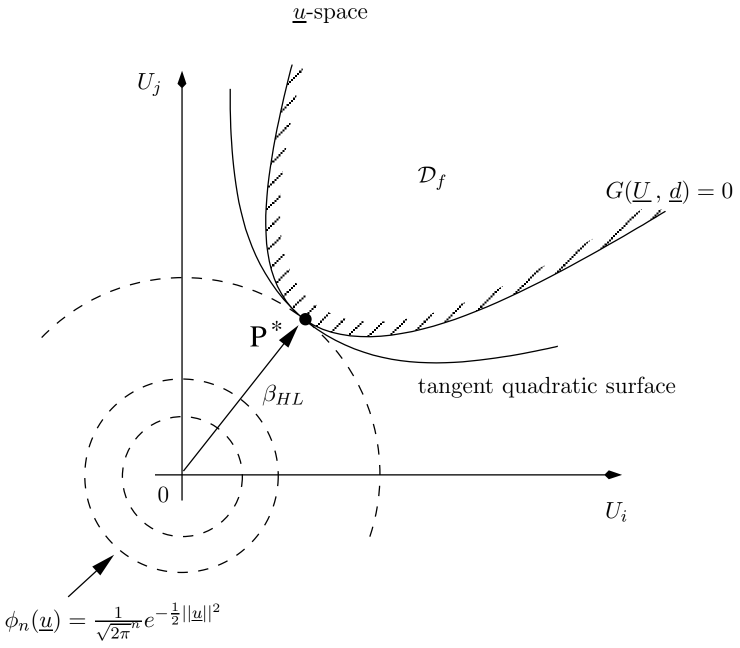
Mathematical description

Goal

The Second Order Reliability Method is used in the same context as the First Order Reliability: refer to [FORM] for further details. The objective of SORM is to evaluate the probability content of the event $\mathcal{D}_f$:

$$P_f = \int_{g(\underline{X}\,,\,\underline{d})\leq 0} f_{\underline{X}}(\underline{x})\,d\underline{x}. \tag{10}$$

Principle


$\underline{u}$-space
U_j
$\mathcal{D}_f$
$G(\underline{U}\,,\,\underline{d}) = 0$
P^*
β_{HL}
tangent quadratic surface
0
U_i
$\phi_n(\underline{u}) = \frac{1}{\sqrt{2\pi}^n} e^{-\frac{1}{2}\|\underline{u}\|^2}$

SORM approximation

The principle is the same as for FORM. In the $\underline{u}$-space, eq. (10) becomes :

$$P_f = \int_{G(\underline{U}\,,\,\underline{d})\leq 0} \phi_n(\underline{u})\,d\underline{u}. \tag{11}$$

The difference with FORM comes from the approximation of the limit state surface at the design point $\underline{P}^*$ in the $\underline{u}$-space : SORM approximates it by a quadratic surface which curvatures are evaluated at the design point.
Let us denote by n the dimension of the random vector $\underline{X}$ and $(\kappa_i)_{1\leq i\leq n-1}$ the $n-1$ main curvatures of the limite state function at the design point in the standard space.
Several approximations are available in the standard version of Open TURNS, detailed here in the case where the origin of the standard space does not belong to the failure domain:

- Breitung's formula is an asymptotic results (Breitug, 1984):

$$P_{Breitung} \simeq (2\pi)^{\frac{n-1}{2}} e^{-\beta^2_{HL}} |J|^{\frac{1}{2}} \tag{12}$$

where

$$\left\{ \begin{array}{lcl} J & = & (\nabla l(\underline{U}))^t.C(\underline{U}).\nabla l(\underline{U}) \\ C(\underline{U}) & = & \text{matrix of cofactors of } H(\underline{U}) \\ H(\underline{U}) & = & (\dfrac{\partial l(\underline{U})}{\partial u_i \partial u_j} - \lambda_i \dfrac{\partial G(\underline{U})}{\partial u_i \partial u_j})_{i,j=1\dots n} \\ l(\underline{U}) & = & \log(\phi_n(\underline{U})) \\ \lambda_i & = & \dfrac{|\nabla l(\underline{U})|}{|\nabla G(\underline{U})|} \end{array} \right. \tag{13}$$

- Hohenbichler's formula is an approximation of equation (12):

$$P_{Hohenbichler} = \Phi(-\beta_{HL}) \prod_{i=1}^{n-1} \left(1 - \frac{\phi(-\beta_{HL})}{\Phi(-\beta_{HL})}\kappa_i\right)^{1/2} \tag{14}$$

  This formula is valid only in case of gaussian copula for the dependance structure of the random vector $\underline{X}$ and if $\forall i, 1 - \frac{\phi(-\beta_{HL})}{\Phi(-\beta_{HL})}\kappa_i > 0$.

- Tvedt's formula (Tvedt, 1988) :

$$\left\{ \begin{array}{lcl} P_{Tvedt} & = & A_1 + A_2 + A_3 \\ A_1 & = & \Phi(-\beta_{HL}) \displaystyle\prod_{i=1}^{i=N-1} (1 + \beta_{HL}\kappa_i)^{-1/2} \\ A_2 & = & [\beta_{HL}\Phi(-\beta_{HL}) - \phi(\beta_{HL})] \left[\displaystyle\prod_{j=1}^{N-1} (1 + \beta_{HL}\kappa_i)^{-1/2} - \prod_{j=1}^{N-1} (1 + (1 + \beta_{HL})\kappa_i)^{-1/2} \right] \\ A_3 & = & (1 + \beta_{HL}) \left[\beta_{HL}\Phi(-\beta_{HL}) - \phi(\beta_{HL})\right] \left[\displaystyle\prod_{j=1}^{N-1} (1 + \beta_{HL}\kappa_i)^{-1/2} \right. \\ & & \left. -\mathcal{R}e \left(\displaystyle\prod_{j=1}^{N-1} (1 + (i + \beta_{HL})\kappa_j)^{-1/2} \right) \right] \end{array} \right. \tag{15}$$

  where $\mathcal{R}e$ is the complex real part and i the complex number such that $i^2 = -1$.
  This formula is valid only in case of gaussian copula for the dependance structure of the random vector $\underline{X}$ and if $\forall i, 1 + \beta\kappa_i > 0$ and $\forall i, 1 + (1 + \beta)\kappa_i > 0$.

Other notations

–

Link with OpenTURNS methodology

Within the global methodology, the Second Order Reliability Method is used in the step C : "Uncertainty propagation" in the case of the evaluation of the probability of an event by an approximation method.
It requires to have fulfilled the following steps beforehand:

- step A1: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest $Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$; identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C: select an appropriate optimisation algorithm among those proposed.

The Second Order Reliability Method provides the following results :

- the SORM probabilities calculated in Eqs. (12),(14), (15)
- the importance factors associated to the event : refer to [Importance Factors] to obtain details,
- if asked by user, the sensitivity factors associated to the event : refer to [Sensitivity Factors] to obtain details.

References and theoretical basics

The motivations for using SORM are similar to the motivations for using FORM. As it takes into account the curvatures of the limi state surface, SORM is usually more accurate than FORM e.g. in case when the event boundary is highly curved.

The quality of the results obtained by the Second Order Reliability Method depends on the same points as the FORM approximation. The shape of the event boundary must be well approximated by a quadratic surface near the design point.

The evaluation of the previous formulas requires that the limit state function be differentiable at the design point.

The Tvedt formula is exact for a quadratic surface and asympototically exact for another types of surfaces. The Hoen-Bichler formula is a vraint as regards to the Breitung one.

Let us note some useful references :

- Breitung K., "Asymptotic approximation for probability integral," Probability Engineering Mechanics, 1989, Vol 4, No. 4.
- Breitung (1984), "Asymptotic Approximation for multinormal Integrals," Journal of Engineering Mechanics, ASCE, 110(3), 357-366.
- Hohenbichler M., Rackwitz R., 1988, "Improvement of second order reliability estimates by importance sampling," Journal of Engineering Mechanics, ASCE,114(12), pp 2195-2199.

- Tvedt L. 1988, “Second order reliability by an exact integral,” proc. of the IFIP Working Conf. Reliability and Optimization of Structural Systems, Thoft-Christensen (Ed), pp377-384.
- Zhao Y. G., Ono T., 1999, “New approximations for SORM : part 1”, Journal of Engineering Mechanics, ASCE,125(1), pp 79-85.
- Zhao Y. G., Ono T., 1999, “New approximations for SORM : part 2”, Journal of Engineering Mechanics, ASCE,125(1), pp 86-93.
- Adhikari S., 2004, “Reliability analysis using parabolic failure surface approximation”, Journal of Engineering Mechanics, ASCE,130(12), pp 1407-1427.

Examples

Let’s apply this method to the following analytical example which considers a cantilever beam, of Young’s modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E, F, L, I).
The input random vector is $\underline{X} = (E, F, L, I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{\begin{array}{lcl} E & = & Normal(50, 1) \\ F & = & Normal(1, 1) \\ L & = & Normal(10, 1) \\ I & = & Normal(5, 1) \end{array}\right.$$

The four random variables are independant.

The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E, F, L, I) \in \mathbb{R}^4 \,/\, y(E, F, L, I) \geq 3\}$ We obtain the following results :

$$\left\{\begin{array}{lcl} P_{Breitung} & = & 2.5491e^{-1}\,\% \\ P_{Hohenbichler} & = & 2.648e^{-1}\,\% \\ P_{Tvedt} & = & 2.601e^{-1} \end{array}\right.$$

These three approximations are coherent between them, which increases confidence in these results.

4.3.7 Step C – Reliability Index

Mathematical description

Goal

The generalised reliability index β is used under the following context : $\underline{X}$ is a probabilistic input vector, $f_{\underline{X}}(\underline{x})$ its joint density probability, $\underline{d}$ a determinist vector, $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \leq 0\}$ the event considered here and g($\underline{\mathrm{X}}$,$\underline{\mathrm{d}}$) = 0 its boundary.
The probability content of the event $\mathcal{D}_f$ is P_f:

$$P_f = \int_{g(\underline{X}\,,\,\underline{d})\leq 0} f_{\underline{X}}(\underline{x})\, d\underline{x}. \tag{16}$$

The generalised reliability index is defined as :

$$\beta_g = \Phi^{-1}(1 - P_f) = -\Phi^{-1}(P_f).$$

As β_g increases, P_f decreases rapidly.

Principle

Open TURNS standard version evaluates :

- β_{FORM} the FORM reliability index, where P_f is obtained with a FORM approximation (refer to [FORM]̃): in this case, the generalised reliability index is equal to the Hasofer-Lindt reliability index β_{HL}, which is the distance of the design point from the origin of the standard space,
- β_{SORM} the SORM reliability index, where P_f is obtained with a SORM approximation : Breitung, Hohen-Bichler or Tvedt (refer to [SORM]),
- β_g the generalised reliability index, where P_f is obtained with another technique : Monte Carlo simulations, importance samplings,... (refer to [Monte Carlo] , [LHS] [Importance samplings] and [Directional Simulation]̃).

Other notations

–

Link with OpenTURNS methodology

Within the global methodology, the reliability index is used in the step C: "Uncertainty propagation" in the case of the evaluation of the probability of an event.
It requires to have fulfilled before the following steps:

- step A1: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest

$Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$,

- step A22: identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\ , \ \underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C3: select a method to evaluate the probability content of the event : the FORM or SORM approximation (step C31) or a simulation method (step C32).

References and theoretical basics

Interesting litterature on the subject is :

- Cornell, “A probability-based structural code,” Journal of the American Concrete Institute, 1969, 66(12), 974-985.
- O. Ditlevsen, 1979, “Generalised Second moment reliability index,” Journal of Structural Mechanics, ASCE, Vol.7, pp. 453-472.
- O. Ditlevsen and H.O. Madsen, 2004, “Structural reliability methods,” Department of mechanical engineering technical university of Denmark - Maritime engineering, internet publication.
- Hasofer and Lind, 1974, “Exact and invariant second moment code format,” Journal of Engineering Mechanics Division, ASCE, Vol. 100, pp. 111-121.

Examples

Let’s apply this method to the following analytical example which considers a cantilever beam, of Young’s modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E, F, L, I).
The input random vector is $\underline{X} = (E, F, L, I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{ \begin{array}{lcl} E & = & Normal(50, 1) \\ F & = & Normal(1, 1) \\ L & = & Normal(10, 1) \\ I & = & Normal(5, 1) \end{array} \right.$$

The four random variables are independant.

The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E, F, L, I) \in \mathbb{R}^4 \,/\, y(E, F, L, I) \geq 3\}$ We obtain the following results :

- design point in the $\underline{x}$-space, $P^* = (E^* = 49.97, F^* = 1.842, l^* = 10.45, I^* = 4.668)$
- generalized and Hasofer-Lind reliability index : $\beta_g = \beta_{HL} = 1.009$
- Breitung generalized reliability index $\beta_{Breitung} = 6.591e^{-1}$
- HohenBichler generalized reliability index $\beta_{HohenBichler} = 6.285e^{-1}$
- Tvedt generalized reliability index $\beta_{Tvedt} = 6.429e^{-1}$

We note here that the three approximations SORM are consistent between them and different from the FORM one. It may signify that the curvatures are not important to take into account in the evaluation of the event probability.

4.3.8 Step C – Sphere sampling method

Mathematical description

Goal

Within the context of the First and Second Order of the Reliability Method (refer to [FORM] and [SORM]), the Strong Maximum Test (refer to [Strong Maximum Test]) helps to check whether the design point computed is :

- the true design point, which means a global maximum point,
- a strong design point, which means that there is no other local maximum verifying the event and associated to a value near the global maximum.

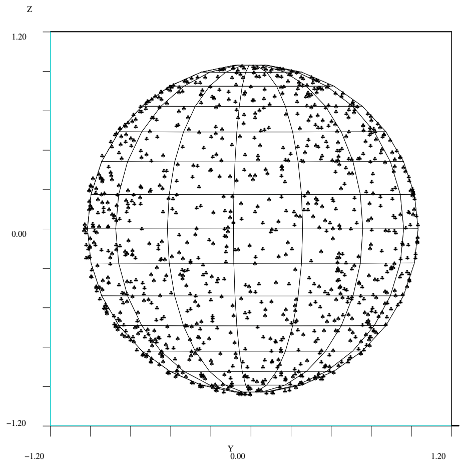
The Strong Maximum Test samples a sphere in the standard space. Open TURNS standard version uses the gaussian random sampling technique described hereafter.

Principle

Open TURNS standard version uses the gaussian random sampling technique:

1. sampling of points in $\mathbb{R}^N$ according to a radial distribution : we generate N independent standard normal samples,
2. projection of the points onto $\mathcal{S}^*$: we map the points different from the origin using the transformation $M \longmapsto m$ such as $\mathbf{Om} = R\dfrac{\mathbf{OM}}{\|\mathbf{OM}\|}$ where R is the radius of the sphere of interest. This transformation does not depend on the angular coordinates. Thus, the generated points follow a uniform distribution on $\mathcal{S}^*$.

A result of such an algorithm is drawn on the following figure 4.3.



Other notations

-

Link with OpenTURNS methodology

Within the global methodology, the sphere sampling method is used in the step C, within the Strong Maximum Test (refer to [Strong Max Test]).
It requires to have fulfilled the following steps beforehand:

- step A: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest $Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$; identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C: select an appropriate optimisation algorithm among those proposed to evaluate the Form or Sorm approximations of P_f; evaluate the quality of the design point resulting from the previous step thanks to the Strong Maximum Test.

References and theoretical basics

Other methods can be used to sample the hypersphere of dimension $N-1$ in $\mathbb{R}^N$: the exclusion method and the parametric method.
The parametric method uses the polar coordinates : the angle parameters are discretized uniformly. It has the inconvenient to generate points principally in the two poles zones.
The exclusion method generates points uniformly within the hypercube containing exactly the sphere. Then we keep only the points located inside the sphere, and we project them on the sphere. This method has the inconvenient to be inefficient for high dimensions : the fraction between the volume of the hypersphere and the volume of the hypercube is less than 0.7% as soon as the dimension is greater than 9. Il means that for a dimension greater than 9, 99.3% of the points generated are rejected.

Let's note some usefull references:

- Luban, Marshall, Staunton, 1988, "An efficient method for generating a uniform distribution of points within a hypersphere," Computer in Physics, 2(6), 55.

4.3.9 Step C – Design point validation : Strong Maximum Test

Mathematical description

Goal

The Strong Maximum Test is used under the following context: $\underline{X}$ denotes a random input vector, representing the sources of uncertainties, $f_{\underline{X}}(\underline{x})$ its joint density probability, $\underline{d}$ a determinist vector, representing the fixed variables $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \leq 0\}$ the event considered here and $g(\underline{X}\,,\,\underline{d}) = 0$ its boundary (also called limit state surface).

The probability content of the event $\mathcal{D}_f$:

$$P_f \;=\; \int_{g(\underline{X}\,,\,\underline{d})\leq 0} f_{\underline{X}}(\underline{x})\, d\underline{x}. \tag{17}$$

may be evaluated with the FORM (refer to [FORM]) or SORM method (refer to [SORM]).

In order to evaluate an approximation of P_f, these analytical methods uses the Nataf isoprobabilistic transformation wich maps the probabilistic model in terms of $\underline{X}$ onto an equivalent model in terms of n independant standard normal random $\underline{U}$ (refer to [Isoprobabilistic Transformation] to have details on the transformation). In that new $\underline{u}$-space, the event has the new expression defined from the transformed limit state function of the model G : $\mathcal{D}_f = \{\underline{U} \in \mathbb{R}^n \,/\, G(\underline{U}\,,\,\underline{d}) \leq 0\}$ and its boundary : $\{\underline{U} \in \mathbb{R}^n \,/\, G(\underline{U}\,,\,\underline{d}) = 0\}$.

These analytical methods rely on the assumption that most of the contribution to P_f comes from points located in the vicinity of a particular point P^*, the design point, defined in the $\underline{u}$-space as the point located on the limit state surface and of maximal likelihood. Given the probabilistic caracteristics of the $\underline{u}$-space, P^* has a geometrical interpretation : it is the point located on the event boundary and at minimal distance from the center of the $\underline{u}$-space. Thus, the design point P^* is the result of a constrained optimisation problem.

The FORM/SORM methods suppose that P^* is unique.

One important difficulty comes from the fact that numerical method involved in the determination of P^* gives no guaranty of a global optimum : the point to which it converges might be a local optimum only. In that case, the contribution of the points in the vicinity of the real design point is not taken into account, and this contribution is the most important one.
Furthermore, even in the case where the global optimum has really been found, there may exist another local optimum $\tilde{P}^*$ which likelihood is slightly inferior to the design point one, which means its distance from the center of the $\underline{u}$-space is slightly superior to the design point one. Thus, points in the vicinity of $\tilde{P}^*$ may contribute significantly to the probability P_f and are not taken into account in the Form and Sorm approximations.
In these both cases, the Form and Sorm approximations are of bad quality because they neglict important contributions to P_f .

The Strong Maximum Test helps to evaluate the quality of the design point resulting from the optimisation algorithm. It checks whether the design point computed is :

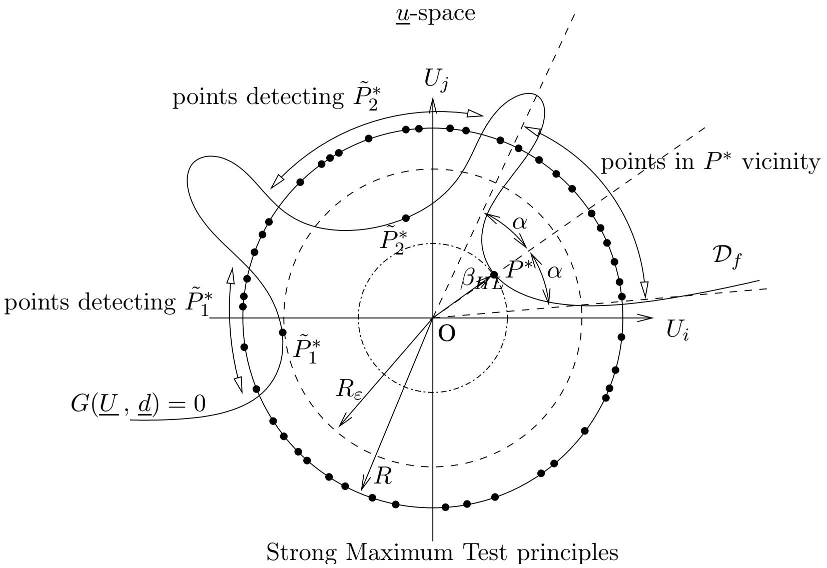
- the *true* design point, which means a global maximum point,

- a *strong* design point, which means that there is no other local maximum located on the event boundary and which likelihood is slightly inferior to the design point one.

This verification is very important in order to give sense to the FORM and SORM approximations .

Principle

The principle of the Strong Maximum Test, which principles are drawn on the figure (4.3) relies on the geometrical definition of the design point.
The objective is to detect all the points $\tilde{P}^*$ in the ball of radius $R_\varepsilon = \beta(1+\delta_\varepsilon)$ which are potentially the real design point (case of $\tilde{P}_2^*$) or which contribution to P_f is not negligeable as regards the approximations Form and Sorm (case of $\tilde{P}_1^*$). The contribution of a point is considered as negligeable when its likelihood in the $\underline{u}$-space is more than ε-times lesser than the design point one. The radius R_ε is the distance to the $\underline{u}$-space center upon which points are considered as negligeable in the evaluation of P_f.
In order to catch the potential points located on the sphere of radius R_ε (frontier of the zone of prospection), it is necessary to go a little further more : that's why the test samples the sphere of radius $R = \beta(1+\tau\delta_\varepsilon)$, with $\tau > 0$.
Points on the sphere sampling (refer to [Sample Sphere] to have details on the generation of the sample) which are in the vicinity of the design point P^* are less interesting than those verifying the event and located *far* from the design point : these last ones might reveal a potential $\tilde{P}^*$ which contribution to P_f has to be taken into account. The vicinity of the design point is defined with the angular parameter α as the cone centered on P^* and of half-angle α.
The number N of the simulations sampling the sphere of radius R is determined to ensure that the test detect with a probability greater than $(1-q)$ any point verifying the event and outside the design point vicinity.



Strong Maximum Test principles

Other notations

–

Link with OpenTURNS methodology

Within the global methodology, the First Order Reliability Method is used in the step C: ”Uncertainty propagation” in the case of the evaluation of the probability of an event by an approximation method.
It requires to have fulfilled the following steps beforehand:

- step A: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest $Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$; identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C: select an appropriate optimisation algorithm among those proposed; select the Strong Maximum Test to validate the design point computed.

The Strong Maximum Test proceeds as follows. The user selects the parameters :

- the importance level ϵ,
- the accuracy level τ,
- the confidence level $(1 - q)$.

The Strong Maximum Test will sample the sphere of radius $\beta(1 + \tau\delta_\epsilon)$, where $\delta_\epsilon = \sqrt{1 - 2\frac{\ln(\epsilon)}{\beta^2}} - 1$.
The test will detect with a probability greater than $(1 - q)$ any point of $\mathcal{D}_f$ which contribution to P_f is not negligeable (i.e. which density value in the $\underline{u}$-space is greater than ϵ times the density value at the design point).

The Strong Maximum Test provides :

- set 1 : all the points detected on the sampled sphere that are in $\mathcal{D}_f$ and outside the design point vicinity, with the corresponding value of the limit state function,
- set 2 : all the points detected on the sampled sphere that are in $\mathcal{D}_f$ and in the design point vicinity, with the corresponding value of the limit state function ,
- set 3 : all the points detected on the sampled sphere that are outside $\mathcal{D}_f$ and outside the design point vicinity, with the corresponding value of the limit state function,
- set 4 : all the points detected on the sampled sphere that are outside $\mathcal{D}_f$ but in the vicinity of the design point, with the corresponding value of the limit state function.

Points are described by their coordinates in the $\underline{x}$-space.

References and theoretical basics

The parameter τ is directly linked to the hypothesis according to which the boundary of the space $\mathcal{D}_f$ is supposed to be well approximated by a plane near the design point, which is primordial for a FORM approximation of the probability content of $\mathcal{D}_f$. Increasing τ is increasing the area where the approximation FORM is applied.

Through the parameter δ, τ also serves as a measure of distance from the design point $\underline{OP}^*$ for a hypothetical local maximum : the greater it is, the further we search for another local maximum.
Numerical experiments show that it is recommanded to take $\tau \leq 4$ (see the given reference below).

The following table helps to quantify the parameters of the test for a problem of dimension 5.

β_g	ε	τ	$1-q$	δ_ε	N
3.0	0.01	2.0	0.9	$4.224e^{-1}$	62
3.0	0.01	2.0	0.99	$4.224e^{-1}$	124
3.0	0.01	4.0	0.9	$4.224e^{-1}$	15
3.0	0.01	4.0	0.99	$4.224e^{-1}$	30
3.0	0.1	2.0	0.9	$2.295e^{-1}$	130
3.0	0.1	2.0	0.99	$2.295e^{-1}$	260
3.0	0.1	4.0	0.9	$2.295e^{-1}$	26
3.0	0.1	4.0	0.99	$2.295e^{-1}$	52
5.0	0.01	2.0	0.9	$1.698e^{-1}$	198
5.0	0.01	2.0	0.99	$1.698e^{-1}$	397
5.0	0.01	4.0	0.9	$1.698e^{-1}$	36
5.0	0.01	4.0	0.99	$1.698e^{-1}$	72
5.0	0.1	2.0	0.9	$8.821e^{-2}$	559
5.0	0.1	2.0	0.99	$8.821e^{-2}$	1118
5.0	0.1	4.0	0.9	$8.821e^{-2}$	85
5.0	0.1	4.0	0.99	$8.821e^{-2}$	169

β_g	ε	τ	N	δ_ε	$1-q$
3.0	0.01	2.0	100	$4.224e^{-1}$	0.97
3.0	0.01	2.0	1000	$4.224e^{-1}$	1.0
3.0	0.01	4.0	100	$4.224e^{-1}$	1.0
3.0	0.01	4.0	1000	$4.224e^{-1}$	1.0
3.0	0.1	2.0	100	$2.295e^{-1}$	0.83
3.0	0.1	2.0	1000	$2.295e^{-1}$	1.0
3.0	0.1	4.0	100	$2.295e^{-1}$	1.0
3.0	0.1	4.0	1000	$2.295e^{-1}$	1.0
5.0	0.01	2.0	100	$1.698e^{-1}$	0.69
5.0	0.01	2.0	1000	$1.698e^{-1}$	1.0
5.0	0.01	4.0	100	$1.698e^{-1}$	1.0
5.0	0.01	4.0	1000	$1.698e^{-1}$	1.0
5.0	0.1	2.0	100	$8.821e^{-2}$	0.34
5.0	0.1	2.0	1000	$8.821e^{-2}$	0.98
5.0	0.1	4.0	100	$8.821e^{-2}$	0.93
5.0	0.1	4.0	1000	$8.821e^{-2}$	0.99

As the Strong Maximum Test involves the computation of N values of the limit state function, which is computationally intensive, it is interesting to have more than just an indication about the quality of $\underline{OP}^*$. In fact, the test gives some information about the trace of the limit state function on the sphere of radius $\beta(1+\delta)$ centered on the origin of the $\underline{u}$-space. Two cases can be distinguished:

- Case 1: set 1 is empty. We are confident on the fact that $\underline{OP}^*$ is a design point verifying the hypothesis according to which most of the contribution of P_f is concentrated in the vicinity of $\underline{OP}^*$. By using the value of the limit state function on the sample $(\underline{U}_1, \ldots, \underline{U}_N)$, we can check if the limit state function is reasonably linear in the vicinity of $\underline{OP}^*$, which can validate the second hypothesis of FORM.
If the behaviour of the limit state function is not linear, we can decide to use an importance sampling version of the Monte Carlo method for computing the probability of failure (refer to [Importance sampling]). However, the information obtained through the Strong Max Test, according to which $\underline{OP}^*$ is the actual design point, is quite essential : it allows to construct an effective importance sampling density, e.g. a multidimensional gaussian distribution centered on $\underline{OP}^*$.

- Case 2: set 1 is not empty. There are two possibilities:

  1. We have found some points that suggest that $\underline{OP}^*$ is not a strong maximum, because for some points of the sampled sphere, the value taken by the limit state function is slightly negative;
  2. We have found some points that suggest that $\underline{OP}^*$ is not even the global maximum, because for some points of the sampled sphere, the value taken by the limit state function is very negative.
  In the first case, we can decide to use an importance sampling version of the Monte Carlo method for computing the probability of failure, but with a mixture of e.g. multidimensional gaussian

distributions centered on the U_i in $\mathcal{D}_f$ (refer to [Importance Sampling]). In the second case, we can restart the search of the design point by starting at the detected U_i.

More details can be found in the following reference:

- A. Dutfoy, R. Lebrun, 2006, “The Strong Maximum Test: an efficient way to assess the quality of a design point,” PSAM8, New Orleans.

4.3.10 Step C – Estimating the probability of an event using Sampling

Mathematical description

Goal

Using the probability distribution of a random vector $\underline{X}$, we seek to evaluate the following probability:

$$P_f = \mathbb{P}\left(g\left(\underline{X}, \underline{d}\right) < 0\right)$$

Here, $\underline{X}$ is a random vector, $\underline{d}$ a deterministic vector, $g(\underline{X}, \underline{d})$ the function known as "limit state function" which enables the definition of the event $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \, / \, g(\underline{X}, \underline{d}) \le 0\}$.

Principle

If we have the set $\{\underline{x}_1, \ldots, \underline{x}_N\}$ of N independent samples of the random vector $\underline{X}$, we can estimate $\widehat{P}_f$ as follows:

$$\widehat{P}_f = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{g(\underline{x}_i, \underline{d}) \geq 0\}}$$

where $\mathbf{1}_{\{g(\underline{x}_i, \underline{d}) \geq 0\}}$ describes the indicator function equal to 1 if $g(\underline{x}_i, \underline{d}) \geq 0$ and equal to 0 otherwise; the idea here is in fact to estimate the required probability by the proportion of cases, among the N samples of $\underline{X}$, for which the event $\mathcal{D}_f$ occurs.

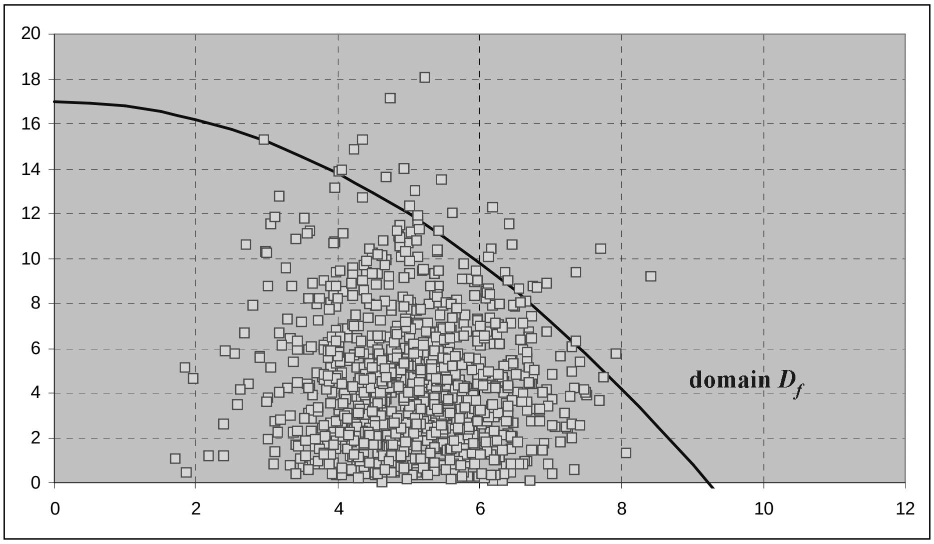
By the law of large numbers, we know that this estimation converges to the required value P_f as the sample size N tends to infinity. A good indicator of the uncertainty of this estimation is the coefficient of variation, which describes the relationship between its standard deviation (i.e its dispersion) and its mean, and this is estimated by:

$$\mathrm{CV}_{\widehat{P}_f} = \left(\frac{1 - \widehat{P}_f}{N \widehat{P}_f}\right)^{0.5}$$

The Central Limit Theorem enables the difference between the estimated value and the sought value to be controlled by means of a confidence interval (especially if N is sufficiently large, typically $N >$ a few dozens even if there is now way to say for sure if the asyptotic behaviour is reached). For a probability α strictly between 0 and 1 chosen by the user, one can, for example, be sure with a confidence α, that the true value of P_f is between $\widehat{P}_{f,\inf}$ and $\widehat{P}_{f,\sup}$ calculated analytically from simple formulae. To illustrate, for $\alpha = 0.95$:

$$\widehat{P}_{f,\inf} = \widehat{P}_f - 1.96 \left(\frac{\widehat{P}_f(1 - \widehat{P}_f)}{N}\right)^{0.5}, \quad \widehat{P}_{f,\sup} = \widehat{P}_f + 1.96 \left(\frac{\widehat{P}_f(1 - \widehat{P}_f)}{N}\right)^{0.5}$$

$$\text{that is to say } \Pr\left(\widehat{P}_{f,\inf} \leq P_f \leq \widehat{P}_{f,\sup}\right) = 0.95$$



Example of Monte-Carlo estimation of the probability of the event Df in dimension 2 :
here, N=1000; the probability estimate is equal to 0.032 because 32 trials out of 1000 are in the domain Df.
The 95% confidence interval is thus [0.021,0.043].

Other notations

Direct sampling, Crude Monte Carlo method, Classical Monte Carlo integration

Link with OpenTURNS methodology

This method is used in step C and enables the probability of exceeding the threshold of an output variable (we refer to the probability of exceeding the threshold (critical region) because the inequality $g(\underline{X}, \underline{d}) \leq 0$ by convention defines a reliability/critical region, and is in the general case the rewritten inequality of type $Z \geq$ threshold where Z is a a random variable function of $\underline{X}$ and $\underline{d}$).
This amounts to calculating the cumulative distribution function of the output variable at a point and thus propagating the uncertainty defined in step B using the model defined in step A.

Input data:

- $\underline{X}$: random vector modelling the unknown variables defined in step A and for which the joint probability density function has been defined in step B,
- $\underline{d}$: vector of deterministic calculation parameters,
- $g(\underline{X}, \underline{d}) < 0$: probabilistic criterion specified in step A,

Parameters:

- N: number of simulations to be carried out (samples to be taken) (maximal in the case where $\left(\mathrm{CV}_{\widehat{P}_f}\right)_{\max}$ is specified, see next parameter),

- $\left(\mathrm{CV}_{\widehat{P}_f}\right)_{\max}$: maximal coefficient of variation of the probability estimator (optional),
- α: confidence level required for the confidence interval.

Outputs:

- $\widehat{P}_f$: estimation of the probability of exceeding the threshold (critical value/region),
- $\mathrm{Var}(\widehat{P}_f)$: estimation of the variance of the probability estimator,
- $\widehat{P}_{f,\mathrm{sup}} - \widehat{P}_{f,\mathrm{inf}}$: lenght of the confidence interval.

References and theoretical basics

The standard Monte-Carlo method requires very few special properties for the function g: it should be measurable and integrable but can be irregular, non-convex…On the other hand, this method is not suitable when the probability to be estimated is small and when the CPU time needed to evaluate the criterion $g(\underline{X}, \underline{d}) \leq 0$ is considerable. In practice, the standard Monte-Carlo method is not recommended except if one has (for $P_f < 10^{-2}$):

$$\frac{t_{\mathrm{CPU}}\left\{g(\underline{X}, \underline{d}) \leq 0\right\}}{P_f \times (\text{estimation precision})^2} \leq \text{available machine time}$$

where:

- $t_{\mathrm{CPU}}\left\{g(\underline{X}, \underline{d}) \leq 0\right\}$: CPU time needed to evaluation the criterion $\left\{g(\underline{X}, \underline{d}) \leq 0\right\}$ for given data values of $\underline{X}$ and $\underline{d}$,
- estimation precision: desired limit for the coefficient of variation of the estimator,
- available machine time: desired limit on the total duration of the estimation.

Readers interested in the problem of estimating the probability of exceeding a threshold are referred to [FORM], [SORM], [LHS], [Importance Sampling] and [Directional Simulation].
The following provide an interesting bibliographical starting point to further study of this method:

- Robert C.P., Casella G. (2004). Monte-Carlo Statistical Methods, Springer, ISBN 0-387-21239-6, 2nd ed.
- Rubinstein R.Y. (1981). Simulation and The Monte-Carlo methods, John Wiley & Sons

4.3.11 Step C – Estimating the probability of an event using Importance Sampling

Mathematical description

Goal

Let us note $\mathcal{D}_f = \{\underline{x} \in \mathbb{R}^{n_X} | g(\underline{x}, \underline{d}) \leq 0\}$. The goal is to estimate the following probability:

$$P_f = \int_{\mathcal{D}_f} f_{\underline{X}}(\underline{x}) d\underline{x} = \int_{\mathbb{R}^{n_X}} \mathbf{1}_{\{g(\underline{x},\underline{d})\ \leq 0\ \}} f_{\underline{X}}(\underline{x}) d\underline{x} = \mathbb{P}\left(\{g(\underline{X}, \underline{d}) \leq 0\}\right)$$

Principles

This methode is a method based on sampling. The main idea of the Importance Sampling method is to replace the initial probability distribution of the input variables by a more "efficient" one. "Efficient" means that more events will be counted in the failure domain $\mathcal{D}_f$ and thus reduce the variance of the estimator of the probability of exceeding a threshold. Let $\underline{Y}$ be a random vector such that its probability density function $f_{\underline{Y}}(\underline{y}) > 0$ almost everywhere in the domain $\mathcal{D}_f$,

$$\begin{aligned} P_f &= \int_{\mathbb{R}^{n_X}} \mathbf{1}_{\{g(\underline{x},\underline{d}) \leq 0\}} f_{\underline{X}}(\underline{x}) d\underline{x} \\ &= \int_{\mathbb{R}^{n_X}} \mathbf{1}_{\{g(\underline{x},\underline{d}) \leq 0\}} \frac{f_{\underline{X}}(\underline{x})}{f_{\underline{Y}}(\underline{x})} f_{\underline{Y}}(\underline{x}) d\underline{x} \end{aligned}$$

The estimator built by Importance Sampling method is:

$$\hat{P}_{f,IS}^N = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{g(\underline{Y}_{\ i}),\underline{d}) \leq 0\}} \frac{f_{\underline{X}}(\underline{Y}_{\ i})}{f_{\underline{Y}}(\underline{Y}_{\ i})}$$

where:

- N is the total number of computations,
- the random vectors $\{\underline{Y}_i, i = 1 \dots N\}$ are independent, identically distributed and following the probability density function $f_{\underline{Y}}$

Confidence Intervals

With the notations,

$$\begin{aligned} \mu_N &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{g(\underline{y}_{\ i}),\underline{d}) \leq 0\}} \frac{f_{\underline{X}}(\underline{y}_{\ i})}{f_{\underline{Y}}(\underline{y}_{\ i})} \\ \sigma_N^2 &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{1}_{\{g(\underline{y}_i),\underline{d}) \leq 0\}} \frac{f_{\underline{X}}(\underline{y}_{\ i})}{f_{\underline{Y}}(\underline{y}_{\ i})} - \mu_N)^2 \end{aligned}$$

The asymptotic confidence interval of order $1 - \alpha$ associated to the estimator $P_{f,IS}^N$ is

$$[\mu_N - \frac{q_{1-\alpha/2} . \sigma_N}{\sqrt{N}}\ ;\ \mu_N + \frac{q_{1-\alpha/2} . \sigma_N}{\sqrt{N}}]$$

where $q_{1-\alpha/2}$ is the $1 - \alpha/2$ quantile from the standard distribution $\mathcal{N}(0, 1)$.

Other notations

This method could also be found under the name "Strategic Sampling", "Ponderated Sampling" or "Biased Sampling" (even if this estimator is not biased as it gives exactly the same result).

Link with OpenTURNS methodology

This method is part of the step C of the global methodology. It requires the specification the joined probability density function of the input variables and the value of the threshold and the comparison operator.

References and theoretical basics

There is no general result concerning the reduction of variance of $\hat{P}^N_{f,IS}$ in comparison with the variance of the initial Monte Carlo estimator $\hat{P}^N_{f,MC}$. Nevertheless, if one knows well the model (regularity, monotoneous,...), it is possible to define a more efficient joined probability density function. Nevertheless, there is a reduction of variance if one chooses a density $f_{\underline{Y}}(\underline{y})$ such that $f_{\underline{Y}}(\underline{y}) > f_{\underline{X}}(\underline{y})$ almost everywhere in the failure space. Indeed, in this case $\frac{f_{\underline{X}}(\underline{y})}{f_{\underline{Y}}(\underline{y})} < 1$ on all the domain, the variance being equal to:

$$\mathrm{Var}\left[\hat{P}_{f,IS}\right] = \int_{\mathcal{D}_f} \left(\frac{f_{\underline{X}}(\underline{y})}{f_{\underline{Y}}(\underline{y})}\right)^2 d\underline{y} - P_f^2 \quad < \quad \mathrm{Var}\left[\hat{P}_{f,MC}\right] = P_f - P_f^2$$

Moreover, one has to pay attention to define the same support for the joined pdf of the input variables to ensure the convergence.
The following references are a first introduction to the Monte Carlo methods:
W.G. Cochran. *Sampling Techniques.* John Wiley and Sons, 1977.
M.H. Kalos et P.A. *Monte Carlo Methods, volume I: Basics.* John Wiley and Sons, 1986.
R.Y. Rubinstein. *Simulation and the Monte Carlo Method.* John Wiley and Sons, 1981.
Autres références à intégrer

4.3.12 Step C – Directional Simulation

Mathematical description

Goal

Using the probability distribution of a random vector $\underline{X}$, we seek to evaluate the following probability:

$$P_f = \mathbb{P}\left(g\left(\underline{X}, \underline{d}\right) < 0\right)$$

Here, $\underline{X}$ is a random vector, $\underline{d}$ a deterministic vector, $g(\underline{X}, \underline{d})$ the function known as "limit state function" which enables the definition of the event $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \, / \, g(\underline{X}, \underline{d}) \le 0\}$.

Principle

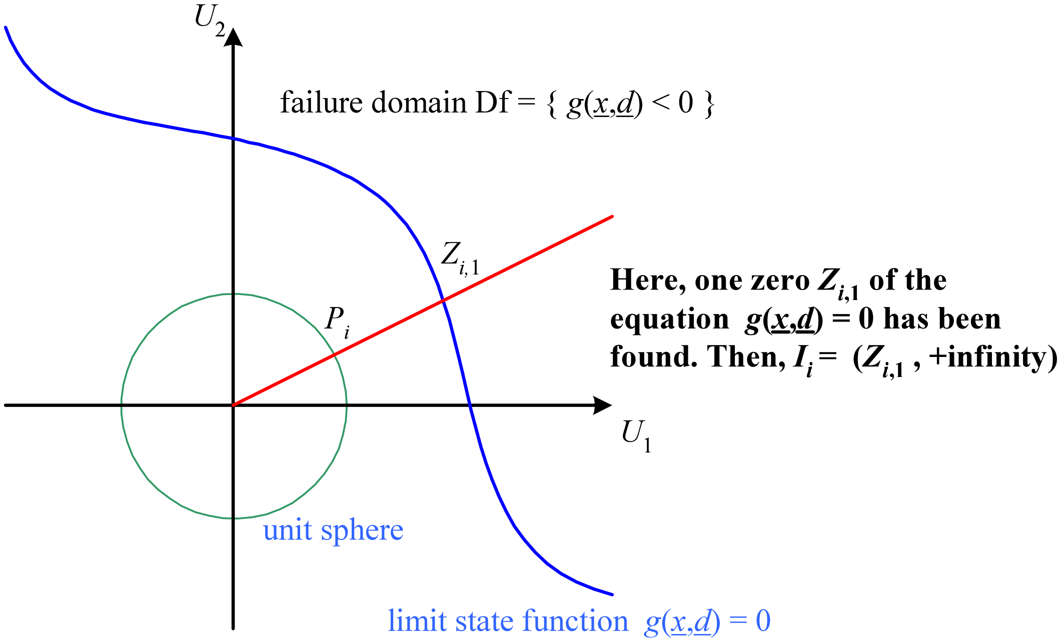
The directional simulation method is an accelerated sampling method. It implies a preliminary [iso-probabilistic transformation] , as for [FORM] and [SORM] methods; however, it remains based on sampling and is thus not an approximation method. In the transformed space, the (transformed) uncertain variables $\underline{U}$ are independant standard gaussian variables (mean equal to zero and standard deviation equal to 1).
Roughly speaking, each simulation of the directional simulation algorithm is made of three steps. For the i^{th} iteration, these steps are the following:

- Let $\mathcal{S} = \left\{\underline{u} \middle| ||\underline{u}|| = 1\right\}$. A point P_i is drawn randomly on $\mathcal{S}$ according to a uniform distribution.
- In the direction starting from the origin and passing through P_i, solutions of the equation $g(\underline{X}, \underline{d}) = 0$ (i.e. limits of $\mathcal{D}_f$) are searched. The set of values of $\underline{u}$ that belong to $\mathcal{D}_f$ is deduced for these solutions: it is a subset $I_i \subset \mathbb{R}$.
- Then, one calculates the probability $q_i = \mathbb{P}\left(||\underline{U}|| \in I_i\right)$. By property of independant standard variable, $||\underline{U}||^2$ is a random variable distributed according to a chi-square distribution, which makes the computation effortless.

Finally, the estimate of the probability P_f after N simulations is the following:

$$\widehat{P}_{f,DS} = \frac{1}{N} \sum_{i=1}^{N} q_i$$

The following figure illustrates the principle of an iteration in dimension 2.


U_2
failure domain Df = { $g(\underline{x},\underline{d}) < 0$ }
$Z_{i,1}$
P_i
Here, one zero $Z_{i,1}$ of the equation $g(\underline{x},\underline{d}) = 0$ has been found. Then, $I_i = (Z_{i,1}$, +infinity)
U_1
unit sphere
limit state function $g(\underline{x},\underline{d}) = 0$

The Central Limit Theorem enables the difference between the estimated value and the sought value to be controlled by means of a confidence interval (if N is sufficiently large, typically $N >$ a few dozens even if there is now way to say for sure if the asyptotic behaviour is reached). For a probability α strictly between 0 and 1 chosen by the user, one can, for example, be sure with a confidence α, that the true value of P_f is between $\widehat{P}_{f,\inf}$ and $\widehat{P}_{f,\sup}$ calculated analytically from simple formulae. To illustrate, for $\alpha = 0.95$:

$$\widehat{P}_{f,\inf} = \widehat{P}_{f,DS} - 1.96\frac{\sigma_q}{\sqrt{N}}, \ \widehat{P}_{f,\sup} = \widehat{P}_{f,DS} + 1.96\frac{\sigma_q}{\sqrt{N}}$$

$$\text{that is to say } \Pr\left(\widehat{P}_{f,\inf} \leq P_f \leq \widehat{P}_{f,\sup}\right) = 0.95$$

where σ_q denotes the empirical standard deviation of the sample $\{q_1, \ldots, q_N\}$.
In practice in Open TURNS, the Directional Sampling simulation requires the choice of:

- a Root Strategy :
  - RiskyAndFast : for each direction, we check whether there is a sign changement of the standard limit state function between the maximum distant point (at distance *MaximumDistance* from the center of the standard space) and the center of the standard space.
    In case of sign changement, we research one root in the segment [origine, maximum distant point] with the selectionned non linear solver.
    As soon as founded, the segment [root, infinity point] is considered within the failure space.
  - MediumSafe : for each direction, we go along the direction by step of lenght *stepSize* from the origin to the maximum distant point (at distance *MaximumDistance* from the center of the standard space) and we check whether there is a sign changement on each segment so formed.
    At the first sign changement, we research one root in the concerned segment with the selectionned

non linear solver. Then, the segment [root, maximum distant point] is considered within the failure space.
If *stepSize* is small enough, this strategy garantees us to find the root which is the nearest from the origine.

  - SafeAndSlow : for each direction, we go along the direction by step of lenght *stepSize* from the origine to the maximum distant point(at distance *MaximumDistance* from the center of the standard space) and we check whether there is a sign changement on each segment so formed.
    We go until the maximum distant point. Then, for all the segments where we detected a the presence of a root, we research the root with the selectionned non linear solver. We evaluate the contribution to the failure probability of each segment.
    If *stepSize* is small enough, this strategy garantees us to find all the roots in the direction and the contribution of this direction to the failure probability is precisely evaluated.

- a Non Linear Solver :
  - Bisection : bisection algorithm,
  - Secant : based on the evaluation of a segment between the two last iterated points,
  - Brent : mix of Bisection, Secant and inverse quadratic interpolation.
- and a Sampling Strategy :
  - RandomDirection : we generate some points on the sphere unity according to the uniform distribution and we consider both opposite directions so formed.
  - OrthogonalDirection : this strategy is parametered by $k \in \mathbb{N}$. We generate one direct orthonormalised base $(e_1, \dots, e_{n_X})$ within the set of orthonormalised bases. We consider all the renormalised linear combinations of k vectors within the n_X vectors of the base, where the coefficients of the linear combinations are equal to $+1, -1$. There are $C_n^k 2^k$ new vectors v_i. We consider each direction defined by each vector v_i.
    If $k = 1$, we consider all the axes of the standard space.

Other notations

Link with OpenTURNS methodology

This method is used in step C and enables the probability of exceeding the threshold of an output variable (we refer to the probability of exceeding the threshold (critical region) because the inequality $g(\underline{X}, \underline{d}) \leq 0$ by convention defines a reliability/critical region, and is in the general case the rewritten inequality of type $Z \geq$ threshold where Z is a a random variable function of $\underline{X}$ and $\underline{d}$).
This amounts to calculating the cumulative distribution function of the output variable at a point and thus propagating the uncertainty defined in step B using the model defined in step A.

Input data:

- $\underline{X}$: random vector modelling the unknown variables defined in step A and for which the joint probability density function has been defined in step B,
- $\underline{d}$: vector of deterministic calculation parameters,
- $g(\underline{X}, \underline{d}) < 0$: probabilistic criterion specified in step A,

Parameters:

- N: number of simulations,
- α: confidence level required for the confidence interval,
- Root Strategy,
- Non-linear Solver,
- Sampling Strategy.

Outputs:

- $\widehat{P}_{f,DS}$: estimation of the probability of exceeding the threshold,
- $\mathrm{Var}(\widehat{P}_f)$: estimation of the variance of the probability estimator,
- $\widehat{P}_{f,\mathrm{sup}} - \widehat{P}_{f,\mathrm{inf}}$: lenght of the confidence interval.

References and theoretical basics

Readers interested in the problem of estimating the probability of exceeding a threshold are also referred to [FORM], [SORM], [LHS], [Importance Sampling] and [Crude Monte-Carlo sampling].
The following provide an interesting bibliographical starting point to further study of this method:

- Robert C.P., Casella G. (2004). Monte-Carlo Statistical Methods, Springer, ISBN 0-387-21239-6, 2nd ed.
- Rubinstein R.Y. (1981). Simulation and The Monte-Carlo methods, John Wiley & Sons
- Bjerager, P. (1988). “Probability integration by Directional Simulation". Journal of Engineering Mechanics, vol. 114, n°8

4.3.13 Step C – Estimating the probability of an event using Latin Hypercube Sampling

Mathematical description

Goal

Let us note $\mathcal{D}_f = \{\underline{x} \in \mathbb{R}^{n_X} \mid g(\underline{x}, \underline{d}) \leq 0\}$. The goal is to estimate the following probability:

$$P_f = \int_{\mathcal{D}_f} f_{\underline{X}}(\underline{x}) d\underline{x} = \int_{\mathbb{R}^{n_X}} \mathbf{1}_{\{g(\underline{x},\underline{d}) \leq 0\}} f_{\underline{X}}(\underline{x}) d\underline{x} = \mathbb{P}(\{\, g(\underline{X}, \underline{d}) \leq 0\})$$

Principles

LHS or Latin Hypercube Sampling is a sampling method enabling to better cover the domain of variations of the input variables, thanks to a stratified sampling strategy. This method is applicable in the case of independent input variables. The sampling procedure is based on dividing the range of each variable into several intervals of equal probability. The sampling is undertaken as follows:

- **Step n°1** The range of each input variable is stratified into isoprobabilistic cells,
- **Step n°2** A cell is uniformly chosen among all the available cells,
- **Step n°3** The random number is obtained by inverting the Cumulative Density Function locally in the chosen cell,
- **Step n°4** All the cells having a common strate with the previous cell are put apart from the list of available cells.

The estimator of the probability of failure with LHS is given by:

$$\hat{P}^N_{f,LHS} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{g(\underline{X}^i, \underline{d}) \leq 0\}}$$

where the sample of $\{\underline{X}^i, i = 1 \ldots N\}$ is obtained as described previously.
One can show that:

$$\text{Var}\left[\hat{P}^N_{f,LHS}\right] \leq \frac{N}{N-1}.\text{Var}\left[\hat{P}^N_{f,MC}\right]$$

where:

- $\text{Var}\left[\hat{P}^N_{f,LHS}\right]$ is the variance of the estimator of the probability of exceeding a threshold computed by the LHS technique,
- $\text{Var}\left[\hat{P}^N_{f,MC}\right]$ is the variance of the estimator of the probability of exceeding a threshold computed by a crude Monte Carlo method.

Confidence Interval

With the notations,

$$\begin{aligned} \mu_N &= \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}_{\{g(\underline{x}_i),\underline{d}) \leq 0\}} \\ \sigma_N^2 &= \frac{1}{N} \sum_{i=1}^{N} (\mathbf{1}_{\{g(\underline{x}^i),\underline{d}) \leq 0\}} - \mu_N)^2 \end{aligned}$$

the asymptotic confidence interval of order $1-\alpha$ associated to the estimator $P^N_{f,LHS}$ is

$$[\mu_N - \frac{q_{1-\alpha/2} \cdot \sigma_N}{\sqrt{N}} \; ; \; \mu_N + \frac{q_{1-\alpha/2} \cdot \sigma_N}{\sqrt{N}}]$$

where $q_{1-\alpha/2}$ is the $1-\alpha/2$ quantile from the reduced standard gaussian law $\mathcal{N}(0,1)$.
It gives an unbiased estimate for P_f (reminding that all input variables must be independent).

Other notations

This method is derived from a more general method called 'Stratified Sampling'.

Link with OpenTURNS methodology

This method is part of the step C of the global methodology. It requires the specification of the joined probability density function of the input variables and the definition of the threshold. The PDF must have an independent copula.

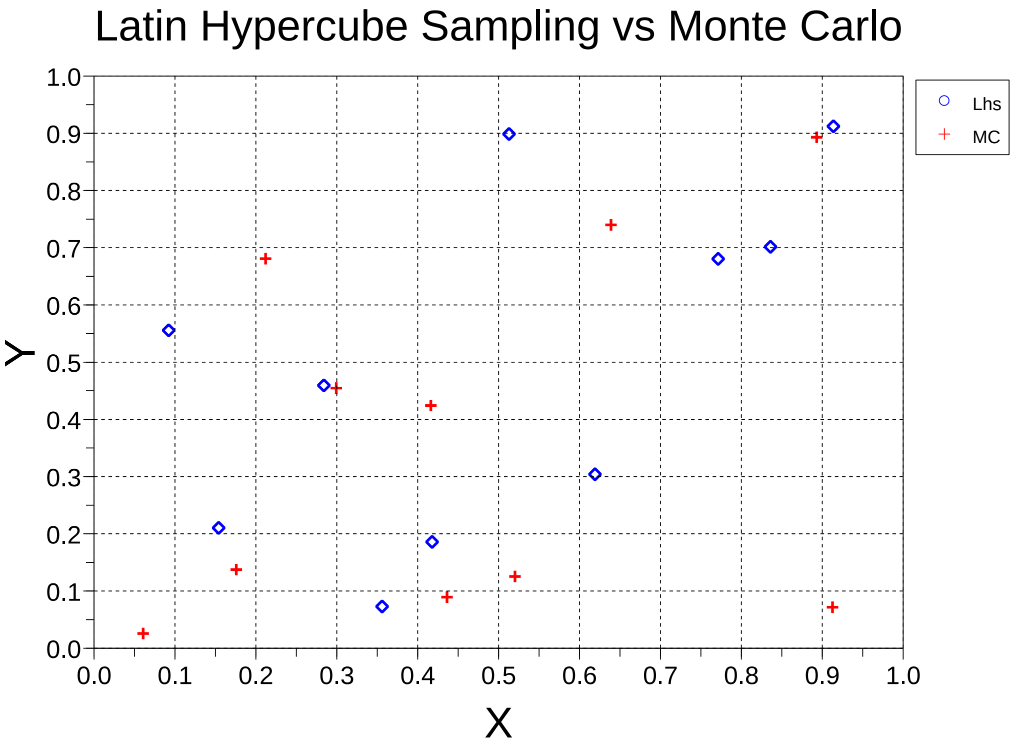
References and theoretical basics

- This method *a priori* enables a better exploration of the domain of variations of the input variables. No general rule can guarantee a better efficiency of the LHS sampling than the classical Monte Carlo sampling. Nvertheless, one can show that the LHS strategy leads to a variance reduction if the model is motoneous over each variable.
- Be careful, this method is valid only if the input random variables are independent!
- Moreover, for reliability problems, when the failure probability is low, the tails of the distributions usually contain the most influent domains in terms of reliability.
- A fruitful link towards the global approach can be established with the files

[Monte Carlo Method to evaluate a probability to exceed a threshold],
[Importance Sampling to evaluate a probability to exceed a threshold] coming from the methodology.

Examples

To illustrate this method, we consider the sampling strategy of an input vector of dimension 2. Both components follow a uniform law $\mathcal{U}(0,1)$. The figure compares the population of 30 points obtained by a Latin Hypercube Sampling and by a Monte Carlo Sampling.



With the LHS sampling strategy, each row and each column is filled by a blue square whereas some row and column do not contain any red cross.

4.3.14 Step C – Estimating a quantile by Sampling / Wilks' Method

Mathematical description

Goal

Let us denote $\underline{Y} = h\left(\underline{X}, \underline{d}\right) = \left(Y^1, \ldots, Y^{n_Y}\right)$, where $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ is a random vector, and $\underline{d}$ a deterministic vector. We seek here to evaluate, using the probability distribution of the random vector $\underline{X}$, the α-quantile $q_{Y^i}(\alpha)$ of Y^i, where $\alpha \in (0, 1)$:

$$\mathbb{P}\left(Y^i \leq q_{Y^i}(\alpha)\right) = \alpha$$

Principle

If we have a sample $\{\underline{x}_1, \ldots, \underline{x}_N\}$ of N independent samples of the random vector $\underline{X}$, $q_{Y^i}(\alpha)$ can be estimated as follows:

- the sample $\{\underline{x}_1, \ldots, \underline{x}_N\}$ of vector $\underline{X}$ is first transformed to a sample $\left\{y^i_1, \ldots, y^i_N\right\}$ of the variable Y^i, using $\underline{y} = h(\underline{x}_i, \underline{d})$,
- the sample $\left\{y^i_1, \ldots, y^i_N\right\}$ is then placed in ascending order, which gives the sample $\left\{y^{(1)}, \ldots, y^{(N)}\right\}$,
- this empirical estimation of the quantile is then calculated by the formula:

  $$\widehat{q}_{y^i}(\alpha) = y^{([N\alpha]+1)}$$

  where $[N\alpha]$ denotes the integral part of $N\alpha$.

For example, if $N = 100$ and $\alpha = 0.95$, $\widehat{q}_Z(0.95)$ is equal to $y^{(96)}$, which is the 5th largest value of the sample $\left\{y^i_1, \ldots, y^i_N\right\}$. We note that this estimation has no meaning unless $1/N \leq \alpha \leq 1 - 1/N$. For example, if $N = 100$, one can only consider values of a to be between 1% and 99%.

It is also possible to calculate an upper limit for the quantile with a confidence level β chosen by the user; one can then be sure with a β level of confidence that the real value of $q_{Y^i}(\alpha))$ is less than or equal to $\widehat{q}_{Y^i}(\alpha)_{\mathrm{sup}}$:

$$\mathbb{P}\left(q_{Y^i}(\alpha) \leq \widehat{q}_{Y^i}(\alpha)_{\mathrm{sup}}\right) = \beta$$

The most robust method for calculating this upper limit consists of taking $\widehat{q}_{Y^i}(\alpha)_{\mathrm{sup}} = y^{(j(\alpha,\beta,N))}$ where $j(\alpha, \beta, N)$ is an integer between 2 and N found by solving the equation:

$$\sum_{k=1}^{j(\alpha,\beta,N)-1} C^k_N \alpha^k \left(1-\alpha\right)^{N-k} = \beta$$

A solution to this does not necessarily exist, i.e. there may be no integer value for $j(\alpha, \beta, N)$ satisfying this equality; one can in this case choose the smallest integer j such that:

$$\sum_{k=1}^{j(\alpha,\beta,N)-1} C^k_N \alpha^k \left(1-\alpha\right)^{N-k} > \beta$$

which ensures that $\mathbb{P}\left(q_{Y^i}(\alpha) \leq \widehat{q}_{Y^i}(\alpha)_{\text{sup}}\right) > \beta$; in other words, the level of confidence of the quantile estimation is greater than that initially required.
This formula of the confidence interval can be used in two ways:

- either directly to determine $j(\alpha, \beta, N)$ for the values α, β, N chosen by the user,
- or in reverse to determine the number N of simulations to be carried out for the values α, β and $j(\alpha, \beta, N)$ chosen by the user; this is known as Wilks' formula.

For example for $\alpha = \beta = 95\%$, we take $j = 59$ with $N = 59$ simulations (that is the maximum value out of 59 samples) or else $j = 92$ with $N = 93$ simulations (that is the second largest result out of the 93 selections). For values of N between 59 and 92, the upper limit is the maximum value of the sample. The following tabular presents the whole results for $N \leq 1000$, still for $\alpha = \beta = 95\%$.

N	Rank of the uper bound of the quantile	Rank of the empirical the quantile
59	59	57
93	92	89
124	122	118
153	150	146
181	177	172
208	203	198
234	228	223
260	253	248
286	278	272
311	302	296
336	326	320
361	350	343
386	374	367
410	397	390
434	420	413
458	443	436
482	466	458
506	489	481
530	512	504
554	535	527
577	557	549
601	580	571
624	602	593
647	624	615
671	647	638
694	669	660
717	691	682
740	713	704
763	735	725
786	757	747
809	779	769
832	801	791
855	823	813
877	844	834
900	866	856
923	888	877
945	909	898
968	931	920
991	953	942

Other notations

$\widehat{q}_{Y^i}(\alpha)$ is often called the “empirical α-quantile" for the variable Y^i.

Link with OpenTURNS methodology

In the overall process, the Monte Carlo simulation method for estimating the variance appears in step C “Propagation of Uncertainty" when the study of uncertainty is concerned with the dispersion of the variable of interest Y^i defined in step A “Specifying Criteria and the Case Study".
Input data:

- $\underline{X}$: random vector modelling the unknown variables defined in step A and for which the joint probability density function has been defined in step B,
- $\underline{d}$: vector of deterministic calculation parameters,
- $\underline{Y} = h(\underline{X}, \underline{d})$: output variable / variable of interest specified in step A

Parameters:

- α: quantile level (α-quantile),
- β: confidence level for the quantile's upper bound,
- N: number of simulations to be carried out (which can be computed by Open TURNS using Wilk's formula)

Outputs:

- $\widehat{q}_Z(\alpha)$: quantile estimate,
- $\widehat{q}_Z(\alpha)_{\text{sup}}$: quantile upper bound with confidence β

References and theoretical basics

The Monte-Carlo standard method does not require the function h to have any special property (it can be non-linear, non-monotonic, non-differentiable, discontinuous, etc.) and the number of necessary simulations does not depend on the number of components of vector $\underline{X}$. On the other hand, this method is not suitable (for the estimation of the quantile) or is very conservative (for the estimation of the upper limit) if N is small and if α and β are very close to 1.
The following references provide an interesting bibliographical starting point for further study of the method described here:

- Wilks, S.S. (1962). “Mathematical Statistics", New York, John Wiley.
- Robert C.P., Casella G. (2004). Monte-Carlo Statistical Methods, Springer, ISBN 0-387-21239-6, 2nd ed.
- Rubinstein R.Y. (1981). Simulation and The Monte-Carlo methods, John Wiley & Sons

5 Open TURNS' methods for Step C': ranking uncertainty sources / sensitivity analysis

Ranking methods can be used to analyse the respective importance of each uncertainty source with respect to a probabilistic criterion. Open TURNS proposes ranking methods for two probabilistic criteria defined in the [global methodology guide]: probabilist criterion on central dispersion (expectation and variance), probability of exceeding a threshold / failure probability.

5.1 Probabilistic criteria

5.1.1 Central dispersion probabilistic criterion

Each propagation method available for this criterion (see step C) leads to one or several ranking methods.

- Approximation methods
  - [Quadratic combination's importance factors] – see page 125

- Sampling methods
  - [Ranking based on Pearson correlation] – see page 127
  - [Ranking based on Spearman rank correlation] – see page 129
  - [Ranking based on Standard Regression Coefficients (SRC)] – see page 131
  - [Ranking based on Partial (Pearson) Correlation Coefficients (PCC)] – see page 133
  - [Ranking based on Partial (Spearman) Rank Correlation coefficients (PRCC)] – see page 135

5.1.2 Probability of exceeding a threshold / failure probability

- Approximation methods
  - FORM-SORM methods
    - [FORM Importance Factors] – see page 137
    - [FORM Sensitivity Factors] – see page 140

5.2 Methods description

5.2.1 Step C' – Importance Factors derived from Quadratic Combination Method

Mathematical description

Goal

The importance factors derived from a quadratic combination method are defined to discriminate the influence of the different inputs towards the output variable for central dispersion analysis.

Principles

The importance factors are derived from the following expression. It can be shown by Taylor expansion of the output variable z ($n_Z = 1$) around $\underline{x} = \underline{\mu}_X$ and computation of the variance that :

$$\mathrm{Var}\left[Z\right] \approx \nabla h(\underline{\mu}_X).\mathrm{Cov}\left[\underline{X}\right].{}^t\nabla h(\underline{\mu}_X)$$

which can be re written :

$$\begin{aligned} 1 &\approx \sum_{i=1}^{n_X} \frac{\partial h(\underline{\mu}_X)}{\partial X^i} \times \frac{\sum_{j=1}^{n_X} \frac{\partial h(\underline{\mu}_X)}{\partial x^j}.(\mathrm{Cov}\left[\underline{X}\right])_{ij}}{\mathrm{Var}\left[Y\right]} \\ &\approx \mathcal{F}_1 + \mathcal{F}_2 + \ldots + \mathcal{F}_{n_X} \end{aligned}$$

Vectorial definition

$$\underline{\mathcal{F}} = \nabla h(\underline{\mu}_X) \times \frac{\mathrm{Cov}\left[\underline{X}\right].{}^t\nabla h(\underline{\mu}_X)}{\mathrm{Var}\left[Z\right]}$$

Scalar definition

$$\mathcal{F}_i = \frac{\partial h(\underline{\mu}_X)}{\partial x^i} \times \frac{\sum_{j=1}^{n_X} \frac{\partial h(\underline{\mu}_X)}{\partial x^j}.(\mathrm{Cov}\left[\underline{X}\right])_{ij}}{\mathrm{Var}\left[Y\right]}$$

where:

- $\nabla h(\underline{x}) = \left(\frac{\partial h(\underline{x})}{\partial x^i}\right)_{i=1,\ldots,n_X}$ is the gradient of the model at the point $\underline{x}$,
- $\mathrm{Cov}\left[\underline{X}\right]$ is the covariance matrix,
- $\underline{\mu}_X$ is the mean of the input random vector,
- $\mathrm{Var}\left[Z\right]$ is the variance of the output variable.

Interpretation of the importance factors obtained with Open TURNS when all X^i are independent the

Each coefficient $\frac{\partial h(\underline{x})}{\partial x^i}$ is a linear estimate of the number of units change in the variable $y = h(\underline{x})$ as a result of a unit change in the variable x^i. This first term depends on the physical units of the variables and is meaningful only when the units of the model are known. In the general case, as the variables have different physical units, it is not possible to compare these sensitivities $\frac{\partial h(\underline{x})}{\partial x^i}$ the one with the others. This is the reason why the importance factor used within Open TURNS are normalized factors. These factors enable to make the results comparable independently of the original units of the inputs of the model. The second term $\frac{\sum_{j=1}^{n_X} \frac{\partial h(\underline{\mu}_X)}{\partial x^j}.(\mathrm{Cov}[\underline{X}])_{ij}}{\mathrm{Var}[Z]}$ is the renormalization factor.

To summarize, the coefficients $(\mathcal{F}_i)_{i=1,\dots,n_X}$ represent a linear estimate of the percentage change in the variable $z = h(\underline{x})$ caused by one percent change in the variable x^i. The importance factors are independent of the original units of the model, and are comparable with each other.

Other notations

Importance Factors derived from Perturbation Methods

Link with OpenTURNS methodology

These computations are part of the step C' of the global methodology. It requires to have performed the steps A, B and C.

References and theoretical basics

The computation of these importance factors enables to rank the influence of the input variables towards the output variable. These factors are computed 'near' the mean value of the output. Thus, it should not be used to evaluate the importance of the input variable around the tail of the output distribution (high level quantile for example).

Examples

5.2.2 Step C' – Uncertainty ranking using Pearson's correlation

Mathematical description

Goal

This method is concerned with analysing the influence the random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ has on a random variable Y^j which is being studied for uncertainty. Here we attempt to measure linear relationships that exist between Y^j and the different components X^i.

Principle

Pearson's correlation coefficient ρ_{Y^j,X^i}, defined in [Pearson's Coefficient], measures the strength of a linear relation between two random variables Y^j and X^i. If we have a sample made up of N pairs (y^j_1, x^i_1), (y^j_2, x^i_2), ..., (y^j_N, x^i_N), we can obtain $\widehat{\rho}_{Y^j,X^i}$ an estimation of Pearson's coefficient. The hierarchical ordering of Pearson's coefficients is of interest in the case where the relationship between Y^j and n_X variables $\left\{X^1, \ldots, X^{n_X}\right\}$ is close to being a linear relation:

$$Y^j \simeq a_0 + \sum_{i=1}^{n_X} a_i X^i$$

To obtain an indication of the role played by each X^i in the dispersion of Y^j, the idea is to estimate Pearson's correlation coefficient $\widehat{\rho}_{X^i,Y^j}$ for each i. One can then order the n_X variables $X^1, \ldots, X^{n_X}$ taking absolute values of the correlation coefficients: the higher the value of $\left|\widehat{\rho}_{X^i,Y^j}\right|$ the greater the impact the variable X^i has on the dispersion of Y^j.

Other notations

-

Link with OpenTURNS methodology

After a propagation of uncertainty (step C) using [Standard Monte Carlo] simulation, a hierarchy of sources of uncertainty can be obtained using Pearson's correlation coefficients. In fact, the N simulations enable the pairs (y_1^j, x_1^i), (y_2^j, x_2^i),..., (y_N^j, x_N^i) to be generated, where:

- $\underline{X} = \left\{X^1, \dots, X^n\right\}$ describes the input vector specified in step A "Specifying Criteria and the Case Study",
- Y^j describes a variable of interest or output variable defined in the same step.

The results produced as output of this method are the estimated Pearson's correlation coefficients $\widehat{\rho}_{X^i,Y^j}$ that the user may use, taking absolute values, to order the variables X^i hierarchically.

References and theoretical basics

This method of uncertainty ranking is particularly useful:

- when the study of uncertainty is concerned with the central dispersion of the variable of interest Y^j and not with its extreme values,
- when the relationships between Y^j and each of the components of $\underline{X}$ are close to linear relationships (so that Pearson's correlation coefficient can be interpreted),
- when this linear relationship is close to $Y^j = a_0 + \sum_{i=1}^{n_X} a_i X^i$ (i.e. no product terms of the type $X^i X^j$), and when the components of vector $\underline{X}$ are statistically independent. If this is not the case, $\left|\widehat{\rho}_{X^i,Y^j}\right|$ reflects not only the influence of X^i on Y^j but equally the influence of other variables X^j related to X^i (e.g. an unimportant variable X^i could have a strong coefficient for the correlation with Y^j only because it is related – statistically or by a product term – to another variable X^j which has enormous impact on Y^j).

Readers interested in other methods of uncertainty ranking that can be applied after Monte-Carlo simulation when the assumptions of linearity and/or independence are violated are also referred to [Uncertainty ranking using Spearman] , [Hierarchical Ordering using SRC] , [Uncertainty ranking with Pearson's Partial Correlation Coefficients] and [Uncertainty ranking using Spearman's Partial Correlation Coefficients].
The following references provide an interesting bibliographic starting point to further study of the method described here:

- Saltelli, A., Chan, K., Scott, M. (2000). "Sensitivity Analysis", John Wiley & Sons publishers, Probability and Statistics series
- J.C. Helton, F.J. Davis (2003). "Latin Hypercube sampling and the propagation of uncertainty analyses of complex systems". Reliability Engineering and System Safety 81, p.23-69
- J.P.C. Kleijnen, J.C. Helton (1999). "Statistical analyses of scatterplots to identify factors in large-scale simulations, part 1 : review and comparison of techniques". Reliability Engineering and System Safety 65, p.147-185

5.2.3 Step C' – Uncertainty ranking using Spearman's correlation

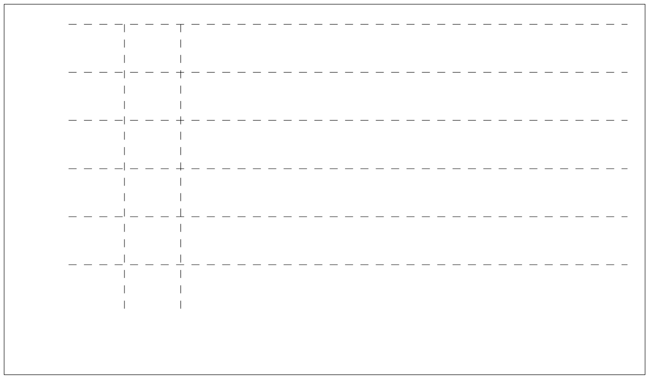
Mathematical description

Goal

This method is concerned with analyzing the influence the random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ has on a random variable Y^j which is being studied for uncertainty. Here we attempt to measure monotonic relationships that exist between Y^j and the different components X^i.

Principle

Spearman's correlation coefficient $\rho^S_{Y^j,X^i}$, defined in [Spearman's Coefficient], measures the strength of a monotonic relation between two random variables Y^j and X^i. If we have a sample made up of N pairs (y^j_1, x^i_1), (y^j_2, x^i_2), …, (y^j_N, x^i_N), we can obtain $\widehat{\rho}^S_{Y^j,X^i}$ an estimation of Spearman's coefficient.
Hierarchical ordering using Spearman's coefficients is concerned with the case where the variable Y^j monotonically depends on the n_X variables $\left\{X^1, \ldots, X^{n_X}\right\}$. To obtain an indication of the role played by each X^i in the dispersion of Y^j, the idea is to estimate the Spearman correlation coefficients $\widehat{\rho}^S_{X^i,Y^j}$ for each i. One can then order the n_X variables $X^1, \ldots, X^{n_X}$ taking absolute values of the Spearman coefficients: the higher the value of $\left|\widehat{\rho}^S_{X^i,Y^j}\right|$, the greater the impact the variable X^i has on the dispersion of Y^j.



Other notations

Link with OpenTURNS methodology

After a propagation of uncertainty (step C) using [Standard Monte Carlo] simulation, a hierarchy of sources of uncertainty can be obtained using Spearman's correlation coefficients. In fact, the N simulations enable the pairs (y_1^j, x_1^i), (y_2^j, x_2^i),..., (y_N^j, x_N^i) to be generated, where:

- $\underline{X} = \left\{X^1, \ldots, X^n\right\}$ describes the input vector specified in step A "Specifying Criteria and the Case Study",
- Y^j describes the final variable of interest or output variable defined in the same step.

The results produced as output of this method are the estimated Spearman's correlation coefficients $\widehat{\rho}^S_{X^i,Y^j}$ that the user may use, taking absolute values, to order the variables X^i hierarchically.

References and theoretical basics

This method of hierarchical ordering is particularly useful:

- when the study of uncertainty is concerned with the central dispersion of the variable of interest Y^j and not with its extreme values,
- when the relationships between Y^j and each of the components of $\underline{X}$ are monotonic relationships (so that Spearman's correlation coefficient can be interpreted),
- when the components of vector $\underline{X}$ are statistically independent. If this is not the case, $\left|\widehat{\rho}^S_{X^i,Y^j}\right|$ reflects not only the influence of X^i on Y^j but equally the influence of other variables X^j related to X^i (e.g. an unimportant variable X^i could have a strong coefficient for the correlation with Y^j only because it is related to another variable X^j which has enormous impact on Y^j).

Readers interested in other methods of uncertainty ranking that can be applied after Monte-Carlo simulation when the assumptions of independence are violated are also referred to [Uncertainty ranking using SRC], [Uncertainty ranking with Pearson's Partial Correlation Coefficients] and [Uncertainty ranking using Spearman's Partial Correlation Coefficients].
The following references provide an interesting bibliographic starting point to further study of the method described here:

- Saltelli, A., Chan, K., Scott, M. (2000). "Sensitivity Analysis", John Wiley & Sons publishers, Probability and Statistics series
- J.C. Helton, F.J. Davis (2003). "Latin Hypercube sampling and the propagation of uncertainty analyses of complex systems". Reliability Engineering and System Safety 81, p.23-69
- J.P.C. Kleijnen, J.C. Helton (1999). "Statistical analyses of scatterplots to identify factors in large-scale simulations, part 1 : review and comparison of techniques". Reliability Engineering and System Safety 65, p.147-185

5.2.4 Step C' – Uncertainty Ranking using Standard Regression Coefficients

Mathematical description

Goal

This method is concerned with analysing the influence the random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ has on a random variable Y^j which is being studied for uncertainty. Here we attempt to measure linear relationships that exist between Y^j and the different components X^i.

Principle

The principle of the multiple linear regression model (see [Linear Regression] for more details) consists of attempting to find the function that links the variable Y^j to the n_x variables X^1,…,X^{n_X} by means of a linear model:

$$Y^j = a_0 + \sum_{i=1}^{n_X} a_i X^i + \varepsilon$$

where ε describes a random variable with zero mean and standard deviation σ independent of the input variables X^i. If the random variables $X^1, \ldots, X^{n_X}$ are independent, the variance of Y^j can be written as follows:

$$\mathrm{Var}\left[Y^j\right] = \sum_{i=1}^{n} a_i^2 \mathrm{Var}\left[X^i\right] + \sigma^2$$

The estimators for the regression coefficients $\widehat{a}_0, \widehat{a}_1, \ldots, \widehat{a}_{n_X}$, and the standard deviation σ are obtained from a sample of $(Y^j, X^1, \ldots, X^{n_X})$, that is a set of N values $(y_1^j, x_1^1, \ldots, x_1^{n_X})$,…,$(y_N^j, x_N^1, \ldots, x_N^{n_X})$. Uncertainty ranking by linear regression uses these estimates, and involves ordering the n_X variables $X^1, \ldots, X^{n_X}$ in terms of the estimated contribution of each X^i to the variance Y^j:

$$\widehat{C}_i = \frac{\widehat{a}_i^2 \widehat{\sigma}_i^2}{\displaystyle\sum_{j=1}^{n_X} a_j^2 \widehat{\sigma}_j^2 + \widehat{\sigma}^2}$$

where $\widehat{\sigma}_j$ describes the empirical standard deviation of the sample $(x_1^j, \ldots, x_N^j)$. This estimated contribution is by definition between 0 and 1. The closer it is to 1, the greater the impact the variable X^i has on the dispersion of Y^j.

Other notations

The contribution to the variance C_i is sometimes described in the literature as the "importance factor", because of the similarity between this approach to linear regression and the method of cumulative variance quadratic which uses the term importance factor (see [Quadratic combination – Perturbation method] and [Importance Factors]).

Link with OpenTURNS methodology

After a propagation of uncertainty (step C) using [Standard Monte Carlo] simulation, a hierarchy of sources of uncertainty can be obtained using Linear Regression. In fact, the N simulations enable the pairs (y_1^j, x_1^i),

(y_2^j, x_2^i),..., (y_N^j, x_N^i) to be generated, where:

- $\underline{X} = \left\{X^1, \ldots, X^n\right\}$ describes the input vector specified in step A "Specifying Criteria and the Case Study",
- Y^j describes the final variable of interest or output variable defined in the same step.

The results produced as output of this method are the estimated variance contributions $\widehat{C}_i$ that the user may use to order the variables X^i hierarchically.

References and theoretical basics

This method of hierarchical ordering is particularly useful:

- when the study of uncertainty is concerned with the central dispersion of the variable of interest Y^j and not with its extreme values, item when the relationships between Y^j and the components of $\underline{X}$ are close to linear relationships, and more generally when all the underlying assumptions of the multiple linear regression model are valid,
- when the components of vector $\underline{X}$ are independent, because if this is not the case the decomposition of the variance of Y^j given here would be no longer exact,
- when the number N of Monte-Carlo simulations is significantly higher than the number n_X of input random variables (it is preferable to have N/n_X at least greater by a factor of 10 so that the estimation of the n_X correlation coefficients provides a reasonable picture of reality).

Readers interested in the assumptions made for multiple linear regression models and in the tests needed to validate these assumptions are referred to [Linear Regression].
Other methods of uncertainty ranking can be applied after Monte-Carlo simulation, requiring a lesser number N of simulations or that can deal with non-linear/non-independent cases, are described in [Uncertainty Ranking using Pearson] , [Uncertainty Ranking using Spearman] , [Uncertainty Ranking using Pearson's Partial Correlation Coefficients] and [Uncertainty Ranking using Pearson's Partial Correlation Coefficients].
The following references provide an interesting bibliographic starting point to further study of the method described here:

- Saltelli, A., Chan, K., Scott, M. (2000). "Sensitivity Analysis", John Wiley & Sons publishers, Probability and Statistics series
- J.C. Helton, F.J. Davis (2003). "Latin Hypercube sampling and the propagation of uncertainty analyses of complex systems". Reliability Engineering and System Safety 81, p.23-69
- J.P.C. Kleijnen, J.C. Helton (1999). "Statistical analyses of scatterplots to identify factors in large-scale simulations, part 1 : review and comparison of techniques". Reliability Engineering and System Safety 65, p.147-185

5.2.5 Step C' – Uncertainty Ranking using Pearson's Partial Correlation Coefficients

Mathematical description

Goal

This method is concerned with analyzing the influence the random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ has on a random variable Y^j which is being studied for uncertainty. Here we attempt to measure linear relationships that exist between Y^j and the different components X^i.

Principle

The basic method of hierarchical ordering using Pearson's coefficients (see [Uncertainty Ranking using Pearson]) is concerned with the case where the variable Y^j linearly depends on n_X variables $\left\{X^1, \ldots, X^{n_X}\right\}$ but this can be misleading when statistical dependencies or interactions between the variables X^i (e.g. a crossed term $X^i \times X^j$) exist. In such a situation, the partial correlation coefficients can be more useful in ordering the uncertainty hierarchically: the partial correlation coefficients PCC_{X^i,Y^j} between the variables Y^j and X^i attempts to measure the residual influence of X^i on Y^j once influences from all other variables X^j have been eliminated.

The estimation for each partial correlation coefficient PCC_{X^i,Y^j} uses a set made up of N values $\left\{(y_1^j, x_1^1, \ldots, x_1^{n_X}), \ldots, (y_N^j, x_N^1, \ldots, x_N^{n_X})\right\}$ of the vector $(Y^j, X^1, \ldots, X^{n_X})$. This requires the following three steps to be carried out:

1. Determine the effect of other variables $\left\{X^j,\ j \neq i\right\}$ on Y^j by linear regression (see [Linear Regression]); when the values of variable $\left\{X^j,\ j \neq i\right\}$ are known, the average forecast for the value of Y^j is then available in the form of the equation:

$$\widehat{Y^j} = \sum_{k \neq i,\ 1 \leq k \leq n_X} \widehat{a}_k X^k$$

2. Determine the effect of other variables $\left\{X^j,\ j \neq i\right\}$ on X^i by linear regression; when the values of variable $\left\{X^j,\ j \neq i\right\}$ are known, the average forecast for the value of Y^j is then available in the form of the equation:

$$\widehat{X}^i = \sum_{k \neq i,\ 1 \leq k \leq n_X} \widehat{b}_k X^k$$

3. PCC_{X^i,Y^j} is then equal to the Pearson's correlation coefficient $\widehat{\rho}_{Y^j-\widehat{Y^j},X^i-\widehat{X}^i}$ estimated for the variables $Y^j - \widehat{Y^j}$ and $X^i - \widehat{X}^i$ on the N-sample of simulations (see [Pearson's Coefficient]).

One can then class the n_X variables $X^1, \ldots, X^{n_X}$ according to the absolute value of the partial correlation coefficients: the higher the value of $\left|\mathrm{PCC}_{X^i,Y^j}\right|$, the greater the impact the variable X^i has on Y^j.

Other notations

-

Link with OpenTURNS methodology

After a propagation of uncertainty (step C) using [Standard Monte Carlo] simulation, a hierarchy of sources of uncertainty can be obtained Partial Pearson's Correlation Coefficients. In fact, the N simulations enable the pairs (y_1^j, x_1^i), (y_2^j, x_2^i),..., (y_N^j, x_N^i) to be generated, where:

- $\underline{X} = \left\{X^1, \ldots, X^n\right\}$ describes the input vector specified in step A "Specifying Criteria and the Case Study",
- Y^j describes the final variable of interest or output variable defined in the same step.

The results produced as output of this method are Pearson's partial correlation coefficients PCC_{X^i,Y^j}, that the user may use, taking absolute values, to order the variables X^i hierarchically.

References and theoretical basics

This method of hierarchical ordering is particularly useful:

- when the study of uncertainty is concerned with the central dispersion of the variable of interest Y^j and not with its extreme values,
- when the relationships between Y^j and each of the components of $\underline{X}$ are close to linear relationships (so that Pearson's correlation coefficient can be interpreted),
- when the number N of Monte-Carlo simulations is significantly higher than the number n_X of input random variables (it is preferable to have N/n_X at least greater than a factor of 10 so that the estimation of the n_X partial correlation coefficients provides a reasonable picture of reality).

Readers interested in the assumptions made for multiple linear regression models and in the tests needed to validate these assumptions are referred to [Linear Regression].
Other methods of uncertainty ranking can be applied after Monte-Carlo simulation, requiring a lesser number N of simulations or that can treat non-linear cases, are described in [Uncertainty Ranking using Pearson] , [Uncertainty ranking using Spearman] , and [Uncertainty Ranking using Spearman's Partial Correlation Coefficients].
The following references provide an interesting bibliographic starting point to further study of the method described here:

- Saltelli, A., Chan, K., Scott, M. (2000). "Sensitivity Analysis", John Wiley & Sons publishers, Probability and Statistics series
- J.C. Helton, F.J. Davis (2003). "Latin Hypercube sampling and the propagation of uncertainty analyses of complex systems". Reliability Engineering and System Safety 81, p.23-69
- J.P.C. Kleijnen, J.C. Helton (1999). "Statistical analyses of scatterplots to identify factors in large-scale simulations, part 1 : review and comparison of techniques". Reliability Engineering and System Safety 65, p.147-185

5.2.6 Step C' – Uncertainty Ranking using Partial Rank Correlation Coefficients

Mathematical description

Goal

This method is concerned with analyzing the influence the random vector $\underline{X} = \left(X^1, \ldots, X^{n_X}\right)$ has on the random variable Y^j which is being studied for uncertainty. Here we attempt to measure monotonic relationships that exist between Y^j and the different components X^i.

Principle

The basic method of hierarchical ordering using Spearman's coefficients (see [Uncertainty Ranking using Spearman]) is concerned with the case where the variable Y^j monotonically depends on n_X variables $\left\{X^1, \ldots, X^{n_X}\right\}$ but this can be misleading when statistical dependencies between the variables X^i exist. In such a situation, the partial rank correlation coefficients can be more useful in ordering the uncertainty hierarchically: the partial rank correlation coefficients PRCC_{X^i,Y^j} between the variables Y^j and X^i attempts to measure the residual influence of X^i on Y^j once influences from all other variables X^j have been eliminated.
The estimation for each partial rank correlation coefficient PRCC_{X^i,Y^j} uses a set made up of N values $\left\{(y^j 1, x_1^1, \ldots, x_1^{n_X}), \ldots, (y^j N, x_N^1, \ldots, x_N^{n_X})\right\}$ of the vector $(Y^j, X^1, \ldots, X^{n_X})$. This requires the following three steps to be carried out:

1. Determine the effect of other variables $\left\{X^j,\ j \neq i\right\}$ on Y^j by linear regression (see [Linear Regression]); when the values of variable $\left\{X^j,\ j \neq i\right\}$ are known, the average forecast for the value of Y^j is then available in the form of the equation:

$$\widehat{Y^j} = \sum_{k \neq i,\ 1 \leq k \leq n_X} \widehat{a}_k X^k$$

2. Determine the effect of other variables $\left\{X^j,\ j \neq i\right\}$ on X^i by linear regression; when the values of variable $\left\{X^j,\ j \neq i\right\}$ are known, the average forecast for the value of Y^j is then available in the form of the equation:

$$\widehat{X}^i = \sum_{k \neq i,\ 1 \leq k \leq n_X} \widehat{b}_k X^k$$

3. PRCC_{X^i,Y^j} is then equal to the Spearman's correlation coefficient $\widehat{\rho}^S_{Y^j - \widehat{Y^j}, X^i - \widehat{X}^i}$ estimated for the variables $Y^j - \widehat{Y^j}$ and $X^i - \widehat{X}^i$ on the N-sample of simulations (see [Spearman's Coefficient]).

One can then class the n_X variables $X^1, \ldots, X^{n_X}$ according to the absolute value of the partial rank correlation coefficients: the higher the value of $\left|\mathrm{PRCC}_{X^i,Y^j}\right|$, the greater the impact the variable X^i has on Y^j.

Other notations

-

Link with OpenTURNS methodology

After a propagation of uncertainty (step C) using [Standard Monte Carlo] simulation, a hierarchy of sources of uncertainty can be obtained Partial Rank Correlation Coefficients. In fact, the N simulations enable the pairs $(y^j 1, x^i_1)$, $(y^j 2, x^i_2)$,..., $(y^j N, x^i_N)$ to be generated, where:

- $\underline{X} = \left\{X^1, \ldots, X^n\right\}$ describes the input vector specified in step A “Specifying Criteria and the Case Study",
- Y^j describes the final variable of interest or output variable defined in the same step.

The results produced as output of this method are partial rank correlation coefficients PRCC_{X^i,Y^j}, that the user may use, taking absolu

References and theoretical basics

This method of hierarchical ordering is particularly useful:

- when the study of uncertainty is concerned with the central dispersion of the variable of interest Y^j and not with its extreme values,
- when the relationships between Y^j and each of the components of $\underline{X}$ are monotonic relationships (so that Spearman's correlation coefficient can be interpreted),
- when the number N of Monte-Carlo simulations is significantly higher than the number n_X of input random variables (it is preferable to have N/n_X at least greater than a factor of 10 so that the estimation of the n_X partial rank correlation coefficients provides a reasonable picture of reality).

Readers interested in the assumptions made for multiple linear regression models and in the tests needed to validate these assumptions are referred to [Linear Regression].
Other methods of uncertainty ranking can be applied after Monte-Carlo simulation, requiring a lesser number N of simulations, are described in [Uncertainty Ranking using Pearson] , [Uncertainty ranking using Spearman].
The following references provide an interesting bibliographic starting point to further study of the method described here:

- Saltelli, A., Chan, K., Scott, M. (2000). “Sensitivity Analysis", John Wiley & Sons publishers, Probability and Statistics series
- J.C. Helton, F.J. Davis (2003). “Latin Hypercube sampling and the propagation of uncertainty analyses of complex systems". Reliability Engineering and System Safety 81, p.23-69
- J.P.C. Kleijnen, J.C. Helton (1999). “Statistical analyses of scatterplots to identify factors in large-scale simulations, part 1 : review and comparison of techniques". Reliability Engineering and System Safety 65, p.147-185

5.2.7 Step C' – Importance Factors from FORM-SORM methods

Mathematical description

Goal

Importance Factors are evaluated in the following context : $\underline{X}$ denotes a random input vector, representing the sources of uncertainties, $f_{\underline{X}}(\underline{x})$ its joint density probability, $\underline{d}$ a determinist vector, representing the fixed variables $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \le 0\}$ the event considered here and $g(\underline{X}\,,\,\underline{d}) = 0$ its boundary (also called limit state surface).
The probability content of the event $\mathcal{D}_f$ is P_f:

$$P_f = \int_{g(\underline{X}\,,\,\underline{d})\le 0} f_{\underline{X}}(\underline{x})\, d\underline{x}. \tag{18}$$

In this context, the probability P_f can often be efficiently estimated by FORM or SORM approximations (refer to [FORM] and [SORM]).

The FORM importance factors offer a way to rank the importance of the input components with respect the realization of the event. They are often interpreted also as indicators of the impact of modeling the input components as random variables rather than fixed values. The FORM importance factors are defined as follows.

Principle

The isoprobabilistic transformation used in the FORM and SORM approximation (refer to [Iso Probabilistic Transformation]) creates the vector of gaussian components $\boldsymbol{Y}$ as a result of step 1 : the probabilistic input vector $\boldsymbol{X}$ is mapped onto a probabilistic input vector $\boldsymbol{Y}$. We suppose here that $\boldsymbol{Y}$ is a gaussian random vector, centered and reduced.
The first step is $T_1 : \mathbb{R}^n \rightarrow \mathbb{R}^n$:

$$\boldsymbol{Y} = T_1(\boldsymbol{X}) = \left(\begin{array}{c} \Phi^{-1}(F_1(X_1)) \\ \Phi^{-1}(F_2(X_2)) \\ \vdots \\ \Phi^{-1}(F_n(X_n)) \end{array}\right). \tag{19}$$

where $\Phi(z) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z} \exp(-\frac{u^2}{2})\, du$.

The design point associated to the event considered in the $\boldsymbol{Y}$-space is noted $Y^* = (y_i^*)_i$.
The importance factor α_i^2 of the variable X_i is defined as the square of the co-factor of the design point in the $\boldsymbol{Y}$-space :

$$\alpha_i^2 = \frac{(y_i^*)^2}{||Y^*||^2}$$

This definition guarantees the relation : $\Sigma_i \alpha_i^2 = 1$.

Other notations

Here, the event considered is explicited directly from the limit state function $g(\underline{X}\,,\,\underline{d})$: this is the classical

structural reliability formulation.
However, if the event is a threshold exceedance, it is useful to explicite the variable of interest $Z = \tilde{g}(\underline{X}\,,\,\underline{d})$, evaluated from the model $\tilde{g}(.)$. In that case, the event considered, associated to the threshold z_s has the formulation : $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, Z = \tilde{g}(\underline{X}\,,\,\underline{d}) > z_s\}$ and the limit state function is : $g(\underline{X}\,,\,\underline{d}) = z_s - Z = z_s - \tilde{g}(\underline{X}\,,\,\underline{d})$. P_f is the threshold exceedance probability, defined as : $P_f = P(Z \geq z_s) = \int_{g(\underline{X}\,,\,\underline{d}) \leq 0} f_{\underline{X}}(\underline{x})\, d\underline{x}$.
Thus, the FORM importance factors offer a way to rank the importance of the input components with respect to the threshold exceedance by the quantity of interest Z. They can be seen as a specific sensitivity analysis technique dedicated to the quantity Z around a particular threshlod rather than to its variance.

Link with OpenTURNS methodology

Within the global methodology, these importance factors are used in the step C': "Ranking sources of uncertainty" in the case of the evaluation of the probability of an event by an approximation method.
It requires to have fulfilled the following steps beforehand:

- step A: identify of an input vector $\underline{X}$ of sources of uncertainties and an output variable of interest $Z = \tilde{g}(\underline{X}, \underline{d})$, result of the model $\tilde{g}()$; identify a probabilistic criteria such as a threshold exceedance $Z > z_s$ or equivalently a failure event $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: identify one of the proposed techniques to estimate a probabilistic model of the input vector $\underline{X}$,
- step C: select an appropriate optimisation algorithm among those proposed to evaluate the event probability : FORM or SORM.

Note that the relevance of FORM importance factors as a means to rank the importance of the sources of uncertainty is closely dependant on the validity of FORM approximation (refer to [FORM] and [SORM]).

The sensitivity factors (refer to [Sensitivity Factors]) indicate the importance on the Hasofer-Lind reliability index (refer to [Reliability Index]) of the value of the parameters used to define the distribution of the random vector $\underline{X}$.

References and theoretical basics

Interesting litterature on the subject is :

- H.O. Madsen, "Omission Sensitivity Factors," 1988, Structural Safety, 5, 35-45.

Examples

Let's apply this method to the following analytical example which considers a cantilever beam, of Young's modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E, F, L, I).
The input random vector is $\underline{X} = (E, F, L, I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{ \begin{array}{lcl} E & = & Normal(50, 1) \\ F & = & Normal(1, 1) \\ L & = & Normal(10, 1) \\ I & = & Normal(5, 1) \end{array} \right.$$

The four random variables are independant.
The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E, F, L, I) \in \mathbb{R}^4 \, / \, y(E, F, L, I) \geq 3\}$.
The importance factors obtained are :

$$\left\{ \begin{array}{lcl} \alpha_E^2 & = & 9.456e^{-2}\,\% \\ \alpha_F^2 & = & 6.959e^{+1}\,\% \\ \alpha_L^2 & = & 1.948e^{+1}\,\% \\ \alpha_I^2 & = & 1.084e^{+1}\,\% \end{array} \right.$$

5.2.8 Step C' – Sensitivity Factors from FORM method

Mathematical description

Goal

Sensitivity Factors are evaluated under the following context : $\underline{X}$ denotes a random input vector, representing the sources of uncertainties, $f_{\underline{X}}(\underline{x})$ its joint density probability, $\underline{d}$ a determinist vector, representing the fixed variables $g(\underline{X}\,,\,\underline{d})$ the limit state function of the model, $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, g(\underline{X}\,,\,\underline{d}) \le 0\}$ the event considered here and $g(\underline{X}\,,\,\underline{d}) = 0$ its boundary (also called limit state surface).
The probability content of the event $\mathcal{D}_f$ is P_f:

$$P_f = \int_{g(\underline{X}\,,\,\underline{d})\le 0} f_{\underline{X}}(\underline{x})\, d\underline{x}. \tag{20}$$

In this context, the probability P_f can often be efficiently estimated by FORM or SORM approximations (refer to [FORM] and [SORM]).

The FORM importance factors offer a way to analyse the sensitivity of the probability the realization of the event with respect to the parameters of the probability distribution of $\underline{\mathrm{X}}$.

Principle

A sensitivity factor is defined as the derivative of the Hasofer-Lind reliability index with respect to the paramater θ. The paramater θ is a parameter in a distribution of the random vector $\underline{X}$.

If $\underline{\theta}$ represents the vector of all the parameters of the distribution of $\underline{X}$ which appear in the definition of the isoprobabilistic transformation T (refer to [IsoProbabiliticFunction]), and $U^*_{\underline{\theta}}$ the design point associated to the event considered in the U-space, and if the mapping of the limit state function by the T is noted $G(\underline{U}\,,\,\underline{\theta}) = g[T^{-1}(\underline{U}\,,\,\underline{\theta}), \underline{d}]$, then the sensitivity factors vector is defined as :

$$\nabla_{\underline{\theta}}\beta_{HL} = +\frac{1}{||\nabla_{\underline{\theta}}G(U^*_{\underline{\theta}}, \underline{d})||}\nabla_{\underline{u}}G(U^*_{\underline{\theta}}, \underline{d}).$$

The sensitivity factors indicate the importance on the Hasofer-Lind reliability index (refer to [Reliability Index]) of the value of the parameters used to define the distribution of the random vector $\underline{X}$.

Other notations

Here, the event considered is explicited directly from the limit state function $g(\underline{X}\,,\,\underline{d})$: this is the classical structural reliability formulation.
However, if the event is a threshold exceedance, it is useful to explicite the variable of interest $Z = \tilde{g}(\underline{X}\,,\,\underline{d})$, evaluated from the model $\tilde{g}(.)$. In that case, the event considered, associated to the threshold z_s has the formulation : $\mathcal{D}_f = \{\underline{X} \in \mathbb{R}^n \,/\, Z = \tilde{g}(\underline{X}\,,\,\underline{d}) > z_s\}$ and the limit state function is : $g(\underline{X}\,,\,\underline{d}) = z_s - Z = z_s - \tilde{g}(\underline{X}\,,\,\underline{d})$. P_f is the threshold exceedance probability, defined as : $P_f = P(Z \ge z_s) = \int_{g(\underline{X}\,,\,\underline{d})\le 0} f_{\underline{X}}(\underline{x})\, d\underline{x}$. Thus, the FORM sensitivity factors offer a way to rank the importance of the parameters of the input components with respect to the threshold exceedance by the quantity of interest Z. They can be seen as a specific sensitity analysis technique dedicated to the quantity Z around a particular threshlod rather than to its variance.

Link with OpenTURNS methodology

Within the global methodology, sensitivity factors are evaluated in the step C': "Ranking sources of uncertainty" in the case of the evaluation of the probability of an event by an approximation method.
It requires to have fulfilled before the following steps:

- step A: input vector $\underline{X}$, final variable of interest (result of a model), probabilistic criteria (the event considered) $g(\underline{X}\,,\,\underline{d}) \leq 0$,
- step B: one of the proposed techniques to describe the probabilistic modelisation of the input vector $\underline{X}$,
- step C: one method to evaluate the probability content of the event : the FORM or SORM approximation

References and theoretical basics

The standard version of Open TURNS takes into account only the sensitivity with respect to the parameters of the distributino of $\underline{X}$ which appear in the definition of the isoprobabilistic transformation T. It does not calculate the sensitivity with respect to the other parameters, in particular those of the limite state function $\underline{d}$.

The FORM importance factors (refer to [Importance Factors]) offer a way to rank the importance of the input components with respect the realization of the event. They are often interpreted also as indicators of the impact of modeling the input components as random variables rather than fixed values.

Let's note some usefull references:

- O. Ditlevsen, H.O. Madsen, 2004, "Structural reliability methods," Department of mechanical engineering technical university of Denmark - Maritime engineering, internet publication.

Examples

Let's apply this method to the following analytical example which considers a cantilever beam, of Young's modulus E, length L, section modulus I. We apply a concentrated bending force at the other end of the beam. The vertical displacement y of the extrême end is equal to :

$$y(E, F, L, I) = \frac{FL^3}{3EI}$$

The objective is to propagate until y the uncertainties of the variables (E, F, L, I).
The input random vector is $\underline{X} = (E, F, L, I)$, which probabilistic modelisation is (unity is not precised):

$$\left\{\begin{array}{lcl} E & = & Normal(50,1) \\ F & = & Normal(1,1) \\ L & = & Normal(10,1) \\ I & = & Normal(5,1) \end{array}\right.$$

The event considered is the threshold exceedance : $\mathcal{D}_f = \{(E,F,L,I) \in \mathbb{R}^4 \,/\, y(E,F,L,I) \geq 3\}$.

If we note μ the mean and σ the standard deviation a the random variable, we obtain the following results, gathered in the following tables.

β_{HL}	μ	σ
E	0.0307508	-0.000954364
F	-0.834221	-0.000954364
L	-0.441319	-0.000954364
I	0.329191	-0.000954364

$P_{f,FORM}$	μ	σ
E	-0.00737194	0.000228791
F	0.199989	0.000228791
L	0.105798	0.000228791
I	-0.0789175	0.000228791

6 Response Surface

In some situations, the model h is too CPU-time consuming to enable the uncertainty analysis defined in step A. A possible approach to overcome this difficulty consists in replacing h with a “simpler" model $\widetilde{h}$, usually called response surface (or meta-model, surrogate model). Open TURNS offers several classical methods to build such a response surface.

6.1 Methods description

6.1.1 Step c' – Response Surface by Taylor Expansion

Mathematical description

Goal

In order to reduce computational costs, we use approximate functions instead of the initial function. When studying uncertainty management problems, one well-established class of method to deal with suitable approximations is the response surface method. The basic idea is to replace the initial model by an approximation, the so called response surface, whose function values can be computed more easily. Hence, there are two steps:

- **Step n°1** Choice of the type of response surface (e.g. polynom,...) caracterized by a set of parameters /degrees of freedom,
- **Step n°2** Estimation of the parameters of the response surface by a finite number of computations.

Within this file, we are dealing with the step n°1. In this case, we describe the response surface by Taylor expansion. The initial model is thus replaced by a polynomial expansion: h becomes p^{Taylor} and $\underline{y}$ becomes $\underline{y}^{Taylor}$. If the response surface is used for the same uncertainty problem, the criterion will be applied not on $\underline{y}$ but on $\underline{y}^{Taylor}$.

Principles

We give the first order and second order Taylor expansions around $\underline{x} = \underline{x}_C$.
First order Taylor Expansion

$$\underline{z}^T = p_1^T(\underline{x}) = h(\underline{x}_C) + \sum_{i=1}^{n_X} \frac{\partial h}{\partial x^i}(\underline{x}_C).\left(x^i - x_C^i\right) + o(\|\underline{x} - \underline{x}_C\|)$$

Second Order Taylor Expansion

$$\underline{y}^T = p_2^T(\underline{x}) = h(\underline{x}_C) + \sum_{i=1}^{n_X} \frac{\partial h}{\partial x_i}(\underline{x}_C).\left(x^i - x_C^i\right) + \frac{1}{2}.\sum_{i,j=1}^{n_X} \frac{\partial^2 h}{\partial x^i \partial x^j}(\underline{x}_C).\left(x^i - x_C^i\right).\left(x^j - x_C^j\right) + o(\|\underline{x} - \underline{x}_C\|^2)$$

Vectorial writing

To synthetize these decompositions in a vectorial way, we can write

$$\underline{y^T} = \underline{y}_C + < \underline{\underline{L}}, \underline{x} - \underline{x}_C > + \frac{1}{2} << \underline{\underline{\underline{Q}}}, \underline{x} - \underline{x}_C >, \underline{x} - \underline{x}_C >$$

where:

- $\underline{y_0}$ is a constant vector,
- $\underline{x}$ is the vector of the input variables,

- $\underline{x_C}$ is a recentring vector dedicated to increase the numerical accuracy,
- $\underline{\underline{L}} = \left(\frac{\partial y^j}{\partial x^i} \right)_{i=1,\ldots,n_Y, \quad j=1,\ldots,n_X}$ is the transposed Jacobian matrix,
- $\underline{\underline{\underline{Q}}} = \left(\frac{\partial^2 y^j}{\partial x^i \partial x^k} \right)_{i=1,\ldots,n_Y, \quad j,k=1\cdots n_X}$ is the transposed hessian matrix,

Other notations

Link with OpenTURNS methodology

This method is used when one wants to build a surface response (before starting the step A). A Taylor expansion polynom is well fitted when one wants to replace 'locally' the model of interest. It means that the model is replaced by the approximate model in a restricted domain of the input variables. One has to pay attention that this is a strong assumption: this approximate model could behave differently from the initial one and thus induce different results towards the criterion which is studied. The accuracy is degraded by this approximation and is usually valid only in a small region of interest. To compute the probability of exceedance of a threshold, the quality of this approximation by Taylor response surface has to be strongly justified. Anyway, to study central tendencies, it can be very useful.

References and theoretical basics

This method is valid in the neighbourhood of $\underline{x_C}$ or in a larger domain but this has to be justified.

6.1.2 Step c' – Response Surface by polynome

Mathematical description

Goal

In order to reduce computational costs, we use approximate functions instead of the initial function. When studying uncertainty management problems, one well-established class of method to deal with suitable approximations is the response surface method. The basic idea is to replace the initial model by an approximation, the so called response surface, whose function values can be computed more easily. Hence, there are two steps:

- **Step n°1** Choice of the type of response surface (e.g. polynom,...) caracterized by a set of parameters /degrees of freedom,
- **Step n°2** Estimation of the parameters of the response surface by a finite number of computations.

Within this file, we are dealing with the step n°1. A classical family of functions is represented by polynomial families.

Principles of Polynomial Surface Responses

The initial model is noted h and links the input variables $\underline{x} = (x^1, \ldots, x^{n_x})$ with the output variables $\underline{z} = (z^1, \ldots, z^{n_z})$. To simplify the notations in this first part of the file, we consider that $n_z = 1$ and use z for z^1. The results obtained for a polynomial response surface in dimension $n_z = 1$ are given below.

Principles in dimension $n_z = 1$

$$z = h(\underline{x})$$

The approximate model q is parametrized by the vector $\underline{a} = (a^1, \cdots, a^{n_a})$ containing n_a coefficients. The approximate values are noted $\hat{z}$ such that:

$$\hat{z} = q\left(\underline{x}, \underline{a}\right)$$

If the response surface is linear in its parameters $\underline{a}$, that is to say if the response surface is defined such that:

$$q(\underline{x}, \underline{a}) = \sum_{i=1}^{n_a} a_i.\Psi_i(\underline{x})$$

As already mentioned, response surfaces are designed such, that a complex relation between the inputs and the outputs, is described by an appropriate, but as simple as possible mathematical model. The term 'simple' means in the context of response surfaces, that the model should be continuous in the basic variables and should have a small number of terms, whose coefficients can be easily estimated. Polynomial response surfaces represent a classic way of building response surfaces. Following the previous notations, it only means that $\Psi_i(\underline{x})$ is a polynom in $\underline{x}$. Different families of polynoms are considered within Open TURNS:

- **Linear Polynom**
  In this case, the family of polynoms which is considered is the following : $(\Psi_i(\underline{x}))_{i=1,\cdots,n_a} = \left(1, x^1, \dots, x^{n_X}\right)$
  $$q^L(\underline{x}) = a_1 + a_2.x^1 + \cdots + a_{n_x+1}.x^{n_X}$$
  The number of parameters to be determined is equal to: $n_a = 1 + n_X$

- **Quadratic Polynoms with cross terms**
  In this case, the family of polynoms which is considered is the following : $(\Psi_i(\underline{x}))_{i=1,\cdots,n_a} = \left(1, x^1, \dots, x^{n_X}, (x^1)^2, \dots, (x^{n_X})^2, x^1.x^2, \dots, x^{n_X-1}.x^{n_X}\right)$
  $$q^{Q_2}(\underline{x}) = a_1+a_2.x^1+\cdots+a^{n_x+1}.x^{n_x}+a_{n_X+2}.(x^1)^2+\cdots+a_{2.n_X+1}.(x^{n_X})^2+\sum_{i,j=1\ ,\ i<j}^{n_X} a_{2.n_X+1+i+j}\,x^i.x^j$$
  The number of parameters to be determined is equal to: $n_a = 1 + n_X + n_X + \frac{n_X.(n_X-1)}{2}$

The coefficients $(\underline{a})$ can be obtained by a Least Square method for example (see []) from a sample of N runs obtained with the initial models: $(\underline{x}_k)_{k=1,\dots,N} \longrightarrow (\underline{z}_k)_{k=1,\dots,N}$.

Principles in dimension $n_z \geq 1$

The surface response is built for each dimension. We obtain the following vectorial writing:

$$\hat{\underline{z}} = \underline{z}_0 + < \underline{\underline{M}}, \underline{x} - \underline{x}_C > + << \underline{\underline{\underline{Q}}}, \underline{x} - \underline{x}_C >, \underline{x} - \underline{x}_C >$$

where:

- $\underline{x}$ is the vector of input variables,
- $\underline{x}_C$ is a remapping vector,
- $< \underline{\underline{M}}, \underline{x} - \underline{x}_C >= \left(\sum_{j=1}^{n_x} M_{ij}.(x^j - x^j_C)\right)_{i=1,\dots,n_z}$,
- $<< \underline{\underline{\underline{Q}}}, \underline{x} - \underline{x}_C >, \underline{x} - \underline{x}_C >= \left(\sum_{i=1}^{n_x}\sum_{j=1}^{n_x} Q_{ijk}.(x^j - x^j_C).(x^k - x^k_C)\right)_{k=1,\dots,n_z}$,

Other notations

Link with OpenTURNS methodology

This method is used when one wants to build a surface response (before starting a new round from the step A for example). It requires a set of output values obtained with the initial model computed at different input values. It enables to create a new 'model' which could be used for the same purpose than the initial model or for other purposes. In any case, be careful when using this approximate model instead of the initial model: it could behave differently from the initial one and thus induce different results towards the criterion which is studied.

References and theoretical basics

The surface response built by this method is fully deterministic.
Link with other files from the documentation of reference : [Response Surface by Taylor decomposition], [Least Square method to build response surface]
Saltelli 'Sensitivity Analysis', Wiley

6.1.3 Step C' – Surface Response obtained by Least Square method

Mathematical description

Goal

In order to reduce computational costs, we use approximate functions instead of the initial function. When studying uncertainty management problems, one well-established class of method to deal with suitable approximations is the response surface method. The basic idea is to replace the initial model by an approximation, the so called response surface, whose function values can be computed more easily. Hence, there are two steps:

- **Step n°1** Choice of the type of response surface (e.g. polynom,...) caracterized by a set of parameters /degrees of freedom,
- **Step n°2** Estimation of the parameters of the response surface by a finite number of computations.

Within this file, we are dealing with the step n°2, using the response surface family obtained by the decomposition over a family of functions (polynomial families, ...). The coefficients of the decomposition are optimal in a certain way. The Least Square Method enables to define these "best" coefficients which minimize the quadratic error between the "true" output values of the model, computed on a finite set of input values, and the approximate output values obtained by the response surface on the same finite set of input values.

Principles

The initial model is noted h and links the input variables $\underline{x} = (x^1, \ldots, x^{n_X})$ with the output variables $\underline{z} = (z^1, \ldots, z^{n_Y})$. To simplify the notations in this first part of the file, we consider that $n_Y = 1$ (meaning that $\underline{z} = z^1$ is scalar in this file) and use z for z^1. The results obtained for a polynomial response surface in dimension $n_y > 1$ are given below.

Principles in dimension $n_z = 1$

$$z = h(\underline{x})$$

The approximate model q is parametrized by the vector $\underline{a} = (a^1, \cdots, a^{n_a})$ containing n_a coefficients. The approximate values are noted $\hat{z}$ such that:

$$\hat{z} = q\left(\underline{x}, \underline{a}\right)$$

We consider $\epsilon(\underline{x}, \underline{a})$ which measures the difference between the initial model h and the response surface model q at a given point $\underline{x}$.

$$\epsilon(\underline{x}, \underline{a}) = z - \hat{y} = h(\underline{x}) - q(\underline{x}, \underline{a})$$

We consider that N computations are realized to build the response surface. These N points of computations are noted:

$$\underline{x}_k = \left(x_k^1, \ldots, x_k^{n_x}\right), \quad k = 1, 2, \cdots, N$$

At each point of computation, it is possible to compute the error $\epsilon(\underline{x}_k, \underline{a})$. It is compiled in the following vector $(\underline{\epsilon})$:

$$\underline{\epsilon}(\underline{a}) = (h(\underline{x}_1) - q(\underline{x}_1, \underline{a}), \ldots, h(\underline{x}_N) - q(\underline{x}_N, \underline{a}))$$

The sum of squares of the differences between the value of the response surface $\hat{z}_k = q\left(\underline{x}_k, \underline{a}\right)$ and the values of the initial model $z_k = h(\underline{x}_k)$ at the N points of computations is equal to:

$$\|\underline{\epsilon}(\underline{a})\|_2^2 = \sum_{k=1}^{N} \left(z^k - \hat{z}^k\right)^2$$

The goal of the Least Square method is to minimize this function $({}^t\underline{\epsilon}.\underline{\epsilon})(\underline{a})$. The problem to be solved is thus the following:

$$\underline{\hat{a}} = \overset{Argmin}{\mathbb{R}^{n_a}} \left({}^t\epsilon.\epsilon\right)$$

Classical optimization methods can be used to obtain the optimal coefficients $\underline{\hat{a}}$.

Particular cases

Within Open TURNS, the response surfaces are considered to be linear in $\underline{a}$ and the decomposition is done on a polynomial basis.

Surface Responses:

If the response surface is linear in its parameters $\underline{a}$, that is to say if the response surface is defined such that:

$$q(\underline{x}, \underline{a}) = \sum_{i=1}^{n_a} a_i.\Psi_i(\underline{x})$$

the previous problem is much more simple and it is possible to show that:

$$\underline{\hat{a}} = \left({}^t\mathcal{Z}.\mathcal{Z}\right)^{-1}.{}^t\mathcal{Z}.\underline{z}$$

where $\underline{z} =^t (z^1, \ldots, z^N)$

$$\mathcal{Z} = \begin{pmatrix} \Psi_1(\underline{x}^1) & \cdots & \Psi_{n_a}(\underline{x}^1) \\ \vdots & & \vdots \\ \Psi_1(\underline{x}^N) & & \Psi_{n_a}(\underline{x}^N) \end{pmatrix}$$

Other notations

Link with OpenTURNS methodology

This method is used when one wants to build a surface response (before starting a new round from the step A for example). It requires a set of output values obtained with the initial model computed at different input values or the set of inputs and the initial model. It enables to create a new 'model' which could be used for the same purpose than the initial model or for other purposes. In any case, be careful when using this approximate model instead of the initial model: it could behave very differently from the initial one and thus induce very different results towards the criterion which is studied.

References and theoretical basics

The surface response built by this method is fully deterministic. The condition $rank(\mathcal{Z}^i) \geq n_a$ on the rank of each matrix $(\mathcal{Z}^i)_{i=1,\ldots,n_Y}$ has to be fulfilled, which induces a minimum number of computation $N \geq n_a$. A fruitful link towards the global approach can be established with the files: [Response Surface by Taylor], [Response Surface by Polynoms of order 1 or 2].
The following reference is a good introduction to the subject: Saltelli and Al., 'Sensitivity Analysis', Wiley.