Global sensitivity analysis for nested and multiscale modelling

Yann Caniou

N° d'ordre : D.U. : 2296
EDSPIC : 587

Université Blaise Pascal - Clermont II

École Doctorale
Sciences pour l'Ingénieur de Clermont-Ferrand

Thèse

présentée par

Yann Caniou
Ingénieur IFMA

le **29 novembre 2012**

en vue d'obtenir le grade de

Docteur d'Université
(Spécialité : Génie Mécanique)

Analyse de sensibilité globale pour les modèles de simulation imbriqués et multiéchelles

soutenue publiquement devant un jury composé de MM.

Pr. Bruno SUDRET	ETH Zürich	Directeur de thèse
Dr. Thierry YALAMAS	PHIMECA Engineering S.A., Paris	Encadrant de thèse
Pr. Zohra CHERFI	Univeristé de Technologie de Compiègne	Rapporteur
Dr. Bertrand IOOSS	EDF R&D MRI, Chatou	Rapporteur
Pr. Clémentine PRIEUR	Université Joseph Fourier, Grenoble	Examinateur
Dr. Nicolas GAYTON	IFMA / Institut Pascal, Clermont-Ferrand	Examinateur
Pr. Maurice LEMAIRE	IFMA / Institut Pascal, Clermont-Ferrand	Examinateur

Laboratoire de Mécanique et Ingénieries (LaMI)
Université Blaise Pascal et Institut Français de Mécanique Avancée

Global sensitivity analysis for nested and multiscale modelling

A thesis submitted by **Yann Caniou**
in partial fulfillment of the requirements for
the degree of doctor of philosophy

Blaise Pascal University – Clermont II

Clermont-Ferrand, France

defended publicly on November 29, 2012 in front of a jury made up of:

Pr. Bruno SUDRET	ETH Zürich	Supervisor
Dr. Thierry YALAMAS	PHIMECA Engineering S.A., Paris	Co-supervisor
Pr. Zohra CHERFI	Univeristé de Technologie de Compiègne	Reviewer
Dr. Bertrand IOOSS	EDF R&D MRI, Chatou	Reviewer
Pr. Clémentine PRIEUR	Université Joseph Fourier, Grenoble	Examiner
Dr. Nicolas GAYTON	IFMA / Institut Pascal, Clermont-Ferrand	Examiner
Pr. Maurice LEMAIRE	IFMA / Institut Pascal, Clermont-Ferrand	Examiner

Abstract

This thesis is a contribution to the nested modelling of complex systems. A global methodology to quantify uncertainties and their origins in a workflow composed of several models that can be intricately linked is proposed. This work is organized along three axes. First, the dependence structure of the model parameters induced by the nested modelling is rigorously described thanks to the copula theory. Then, two sensitivity analysis methods for models with correlated inputs are presented: one is based on the analysis of the model response distribution and the other one is based on the decomposition of the covariance. Finally, a framework inspired by the graph theory is proposed for the description of the imbrication of the models. The proposed methodology is applied to different industrial applications: a multiscale modelling of the mechanical properties of concrete by homogenization method and a multiphysics approach of the damage on the cylinder head of a diesel engine. The obtained results provide the practitioner with essential informations for a significant improvement of the performance of the structure.

Keywords: Global sensitivity analysis, correlation, copula theory, graph theory, nested modelling, multiscale modelling.

Résumé

Cette thèse est une contribution à la modélisation imbriquée de systèmes complexes. Elle propose une méthodologie globale pour quantifier les incertitudes et leurs origines dans une chaîne de calcul formée par plusieurs modèles pouvant être reliés les uns aux autres de façon complexe. Ce travail est organisé selon trois axes. D'abord, la structure de dépendance des paramètres du modèle, induite par la modélisation imbriquée, est modélisée de façon rigoureuse grâce à la théorie des copules. Puis, deux méthodes d'analyse de sensibilité adaptées aux modèles à paramètres d'entrée corrélés sont présentées : l'une est basée sur l'analyse de la distribution de la réponse du modèle, l'autre sur la décomposition de la covariance. Enfin, un cadre de travail inspiré de la théorie des graphes est proposé pour la description de l'imbrication des modèles. La méthodologie proposée est appliquée à des exemples industriels d'envergure : un modèle multiéchelles de calcul des propriétés mécaniques du béton par une méthode d'homogénéisation et un modèle multiphysique de calcul de dommage sur la culasse d'un moteur diesel. Les résultats obtenus fournissent des indications importantes pour une amélioration significative de la performance d'une structure.

Mots-clés: Analyse de sensibilité globale, corrélation, théorie des copules, théorie des graphes, modélisation imbriquée, modélisation multiéchelles.

Acknowledgements

Yann Caniou, March 13, 2012

Although my PhD officially started in September 2009 in Paris, the idea came up much earlier during my third year at the French Institute for Advanced Mechanics where I achieved research oriented courses in parallel at the Blaise Pascal University, Clermont-Ferrand. Then came two research internships during my International Year, both proposed and supervised by Dr. Jean-Marc Bourinet, one semester at the University of Arizona with Dr. Samy Missoum and one semester at Audi A.G. in Ingolstadt, Germany, with Dr. Lars Hinke and Dr. Paul Heuler. Once back in France, I achieved the last step of my curriculum with my graduation project in a local company named PHIMECA Engineering and founded by former IFMA student and professor with the objective to get funding for a PhD thesis. It seems that my work in optimization has paid off.

In the long list of persons to thank, I would like to start with the board of PHIMECA Engineering, Maurice Pendola, CEO, and Thierry Yalamas, Associate director, who hired me as a research engineer and offered me perfect conditions to carry out my research.

During these three years, Bruno Sudret has been my supervisor. At the beginning, we worked together in Paris and then I moved back to Clermont-Ferrand but we managed to have regular webmeetings from Paris and Zürich in the very last months. Bruno guided me in my research, gave me directions and advices, pushed me to develop my skills and improve my work and for all of this I would like to thank him gratefully. Once again I would like to thank Thierry Yalamas as my co-supervisor for directing my work to potential commercial applications. To end up with my supervision team, I also sincerely thank Pr. Maurice Lemaire for his wisdom and our scientific discussions.

Although a thesis is an individual work and experience, the work I carried out would never have been so far without my colleagues. The first of them that I would like to thank is Vincent. Since he started his PhD one year before I did start mine, he was always good advice and taught me many tips on Python and LaTeXprogramming. Among the team of coworkers and former coworkers, I also sincerely thank François, Gilles, Alexandre, Pierre, Emmanuel, Julien and Mathieu.

The contributions of a PhD thesis may also be measured by the quality of its industrial applications. For providing me the test cases to improve and demonstrate the scope of my

work, I would like to thank Marc Berveiller (EDF), Anthony Hähnel (Renault), Nicolas Gayton (IFMA) and Laurent Gauvrit (Radiall).

The level of a PhD work may also be assessed by the expertise of its jury. Consequently, I would like to thank Pr. Zohra Cherfi-Boulanger, Dr. Bertrand Iooss who carefully reviewed my manuscript and Pr. Clémentine Prieur, Dr. Nicolas Gayton and Pr. Maurice Lemaire for their positive comments as examiners.

Finally, although they did not contribute directly to this work, I must thank my parents and girlfriend, Emilie, for their unconditional support.

Contents

Introduction

"If I can see further than anyone else,
it is only because I am standing on the shoulders of giants."
Isaac Newton

Context

Modern engineering aims at designing increasingly complex structures. The diversity of the physical fields that are involved (mechanics, electronics, thermodynamics) and the costs policy lead the designer to build numerical models that mimic physical phenomena in the most rigorous way. Thanks to the recent improvements in the computing performance, one is able to address problems that are always bigger in dimension and precision. Nonetheless, differences may be observed between simulations and experiments.

Simulations are basically *mathematical representations* of the response of a model that is set by a collection of input parameters, *e.g.* dimensions, material properties, load cases, environment parameters, etc. The list of possible input parameters may be endless but for the sake of feasibility, one has to neglect the most probable insignificant parameters to only retain the leading ones. Therefore, the model may lack accuracy since minor phenomena, physics or interactions are not taken into account. Although the accuracy of the model may increase with the number of input parameters, one may not be able to describe the possible discrepancies between the model response and experimental observations.

Classical modelling of complex structures has to be pushed a step forward where the deterministic nature of parameters have to give the way to *probabilistic modelling*. The principle of this advanced framework is to consider that the exact value of a parameter is not known but can be described by a probability distribution with central tendency and dispersion. The arising issue consists in building the probabilistic model of the parameters, either from a set of observed data, or from the expertise and knowledge in the physical field of interest.

Over the last decades, probabilistic engineering has become an essential framework for whom may be concerned by the robustness of his applications. Engineers are now aware that more than the nominal value of the quantity, the interval (and its boundaries) it may belong to represents a crucial information to ensure the reliability of the design. Quantifying the uncertainties in the input parameters of a model is the first step for the

robust design of a structure. The second one is the *propagation* of these uncertainties through the model(s) in order to characterize the output possible variability.

Problem statement

The design of complex structures consists in the implementation of *virtual testing* platforms that allow one to aggregate design schemes, models or data at different steps of the project life. These different models may correspond to different scales of modelling, physical fields, components of the structure but they all are aimed at being related with each other in the form of an imbrication tree, output variables of one model being the input variables of a second one which also predict quantities for a third one, etc. The models form a *workflow* whose different scales are often related in multiple ways. In this context, the well-established uncertainty propagation methods cannot be applied directly.

Global sensitivity analysis aims at identifying and prioritizing the input variables of a model that contribute the most to the variability of its output. These methods are today well-established when dealing with a single model, especially when they are implemented in association with surrogate modelling techniques that reduce their computational costs. However, their application to nested models is neither direct, nor trivial since they are based on requirements that are not fulfilled by this type of modelling. In particular, the complex input-output relationships between models sharing input parameters involve correlation between the intermediate variables of the modelling. Thus, the notion of sensitivity for dependent input variables constitutes an open problem which is addressed in this work.

First of all, since no conventions have been established for the nested modelling of complex structures, a theoretical framework is required with the aim of easing uncertainty propagation in such models. The derived methodology has to meet the expectations of the decision maker who needs to quantify which design parameters are the most influent in the whole workflow and which level of modelling penalizes the most the global performance of the structure.

Objectives and outline of the thesis

The methodology for addressing global sensitivity analyses in nested models proposed in this work is designed to meet the following objectives:

(*i*) Achieving methodological advances in the field of uncertainty propagation using surrogate models (polynomial chaos expansions, HDMR, SVR, Kriging) in the presence of *dependent variables*. The copula theory will be used as a mathematical framework to describe the dependence that may appear between the outputs of nested models;

(*ii*) Defining sensitivity measures that are computable and interpretable in the context of input variables having a complex dependence structure, which is typical of nested

models;

(*iii*) Validating the proposed methodology on significant industrial applications.

These objectives are fulfilled through the six chapters of this thesis whose content is now detailed.

Chapter 1 first introduces the basics of the *probability theory* and the notations used in this thesis. The emphasis is put on the joint probabilistic modelling of random variables. After several types of correlation measures are defined, a mathematical framework for modelling the dependence structure of variables, namely the *copula theory*, is presented. For practical applications, methods for simulating and identifying copula functions are presented.

The statistical field of sensitivity analysis is introduced in **Chapter 2**. In its first section, correlation- and regression-based methods are introduced. Then the well-established ANOVA decomposition that divides the variance of the model response into shares attributable to the input variables is presented. Due to the usage restriction of these methods to models with independent inputs, techniques for models with correlated inputs are presented. After a short review on existing techniques, two major methods are detailed. The first approach is based on the impact of a variable on the entire distribution of the model response. The second method, referred to as ANCOVA, proposes to study the covariance of the model response with the contribution of the input parameters as a generalization of the variance decomposition. Finally, the generalization of sensitivity indices to models with correlated input parameters being a topic broadly studied in the recent literature, a discussion on several competing approaches is proposed.

Like many statistical techniques, sensitivity analyses require a large number of simulations which are often not affordable when one call to the considered model takes more than a few seconds. To circumvent the issue of computational cost, surrogate modelling techniques are presented in **Chapter 3**. Surrogate models are basically mathematical representations of the physical models built from a limited number of well-chosen design points and whose computing cost is way cheaper. In the first section, existing techniques, namely, the Support Vector regression, Gaussian processes (or *Kriging*) and the high-dimensional model representation, are presented. In the second section, the focus is put on a well-established technique referred to polynomial chaos expansions where the model response is decomposed onto a suitable polynomial basis.

In **Chapter 4**, methodologies to compute the indices from the two major sensitivity analysis methods for models with correlated inputs are proposed. They both use the polynomial chaos expansions to build the corresponding surrogate models. If the polynomial chaos expansions is only used as a response surface for the computation of the distribution-based indices, one benefits from the functional decomposition it offers to compute the so-called ANCOVA indices. A particular attention is given to the accuracy of the estimation procedure and the corresponding computational cost. Since the practitioner is not used to such indices, their interpretation is discussed to help making decisions.

This work aims at addressing sensitivity analysis for problems involving nested models. The sensitivity analysis part of the work is treated throughout the first four chapters. In **Chapter 5**, nested models are put in the framework of the *graph theory*. This mathematical field helps formalizing the complex relationships between the variables at the different scales of such a modelling; in particular the notion of kinship of one variable onto another is defined. A section is also dedicated to the computational issues implied by such a study. A coupling technique between two softwares, namely OpenTURNS for the probabilistic modelling and YACS for the physical nested modelling, is carried out.

Chapter 6 is devoted to industrial applications of the methodology. After validating the sensitivity analysis methods on numerical functions and academical mechanical problems from the literature, the global methodology is applied to two major industrial problems. The homogenization of concrete is first treated from a multiscale point of view. The sensitivity of the mechanical properties of concrete to those of its constituents and respective proportions are computed and results are discussed. Then the complex phenomena resulting from the fuel combustion in a diesel engine are modelled by a multiphysics approach with the aim of determining which design parameters are the most responsible for the damage in the cylinder head.

1

Probabilistic modelling

Contents

1.1 Introduction

Modern engineering has to take the uncertainty in the parameters of model into account in order to ensure the robustness of designed systems. These uncertainties are typically the dispersion of the properties of a material, the tolerances on one dimension of a part or the lack of knowledge on the load that is applied to the structure. Then, the uncertainties are propagated through a computational model so that its output is also represented by a random variable with possible post-processing such as analysis of dispersion, probability of exceeding a threshold, *etc.*

From a mathematical point of view, the input parameters of the model are represented by random variables that are associated to probability distributions. The probability distribution describes the very behaviour of a random variable, the range it belongs to and the probability it has to take one value more than another. In the case of a multiscale modelling of a structure, the main model is decomposed into submodels that may share input variables. Consequently their output variables are linked. Thus, the dependence structure of the probabilistic model has to be considered in order to define an accurate probabilistic description of the parameters.

This first chapter introduces the basics of probability theory and statistics that are necessary for the understanding of this thesis. It also defines the notations that will be used all along the document. For a complete overview on the topic, the reading of Saporta (2006) is recommended.

While the first section introduces the basics of probability and statistics, the second section presents tools that are useful for modelling the dependence between the input parameters of a computational model. Finally, an overview of the copula theory and a mathematical framework for the dependence modelling are proposed.

1.2 Probabilistic modelling,

1.2.1 Probability space

In probability theory, a probability space is a triplet $(\Omega, \mathcal{F}, \mathbb{P})$ composed of a sample set Ω, a σ-algebra $\mathcal{F}$ and a probability measure $\mathbb{P}$ such that $\mathbb{P} : \mathcal{F} \to [0, 1]$. More precisely:

- Ω is the set of all possible outcomes, (x);
- $\mathcal{F}$ is the set of events, where each event is a set containing 0 or more outcomes;
- Function $\mathbb{P}$ assigns a probability to the events.

1.2.2 Random variable, random vector

A real-valued random variable X is function that associates to any elementary event ω a real value:

$$X : \omega \to X(\omega) \in \mathrm{D}_X \subset \mathbb{R} \tag{1.1}$$

where ω is the elementary event in the space of all the possible outcomes Ω and where D_X is the support of X.

A real-valued random vector is a n-dimensional generalization of a real-valued random variable. As a real-valued random variable is a function X that associates to any elementary event a real value, a random vector is a function $\boldsymbol{X}$ that associates to any elementary event a real vector of $\mathbb{R}^n$:

$$\boldsymbol{X} : \omega \to \boldsymbol{X}(\omega) = [X_1(\omega), \ldots, X_n(\omega)]^{\mathsf{T}} \in \mathbb{R}^n \tag{1.2}$$

The applications $X_1, \ldots, X_n$ are random variables referred to as the *components* of the random vector $\boldsymbol{X}$. The following notation will be used : $\boldsymbol{X} = [X_1, \ldots, X_n]^{\mathsf{T}}$.

1.2.3 Probability distribution

A probability distribution is a function that describes the set of values a random variable can take and more specifically the probability for this value to belong to any measurable subset of its support D_X. In the case of real-valued random variables, the probability distribution is fully described by its cumulative distribution function.

1.2.3.1 Cumulative distribution fonction

The cumulative distribution function (or CDF) of a real-valued random variable X is the function F_X that associates to any real number x:

$$F_X(x) = \mathbb{P}\left(X \leqslant x\right) \tag{1.3}$$

where the right member of the equation represents the probability that the variable takes a value smaller than or equal to x. The probability that X belongs to the interval $]a, b]$, $a < b$ then reads:

$$\mathbb{P}\left(a < X \leqslant b\right) = F_X(b) - F_X(a). \tag{1.4}$$

1.2.3.2 Probability density fonction

The probability density function (or PDF) is a function that enables one to write the distribution function with integrals. The PDF of a real-valued random variable is defined as follows.

Definition 1 *A function f is the probability density function of a real-valued random variable X if, for any real number x:*

$$\mathbb{P}(X \leqslant x) = \int_{\inf \mathrm{D}_X}^{x} f(u)\,\mathrm{d}u. \tag{1.5}$$

where $\inf \mathrm{D}_X$ *is the lower bound of the support of X, possibly equal to $-\infty$.*

Consequently, for any $a, b \in \mathrm{D}_X$, the probability $\mathbb{P}(a < X \leqslant b)$ is given by:

$$\mathbb{P}(a < X \leqslant b) = \int_{a}^{b} f(u)\,\mathrm{d}u. \tag{1.6}$$

By plotting the graphical representation of the probability density function, the probability $\mathbb{P}(a < X \leqslant b)$ is given by the area under the curve on the interval $[a, b]$. As a consequence, the cumulative distribution function F_X of X is continuous and $\mathbb{P}(X = a) = 0$ for all real number a.

1.2.4 Joint distribution function

The following definitions are given for two random variables for the sake of simplicity but they are generalizable for any higher dimension $n \in \mathbb{N}$.

Let us consider two real-valued random variables X_1 and X_2. The probability that X_1 takes a numerical value smaller than x_1 and that X_2 takes a numerical value smaller than x_2 defines the joint cumulative distribution function:

$$F_{X_1,X_2}(x_1, x_2) = \mathbb{P}(X_1 \leqslant x_1, X_2 \leqslant x_2) \tag{1.7}$$

This function is increasing with x_1 (respectively x_2) between 0 when both variables respectively tend towards $\inf \mathrm{D}_{X_1}$ and $\inf \mathrm{D}_{X_2}$, and 1 when both variables tend towards $\sup \mathrm{D}_{X_1}$ and $\sup \mathrm{D}_{X_2}$ (in the case of continuous random variables). The joint probability density function is obtained by partial differentiation:

$$f_{X_1,X_2}(x_1, x_2) = \frac{\partial F_{X_1,X_2}(x_1, x_2)}{\partial x_1 \partial x_2} \tag{1.8}$$

1.2.4.1 Marginal distribution

Let us consider two random variables X_1 and X_2 with joint probability density function f_{X_1,X_2}. The marginal distribution of X_1 is the probability distribution of X_1 ignoring the information on X_2. The marginal probability density function of X_1 (resp. X_2) is obtained by integrating the joint distribution function along x_2 (resp. x_1):

$$f_{X_1}(x_1) = \int_{\mathrm{D}_{X_2}} f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_2 \quad , \quad f_{X_2}(x_2) = \int_{\mathrm{D}_{X_1}} f_{X_1,X_2}(x_1, x_2)\,\mathrm{d}x_1 \tag{1.9}$$

Two random variables X_1 and X_2 are independent if and only if:

$$f_{X_1,X_2}(x_1, x_2) = f_{X_1}(x_1) f_{X_2}(x_2) \tag{1.10}$$

1.2.4.2 Conditional distribution

The conditional probability density function of X_1 given X_2, is defined by the ratio between the joint probability density function and the marginal probability density function of X_2:

$$f_{X_1|X_2}(x_1, x_2) = \frac{f_{X_1,X_2}(x_1, x_2)}{f_{X_2}(x_2)} = \frac{f_{X_1,X_2}(x_1, x_2)}{\int_{\mathrm{D}_{X_1}} f_{X_1,X_2}(x_1, x_2)\, dx_1} \tag{1.11}$$

1.2.5 Moments of a random variable

Apart from its probability distribution function, which can be given, approximated or unknown, a real-valued random variable can be described by its moments.

1.2.5.1 Mathematical expectation

The mathematical expectation (or expected value or first order moment) $\mathrm{E}[X]$ of a random variable X is the weighted average of all the possible values X can take. For continuous real-valued random variables, the mathematical expectation, also denoted by the greek letter μ, reads:

$$\mathrm{E}[X] = \mu_X = \int_{\mathrm{D}_X} x f_X(x)\, \mathrm{d}x \tag{1.12}$$

1.2.5.2 Variance and standard deviation

The variance of a real-valued random variable describes the dispersion of the possible values it can take. A high variance indicates that the range of possible values is wide whereas a low variance indicates that the realizations are more likely close to the expected value.

$$\mathrm{Var}[X] = \int_{\mathrm{D}_X} (x - \mu_X)^2 f_X(x)\, \mathrm{d}x = \mathrm{E}\left[(X - \mu_x)^2\right] \tag{1.13}$$

A random variable is more usually defined by its *standard deviation*, denoted by the greek letter σ, which is nothing but the square root of the variance $\sigma_X = \sqrt{\mathrm{Var}[X]}$. Another representation of the dispersion is given by the standard deviation normalized by the expected value, namely the *coefficient of variation* given by:

$$CV_X = \frac{\sigma_X}{\mu_X} \tag{1.14}$$

As μ_X and σ_X are homogeneous in dimension, CV_X can be expressed in terms of percentage.

A single realization of X is denoted by $x_0 \equiv X(\omega_0)$. A continuous random variable is defined by its cumulative distribution function. In this work, finite variance random variables $\mathrm{Var}[X] < \infty$ are considered. The corresponding Hilbert space is denoted by $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ when equipped with the inner product $\langle X, Y \rangle = \mathrm{E}[XY]$, where $\mathrm{E}[.]$ indicates the expected operator.

1.2.5.3 n-th order moments

In practice, μ_X and σ_X are mostly used but the third and fourth order moments can provide additional information concerning the shape of the probability density function of X. The general definition of the n-th order centered moment of a real-valued random variable reads:

$$\mu_X^n = \mathrm{E}\left[(X - \mu_X)^n\right] = \int_{\mathrm{D}_X} (x - \mu_X)^n f_X(x)\,\mathrm{d}x \tag{1.15}$$

Consequently, the normalized third and fourth order centered moments, namely the *skewness* coefficient δ_X and the *kurtosis* coefficient κ_X, are respectively defined by:

$$\delta_X = \frac{1}{{\sigma_X}^3} \int_{\mathrm{D}_X} (x - \mu_X)^3 f_X(x)\,\mathrm{d}x \tag{1.16}$$

and:

$$\kappa_X = \frac{1}{{\sigma_X}^4} \int_{\mathrm{D}_X} (x - \mu_X)^4 f_X(x)\,\mathrm{d}x \tag{1.17}$$

They respectively describe the asymmetry and the peakedness of the probability density function of X.

1.2.5.4 Gaussian variables

The normal distribution is one of the main distribution in the probability theory. It has been brought out by Karl Friedrich Gauss in the 19^{th} century with the aim of modelling biometric parameters. Thus, it is also referred to as the Gaussian distribution. The probability density function of a Gaussian random variable X with mean μ and variance σ^2 is denoted by $X \sim \mathcal{N}(\mu, \sigma)$ and reads:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] \tag{1.18}$$

A standard Gaussian variable $X \sim \mathcal{N}(0, 1)$ is a Gaussian variable with zero mean and unity variance. The probability density function, denoted by $\varphi(x)$, draws a symmetric bell curve (or Gaussian curve) centered on 0. Its cumulative distribution function $\Phi(x)$ reads:

$$\Phi(x) = \int_{-\infty}^{x} \varphi(u)du = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) du \tag{1.19}$$

$\Phi(x)$ cannot be expressed in termes of usual functions. Actually, $\Phi(x)$ is a usual function, unavoidable for anyone who deals with probabililty and statistics. It can written using the *error function* erf:

$$\Phi(x) = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right). \tag{1.20}$$

Approximated values of $\Phi(x)$ are generally given in tables. Most scientific computing packages such as Matlab, Python, R, etc. provide an implementation of this function.

This section has recalled the basics of the probability theory and introduced the mathematical notations for the probabilistic modelling. Before setting the concept of copulas, different measures of correlation are now presented.

1.3 Correlation

In statistics, the *dependence* describes any kind of relationship between two (or more) data sets. More generally, the dependence corresponds to any link between two random variables that are not independent. For example, the size and the weights of individuals in the human population, the risk of observing both habitation and vehicle damage due to a single natural disaster in insurance or the price of gold and the number of the *price of gold* keyword researches in Google are dependent data. The correlation refers to any specific relationship between two random variables. The correlation can be linear or non linear. The intensity of the link is measured through several correlation coefficients. The most common of these is the Bravais-Pearson correlation (or linear correlation) coefficient.

1.3.1 Linear correlation coefficient

The Bravais-Pearson correlation coefficient, denoted by ρ, is the most common correlation measure. The linear correlation coefficient ρ_{X_1,X_2} between two random variables X_1 and X_2 is defined as the covariance of these variables divided by the product of their standard deviations σ_{X_1} and σ_{X_2}, namely:

$$\rho_{X_1,X_2} = \frac{\mathrm{Cov}\left[X_1, X_2\right]}{\sigma_{X1}\sigma_{X_2}} \tag{1.21}$$

where:

$$\mathrm{Cov}\left[X_1, X_2\right] = \mathrm{E}\left[(X_1 - \mathrm{E}\left[X_1\right])(X_2 - \mathrm{E}\left[X_2\right])\right] \tag{1.22}$$

This coefficient is only defined if both random variables have finite and non zero standard deviation. According to the Cauchy-Bunyakovsky-Schwarz inequality given in Eq. (1.23), the linear correlation coefficient is smaller than 1 in absolute value since:

$$\left|\mathrm{Cov}\left[X_1, X_2\right]\right| \leqslant \left(\mathrm{E}\left[(X_1 - \mathrm{E}\left[X_1\right])\right]^2 \mathrm{E}\left[(X_2 - \mathrm{E}\left[X_2\right])\right]^2\right)^{\frac{1}{2}} \tag{1.23}$$

If there is an increasing linear relationship between X_1 and X_2, *i.e.* $X_2 = \alpha X_1 + \beta$, $\alpha > 0$, then $\rho(X_1, X_2) = 1$. On the contrary, if the variables have a perfect decreasing linear relationship, $\rho(X_1, X_2) = -1$. This last extreme case is sometimes referred to as *anti-correlation*. Finally, the linear correlation coefficient is 0 if X_1 and X_2 are independent. It is to be mentioned that the inverse proposition is not true because the linear correlation only evaluates linear relationships between variables. The correlation coefficient ρ indicates the strength of the link between the variables. As illustrated in Figure 1.1, the higher the absolute value of the linear correlation coefficient, the stronger the variables are linked.

Let us now consider a sample $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \mathcal{X}_2\}$ of dimension $n = 2$ and size N, *i.e.* $x_k = \left\{x_k^{(i)}, i = 1, \ldots, N\right\}$, $k = 1, 2$. An estimator of the linear correlation coefficient

between X_1 and X_2 is given by:

$$\hat{\rho}(X_1, X_2) = \frac{\sum_{i=1}^{N}(x_1^{(i)} - \bar{x}_1)(x_2^{(i)} - \bar{x}_2)}{\sqrt{\sum_{i=1}^{N}(x_1^{(i)} - \bar{x}_1)^2 \sum_{i=1}^{N}(x_2^{(i)} - \bar{x}_2)^2}} \tag{1.24}$$

where $\bar{x}_1$ and $\bar{x}_2$ are the sample means of $\mathcal{X}_1$ and $\mathcal{X}_2$ respectively. In practice, especially if N is small, the computed linear correlation coefficient might be non zero for independent variables.

It is always possible to compute the linear correlation coefficient between two variables with finite variance. However it must be taken into account that the latter only describes a linear relationship between the variables. In case of a non linear relationship, its value can be misinterpreted. In order to avoid erroneous analyses, more appropriate correlation coefficients such as the Spearman *rank correlation* coefficient or the Kendall *pair correlation* coefficient must be used instead.

1.3.2 Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient, also referred to as Spearman's rho (or Spearman's ρ), was introduced in Spearman (1904). It is a correlation coefficient no more based on the value of the individuals but on their rank in the bivariate sample. Therefore, it is non parametric because the joint distribution of the sample is not taken into account. It is denoted by ρ_S where the greek letter ρ is highlighted by the subscript S (for Spearman).

The Spearman's rho can be described as the Pearson's correlation coefficient between the ranks of the observations. Let us denote by $r_1^{(i)}$ and $r_2^{(i)}$ the rank of $x_1^{(i)}$ and $x_2^{(i)}$ in the sample $\boldsymbol{\mathcal{X}}$. An estimator of the Spearman's rho between X_1 and X_2 is given by:

$$\hat{\rho}_S(X_1, X_2) = \frac{\sum_{i=1}^{N}(r_1^{(i)} - \bar{r}_1)(r_2^{(i)} - \bar{r}_2)}{\sqrt{\sum_{i=1}^{N}(r_1^{(i)} - \bar{r}_1)^2 \sum_{i=1}^{N}(r_2^{(i)} - \bar{r}_2)^2}} \tag{1.25}$$

where the means of the ranks read $\bar{r}_k = \dfrac{N+1}{2}$, $k = 1, 2$. In case of tied (equal) numerical values, which may happen when a few number of decimals are used, the rank assigned to the two (or more) observations is the mean of their positions in the ascending order. In the absence of tied values in the sample, another expression of the Spearman's rho is given by:

$$\hat{\rho}_S(X_1, X_2) = 1 - \frac{6\sum_{i=1}^{N} {d_i}^2}{N(N^2 - 1)} \tag{1.26}$$

where $d_i = r_1^{(i)} - r_2^{(i)}$.

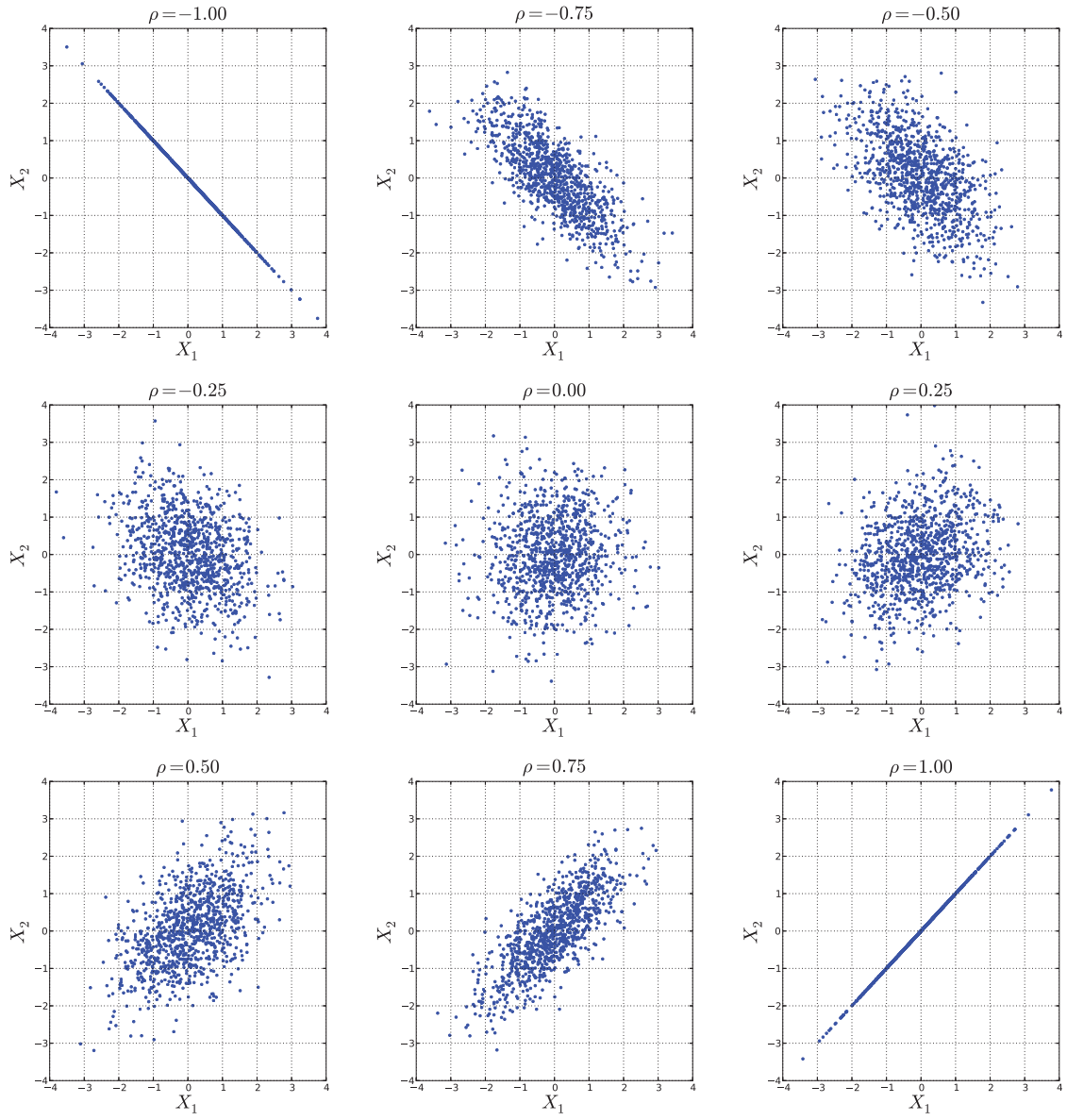


Figure 1.1: *Scatterplots for samples of size* $N = 1000$ *from standard normal random variables* X_1, X_2 *with increasing linear correlation coefficients.*

One advantage of the rank-based correlation coefficient is that it is no longer restricted to linear effect but handles any monotonic relationship between variables. Indeed, the sign of the Spearman's rho indicates the direction of the relationship between the variables and its value quantifies the strength of the link. According to Eqs. (1.25) and (1.26), if the second variable is a monotonic function m of the first one, thus $\rho_S(X_1, X_2 = m(X_1))$ is equal to 1 for an increasing function and -1 for a decreasing function.

1.3.3 Kendall's pair correlation coefficient

The Kendall's pair correlation coefficient, also referred to as Kendall's tau (or Kendall's τ) was introduced by Kendall (1955). It is a correlation coefficient based on pairwise relationships between the individuals of two samples. It is denoted by the greek letter τ.

Let us first consider two pairs of joint observations $(x_1^{(i)}, x_2^{(i)})$ and $(x_1^{(j)}, x_2^{(j)})$ from the same sample $\boldsymbol{\mathcal{X}}$ where i and j denotes the position of the observations in the sample. Two pairs are said *concordant* if and only if the order of both variables is the same, *i.e.* $x_1^{(i)} > x_1^{(j)}$ and $x_2^{(i)} > x_2^{(j)}$, or $x_1^{(i)} < x_1^{(j)}$ and $x_2^{(i)} < x_2^{(j)}$. In the opposite situation, *i.e.* $x_1^{(i)} > x_1^{(j)}$ and $x_2^{(i)} < x_2^{(j)}$, or $x_1^{(i)} < x_1^{(j)}$ and $x_2^{(i)} > x_2^{(j)}$, pairs are said *discordant*. An estimator of the Kendall's tau is given by the following equation:

$$\hat{\tau} = \frac{N_c - N_d}{\frac{1}{2}N(N-1)} \tag{1.27}$$

where N_c and N_d are respectively the numbers of concordant and discordant pairs. The denominator is equal to the total number of pairs that it is possible to compare. Consequently, the Kendall's tau belongs to the interval $[-1, 1]$. A Kendall's tau equal to 1 corresponds to a perfectly ordered bidimensional sample. On the contrary, a Kendall's tau equal to -1 indicates that the order of the second sample is the perfect reverse of the first one. For two independent variables, the Kendall's tau is close to 0.

In case of ties, *i.e.* $x_1^{(i)} = x_1^{(j)}$ and $x_2^{(i)} = x_2^{(j)}$ for some $i, j \in \{1, \ldots, N\}$, pairs are said to be neither concordant nor discordant. Variations of Kendall's tau allows one to treat the problem (Laurencelle, 2009):

1. τ_A: tied pairs count neither for N_c nor N_d;
2. τ_B: adjustments are made for the correlation coefficient:

$$\tau_B = \frac{N_c - N_d}{\sqrt{(N_0 - N_1)(N_0 - N_2)}} \tag{1.28}$$

where:

- $N_0 = N(N-1)/2$,
- $N_1 = \sum_{i=1}^{N} t_i(t_i - 1)/2$,
- $N_2 = \sum_{j=1}^{N} u_j(u_j - 1)/2$,

- t_i is the number of tied values in the i^{th} group of ties for the first parameter,
- u_j is the number of tied values in the j^{th} group of ties for the second parameter.

1.3.4 Correlation matrices

Several correlation coefficients have been defined for couples of variables. A most general framework to describe the relationship between two or more variables is to use *correlation matrices*. A correlation matrix $\boldsymbol{R}$ of n random variables is a $n \times n$ positive semi-definite matrix where $R_{i,j}$ is the correlation coefficient between the variables X_i and X_j. In the case of standard normal variables, the Bravais-Pearson correlation matrix corresponds the covariance matrix $\boldsymbol{\Sigma}$. The Bravais-Pearson linear correlation matrix $\boldsymbol{\rho}$ reads:

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho_{1,1} & \cdots & \rho_{1,n} \\ \rho_{1,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{n-1,n} \\ \rho_{n,1} & \cdots & \rho_{n,n-1} & 1 \end{bmatrix} \tag{1.29}$$

Similarly, the Spearman's rank correlation matrix $\boldsymbol{\rho_S}$ and the Kendall's pair correlation matrix $\boldsymbol{\tau}$ are respectively defined by:

$$\boldsymbol{\rho_S} = \begin{bmatrix} 1 & \rho_{S1,1} & \cdots & \rho_{S1,n} \\ \rho_{S1,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \rho_{Sn-1,n} \\ \rho_{Sn,1} & \cdots & \rho_{Sn,n-1} & 1 \end{bmatrix} \quad \boldsymbol{\tau} = \begin{bmatrix} 1 & \tau_{1,1} & \cdots & \tau_{1,n} \\ \tau_{1,1} & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tau_{n-1,n} \\ \tau_{n,1} & \cdots & \tau_{n,n-1} & 1 \end{bmatrix} \tag{1.30}$$

1.3.5 Association measures

An association measure, or measure of association, is a scalar function r which satisfies the following properties:

- r is defined for any couple of variables (X_1, X_2).
- $r(X_1, X_2) : \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \times \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) \mapsto [-1, 1]$.
- $r(X_1, X_2) = 0$ if X_1 and X_2 are independent.
- $r(X_1, X_1) = 1$ and $r(X_1, -X_1) = -1$.
- $r(X_1, X_2) = r\left(g(X_1), h(X_2)\right)$ for any strictly increasing functions g and h.

This last property shows that the linear correlation coefficient ρ is not an association measure. Indeed it does not satisfy the equality as illustrated by the example in Figure 1.2. For example, let us consider a pair of standard normally distributed random variables $(X_1,\, X_2) \sim \mathcal{N}(0, 1)$ with linear correlation coefficient $\rho = 0.52$. Let us now transform the

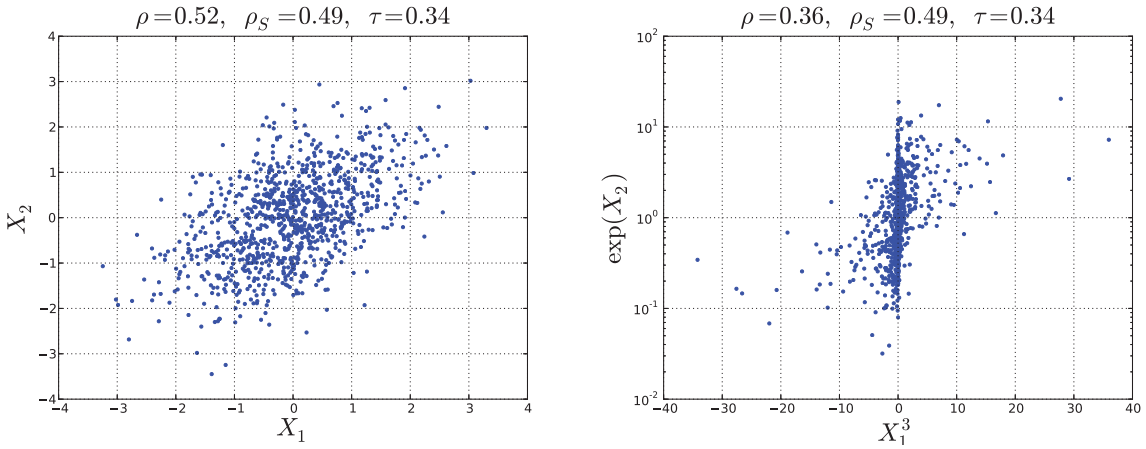


Figure 1.2: *Linear, rank and pairwise correlation coefficients for samples of size* $N = 1000$ *from couples of variables* (X_1, X_2) *(left) and* $(X_1^3, \exp(X_2))$ *(right).*

variables (X_1, X_2) using strictly increasing functions $g(X) = X^3$ and $h(X) = \exp(X)$. The transformed couple $(X_1^3, \exp(X_2))$ has a linear correlation coefficient $\rho = 0.35$ whereas both couples have exactly the same rank correlation coefficient $\rho_S \approx 0.49$ and Kendall's tau $\tau \approx 0.34$.

1.3.6 The Fisher transform

The Fisher transformation F was introduced in Fisher (1915). It is a powerful tool to estimate a confidence interval on correlation coefficients. Let us consider a normally distributed bivariate sample $\boldsymbol{\mathcal{X}} = \{\mathcal{X}_1, \mathcal{X}_2\}$ composed of N independent pairs $(X_1^{(i)}, X_2^{(i)})$ with $\rho\left(X_1^{(i)}, X_2^{(i)}\right) = \rho_0$. The Fisher transform F of the linear correlation coefficient ρ reads:

$$z = F(\rho) = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} = \operatorname{arctanh}(\rho) \tag{1.31}$$

In Fisher (1921), the author identifies the exact distribution of z for data from a bivariate normal sample: z is normally distributed with mean μ_z:

$$\mu_z = \frac{1}{2} \ln \frac{1+\rho}{1-\rho} \tag{1.32}$$

and standard deviation σ_z:

$$\sigma_z = \frac{1}{\sqrt{N-3}} \tag{1.33}$$

Then, a α confidence interval for ρ can be described by:

$$\mu_z - t_\alpha \frac{\sigma_z}{\sqrt{N}} \leqslant \operatorname{arctanh}(\rho) \leqslant \mu_z + t_\alpha \frac{\sigma_z}{\sqrt{N}} \tag{1.34}$$

and:

$$\tanh\left(\mu_z - t_\alpha \frac{\sigma_z}{\sqrt{N}}\right) \leqslant \rho \leqslant \tanh\left(\mu_z + t_\alpha \frac{\sigma_z}{\sqrt{N}}\right) \tag{1.35}$$

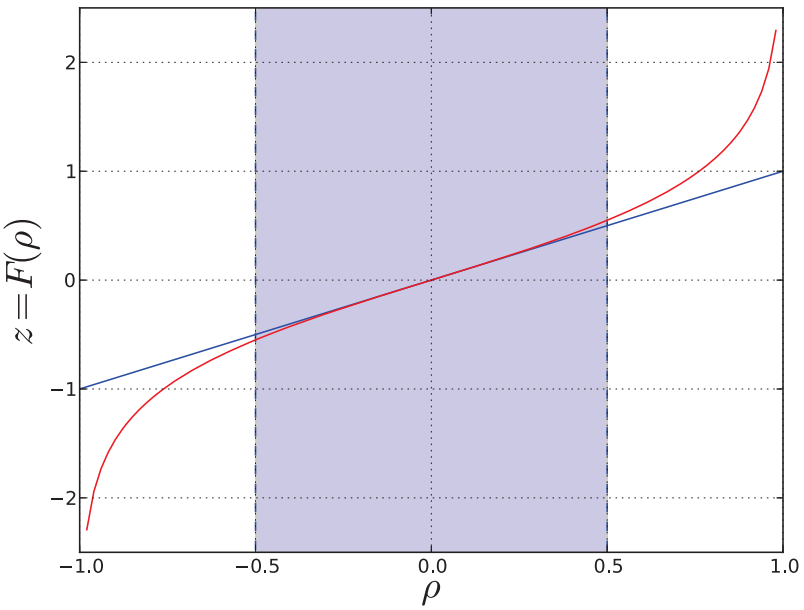


Figure 1.3: *The Fisher transformation of the linear correlation coefficient (red) is approximately the identity function (blue) for* $|\rho| \leqslant \frac{1}{2}$.

where $t_\alpha = \Phi(1 - \frac{1-\alpha}{2})$, that is the quantile of order α of the standard Gaussian distribution. Consequently, a 95% confidence interval for ρ is given by:

$$\left[\tanh\left(\mu_z - 1.96\frac{\sigma_z}{\sqrt{N}}\right) \ , \ \tanh\left(\mu_z + 1.96\frac{\sigma_z}{\sqrt{N}}\right)\right]. \tag{1.36}$$

where $-1.96 = \Phi^{-1}(0.025)$ and $1.96 = \Phi^{-1}(0.975)$ are approximately the 2.5% and 97.5% quantiles of the standard normal distribution $\mathcal{N}(0, 1)$.

The Fisher transformation is mainly related to the Bravais-Pearson linear correlation coefficient but the same transformation can also be used for the Spearman rank correlation coefficient using a few adjustments (Fieller et al., 1957; Choi, 1977). Let us first recall the Fisher transformation for the Spearman's rho:

$$F(\rho_S) = \frac{1}{2}\ln\left(\frac{1+\rho_S}{1-\rho_S}\right) = \operatorname{arctanh}(\rho_S) \tag{1.37}$$

Two variables can be defined in order to study the confidence interval of ρ_S. The first variable z defined by:

$$z = \sqrt{\frac{n-3}{1.06}}F(\rho_S) \tag{1.38}$$

follows a standard normal distribution under the null hypothesis of independence between variables X_1 and X_2. The second variable t defined by:

$$t = \rho_S\sqrt{\frac{n-1}{1-{\rho_S}^2}} \tag{1.39}$$

approximately follows a Student's t distribution with $n-2$ degrees of freedom under the null hypothesis.

1.4 Why should dependence be taken into account?

Through the following short example, the influence of dependence is illustrated. Let us consider a 2−dimensional model $Y = \mathcal{M}(X_1, X_2)$ defined by :

$$Y = X_1 + X_2 \tag{1.40}$$

with $X_i \sim \mathcal{U}(0,1)$, $i = 1, 2$. The probability that Y exceeds a threshold $t = 1.5$ is studied. In the sequel, $\mathbb{P}(Y > t)$ is also referred to as the *probability of failure* (with respect to exceeding threshold t).

Three correlation configurations are treated, namely $\rho_{1,2} = \{0.0, 0.6, -0.6\}$. In the first configuration, the input variables X_1 and X_2 are independent, *i.e.* X_2 is sampled regardless of the values of X_1. The probability that Y exceeds t estimated using $N = 10^3$ Monte Carlo simulations is $\widehat{\mathbb{P}}[Y > t] = 0.12$.

In the second configuration, suppose $\rho_{1,2} = 0.6$. The variables are positively correlated meaning that a strong value of X_1 implies a strong value of X_2. This relationship leads to a higher probability $\widehat{\mathbb{P}}[Y > t] = 0.21$. Finally, in the third case, the variables are assumed negatively correlated, namely $\rho_{1,2} = -0.6$. X_1 and X_2 are sampled in an opposite direction and this results in a lower probability $\widehat{\mathbb{P}}[Y > t] = 0.03$. Scatterplots of the three configurations are presented in Figure 1.4 for the sake of illustration.

It is to be noticed that in the second case, positive correlation is penalizing because in comparison with the independence case, it significantly increases the probability of failure. On the contrary, the negative correlation in the third case benefits to the probability of failure which is substantially reduced. From this simple example, it is shown that ignoring the dependence between input variables of a model can lead either to underestimating the risk or, on the contrary, to oversizing a structure and its related cost.

1.5 The copula theory

It is sometimes said that when the variables are not independent, it is very hard or even vain to try to identify their joint distribution function. In practice, *black-box* softwares are able to characterize the linear correlation between two variables so that the dependence structure of the data is approximated. Nevertheless, a suitable framework for modelling the dependence structure of data exists, namely the *copula theory*. The main concepts behind this theory are now introduced. For a complete overview on copula theory, the reading of the reference book by Nelsen (1999) is recommended.

After a brief history, the basics of copula theory are introduced in order to fit the engineering needs. Then several copulas are presented and methods to identify copulas from samples of data are described.

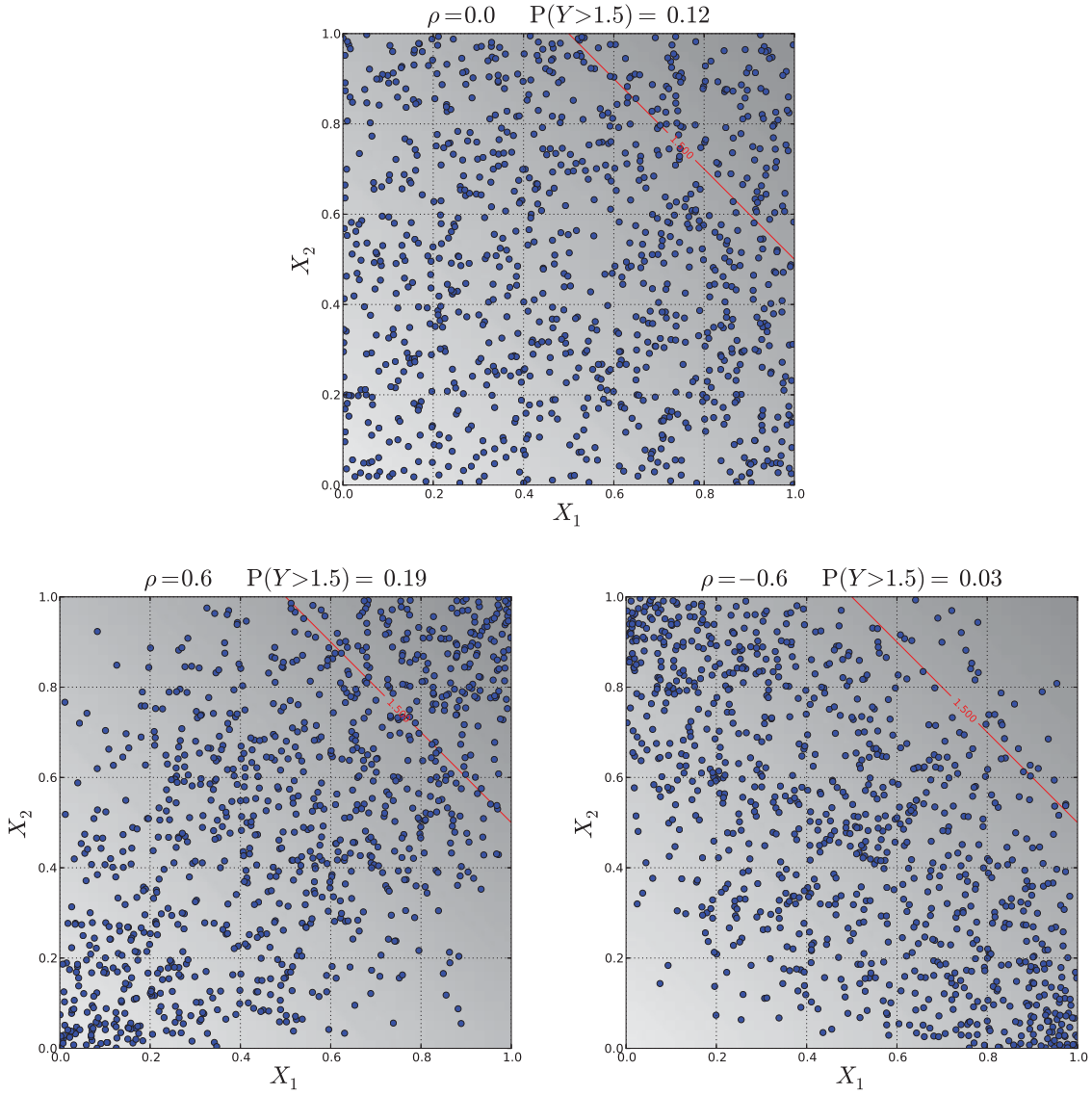


Figure 1.4: *Probability of exceeding a threshold in case of independence (top), positive correlation (bottom left) or negative correlation (bottom right) of the input variables.*

1.5.1 A brief history

Early works on the dependence measures with linear correlation and uniform marginal distribution functions defined on $\left[-\frac{1}{2}, \frac{1}{2}\right]$ are due to Hoeffding (1940). Except for the support of the marginal distribution functions, Hoeffding invented copulas which are defined in the modern way by means of uniform marginal distribution functions defined on $[0, 1]$. At the same time, Fréchet (1951) obtained similar and significant results that led to the *Fréchet bounds* and the *Fréchet classes*. For the recognition of both contributions, these objects are usually referred to as *Fréchet-Hoeffding bounds* and *classes*. The word

copula first appeared with a mathematical purpose in Sklar (1959), more specifically in the theorem currently named after him. This theorem describes the function that *couples* unidimensional distribution functions to build a multivariate distribution function.

The copula theory is exposed in a very comprehensive way in the book by Nelsen (1999). Since the early 2000s, copulas are famous because of the improvements they allow in finance, insurance, and more generally in risk analysis. This democratization in mathematics-related fields is due to Embrechts et al. (1999, 2001), among others. In his work, Embrechts shows the importance of modelling the dependence structure of financial model parameters and more particularly for their extreme behaviour. Finally, Fermanian (2005), Lambert (2007) and Genest et al. (2007) provide methodologies for the inference of copulas.

The use of copula theory for engineering applications has emerged a few years ago, especially for reliability and sensivity analysis thanks to the work of Lebrun and Dutfoy (2009a), Lebrun and Dutfoy (2009c), Lebrun and Dutfoy (2009b) dealing with the Nataf transformation. The most important steps in the brief history of copula are summarized in Table 1.1.

Year	Author	Improvements
1940	Hoeffding	Dependence measures, linear correlation, multivariate distributions with uniform marginals on $\left[-\frac{1}{2}, \frac{1}{2}\right]$.
1951	Fréchet	Multivariate distributions with fixed marginals.
1959	Sklar & Schweizer	Probabilistic metric spaces, first use of the term *copula*.
1999	Nelsen	Reference book on the copula theory.
1999	Embrechts	Risk engineering for finance and insurance.
2007	Genest	Inference methods for copulas.

Table 1.1: *Important dates in the history of copula theory.*

1.5.2 Definitions

A function C has to satisfy some properties in order to be defined as a copula. For the sake of simplicity, the following properties are given in two dimensions but they can be generalized for $n \in \mathbb{N}$.

Definition 2 *A copula is a function C defined on $[0,1]^2$ and verifying:*

1. $\forall (x,y) \in [0,1]^2$, $C(x,0) = C(0,y) = 0$,
2. $\forall (x,y) \in [0,1]^2$, $|C(x,y)| \leqslant M$,
3. $C(x,1) = x$ *and* $C(1,y) = y$,

4. *Let us consider a rectangle with summits* (a_1, b_1), (a_1, b_2), (a_2, b_2), (a_2, b_1) *verifying* $a_1 < a_2$ *and* $b_1 < b_2$, *then:*

$$C(a_1, b_1) - C(a_1, b_2) - C(a_2, b_1) + C(a_2, b_2) \geqslant 0 \tag{1.41}$$

The first property indicates that C is zero on the lower bounds of its domain of definition. The second property implies that C is bounded. The third property validates the uniformity of the marginals. The fourth property shows that C is a 2−increasing function, *i.e.* that the probability measure of any rectangle embedded in the unit square shall be positive. The properties are summarized in Figure 1.5.

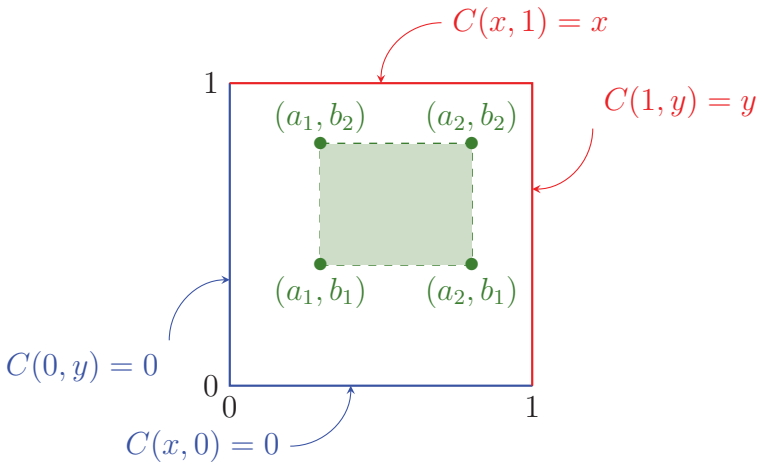


Figure 1.5: *Illustration of the copula properties.*

Let us consider two random variables U_1 and U_2 with uniform distributions on $[0, 1]$ and let C be the function defined by:

$$C(u_1, u_2) = \mathbb{P}\left(U_1 \leqslant u_1, U_2 \leqslant u_2\right), \forall (u_1, u_2) \in [0, 1]^2 \tag{1.42}$$

This function which has all the properties previously enounced is a copula. Reciprocally, a copula can be considered as what is left of a multivariate distribution once the influence of the marginals has been removed.

1.5.3 Properties

1.5.3.1 Sklar's theorem

Sklar's theorem is a fundamental property of the copula theory. This theorem draws the link between the joint distribution F of two random variables (X_1, X_2) and their marginal distributions F_1 and F_2.

Theorem 1 (Sklar's theorem) *Let $\boldsymbol{X} = [X_1, X_2]^T$ be a random vector with joint cumulative distribution function $F_{\boldsymbol{X}}$ and marginal distributions F_1 and F_2. There exists a $2-$dimensional copula C such that:*

$$\forall \boldsymbol{x} \in \mathbb{R}^2, F_{\boldsymbol{X}}(x_1, x_2) = C\left(F_1(x_1), F_2(x_2)\right). \tag{1.43}$$

If the marginal distributions F_1 and F_2 are continuous, then the copula C is unique and reads:

$$C(u_1, u_2) = F_{\boldsymbol{X}}\left(F_1^{-1}(u_1), F_2^{-1}(u_2)\right) \tag{1.44}$$

Otherwise, C is uniquely determined on $\mathrm{D}_{X_1} \times \mathrm{D}_{X_2}$, where D_{X_i} is the support of the marginal distribution F_i.

This theorem has been first proven in Sklar (1959). It illustrates how the term *copula* has been chosen to highlight the way the copula *couples* the marginal distributions to build the joint distribution.

1.5.3.2 Fréchet-Hoeffding bounds

Let us first introduce two important copulas M and W. They are respectively defined by:

$$M(x_1, x_2) = \min(x_1, x_2) \tag{1.45}$$

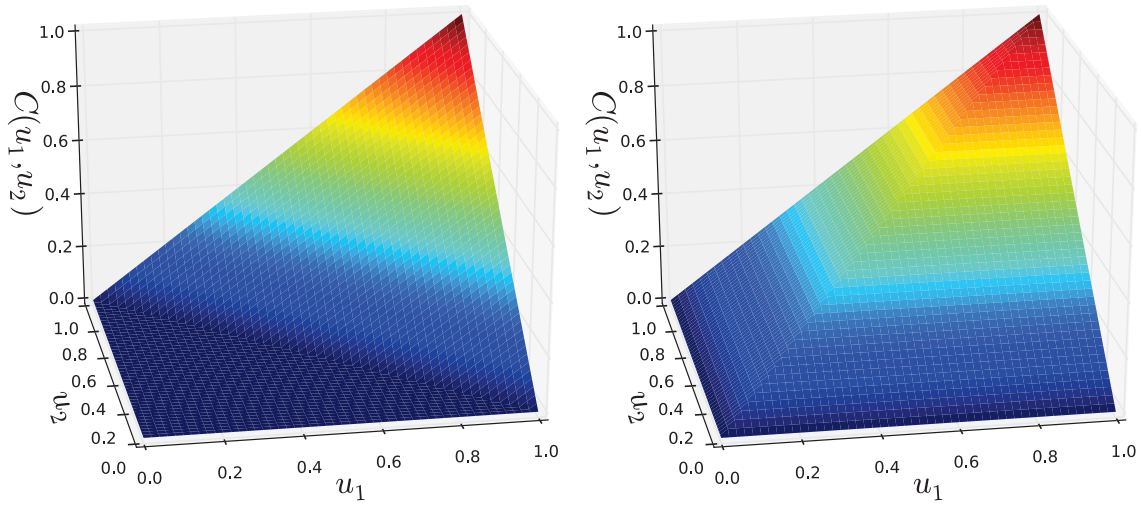
and:

$$W(x_1, x_2) = \max(x_1 + x_2 - 1, 0) \tag{1.46}$$

M and W, as pictured in Figure 1.6 are essential because they define the extremal values of all copulas. They are referred to as Fréchet-Hoeffding bounds. They verify the following property.

Property 1 *For any copula C and any vector $\boldsymbol{x} = [x_1, x_2]^T \in [0,1]^2$, C satisfies:*

$$W(x_1, x_2) \leqslant C(x_1, x_2) \leqslant M(x_1, x_2) \tag{1.47}$$



(a) $W(x_1, x_2) = \max(x_1 + x_2 - 1, 0)$ (b) $M(x_1, x_2) = \min(x_1, x_2)$

Figure 1.6: *Fréchet-Hoeffding bounds.*

1.5.3.3 Invariance theorem

A second important theorem of the copula theory is the invariance theorem.

Theorem 2 *Let X_1 and X_2 be two continuous random variables with respective marginal distributions F_1 and F_2 and copula C_{X_1,X_2}. If h_1 and h_2 are two strictly increasing functions on D_{X_1} and D_{X_2} respectively, then:*

$$C_{h_1(X_1),h_2(X_2)} = C_{X_1,X_2} \tag{1.48}$$

It means that the copula is invariant under strictly increasing transformation of the random variables.

1.5.3.4 Tail dependence

The concept of *tail dependence* is related to the probability of getting simultaneously extreme (small or large) outcomes. Lower and upper tail dependences are seperately studied. Let us consider in the sequel a pair of random variables (X_1, X_2) with uniform marginals $\mathcal{U}[0, 1]$.

Definition 3 *A copula C has a lower tail dependence if:*

$$\lambda_L = \lim_{x \to 0^+} \frac{C(x,x)}{x} = \lim_{x \to 0^+} \frac{\mathbb{P}\left(X_1 \leqslant x, X_2 \leqslant x\right)}{x} \tag{1.49}$$

exists and $\lambda_L \in]0,1]$. If $\lambda_L = 0$, then C has no lower tail dependence.

Definition 4 *A copula C has an upper tail dependence if:*

$$\lambda_U = \lim_{x \to 1^-} \frac{1 - 2x + C(x,x)}{1-x} = \lim_{x \to 1^-} \frac{\mathbb{P}\left(X_1 \geqslant x, X_2 \geqslant x\right)}{1-x} \tag{1.50}$$

exists and $\lambda_U \in]0,1]$. If $\lambda_U = 0$, then C has no upper tail dependence.

1.5.3.5 Marginal and conditional copulas

Before presenting the marginal and conditional copulas, let us first introduce the notion of copula density.

Definition 5 *The copula density c if it exists, is defined by:*

$$c(u_1, u_2) = \frac{\partial^2}{\partial u_1 \partial u_2} C(u_1, u_2) \tag{1.51}$$

Conditional copulas are necessary for the building of multidimensional copulas ($n > 2$). Let us now introduce the following notations:

$$\forall k \in [1, \ldots, n], \quad C_k(u_1, \ldots, u_k) = C_n(u_1, \ldots, u_k, 1, \ldots, 1) \tag{1.52}$$

For a random vector with uniform marginals on $[0,1]$, $C_k(u_1, \ldots, u_k)$ is the cumulative distribution function of the subvector $[u_1, \ldots, u_k] \subset [u_1, \ldots, u_n]$. Then, the conditional copula of $[u_1, \ldots, u_k]$ knowing $[u_1, \ldots, u_{k-1}]$ reads:

$$C_k(u_k | u_1, \ldots, u_{k-1}) = \frac{\partial^{k-1} C_k(u_1, \ldots, u_k)}{\partial^{k-1} C_{k-1}(u_1, \ldots, u_{k-1})} \tag{1.53}$$

1.5.3.6 Composition of copulas

Let us consider two random vectors $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ with respective dimension n and m and respective copula C_1 and C_2 that are independent, *i.e.* $\forall X_{1,i} \in \boldsymbol{X}_1$, $\forall X_{2,j} \in \boldsymbol{X}_2$, $X_{1,i}$ and $X_{2,j}$ are independent. Then, the copula C of the random vector made of the union $\boldsymbol{X} = [\boldsymbol{X}_1, \boldsymbol{X}_2]^\mathsf{T}$ is the product of the copula C_1 and C_2:

$$C\left(\boldsymbol{X}_1, \boldsymbol{X}_2\right) = C_1\left(X_{1,1}, \ldots, X_{1,n}\right) \times C_2\left(X_{2,1}, \ldots, X_{2,m}\right) \tag{1.54}$$

The resulting copula C is of dimension $n + m$. This property enables one to build easily the copula of a random vector with independent subvectors.

1.5.3.7 Copulas and dependence measures

It has been shown that the dependence between two random variables can be characterized by dependence measures such as the Spearman's ρ or the Kendall's τ. Nelsen (1999) shows that these measures can be deduced from the copula of the joint distribution. The Spearman's ρ or the Kendall's τ between two random variables X_1 and X_2 respectively read:

$$\rho_S(X_1, X_2) = 12 \iint_{[0,1]^2} C(u, v)\, \mathrm{d}u\, \mathrm{d}v - 3 \tag{1.55}$$

and:

$$\tau(X_1, X_2) = 4 \iint_{[0,1]^2} C(u, v)\, \mathrm{d}C(u, v) - 1 \tag{1.56}$$

Considering all the enounced properties and definitions, copulas enable one to model the dependence structure of any random vector. The next subsection shows a variety of copulas commonly used in engineering applications.

1.5.4 Classes of copulas

Many types of copulas have been defined, especially in the past few years, in order to model extreme values. This subsection presents the most common copula families and their main properties.

1.5.4.1 The independent copula

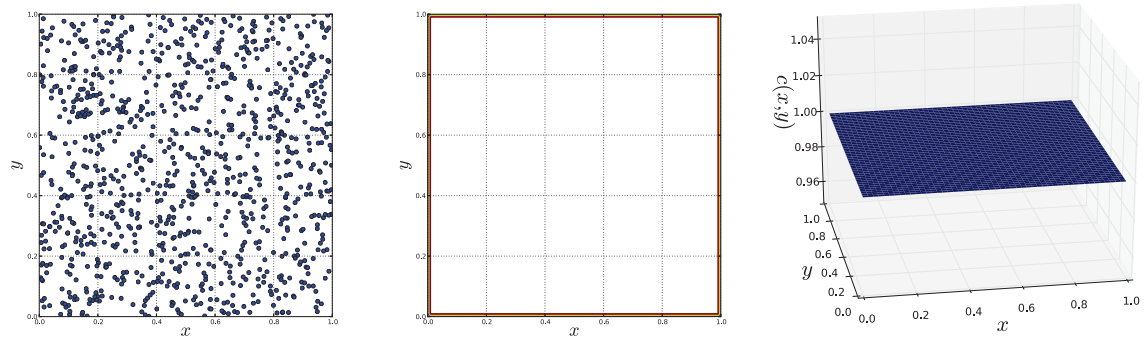
The copula theory represents a global framework to model any kind of dependence between random variables, including the independent case. The independent copula (or product copula) is defined by:

$$C(u_1, u_2) = u_1 \cdot u_2 \tag{1.57}$$

It corresponds in the probability theory to the probability product:

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\, \mathbb{P}(B) \tag{1.58}$$

when A and B are two independent events. The independent copula is displayed in Figure 1.7. The first picture (a) represents a scatterplot of a sample with $\mathcal{U}[0, 1]$ marginals and independent copula. The second picture shows a contourplot of the copula density $c(u_1, u_2)$. The third picture exhibits the copula distribution $c(u_1, u_2)$. The same three representations will be used to illustrate all the copula types.



(a) (b) (c)

Figure 1.7: *Independent copula, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

1.5.4.2 Elliptical copulas

An elliptical copula $C^{\mathcal{E}}_{\mathbf{R},\psi}$ is the copula of an elliptical distribution $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{R},\psi}$. In general, the copula $C^{\mathcal{E}}_{\mathbf{R},\psi}$ is not the CDF of $\mathcal{E}_{\boldsymbol{\mu},\boldsymbol{\sigma},\mathbf{R},\psi}$. The elliptical copula family corresponds to two well-known copulas: the Gaussian copula and the Student copula (or $t-$copula).

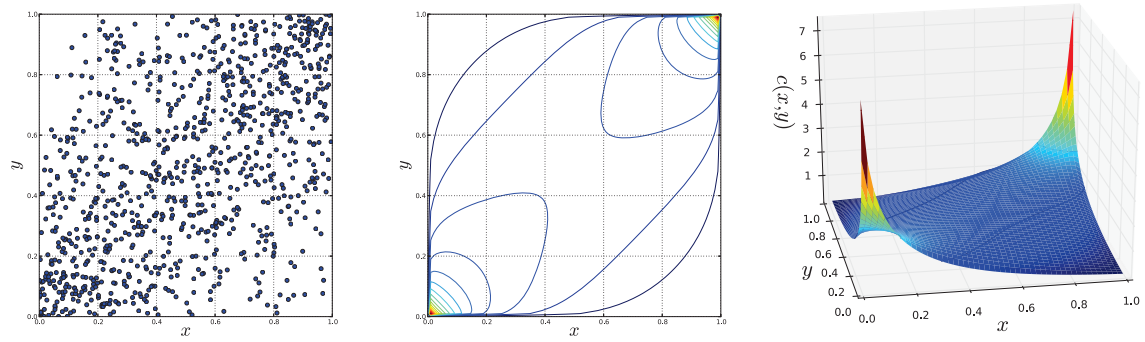
Gaussian copula The Gaussian copula is defined by:

$$C(u_1, u_2; \rho) = \Phi_2\left(\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho\right) \tag{1.59}$$

where ρ is the linear correlation coefficient and $\Phi_2(x, y; \rho)$ is the cumulative distribution function of the bivariate standard Gaussian distribution with correlation coefficient ρ:

$$\Phi_2(x, y; \rho) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{x^2 - 2\rho xy + y^2}{2(1-\rho^2)}\right] \tag{1.60}$$

The Gaussian copula, pictured in Figure 1.8, is asymptotically independent in both upper and lower tails. This means that no matter how high the correlation coefficient ρ is, there will be no tail dependence ($\lambda_L = \lambda_U = 0$) from a Gaussian copula (except if $\rho = 1$). Modelling the dependence structure of a random vector using a Gaussian copula is consistent with the measure of this dependence with the linear correlation coefficient. When non linear correlation dependence or extreme value events have to be modelled, other types of copulas (presented hereinafter) have to be used.



(a) (b) (c)

Figure 1.8: *Gaussian copula with $\rho = 0.5$, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

Remark : It is said that the Gaussian copula, presented in the financial world as *Li's formula* (Li, 1999) is partly responsible for the 2008 financial crisis (Salmon, 2009) (article available here). In that case, the dependence between events was modelled with a single number, ignoring both the real-world context and the limitations of this supposed breakthrough in financial models.

Student copula The second elliptical copula is the Student copula (or $t-$copula) that is derived from the bivariate Student distribution. It is defined by:

$$C(u_1, u_2; \rho, k) = t_{\rho,k}(t_\rho^{-1}(u_1), t_\rho^{-1}(u_2)) \tag{1.61}$$

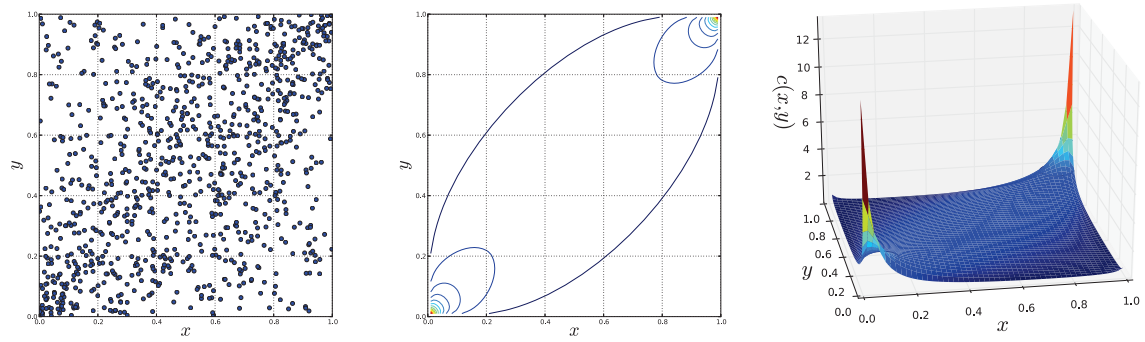
where:

$$t_k(x) = \frac{1}{\sqrt{k\pi}} \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{k}{2}\right)} \left(1 + \frac{x^2}{k}\right)^{-\left(\frac{k+2}{2}\right)}, \quad k \geqslant 1 \tag{1.62}$$

is the probability density function of the Student distribution with k degrees of freedom and where:

$$t_{\rho,k}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} \frac{1}{2\pi\sqrt{1-\rho^2}} \left(1 + \frac{x^2 - 2\rho xy + y^2}{k(1-\rho^2)}\right)^{-\left(\frac{k+2}{2}\right)} \mathrm{d}x\mathrm{d}y \tag{1.63}$$

is the bivariate probability density function with linear correlation ρ. The Student copula is pictured in Figure 1.9.



(a) (b) (c)

Figure 1.9: *Student copula with $\rho = 0.5$, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

Unlike the Gaussian copula, the Student copula has both lower and upper tail dependences. They are equal (symmetrical copula) and read:

$$\lambda_L = \lambda_U = 2t_{k+1}\left(\frac{(k+1)(1-\rho)}{1+\rho}\right) \tag{1.64}$$

1.5.4.3 Archimedean copulas

The Archimedean copulas (Genest and MacKay, 1986) are a class of copulas characterized by a generator function φ. The general definition reads:

Definition 6 *C is an Archimedean copula if:*

$$C(u_1, u_2) = \begin{cases} \varphi^{-1}(\varphi(u_1) + \varphi(u_2)) & \text{if } \varphi(u_1) + \varphi(u_2) \geqslant \varphi(0) \\ 0 & \text{either} \end{cases} \tag{1.65}$$

with φ verifying $\varphi(1) = 0$, $\varphi'(u) < 0$ et $\varphi''(u) > 0$, $\forall\, 0 \leqslant u \leqslant 1$.

In the next pararaphs, three Archimedean copulas are presented, namely, the Gumbel copula, the Clayton copula and the Frank copula.

Gumbel copula The Gumbel copula corresponds to the generator function $\varphi(t) = [-\ln t]^\theta$ (see Figure 1.10). It is defined by:

$$C(u_1, u_2) = \exp\left[-((-\ln u_1)^\theta + (-\ln u_2)^\theta)\right]^{\frac{1}{\theta}} \tag{1.66}$$

where $\theta \geqslant 1$ is the copula parameter.

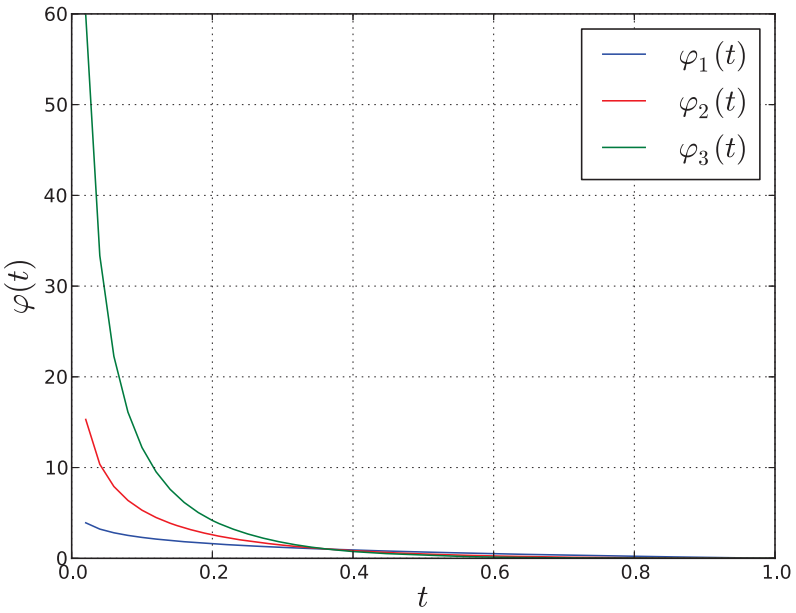
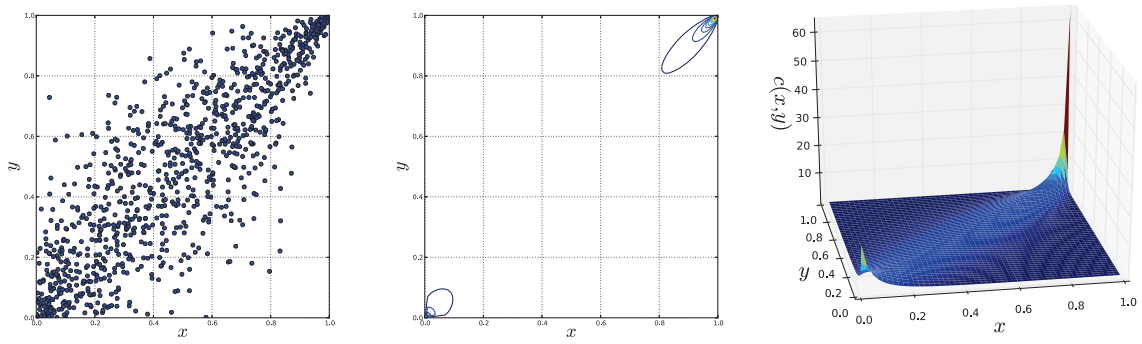


Figure 1.10: *Generator function of the Gumbel copula with $\theta = 1$ (blue), $\theta = 2$ (red) et $\theta = 3$ (green).*

The Gumbel copula is pictured in Figure 1.11. Unlike the elliptical copulas which are symmetrical, there is a noticeable difference in the behaviour for lower and upper tails. The Gumbel copula has a non zero upper tail dependence:

$$\lambda_U = 2 - 2^{\frac{1}{\theta}} \tag{1.67}$$

which, according to this definition, appears only when $\theta > 1$. The lower tail dependence λ_L is zero.



(a) (b) (c)

Figure 1.11: *Gumbel copula with $\theta = 3$, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

Clayton copula The Clayton copula corresponds to the generator function $\varphi(t) = t^{-\theta} - 1$ (see Figure 1.12). It is defined by:

$$C(u_1, u_2) = (u_1^{-\theta} + u_2^{-\theta} - 1)^{-\frac{1}{\theta}} \tag{1.68}$$

where $\theta > 0$ is the copula parameter.

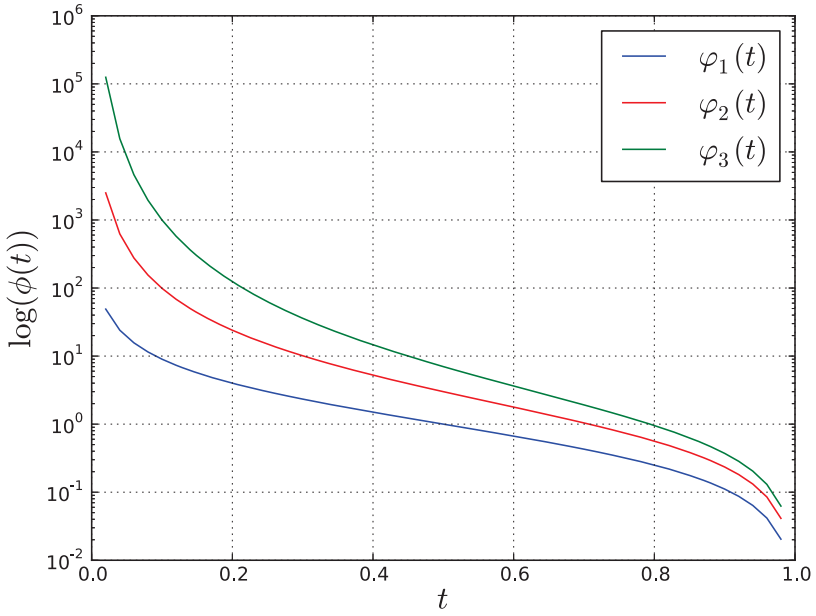
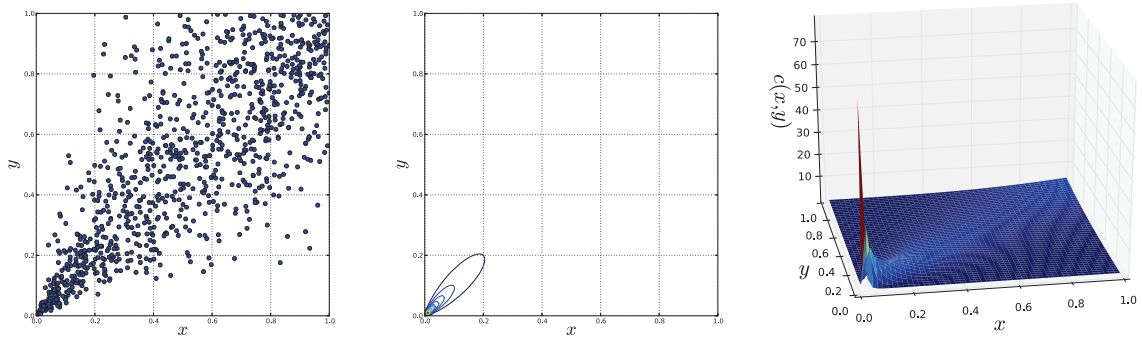


Figure 1.12: *Generator function of the Clayton copula with $\theta = 1$ (blue), $\theta = 2$ (red) et $\theta = 3$ (green).*



(a) (b) (c)

Figure 1.13: *Clayton copula with $\theta = 3$, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

The Clayton copula is pictured in Figure 1.13. A lower tail dependence can be graphically noticed. It reads :

$$\lambda_L = 2^{-\frac{1}{\theta}} \tag{1.69}$$

The upper tail dependence λ_U is zero.

Frank copula The Frank copula corresponds to the generator function $\varphi(t) = -\ln \dfrac{e^{-\theta t} - 1}{e^{-\theta} - 1}$ (see Figure 1.14). It is defined by:

$$C(u_1, u_2) = -\frac{1}{\theta} \ln \left(1 + \frac{(e^{-\theta u_1} - 1)(e^{-\theta u_2} - 1)}{e^{-\theta} - 1} \right) \tag{1.70}$$

where $\theta \in \mathbb{R}^*$ is the copula parameter.

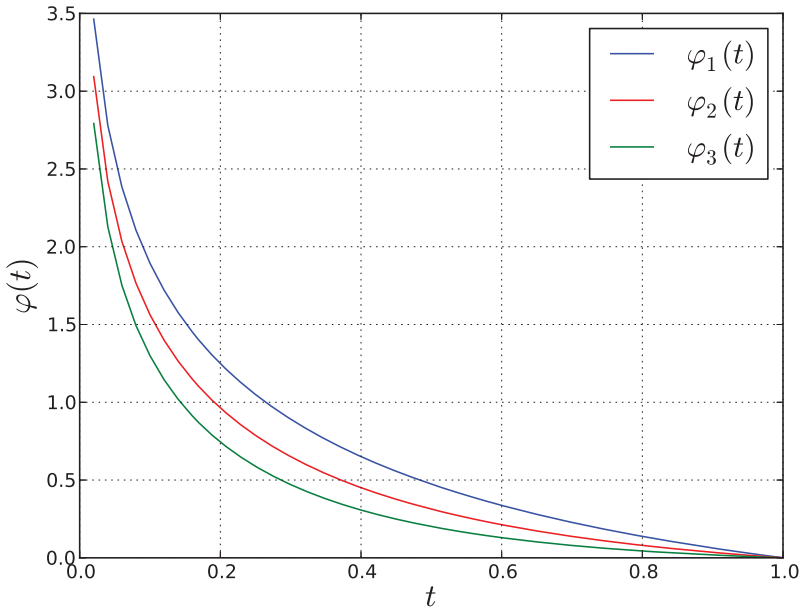
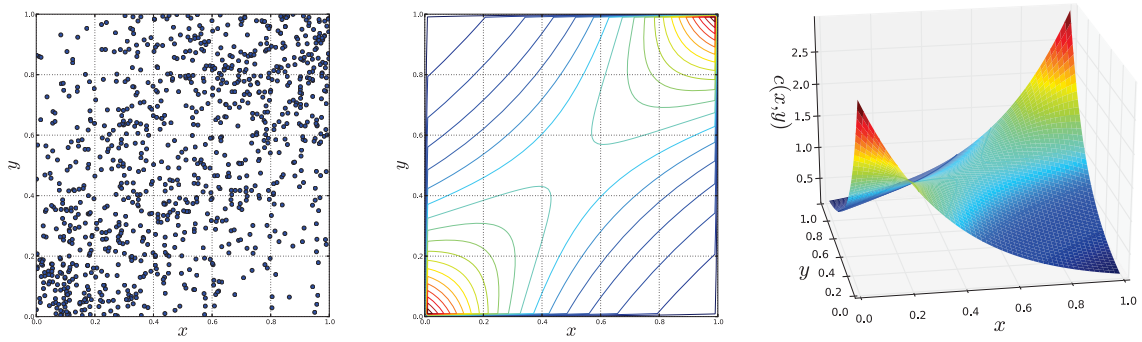


Figure 1.14: *Generator function for the Frank copula with $\theta = 1$ (blue), $\theta = 2$ (red) et $\theta = 3$ (green).*

The Frank copula is pictured in Figure 1.15.



(a) (b) (c)

Figure 1.15: *Frank copula with $\theta = 3$, scatterplot (a), contourplot (b) and 3D-view (c) of the copula density $c(u_1, u_2)$.*

The basics of the copula theory have been presented. In the next subsection, copulas are seen from a practical point of view. The purpose of the copulas is to model the dependence structure of a random vector. In order to use it, one must be able to simulate and identify the copula from a given data set.

1.5.5 Simulation of a copula

1.5.5.1 Simulation of a joint distribution

The simulation of a random vector $\boldsymbol{X}$ with marginal distributions $F_1, \ldots, F_n$ and copula C_n can be achieved in two steps:

1. Simulate a sample $\mathcal{U}$ from a random vector $\boldsymbol{U}$ with copula C and uniform margins on $[0, 1]$.

2. Transform the sample $\mathcal{U}$ into $\mathcal{X}$ by applying the effects of the marginal distributions:

$$x_i = F_i^{-1}(u_i), \quad i = 1, \ldots, n \tag{1.71}$$

Consequently, the difficulty relies in simulating realizations from the copula C.

1.5.5.2 Simulation of a copula

The simulation procedure depends on the type of the copula. Here, the case of a bivariate Gaussian copula is considered. For a given correlation matrix $\boldsymbol{\Sigma}$, the algorithm reads:

1. Perform a Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}^\top \mathbf{L}$.

2. Generate independent identically distributed (iid) standard normal random variables X_1' and X_2'.

3. Compute $(X_1, X_2)^\mathsf{T} = \boldsymbol{X} = \mathbf{L}\boldsymbol{X'}$ where $\boldsymbol{X'} = (X_1', X_2')^\mathsf{T}$.

4. Return $U_i = \Phi(X_i)$, $i = 1, 2$ where Φ is the standard normal cumulative distribution function.

For other types of copula, the global algorithm is quite similar but there are changes in the distributions used to simulate and transform the intermediate samples.

1.5.6 Identification of a copula

Building the probabilistic model of a random vector $\boldsymbol{X}$ from a data sample $\boldsymbol{\mathcal{X}}$ corresponds to identifying:

1. the marginal distributions of each component X_i,

2. the dependence structure, namely the copula C.

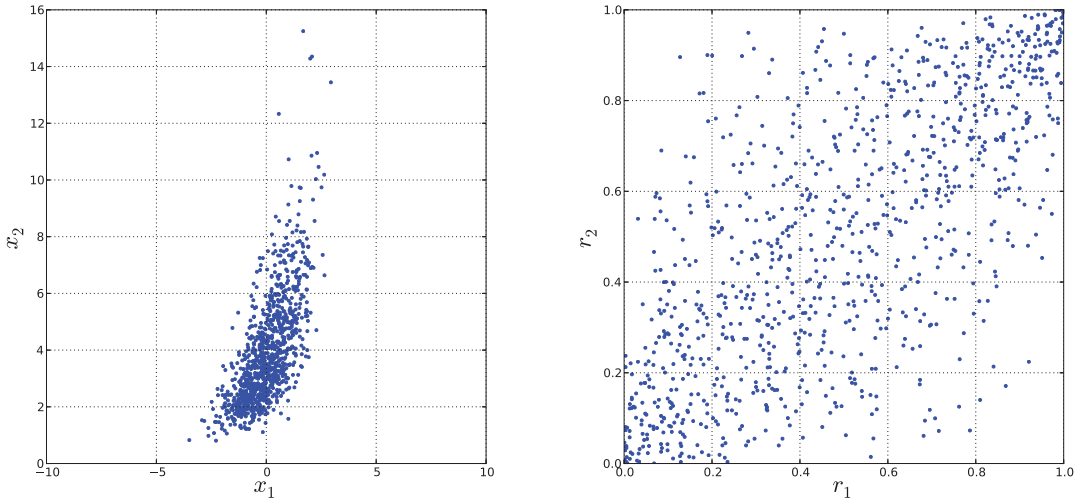
The identification techniques for the marginal distributions are well-established ($QQ-$plot, maximum-likelihood or kernel smoothing estimation or goodness-of-fit tests such as the Kolmogorov-Smirnov test). Those for copulas are inspired from them. But, for a first intuition, it is necessary to *visualize* the copula.

1.5.6.1 Dependogram of a copula

The *dependogram* of a bivariate sample $\boldsymbol{\mathcal{X}}$ of size N is a modified scatterplot for which the numerical values of the realizations $x_k^{(i)}$ are replaced by their normed position $r_k^{(i)}$ in the positively ranked marginal sample X_1:

$$r_k^{(i)} = \frac{\text{rank}(x_k^{(i)})}{N},\ k = 1, \ldots, n,\ i = 1, \ldots, N \tag{1.72}$$

The marginal distributions are transformed into uniform distributions on $[0, 1]$. The goal of this operation is to erase the effects of the margins so that only the copula is preserved. An illustration of the transformation is given in Figure 1.16. The joint distribution $F(\boldsymbol{x}) = C\left(F_1(x_1), F_2(x_2)\right)$ is composed of a standard normal distribution $F_1 \sim \mathcal{N}(0, 1)$ and a lognormal distribution $F_2 \sim \mathcal{LN}\left(\mu = 4, \sigma = 2, \gamma = 0\right)$ coupled by a Gaussian copula $C_{\theta=0.7}$.



(a) (b)

Figure 1.16: *Scatterplot (a) and dependogram (b) of a bivariate sample. On the dependogram, the effects of the margins have been removed and the copula appears through the scatterplot.*

1.5.6.2 Parametric estimation of a Gaussian copula

Assuming a multivariate sample has a Gaussian copula, the methodology consists in computing the rank correlation matrix $\boldsymbol{\rho}_S$ or the Kendall $\boldsymbol{\tau}$ matrix from the considered sample $\boldsymbol{\mathcal{X}}$ in order to parameterize the corresponding Gaussian copula. Then the copula parameter matrix $\mathbf{R}$ is defined by:

$$R_{ij} = 2\sin\left(\frac{\pi}{6}\rho_{S,ij}\right) = \sin\left(\frac{\pi}{2}\tau_{ij}\right) \tag{1.73}$$

It is comparable to the so-called *method of moments* for estimating the parameters of univariate distributions.

1.5.6.3 Kendall plot

The *Kendall plot* is an easy graphical tool whose goal is to test the structure of dependence of the data (Genest and Boies, 2003). The considered hypotheses are:

$$H_0 : C = C_0 \quad \text{versus} \quad H_1 : C \neq C_0 \tag{1.74}$$

where C is the copula of the data and C_0 is the tested copula. Like the $QQ-$plot for distributions, it compares quantiles. The Kendall plot for any bivariate sample requires to transform the original sample by removing the effects of the marginals. The numerical

value of the realizations x_{1^i} of X_1 (respectively X_2) is replaced by its normed rank in $\mathcal{X}_1$ (respectively $\mathcal{X}_2$). The normed ranks are then referred to as pseudo-observations.

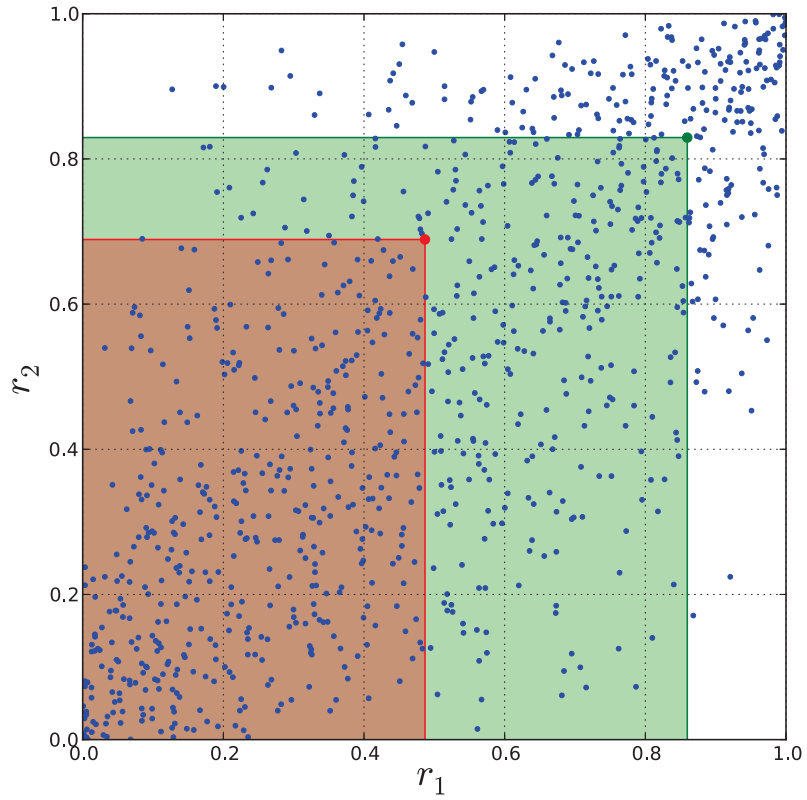


Figure 1.17: *Principle of estimation for the Kendall plot's quantiles. For any pseudo-observation x_i, one has to count the number of pseudo observations $x_{i,i\neq j}$ that respect $r_{1,i} > r_{1,j}$ and $r_{2,i} > r_{2,j}$. Examples are given for two pseudo-observations. The empirical quantiles $\hat{H}_i$ are respectively the number of points that belong to the green-filled or red-filled areas.*

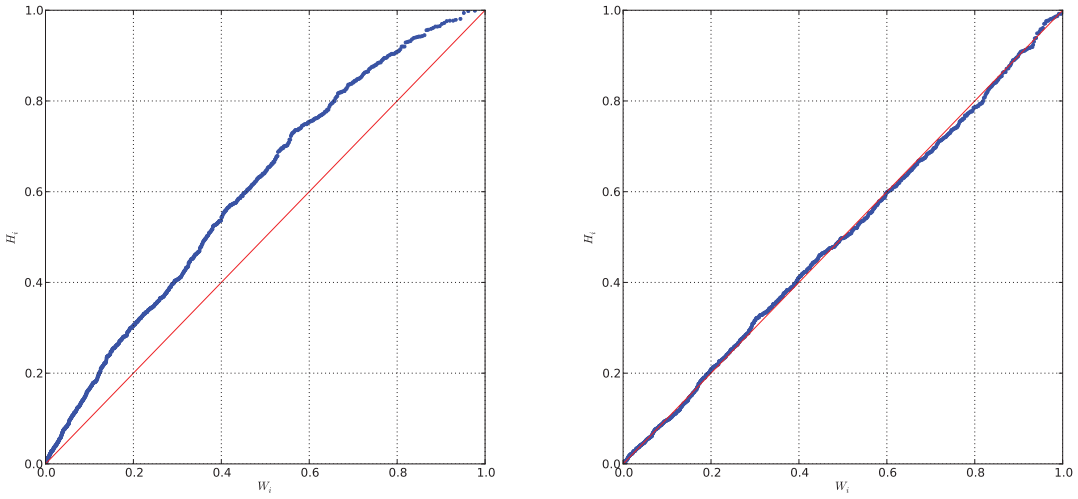
Let us consider a bivariate data sample (X_1, X_2) of size N with uniform margins on $[0, 1]$, *i.e.* a sample from a copula. The empirical quantiles $\hat{H}_i$, calculated from the data are defined by:

$$\hat{H}_i \equiv H(X_1^i, X_2^i) = \frac{1}{N-1} \# \left\{ i \neq j, X_1^j \leqslant X_1^i, X_2^j \leqslant X_2^i \right\} \tag{1.75}$$

where # denotes the size of the set $\{.\}$. The estimation of the quantile $\hat{H}_i$ consists in counting the realizations for which *both* coordinates are smaller than the X_i's as it is shown on Figure 1.17.

The theoretical quantiles (corresponding to the copula C_0) are defined by $W_i = \mathrm{E}\left[H^0(X_1^0, X_2^0)\right]$. In other words, the empirical quantiles $\hat{H}_i$ calculated from a data sample

with copula C are compared to the expected quantiles of a sample with copula C_0 of the same size. The expected quantiles are estimated from k samples from the copula C_0. In practice, $k = 100$ samples are used for the estimation of the synthetic theoretical quantiles W_i. The Kendall plot is the graph of the points $(\hat{H}_i, W_i)$, $i = 1, \ldots, N$. These pairs tends to concentrate along the main diagonal under the null hypothesis ($C = C_0$), see Figure 1.18.



(a) (b)

Figure 1.18: *Two examples of* Kendall Plots. *In the subfigure (a), a Gaussian copula $C_{\theta=0.2}$ is tested (H_0 rejected), whereas in the subfigure (b), the true copula $C_{\theta=0.7}$ is compared to the sample (H_0 accepted).*

1.5.6.4 Semi-parametric estimation

Here, two *blanket tests* are considered. The term *blanket* refers to the fact that neither parameter tuning nor strategic choices are required. For a given class of copula $\mathcal{C}$ (Gaussian, Gumbel, Clayton), the goal is to determined the copula parameter θ so that the empirical copula C_θ fits the data. In Genest et al. (2007), the author proposes a review on a large variety of goodness-of-fit tests for copulas. A particular attention is given to ranked versions of Cramér-von Mises and Kolmogorov-Smirnov. The associated statistics S_N (resp. T_N) is given in Eq. (1.76) (resp. Eq. (1.77)). Given a $n-$variate sample $\boldsymbol{\mathcal{U}} = \{\boldsymbol{U}_1, \ldots, \boldsymbol{U}_N\}$ of size N, the idea is to determine a *distance* between the tested copula C_N and an estimation C_{θ_N} of the real copula $C \in \mathcal{C}$. These statistics respectively read:

$$S_N = \int_{[0,1]^n} \mathbb{C}_N(\boldsymbol{u})^2 dC_N(\boldsymbol{u}) \tag{1.76}$$

and

$$T_N = \sup_{\boldsymbol{u}\in[0,1]^n} |\mathbb{C}_N(\boldsymbol{u})| \tag{1.77}$$

with:

$$\mathbb{C}_N = \sqrt{N}\,(C_N - C_{\theta_N}) \tag{1.78}$$

and:

$$C_N(\boldsymbol{u}) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\,(U_{i1} \leqslant u_1, \ldots, U_{in} \leqslant u_n) \tag{1.79}$$

Large values of these statistics lead to the rejection of H_0. Approximate $p-$values can be estimated by bootstraping. Genest and Rémillard (2008) show that these tests are consistent and the larger the sample is (at least 10^2 points for a bivariate sample), the better the results are. For more tests, the reader is referred to Kostadinov (2005), Fermanian (2005).

1.5.6.5 Non parametric estimation

The main drawback of the previous estimation techniques is that they make hypotheses on the class of copula. In other words, for a given sample, one has to run one test for each class of copula that could fit. For instance, a sample with a symmetrical dependogram, Gaussian, Student, Frank could fit but potentially also any other symmetrical copulas. To circumvent this difficulty, non parametric techniques have been developped by Charpentier (2006) (kernel smoothing technique) and Lambert (2007) (Bayesian splines smoothing techniques for Archimedean copulas).

1.5.7 Copula and isoprobabilistic transformations

In this section, it is shown how isoprobabilistic transformations (*i.e.* Nataf and Rosenblatt transform) can be seen from a copula point of view. This observation comes from the articles by Lebrun and Dutfoy (2009a), Lebrun and Dutfoy (2009c), Lebrun and Dutfoy (2009b) which is a comprehensive work on isoprobabilistic transformations in three parts.

An isoprobabilistic transformation T is a diffeomorphism from the support D_X to $\mathbb{R}^n$ so that $\mathbf{U}$ and $\mathbf{RU}$ have the same distribution for all rotation $\mathbf{R} \in s\mathcal{O}_n(\mathcal{R})$. The Nataf and Rosenblatt transform have this property.

1.5.7.1 Nataf transformation

The Nataf transformation (Nataf, 1962) allows one to build a multivariate distribution that fits a collection of marginal distributions and a correlation matrix.

Definition 7 (Nataf tranformation) *Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with a joint distribution $F_{\boldsymbol{X}}$ defined by its marginal cumulative distribution functions*

$F_1, \ldots, F_n$ and its normal copula $C_{\mathbf{R}}$. The Nataf tranformation T_N of $\boldsymbol{X}$ reads:

$$\boldsymbol{U} = T_N(\boldsymbol{X}) = T_2 \circ T_1(\boldsymbol{X}) \tag{1.80}$$

where T_1 and T_2 are respectively defined by:

$$\begin{aligned} T_1 : \quad & \mathbb{R}^n \to \mathbb{R}^n \\ & \boldsymbol{X} \mapsto \boldsymbol{Y} = \begin{pmatrix} \Phi^{-1} \circ F_1(X_1) \\ \vdots \\ \Phi^{-1} \circ F_n(X_n) \end{pmatrix} \end{aligned} \tag{1.81}$$

$$\begin{aligned} T_2 : \quad & \mathbb{R}^n \to \mathbb{R}^n \\ & \boldsymbol{Y} \mapsto \boldsymbol{U} = \boldsymbol{\Gamma Y} \end{aligned} \tag{1.82}$$

where Φ is the cumulative distribution function of the univariate standard normal distribution, $\boldsymbol{\Gamma} = \mathbf{L}^{-1}$ and $\mathbf{L}$ is the Cholesky decomposition of $\mathbf{R}$ so that $\mathbf{R} = \mathbf{L}\mathbf{L}^T$. $\mathbf{R}_0 \cdot Y$ is a Gaussian vector with standard Gaussian marginals and correlation matrix $\mathbf{R}_0$. Consequently, $\boldsymbol{U}$ is a Gaussian vector with the same marginals as $\boldsymbol{X}$ but independent.

The space $\boldsymbol{U}$ is defined in Eq. (1.82) is referred to as *standard space* where all the variables are independent with standard Gaussian distributions whereas the space of $\boldsymbol{X}$ is called the *physical space*. The generalized Nataf tranformation extends the above definition to any elliptical copula.

1.5.7.2 Rosenblatt transformation

The Rosenblatt tranformation offers an alternative for building a multivariate distribution with any copula.

Definition 8 (Rosenblatt transformation) *Let $\boldsymbol{X} = (X_1, \ldots, X_n) \in \mathbb{R}^n$ be a random vector with joint distribution $F_{\boldsymbol{X}}$ defined by its marginal cumulative distribution functions $F_1, \ldots, F_n$ and its copula C. The Rosenblatt tranformation T_R of $\boldsymbol{X}$ reads:*

$$\boldsymbol{U} = T_R(\boldsymbol{X}) = T_2 \circ T_1(\boldsymbol{X}) \tag{1.83}$$

where T_1 and T_2 are respectively defined by:

$$\begin{aligned} T_1 \quad : \quad & \mathbb{R}^n \to \mathbb{R}^n \\ & \boldsymbol{X} \mapsto \boldsymbol{Y} = \begin{pmatrix} F_1(x_1) \\ \vdots \\ F_{k|1,\ldots,k-1}(X_k|X_1, \ldots, X_{k-1}) \\ \vdots \\ F_{n|1,\ldots,n-1}(X_n|X_1, \ldots, X_{n-1}) \end{pmatrix} \end{aligned} \tag{1.84}$$

$$T_2 \quad : \quad \mathbb{R}^n \to \mathbb{R}^n$$
$$\boldsymbol{Y} \mapsto \boldsymbol{U} = \begin{pmatrix} E^{-1}(Y_1) \\ \vdots \\ E^{-1}(Y_n) \end{pmatrix} \tag{1.85}$$

where E is the univariate cumulative distribution function of an elliptical distribution.

This transformations has been defined for Gaussian copulas and may be extended to any elliptical copulas.

One drawback of the Rosenblatt transform is that it is not unique. The transformation depends on the order of the conditional simulations in T_1 (Eq. (1.84)).

1.5.7.3 Link between copulas and isoprobabilistic transforms

The Nataf transform is one way of modelling the dependence structure of a random vector by a Gaussian copula parametrized by its correlation matrix $\boldsymbol{\rho}$. Indeed, it is shown that the Nataf transform is one particular way of modelling the stochastic dependence using the Gaussian copula. The demonstration is based on the invariance properties of the copula by increasing transformation of the components of random vector. The generalized Nataf transform enables one to generalize this principle to any multivariate distribution with an elliptical copula. However, it is important to pay attention to the parametrization using the linear correlation matrix because the latter does not take non linear aspects such as tail dependences into account. Then, the practitioner should rather use the other dependence measures such as the Spearman's ρ or the Kendall's τ. Concerning the Rosenblatt transform, Lebrun and Dutfoy (2009c) show that it is equivalent to the Nataf transform in the case of a Gaussian copula. In other cases, that is when the copula is not Gaussian, the two probabilistic transforms differ in the sense that their standard spaces are different.

1.6 Conclusion

This chapter has recalled some basics of the probability theory that will be used all along this manuscript. The mathematical notations have been set up. The tools presented here put together the mathematical framework to build the probabilistic modelling of data. First the intrinsic behaviour of the parameters is described by the marginal distributions. Then interactions between the parameters are taken into account. Therefore, an overview on correlation measures and the copula theory is proposed. For a more global vision on copula, Embrechts (2009) recommends the reading of three *must-read* articles: Embrechts et al. (2002) on the hazards of using simple dependence measures, Genest and Favre (2007) on ranked-based inference methods for copulas and Genest and Neslehova (2007) on the use of copulas for count data. The same author also advises the reading of Genest and MacKay (1986) on the geometrical interpretation of the Kendall's τ and Mikosch (2006)

who proposes an incisive analogy between the sudden craze for copulas and the tale of Andersen *The Emperor's New Clothes*.

Now that all the tools to build a joint probability distribution function have been developed and that uncertainties in the input parameters are propagated through a model $\mathcal{M}$, it is time to consider the uncertainties in the output parameter. A particular attention is given to the identification of the sources of this dispersion. In the next chapter, global sensitivity analysis is studied.

2

Global sensitivity analysis

Contents

2.1 Introduction

Sensitivity analysis (SA) aims at identifying how the uncertain input parameters X_i, $i = 1, \ldots, n$ of a model $\mathcal{M}$ contribute to the variability of its output Y. The objectives of this type of analysis are multiple (Saltelli et al., 2000):

- Analyzing the role of the parameters in the model, *i.e.* self-contributions and interactions;
- Identifying the less influential parameters in the model in order to reduce the dimension of the problem;
- Reducing the dispersion of the model output by minimizing the variability of the most influent parameters.

Motivations and applications of sensitivity analysis are originally found in biology and chemistry where the number of possible experiments is limited in time and by their costs. Nowadays, SA is used in many engineering fields, including mechanical engineering, where there is a need for optimizing numerical simulations.

Three approaches have to be distinguished. *Screening* methods (Morris, 1991) qualitatively analyze the importance of an input parameter on the model output. They allow one to rank the input parameters according to their contribution to the output variability and consequently to easily identify which ones must be studied more precisely and which ones can be neglected.

Local sensitivity analysis (LSA) is a quantitative method. In addition to an input parameter ranking, it also defines a parameter importance as a quantity, namely a *sensitivity index*. LSA methods focus on the effects of an input parameter taking one fixed value $\boldsymbol{x}^* = [x_1^*, \ldots, x_n^*]^\intercal$, which is often its mean value $\bar{\boldsymbol{x}}$. The so-called *one at a time* (OAT method) consists in computing the following sensitivity indices:

$$S_i^* = \frac{\partial \mathcal{M}}{\partial x_i}(x_1^*, \ldots, x_n^*) \tag{2.1}$$

The second approach consists in a more general identification of the contribution of the parameters. One first family of methods study the simple (linear) relationship that exists between the input parameter and the output of the model. Therefore, they are reserved for linear or at least monotonic models. When no hypothesis on the structure of the model is made, the analysis must consider all the values each input parameter can take. This approach is referred to as *global sensitivity analysis* (GSA). The most popular method is based on the *decomposition of the variance* of the model output. This powerful technique provides useful results but requires a large number of calls to the numerical model.

Sensitivity analysis methods are well defined when the variables are independent but in the presence of correlation, the results they provide are either erroneous or simply not calculable. To circumvent this problem, methods derived from the classical ones have been

recently developped. Different options are explored. As shown later on, authors consider the independence of the variables as a particular case and search for a generalization of the methods while others define new quantities to characterize the input parameter contributions.

This chapter mostly deals with global sensitivity anlaysis. The two first sections describe general methods in the case of independence of the input parameters while the next three sections propose new techniques to treat problems with correlated input parameters. For a general overview on sensitivity analysis, the reading of the book by Saltelli et al. (2004) is advised.

2.2 Correlation and regression-based methods

In this first section, global sensitivity analysis methods for linear or monotonic models are presented. They are based on the study of the relation between an input parameter X_i and the model output $Y = \mathcal{M}(\boldsymbol{X})$ where $\boldsymbol{X}$ is a random vector with independent components.

2.2.1 Linear models

First, two global sensitivity indices, namely the *Standard Regression Coefficients SRC* and the *Partial Correlation Coefficients PCC*, are introduced.

2.2.1.1 Standard regression coeffcients

Let us first consider the model $\mathcal{M}$ as linear. Thus, the model response Y reads:

$$Y = \beta_0 + \sum_{i=1}^{n} \beta_i X_i \tag{2.2}$$

Thanks to the independence of the X_i's, the variance of Y can be decomposed as follows:

$$\mathrm{Var}[Y] = \sum_{i=1}^{n} \beta_i^2 \mathrm{Var}[X_i] \tag{2.3}$$

where $\beta_i^2 \mathrm{Var}[X_i]$ is the share of the variance of Y due to X_i. Then, the sensibility of Y to input variable X_i is given by the *Standard Regression Coefficient* SRC_i:

$$SRC_i = \frac{\beta_i^2 \mathrm{Var}[X_i]}{\mathrm{Var}[Y]} \tag{2.4}$$

As the SRC_i index represents a share of variance, its value belongs to the interval $[0, 1]$. A value close to 1 indicates that the variable X_i has a major contribution to the variability of Y whereas a value that tends to 0 shows that whatever dispersed X_i is, it has no influence on the variability of Y.

Remark: The SRC_i index corresponds to the squared linear correlation coefficient ρ_{Y,X_i} between the input parameter X_i and the output parameter Y. Because $\mathcal{M}$ is linear, $\mathrm{Cov}\left[Y, X_i\right] = \beta_i \mathrm{Var}\left[X_i\right]$ and:

$$\rho_{Y,X_i} = \frac{\mathrm{Cov}\left[Y, X_i\right]}{\sqrt{\mathrm{Var}\left[Y\right]\mathrm{Var}\left[X_i\right]}} = \beta_i \sqrt{\frac{\mathrm{Var}\left[X_i\right]}{\mathrm{Var}\left[Y\right]}} \tag{2.5}$$

that is $SRC_i = \rho^2_{Y,X_i}$.

2.2.1.2 Partial correlation coefficients

One limitation of the SRC indices relies in the correlation that might occur between the variables due to the simulations and might cause a misinterpretation of the results. Therefore, another sensitivity index, namely the *Partial Correlation Coefficient*, has been proposed to compute the sensitivity of Y to X_i without any effects of the input parameters $X_{j,j\neq i}$. It reads:

$$\begin{aligned} PCC_i &= \rho_{Y,X_i|\boldsymbol{X}_{\sim i}} \\ &= \frac{\mathrm{Cov}\left[Y, X_i|\boldsymbol{X}_{\sim i}\right]}{\sqrt{\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]\mathrm{Var}\left[X_i|\boldsymbol{X}_{\sim i}\right]}} \\ &= \frac{\mathrm{Cov}\left[Y, X_i|\boldsymbol{X}_{\sim i}\right]}{\sqrt{\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]\mathrm{Var}\left[X_i\right]}} \end{aligned} \tag{2.6}$$

where $\boldsymbol{X}_{\sim i}$ is the input vector derived by cancelling the i^{th} component of $\boldsymbol{X}$. In practice, a $N-$sample $\mathcal{Y}|\boldsymbol{\mathcal{X}}_{\sim i}$ is computed from a input sample where the j^{th} components realizations, $j \neq i$, are fixed at a given value x_j^*, namely:

$$\boldsymbol{\mathcal{X}}|\boldsymbol{\mathcal{X}}_{\sim i} = \begin{bmatrix} x_1^* & x_2^* & \cdots & x_{i-1}^* & x_i^{(1)} & x_{i+1}^* & \cdots & x_n^* \\ x_1^* & x_2^* & \cdots & x_{i-1}^* & x_i^{(2)} & x_{i+1}^* & \cdots & x_n^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^* & x_2^* & \cdots & x_{i-1}^* & x_i^{(k)} & x_{i+1}^* & \cdots & x_n^* \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_1^* & x_2^* & \cdots & x_{i-1}^* & x_i^{(N)} & x_{i+1}^* & \cdots & x_n^* \end{bmatrix} \tag{2.7}$$

At the numerator, the quantity $\mathrm{Cov}\left[Y, X_i|\boldsymbol{X}_{\sim i}\right]$ represents the covariance of Y and X_i when $\boldsymbol{X}_{\sim i}$ is know, *i.e.* fixed at given values and reads:

$$\begin{aligned} \mathrm{Cov}\left[Y, X_i|\boldsymbol{X}_{\sim i}\right] &= \mathrm{E}\left[(Y|\boldsymbol{X}_{\sim i} - \mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right])(X_i|\boldsymbol{X}_{\sim i} - \mathrm{E}\left[X_i|\boldsymbol{X}_{\sim i}\right])\right] \\ &= \mathrm{E}\left[(Y|\boldsymbol{X}_{\sim i} - \mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right])(X_i - \mathrm{E}\left[X_i\right])\right] \end{aligned} \tag{2.8}$$

At the denominator, the conditional variance $\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]$ represents the variance of Y when all the parameters but the i^{th} of $\boldsymbol{X}$ are fixed. Only the uncertainty brought by X_i is taken into account in the calculation of Y.

As it is expressed as a correlation coefficient, the sensitivity index PCC_i belongs to the interval $[-1, 1]$ (correlation can be negative). Therefore, the ranking has to be done on the indices absolute values. According to Eqs. (2.4) and (2.6), SRC and PCC indices are not equal. However, they provide the same type of parameter ranking.

2.2.2 Monotonic models

When the model $\mathcal{M}$ is no longer linear but still monotonic, the previously described sensivity indices cannot be used directly, but they can be modified so that they fit the model structure. By carrying out a rank transformation on the realizations in the $N-$sample $\boldsymbol{\mathcal{X}}$ (each realizations x_i is replaced by its rank r_i in the increasing ordered sample R_{X_i}), one obtains the *Standard Rank Regression Coefficients* and the *Partial Rank Correlation Coefficients*. Thus, the $SRRC$ and $PRCC$ indices respectively read:

$$SRRC_i = \frac{\beta_i^2 \operatorname{Var}[R_{X_i}]}{\operatorname{Var}[R_Y]} \tag{2.9}$$

and:

$$PRCC_i = \rho_{R_Y, R_{X_i} | \boldsymbol{R}_{\boldsymbol{X}_{\sim i}}} \tag{2.10}$$

When no hypothesis can be made on the model structure (linearity, monotonicity), a more general method is required.

2.3 Variance-based methods

The goal of GSA is to identify the main contributors to the dispersion of the model response Y. Consequently, the variance of Y is the quantity to be studied.

2.3.1 ANOVA decomposition

2.3.1.1 Introduction

The ANOVA (*ANalysis Of VAriance*) decomposition has been introduced in Efron and Stein (1981). In order to appreciate the contribution of the variable X_i to the variance of the model response Y, let us study how much the variance of Y would decrease if the value of X_i were known. The reduced variance of Y when X_i is fixed at a value x_i^* is:

$$\operatorname{Var}[Y | X_i = x_i^*] \tag{2.11}$$

As this quantity depends on the value x_i^*, it is necessary to compute its expected value $\mathrm{E}_{X_i}[\operatorname{Var}[Y|X_i]]$ in order to take all the possible values of X_i into account. The higher the contribution of X_i, the lower the expected variance of $Y|X_i$. According to the *total variance theorem*:

$$\operatorname{Var}[Y] = \operatorname{Var}_{X_i}[\mathrm{E}[Y|X_i]] + \mathrm{E}_{X_i}[\operatorname{Var}[Y|X_i]] \tag{2.12}$$

The share of variance of Y that is due to X_i is the quantity $\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|X_i\right]\right]$, also denoted by V_i. $\mathrm{E}_{X_i}\left[\mathrm{Var}\left[Y|X_i\right]\right]$ represents what is left of $\mathrm{Var}\left[Y\right]$ when X_i is known (deterministic). A sensitivity index of Y to X_i then reads:

$$S_i = \frac{\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|X_i\right]\right]}{\mathrm{Var}\left[Y\right]} \tag{2.13}$$

Because the conditional variance is normalized by the total variance, the sensitivity index $S_i \in [0,1]$. Another popular measure of the contribution of X_i to the variance of Y is the total effect (first order and interactions) index S_{Ti} first introduced in Homma and Saltelli (1996), namely:

$$S_{Ti} = \frac{\mathrm{E}_{\boldsymbol{X}_{\sim i}}\left[\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]\right]}{\mathrm{Var}\left[Y\right]} = 1 - \frac{\mathrm{Var}_{\boldsymbol{X}_{\sim i}}\left[\mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right]\right]}{\mathrm{Var}\left[Y\right]} \tag{2.14}$$

where $\mathrm{E}_{X_i}\left[\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]\right]$ is the expected variance that would be left if all input variables but X_i were known. According to Eq. (2.14), S_{Ti} can also been expressed using the quantity $\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right]\right]$ which is the expected reduction of the variance of Y if all the terms containing X_i were known.

2.3.1.2 Sobol' decomposition

The indices S_i (or S_{1i}) and S_{Ti} are often referred to as *Sobol first order indices* and *Sobol total indices*. Sobol' (1993) introduces these indices by decomposing the model $\mathcal{M}$ in a sum of functions of increasing dimension . Let us first consider that the input variables X_i are independent and uniform over $[0,1]$. Providing $\mathcal{M}$ is square-integrable in $[0,1]^n$, the model admits a unique decomposition:

$$\mathcal{M}(\boldsymbol{x}) = \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(x_i) + \sum_{1 \leqslant i < j \leqslant n} \mathcal{M}_{i,j}(x_i, x_j) + \ldots + \mathcal{M}_{i,\ldots,n}(x_1, \ldots, x_n) \tag{2.15}$$

where $\mathcal{M}_0$ is a constant and where the other functions verify:

$$\int_0^1 \mathcal{M}_{i_1,\ldots,i_s}\left(x_{i_1}, \ldots, x_{i_s}\right) \mathrm{d}x_{i_k} = 0 \tag{2.16}$$

and:

$$\int_0^1 \mathcal{M}_{i_1,\ldots,i_s}\left(x_{i_1}, \ldots, x_{i_s}\right) \mathcal{M}_{j_1,\ldots,j_t}\left(x_{j_1}, \ldots, x_{j_t}\right) \mathrm{d}x = 0 \tag{2.17}$$

$\forall k \in \{1, \ldots, n\}$ and $\{i_1, \ldots, i_s\} \subseteq \{1, \ldots, n\}$. The integral of a component of the decomposition over one of its variable x_{i_k} is zero and two components are orthogonal if at least

one variable is not shared. According to Eqs. (2.16) and (2.17), one obtains:

$$\begin{aligned}
\mathcal{M}_0 &= \int_0^1 \mathcal{M}(\boldsymbol{x})\mathrm{d}x \\
\mathcal{M}_i(x_i) &= \int_0^1 \mathcal{M}(\boldsymbol{x})\mathrm{d}x_{\sim i} - \mathcal{M}_0 \\
\mathcal{M}_{i,j}(x_i, x_j) &= \int_0^1 \mathcal{M}(\boldsymbol{x})\mathrm{d}x_{\sim i,j} - \mathcal{M}_0 - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j) \\
&\dots \\
\mathcal{M}_{1,\dots,n}(x_1, \dots, x_n) &= \mathcal{M}(\boldsymbol{x}) - \mathcal{M}_0 - \sum_{i=1}^{n} \mathcal{M}_i(x_i) - \dots \\
&\quad - \sum_{1 \leqslant i_1 < \dots < i_{n-1} \leqslant n} \mathcal{M}_{i_1,\dots,i_{n-1}}(x_{i_1}, \dots, x_{i_{n-1}})
\end{aligned} \tag{2.18}$$

where the last component $\mathcal{M}_{i_1,\dots,i_n}(x_1, \dots, x_n)$ verifies the decomposition Eq. (2.15). Consequently, the variance of Y can be decomposed according the following theorem.

Theorem 3 *The variance of the model described in Eq. (2.15) can be decomposed as:*

$$\mathrm{Var}\left[Y\right] = \sum_{i=1}^{n} V_i + \sum_{1 \leqslant i < j \leqslant n} V_{ij} + \dots + V_{1\dots n} \tag{2.19}$$

where:

$$\begin{aligned}
V_i &= \mathrm{Var}\left[\mathrm{E}\left[Y|X_i\right]\right] \\
V_{ij} &= \mathrm{Var}\left[\mathrm{E}\left[Y|X_i, X_j\right]\right] - V_i - V_j \\
&\dots \\
V_{1\dots n} &= \mathrm{Var}\left[Y\right] - \sum_{i=1}^{n} V_i - \sum_{1 \leqslant i < j \leqslant n} V_{ij} - \dots - \sum_{1 \leqslant i_1, \dots, i_{n-1} \leqslant n} V_{i_1 \dots i_{n-1}}
\end{aligned} \tag{2.20}$$

The components of the decomposition of the variance of Y are the variances of the components of the decomposition of $Y = \mathcal{M}(\boldsymbol{X})$:

$$V_{i_1\dots,i_s} = \mathrm{Var}\left[\mathcal{M}_{i_1\dots i_s}\left(x_{i_1}, \dots x_{i_s}\right)\right] \quad \{i_1, \dots, i_s\} \subseteq \{1, \dots, n\} \tag{2.21}$$

Thus, the decomposed effects of the variables (or groups of variables) resulting from Eq. (2.15) are well transmitted to the variance decomposition in Eq. (2.19).

In this subsection, sensitivity indices for a model $Y = \mathcal{M}(\boldsymbol{X})$ have been defined. In the next subsection, attention will be given on the computational aspects of GSA.

2.3.2 Computational aspects

The computation of sensitivity indices, and more particularly the evaluation of the conditional moments, involves a very high number of calls to the numerical model $\mathcal{M}$, let us say 10^4 to 10^6. Because one evaluation of $\mathcal{M}$ may last more than a second in engineering applications, there is a need for new methods to compute these indices more efficiently.

2.3.2.1 Computing sensitivity indices

The most intuitive way of computing Sobol' sensitivity indices consists in a *brute force Monte Carlo* approach. For each index S_i given in Eq. (2.13), one has to evaluate as many conditional expectations $\mathrm{E}[Y|X_i]$ as it is needed to estimate their variance. Consequently, this technique leads to a catastrophically slow convergence.

A summary review of the most efficient ways to compute both first order and total indices is proposed in Saltelli et al. (2010). Let us introduce two independent sampling matrices $\mathbf{A}$ and $\mathbf{B}$ with components a_{ji} and b_{ji} where $i = 1, \ldots, n$ and $j = 1, \ldots, N$, denotes j^{th} realization of the i^{th} input variable. Let us now denote by $\mathbf{A}_{\mathbf{B}}^{(i)}$ (resp. $\mathbf{B}_{\mathbf{A}}^{(i)}$) the sampling matrix $\mathbf{A}$ (resp. $\mathbf{B}$) where the i^{th} column as been replaced by the one from $\mathbf{B}$ (resp. $\mathbf{A}$) as shown in Eq. (2.23).

$$\mathbf{A} = \begin{array}{c} \\ 1 \\ \vdots \\ j \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{ccccc} X_1 & \ldots & X_i & \ldots & X_n \end{array} \\ \begin{pmatrix} a_{11} & \ldots & a_{1i} & \ldots & a_{1n} \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ a_{j1} & \ldots & a_{ji} & \ldots & a_{jn} \\ \vdots & \ldots & \vdots & \ldots & \vdots \\ a_{N1} & \ldots & a_{Ni} & \ldots & a_{Nn} \end{pmatrix} \end{array} \tag{2.22}$$

$$\mathbf{A}_{\mathbf{B}}^{(i)} = \begin{array}{c} \\ 1 \\ \vdots \\ j \\ \vdots \\ N \end{array} \begin{array}{c} \begin{array}{ccccccc} X_1 & \ldots & X_{i-1} & X_i & X_{i+1} & \ldots & X_n \end{array} \\ \begin{pmatrix} a_{11} & \ldots & a_{1i-1} & b_{1i} & a_{1i+1} & \ldots & a_{1n} \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ a_{j1} & \ldots & a_{ji-1} & b_{ji} & a_{ji+1} & \ldots & a_{jn} \\ \vdots & \ldots & \vdots & \vdots & \vdots & \ldots & \vdots \\ a_{N1} & \ldots & a_{Ni-1} & b_{Ni} & a_{Ni+1} & \ldots & a_{Nn} \end{pmatrix} \end{array} \tag{2.23}$$

The vector $\mathcal{M}(\boldsymbol{A})$ denotes the output sample $\{\mathcal{M}(\boldsymbol{A})_j \equiv \mathcal{M}(A_j),\ j = 1, \ldots, N\}$ where A_j is the j^{th} row of the matrix $\boldsymbol{A}$. Then, the first order index can be computed from the couple of matrices $\mathbf{A}$, $\mathbf{B}_{\mathbf{A}}^{(i)}$ (resp. $\mathbf{B}$, $\mathbf{A}_{\mathbf{B}}^{(i)}$):

$$\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|X_i\right]\right] = \frac{1}{N}\sum_{j=1}^{N} \mathcal{M}(\mathbf{A})_j \, \mathcal{M}(\mathbf{B}_{\mathbf{A}}^{(i)})_j - \mathcal{M}_0^2 \tag{2.24}$$

The computation of the total index S_{Ti} reads:

$$\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right]\right] = \frac{1}{N}\sum_{j=1}^{N} \mathcal{M}(\mathbf{A})_j \, \mathcal{M}(\mathbf{A}_{\mathbf{B}}^{(i)})_j - \mathcal{M}_0^2 \tag{2.25}$$

where $\mathcal{M}_0$ is the expected value of Y, namely:

$$\mathcal{M}_0 = \mathrm{E}\left[Y\right] \approx \frac{1}{N}\sum_{j=1}^{N} \mathcal{M}(\boldsymbol{X}_j) \tag{2.26}$$

Prooves of Eqs. (2.24) and (2.25) can be found in Saltelli et al. (2010).

Improvements of the estimator of S_i in Eq. (2.24) have been proposed by Saltelli (2002) and Sobol' et al. (2007) where:

$$\mathrm{Var}_{X_i}\left[\mathrm{E}\left[Y|X_i\right]\right] = \frac{1}{N}\sum_{j=1}^{N}\mathcal{M}(\mathbf{A})_j\left(\mathcal{M}(\mathbf{B}_{\mathbf{A}}^{(i)})_j - \mathcal{M}(\mathbf{B})_j\right) \tag{2.27}$$

The estimator of S_{Ti} in Eq. (2.25) has also been numerically improved in Sobol' (2007) and now reads:

$$\mathrm{Var}_{\boldsymbol{X}_{\sim i}}\left[\mathrm{E}\left[Y|\boldsymbol{X}_{\sim i}\right]\right] = \mathrm{Var}\left[Y\right] - \frac{1}{N}\sum_{j=1}^{N}\mathcal{M}(\mathbf{A})_j\left(\mathcal{M}(\mathbf{A})_j - \mathcal{M}(\mathbf{A}_{\mathbf{B}}^{(i)})_j\right) \tag{2.28}$$

Finally, Jansen et al. (1994) and Jansen (1999) offered alternative estimators for S_i and S_{Ti}. The so-called *Jansen's formulae* respectively read:

$$\mathrm{Var}\left[\mathrm{E}\left[Y|X_i\right]\right] = \mathrm{Var}\left[Y\right] - \frac{1}{2N}\sum_{j=1}^{N}\left(\mathcal{M}(\mathbf{B})_j - \mathcal{M}(\mathbf{A}_{\mathbf{B}}^{(i)})_j\right)^2 \tag{2.29}$$

and:

$$\mathrm{E}\left[\mathrm{Var}\left[Y|\boldsymbol{X}_{\sim i}\right]\right] = \frac{1}{2N}\sum_{j=1}^{N}\left(\mathcal{M}(\mathbf{A})_j - \mathcal{M}(\mathbf{A}_{\mathbf{B}}^{(i)})_j\right)^2 \tag{2.30}$$

More Recently, Janon et al. (2012) proposed a new estimator T_N of S_i. Let us first redefine $\mathcal{M}$ as a function of two multidimensional random variables $\boldsymbol{X} \in \mathbb{R}^{n_1}$ and $\boldsymbol{Z} \in \mathbb{R}^{n_2}$ so that $n = n_1 + n_2$. It comes:

$$Y = \mathcal{M}\left(\boldsymbol{X}, \boldsymbol{Z}\right) \tag{2.31}$$

$\boldsymbol{X}'$ denotes an independent copy of $\boldsymbol{X}$ and $Y^{\boldsymbol{X}} = \mathcal{M}\left(\boldsymbol{X}, \boldsymbol{Z}'\right)$. The author shows that:

$$S^{\boldsymbol{X}} = \frac{\mathrm{Var}\left[\mathrm{E}\left[Y|\boldsymbol{X}\right]\right]}{\mathrm{Var}\left[Y\right]} = \frac{\mathrm{Cov}\left[Y, Y^{\boldsymbol{X}}\right]}{\mathrm{Var}\left[Y\right]} \tag{2.32}$$

A first estimator $S_N^{\boldsymbol{X}}$ of $S^{\boldsymbol{X}}$ introduced in Homma and Saltelli (1996) reads:

$$S_N^{\boldsymbol{X}} = \frac{\frac{1}{N}\sum_i Y_i Y_i^{\boldsymbol{X}} - \left(\frac{1}{N}\sum_i Y_i\right)\left(\frac{1}{N}\sum_i Y_i^{\boldsymbol{X}}\right)}{\frac{1}{N}\sum_i Y_i^2 - \left(\frac{1}{N}\sum_i Y_i\right)^2} \tag{2.33}$$

This *natural* estimator consists in the simpliest expression of the covariance $\mathrm{Cov}\left[Y, Y^{\boldsymbol{X}}\right]$ and the variance $\mathrm{Var}\left[Y\right]$.

A second estimator $T_N^{\boldsymbol{X}}$ of $S^{\boldsymbol{X}}$ is now presented. The improvement consits in involving the $Y_i^{\boldsymbol{X}}$, $i = 1, \ldots, N$ in the computation of $\mathrm{Var}[Y]$. That way, the estimator of $\mathrm{Var}[Y]$ is expected to perform better than using only the Y_i, $i = 1, \ldots, N$. $T_N^{\boldsymbol{X}}$ reads:

$$T_N^{\boldsymbol{X}} = \frac{\frac{1}{N}\sum_i Y_i Y_i^{\boldsymbol{X}} - \left(\frac{1}{N}\sum_i \left[\frac{Y_i + Y_i^{\boldsymbol{X}}}{2}\right]\right)^2}{\frac{1}{N}\sum_i \left[\frac{Y_i^2 + \left(Y_i^{\boldsymbol{X}}\right)^2}{2}\right] - \left(\frac{1}{N}\sum_i \left[\frac{Y_i + Y_i^{\boldsymbol{X}}}{2}\right]\right)^2} \tag{2.34}$$

The efficiency of these estimators is now studied through the definitions of confidence intervals.

2.3.2.2 Confidence intervals for sensitivity indices

Let us first introduce the notion of *confidence interval*. A confidence interval I_α is a powerful tool to measure the accuracy of an estimator with a confidence degree α, that is the probability that x belongs to interval $[a, b]$ is $\mathbb{P}(a \leqslant x \leqslant b) = 1 - \alpha$. Considering that X has an Gaussian asymptotic behaviour, I_α is defined by:

$$I_\alpha = \left[\bar{x} - \Phi^{-1}\left(\frac{\alpha}{2}\right)\frac{s}{\sqrt{N}}\ ;\ \bar{x} + \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)\frac{s}{\sqrt{N}}\right] \tag{2.35}$$

where $\bar{x}$ and s are the estimators of the expected value and standard deviation of the $N-$sample $\mathcal{X}$. $\Phi^{-1}(q)$ is the q^{th} quantile of the standard Gaussian distribution. For example, a 95% confidence interval of X, that is $\alpha = 5\%$, reads:

$$I_{0.05} = \left[\bar{x} - 1.96\frac{s}{\sqrt{N}}\ ;\ \bar{x} + 1.96\frac{s}{\sqrt{N}}\right] \tag{2.36}$$

More generally, I_α reads:

$$I_\alpha = \left[\bar{x} - t_\alpha\frac{s}{\sqrt{N}}\ ;\ \bar{x} + t_\alpha\frac{s}{\sqrt{N}}\right] \tag{2.37}$$

where t_α is the $\left(\frac{1-\alpha}{2}\right)^{\text{th}}$ quantile of the *asymptotic distribution* of X. Consequently, the issue consists in identifying the asymptotic distribution of the estimator to be studied.

Since different estimators have been proposed to compute the indices S_i and S_{Ti}, it is gainful to know which one performs best in terms accuracy to number of calls to the model ratio. Several attempts have been proposed to define confidence interval for sensitivity indices estimators. Martinez (2011) identifies the sensitivity index S_i as a linear

correlation coefficient:

$$S_N^{\boldsymbol{X}} = \frac{\mathrm{Var}\left[\mathrm{E}\left[Y|\boldsymbol{X}\right]\right]}{\mathrm{Var}\left[Y\right]} \tag{2.38}$$

$$= \frac{\mathrm{Cov}\left[Y, Y^{\boldsymbol{X}}\right]}{\sqrt{\mathrm{Var}\left[Y\right]\mathrm{Var}\left[Y^{\boldsymbol{X}}\right]}} \tag{2.39}$$

$$= \rho\left(Y, Y^{\boldsymbol{X}}\right) \tag{2.40}$$

Thus, the methods for computing the confidence interval of a linear correlation coefficient can be applied. Let us denote by $\hat{\rho}_N$ the estimator of $\rho\left(Y, Y^{\boldsymbol{X}}\right)$ from a bivariate $N-$sample $[\mathcal{Y}, \mathcal{Y}^{\boldsymbol{X}}]$. According to the *Fisher transformation* (Fisher, 1915) z_N of $\hat{\rho}_N$, the asymptotic behaviour of $\hat{\rho}_N$ reads:

$$z_N = \frac{1}{2}\ln\left(\frac{1+\hat{\rho}_N}{1-\hat{\rho}_N}\right) = \tanh^{-1}\left(\hat{\rho}_N\right) \sim \mathcal{N}\left(\frac{1}{2}\ln\left(\frac{1+\rho\left(Y, Y^{\boldsymbol{X}}\right)}{1-\rho\left(Y, Y^{\boldsymbol{X}}\right)}\right), \frac{1}{N-3}\right) \tag{2.41}$$

Thus, a $\alpha\%$ confidence interval of $\hat{\rho}_N$, that is the first order index $S_N^{\boldsymbol{X}}$ is given by:

$$I_\alpha = \left[\tanh\left(z_N - \frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{N-3}\right) \;;\; \tanh\left(z_N + \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{N-3}\right)\right] \tag{2.42}$$

$$= \left[\tanh\left(\frac{1}{2}\ln\left(\frac{1+S_N^{\boldsymbol{X}}}{1-S_N^{\boldsymbol{X}}}\right) - \frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{N-3}\right) \;;\; \tanh\left(\frac{1}{2}\ln\left(\frac{1+S_N^{\boldsymbol{X}}}{1-S_N^{\boldsymbol{X}}}\right) + \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{N-3}\right)\right] \tag{2.43}$$

In the same way, a confidence interval for the total index $S_N^{T\boldsymbol{X}}$ reads:

$$I_\alpha = \left[1 - \tanh\left(\frac{1}{2}\ln\left(\frac{1+S_N^{\boldsymbol{Z}}}{1-S_N^{\boldsymbol{Z}}}\right) - \frac{\Phi^{-1}\left(\frac{\alpha}{2}\right)}{N-3}\right); 1 - \tanh\left(\frac{1}{2}\ln\left(\frac{1+S_N^{\boldsymbol{Z}}}{1-S_N^{\boldsymbol{Z}}}\right) - \frac{\Phi^{-1}\left(1-\frac{\alpha}{2}\right)}{N-3}\right)\right] \tag{2.44}$$

This asymptotic behaviour holds only if Y and $Y^{\boldsymbol{X}}$ have a Gaussian distribution, as explained in subsection 1.3.6.

To circumvent this issue, Janon et al. (2012) introduces two asymptotic distributions for the sensitivity indices estimators in Eqs. (2.33) and (2.34). The estimators are consistent, thanks to the strong law of large numbers and noticing that $\mathrm{E}\left[Y^{\boldsymbol{X}}\right] = \mathrm{E}\left[Y\right]$ and $\mathrm{Var}\left[Y^{\boldsymbol{X}}\right] = \mathrm{Var}\left[Y\right]$, namely:

$$S_N^{\boldsymbol{X}} \underset{N\to\infty}{\longrightarrow} S^{\boldsymbol{X}} \tag{2.45}$$

$$T_N^{\boldsymbol{X}} \underset{N\to\infty}{\longrightarrow} S^{\boldsymbol{X}} \tag{2.46}$$

and their asymptotic distributions respectively read:

$$\sqrt{N}\left(S_N^{\boldsymbol{X}} - S^{\boldsymbol{X}}\right) \sim \mathcal{N}\left(0, \frac{\mathrm{Var}\left[(Y - \mathrm{E}[Y])\left[\left(Y^{\boldsymbol{X}} - \mathrm{E}[Y]\right) - S^{\boldsymbol{X}}(Y - \mathrm{E}[Y])\right]\right]}{\mathrm{Var}[Y]^2}\right) \tag{2.47}$$

and:

$$\sqrt{N}\left(T_N^{\boldsymbol{X}} - S^{\boldsymbol{X}}\right) \sim \mathcal{N}\left(0, \sigma_T^2\right) \tag{2.48}$$

with:

$$\sigma_T^2 = \frac{\mathrm{Var}\left[(Y - \mathrm{E}[Y])\left(Y^{\boldsymbol{X}} - \mathrm{E}[Y]\right) - \frac{S^{\boldsymbol{X}}}{2}\left((Y - \mathrm{E}[Y])^2 + \left(Y^{\boldsymbol{X}} - \mathrm{E}[Y]\right)^2\right)\right]}{\mathrm{Var}[Y]^2} \tag{2.49}$$

Despite encouraging results in terms of convergence rate, the estimator T_N appears to be slightly biased (Owen, 2012b) because identical realizations are contained in both Y and $Y^{\boldsymbol{X}}$. Therefore, a bias corrected version of the estimator is proposed. Owen (2012a) recently introduced a new computation scheme for the estimation of Sobol' sensitivity indices that now makes use of three independent sampling matrices instead of two. This new estimator appears to perform well for small Sobol' indices but is outperformed by the 2-matrice scheme for large Sobol' indices.

Global sensitivity analysis methods such as ANOVA are well-established for models with independent inputs. When the input parameters are correlated, other techniques have to be developed.

2.4 Sensitivity analysis for models with correlated inputs

From now on in this chapter, attention is given to GSA methods for models with *correlated* input parameters. Let us first explain why the dependence structure of the parameters influences the results of global sensitivity analysis.

2.4.1 Problem statement

When the input variables of a model $\mathcal{M}$ are independent, the Sobol' (1993) functional decomposition of the variance output allows one to identify the contribution of each input parameters or group of input parameters to the variance output. For instance, in case of an additive model, that is in the form of :

$$Y = \mathcal{M}(\boldsymbol{X}) = \sum_{1=1}^{n} a_i X_i \tag{2.50}$$

the variance of the model response is entirely decomposed into first order Sobol' indices. This case is illustrated in Figure 2.1 (a) with $n = 3$, $\boldsymbol{a} = [3, 2, 1]$ and $X_i \sim \mathcal{N}(0., 1.)$,

$i = 1, \ldots, n$. When the input parameters are no longer independent, let us take for instance the following linear correlation matrix :

$$\mathbf{R} = \begin{pmatrix} 1 & 0.8 & 0.4 \\ 0.8 & 1 & 0.7 \\ 0.4 & 0.7 & 1 \end{pmatrix} \tag{2.51}$$

the Sobol' sensitivity indices are still *computable* but it becomes difficult to interpret the results as shown in Figure 2.1 (b). Indeed, it is hard to know if the contribution of a variable X_i is due to its importance in the model structure or to its genuine correlation with other influent variables $X_{j,j\neq i}$.

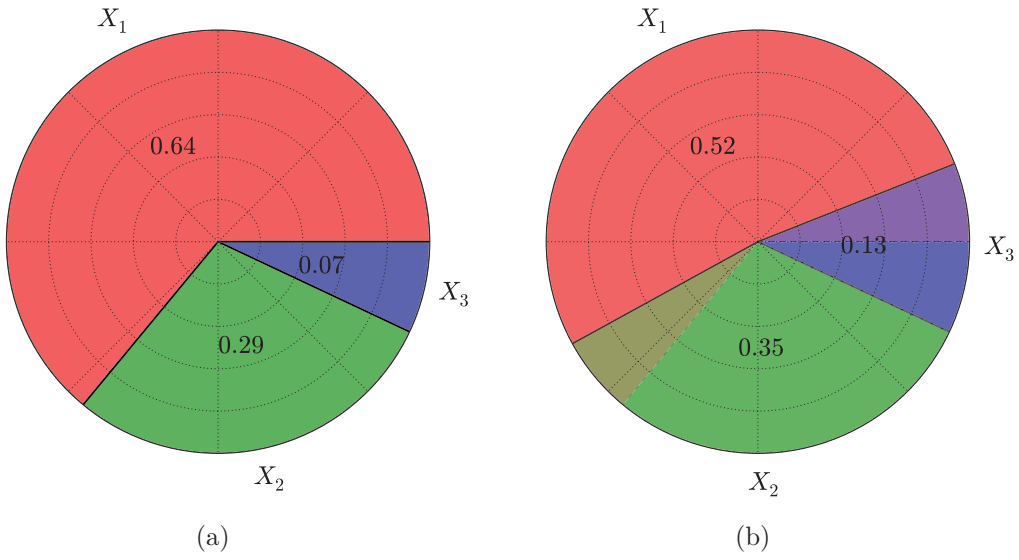


Figure 2.1: *Variance decomposition of an additive model with independent (a) and correlated input variables (b). In the second case, the dependence structure of the input random vector implies a mixture of the contributions.*

When the dependence structure of the input random vector is ignored, the results of a GSA performed with a classic method can be misinterpreted and may lead to bad decisions. Therefore, new techniques have been developed. They are reviewed in the next subsection.

2.4.2 Short review on existing methods

Early work on this topic is due to Iman and Hora (1990). In order to study uncertainty propagation in system fault trees, three *measures of contribution* have been proposed to quantify how low levels events may influence both the frequency of the top event and the uncertainty in this frequency. These measures are respectively based on the expected

reduction in variance of the top event frequency and log frequency and on shifts in the quantiles of the top event frequency.

The first GSA method for models with correlated inputs is due to Chun et al. (2000). An *importance measure* based on a distance metric between cumulative distribution functions is defined. The general form of a metric distance D_k between two functions f_1 and f_2, referred to as the *Minkowski class of distance*, is:

$$D_k\left(f_1, f_2\right) = \left(\sum_{\boldsymbol{x}\in\boldsymbol{X}} \left|f_1(x) - f_2(x)\right|^k\right)^{\frac{1}{k}} \tag{2.52}$$

When $k = 2$, D is the *Euclidean distance*. The so-called *Chun-Han-Tak index* of the variable X_i, denoted by CHT_i, reads:

$$CHT_i = \frac{MD(i:o)}{\mathrm{E}\left[Y^o\right]} \tag{2.53}$$

where:

$$MD(i:o) = \left(\int_0^1 \left[y_p^i - y_p^o\right]^2 \mathrm{d}p\right)^{\frac{1}{2}} \tag{2.54}$$

is the quantile-based metric distance measure between the *base case* and the *sensitivity case*. y_p^o and y_p^i are respectively the p^{th} quantiles of the base case, that is the distribution of the output computed with all inputs, and of the sensitivity case, that is the distribution of the output when the input variable X_i is fixed. $\mathrm{E}\left[Y^o\right]$, the mean of the base case distribution, is introduced for normalization and adimensionality. $MD(i:o)$ represents the Euclidian metric distance between two CDFs normalized by the mean of the base case distribution.

In Jacques (2005), an alternative method consisting in building multidimensional indices is exposed. Starting form the observation that in a $n-$dimensional random vector, the input parameters are often correlated two by two or three by three, one can decomposed the input random vector into *independent subvectors*, for instance:

$$\boldsymbol{X} \equiv \left[\left[X_1, X_2\right], \left[X_3, X_4, X_5\right], \ldots, \left[X_{n-1}, X_n\right]\right] \tag{2.55}$$

where the independent subvectors contains correlated components. This approach is related to the concept of *composed copula* described in 1.5.3.6, that is the product (independence) of multidimensional copulas (dependence). Unfortunately, this solution does not allow one to decompose the contribution of the variables inside a subvector.

Later, Xu and Gertner (2008) introduced a method based on the variance decomposition where the principle is to seperate the shares of variance reduction due to *uncorrelated* and *correlated* effects. Considering the share of variance of the model output Y due to X_i is V_i, it can be decomposed as:

$$V_i = V_i^U + V_i^C \tag{2.56}$$

where V_i^U is the share of variance due to the variable X_i itself and V_i^C is the share of variance due the correlation between X_i and $X_{j,j\neq i}$. If the model is approximately linear,

the share of variance of $\mathrm{Var}\,[Y]$ due to X_i can be derived by regressing Y on X_i only:

$$y = \theta_0 + \theta_i\, x_i + e \tag{2.57}$$

and consequently:

$$\hat{V}_i = \frac{1}{N-1} \sum_{j=1}^{N} \left(\hat{y}_i^{(j)} - \bar{y} \right)^2 \tag{2.58}$$

with:

$$\hat{y}_i^{(j)} = \hat{\theta}_0 + \hat{\theta}_i\, x_i^{(i)} \tag{2.59}$$

Since X_i contains both uncorrelated and correlated effects, the partial variance in Eq. (2.58) is the total partial variance due to X_i. In order to seperate the different effects, the regression of X_i to all the other input parameters X_j's, $j \neq i$ is processed. The residuals z_i of this regression defined by:

$$\hat{z}_i = x_i - \hat{x}_i = x_i - \left(\hat{\eta}_0 + \sum_{j, j \neq i} \hat{\eta}_j\, x_j \right) \tag{2.60}$$

where the $\hat{\eta}_j$ are the least-square estimates of the regression coefficients of X_i on the X_j's, $j \neq i$. Then, the partial variance representing the uncorrelated effects of X_i reads:

$$\hat{V}_i^U = \frac{1}{N-1} \sum_{j=1}^{N} \left(\hat{y}_{\sim i}^{(j)} - \bar{y} \right)^2 \tag{2.61}$$

where $\hat{y}_{\sim i}^{(j)} = \hat{r}_0 + \hat{r}_i\, \hat{z}_i$ is the model response regressed on the residuals in Eq. (2.60). The partial variance due to correlation between X_i and all the other parameters, denoted by V_i^C is:

$$\hat{V}_i^C = \hat{V}_i - \hat{V}_i^U \tag{2.62}$$

According to Eqs. (2.58), (2.61) and (2.62), the partial variance of Y due to X_i can be decomposed by the uncorrelated and correlated effects of the parameter X_i. Thus, using the ratio between the partial variances and the total variance, a collection of three first-order sensitivity indices is defined, namely:

$$S_i = \frac{V_i}{\mathrm{Var}\,[Y]} \tag{2.63}$$

$$S_i^U = \frac{V_i^U}{\mathrm{Var}\,[Y]} \tag{2.64}$$

$$S_i^C = \frac{V_i^C}{\mathrm{Var}\,[Y]} \tag{2.65}$$

S_i, S_i^U and S_i^C respectively represent the total contribution, the uncorrelated contribution and the correlated contribution of X_i to the variance of the model output Y. The major limitation of this approach is the hypothesis of linearity of $\mathcal{M}$ in Eq. (2.57). The results remain interpretable as long as $\mathcal{M}$ is approximately linear but they cannot be used in all cases.

Recent improvement in GSA for models with correlated inputs have been proposed Mara and Tarantola (2012). Let us first write the joint PDF $f_{\boldsymbol{X}}$ of three correlated input variables X_1, X_2 and X_3 in terms of conditional PDFs:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_1(x_1) f_{2|1}(x_2|x_1) f_{3|12}(x_3|x_1, x_2) \tag{2.66}$$

where $f_{2|1}$ is the conditional marginal PDF of X_2 knowing X_1. X_1 and X_2 are dependent because $\mathrm{E}[X_2|X_1] \neq \mathrm{E}[X_2]$. Consequently:

$$f(x_1, x_{2-1}, x_{3-12}) = f_1(x_1) f_{2-1}(x_{2-1}) f_{3-12}(x_{3-12}) \tag{2.67}$$

where:

$$x_{2-1} = x_2 - \mathrm{E}[x_2|x_1] \quad \text{and} \quad x_{3-12} = x_3 - \mathrm{E}[x_3|x_1, x_2] \tag{2.68}$$

This transformation is the so-called *Rosenblatt transformation* (Rosenblatt, 1952) where the random vector $[X_1, X_{2-1}, X_{3-12}]$ is independent. Then, the so-called ANOVA decomposition can be applied. and the following sensitivity indices are computed:

$$\bar{S}_1 = \frac{\mathrm{Var}[\mathrm{E}[Y|X_1]]}{\mathrm{Var}[Y]} = S_1 \tag{2.69}$$

$$\bar{S}_2 = \frac{\mathrm{Var}[\mathrm{E}[Y|X_{2-1}]]}{\mathrm{Var}[Y]} = S_{2-1} \tag{2.70}$$

$$\bar{S}_3 = \frac{\mathrm{Var}[\mathrm{E}[Y|X_{3-12}]]}{\mathrm{Var}[Y]} = S_{3-12} \tag{2.71}$$

The index $\bar{S}_1$ is equivalent to the first order Sobol' index S_1 because it corresponds to the *full marginal contribution* of X_1 to the variance of Y. On the contrary, the indices $\bar{S}_2$ and $\bar{S}_3$ respectively represents the contribution of X_2 knowing X_1 and the contribution X_3 knowing X_1 and X_2. In other words, it corresponds to the *conditional marginal contributions* of X_2 and X_3. The Rosenblatt transformation can be applied in a different order, for instance by writing $f_{\boldsymbol{X}}(\boldsymbol{x}) = f_2(x_2) f_{3|2}(x_3|x_2) f_{1|23}(x_1|x_2, x_3)$, in order to get the full marginal contribution of X_2. Thereby, all the sensitivity indices can be calculated. The main drawback remains that getting all the full marginal contributions requires as many analyses as the number n of input parameters.

2.4.3 Conclusion

Along this short review, different GSA methods for dealing with models with correlated inputs have been presented. The two main ideas are, on the one hand, to build an *importance measure* that describes the sensitivity of the model response Y to the input variable X_i and, on the other hand, an attempt to generalize the Sobol' decomposition to models with correlated inputs and distinguish the uncorrelated and correlated contributions. Two methods, one of each kind, that appeared to be more effective are presented in details in the next two sections.

2.5 A distribution-based method

An *importance measure* characterizes the modification in the model output distribution when an input parameter is perfectly known (deterministic). This approach has already been studied in Iman and Hora (1990) and Chun et al. (2000), where a metric distance between two distributions has been defined. Another solution would have been to use the so-called Kullback-Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between two functions p and q of x reads :

$$D_{KL}(p\|q) = \int_{\mathrm{D}_X} p(x) \log\left(\frac{p(x)}{q(x)}\right) \mathrm{d}x \tag{2.72}$$

According to Eq. (2.72), D_{KL} is a divergence and not a distance because it is not symmetrical. This idea has been originally developed in Park and Ahn (1994). One limitation is also that when $q(x) = 0$, the logarithm is undefined and it makes the use of D_{KL} more difficult. The method that is now presented proposes a new importance measure.

2.5.1 Principle

This uncertainty importance measure has been first introduced in Borgonovo (2007). The objective of this work was to develop an importance measure that would be adapted to any kind of dependence structure of the input random vector. In order to alleviate the numerical cost of the analysis, this important measure is *moment-free*, that is, it does not require any computation of the moments of the model output $Y = \mathcal{M}(\boldsymbol{X})$. The principle here is to quantify how fixing one input parameter X_i to a value x_i^* is affecting the entire distribution of Y and not only its variance, in contrast to the so-called ANOVA-based methods. Therefore, the modification of the output PDF f_Y is represented by its *shift* $s(x_i^*)$, namely:

$$s(x_i^*) = \int_{\mathrm{D}_Y} \left| f_Y(y) - f_{Y|X_i=x_i^*}(y) \right| \mathrm{d}y \tag{2.73}$$

where $f_{Y|X_i=x_i^*}$ is the PDF of Y conditional to $X_i = x_i^*$. The shift between f_Y and $f_{Y|X_i=x_i^*}$ corresponds to the area between the two PDFs illustrated in Figure 2.2.

Eq. (2.73) shows that the shift $s(x_i^*)$ strongly depends on the value of x_i^*. In order to take the whole range of values X_i can take into account, the *expected shift*, namely:

$$\begin{aligned} \mathrm{E}\left[s(x_i)\right] &= \int_{\mathrm{D}_{X_i}} f_{X_i}(x_i)\, s(x_i)\, \mathrm{d}x_i \\ &= \int_{\mathrm{D}_{X_i}} f_{X_i}(x_i) \left[\int_{\mathrm{D}_Y} \left| f_Y(y) - f_{Y|X_i}(y) \right| \mathrm{d}y \right] \mathrm{d}x_i \end{aligned} \tag{2.74}$$

is defined. Then the moment-free important measure δ_i is given by the following definition.

Definition 9 *The quantity δ_i, namely:*

$$\delta_i = \frac{1}{2}\mathrm{E}\left[s(x_i)\right] \tag{2.75}$$

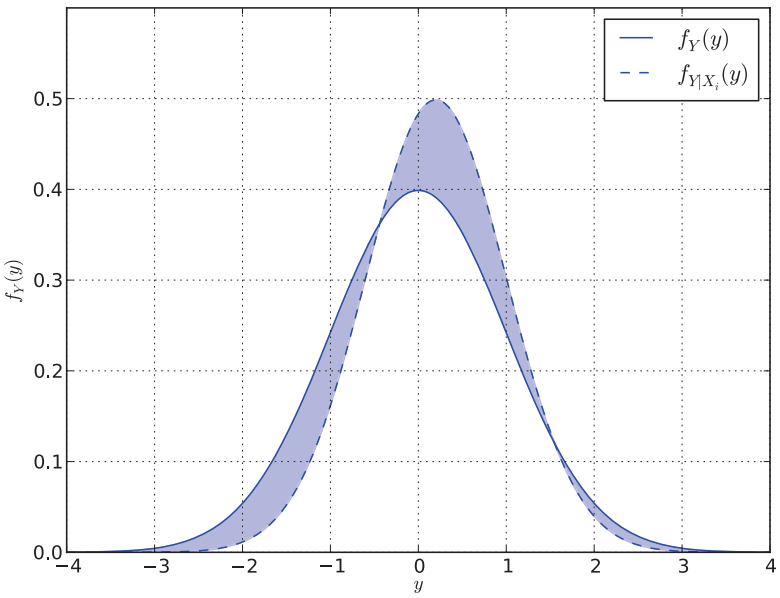


Figure 2.2: *The* shift *between the unconditional PDF f_Y and the PDF $f_{Y|X_i}$ conditional to X_i is represented by the blue area between the distributions.*

is a moment-free important measure of the sensitivity of the model output Y to the input parameter X_i. δ_i is the normalized expected shift of the PDF of Y due to X_i.

The quantity $\frac{1}{2}$ is introduced for normalization. The area under one distribution over its whole support is 1. Consequently, in the case of no intersection between the unconditional and conditional output PDF, the shift would be 2. One also defines a 2−dimensional sensitivity index $\delta_{i,j}$ representing the joint contribution of a pair of variables (X_i, X_j) to the modification of the output PDF:

$$\delta_{i,j} = \frac{1}{2} \int_{\mathrm{D}_{X_i} \times \mathrm{D}_{X_j}} f_{X_i,X_j}(x_i, x_j) \left[\int_{\mathrm{D}_Y} \left| f_Y(y) - f_{Y|X_i,X_j}(y) \right| \mathrm{d}y \right] \mathrm{d}x_i \, \mathrm{d}x_j \tag{2.76}$$

This definition is extensible for a higher number of variables. The multi-dimensional index definition is presented below.

Definition 10 *Let us consider a m−dimensional subvector $\boldsymbol{X}' = [X_{i_1}, \ldots, X_{i_m}]$ of $\boldsymbol{X}$, $\boldsymbol{X}' \subset \boldsymbol{X}$, $m < n$. The normalized expected shift in f_Y due to $\boldsymbol{X}'$ reads:*

$$\begin{aligned} \delta_{i_1,\ldots,i_m} &= \frac{1}{2} \mathrm{E}\left[s(\boldsymbol{X}') \right] \\ &= \frac{1}{2} \int_{\mathrm{D}_{\boldsymbol{X}'}} f_{\boldsymbol{X}'}(\boldsymbol{x}') \left[\int_{\mathrm{D}_Y} \left| f_Y(y) - f_{Y|\boldsymbol{X}'}(y) \right| \mathrm{d}y \right] \mathrm{d}\boldsymbol{x}' \end{aligned} \tag{2.77}$$

with $\mathrm{d}\boldsymbol{x}' = \mathrm{d}x_{i_1} \times \ldots \times \mathrm{d}x_{i_m}$.

In the same way, the conditional δ index is given by the following definition.

Definition 11 *The sensitivity measure of Y to X_j conditionally to X_i reads:*

$$\delta_{j|i} = \frac{1}{2} \int_{\mathrm{D}_{X_i} \times \mathrm{D}_{X_j}} f_{X_i}(x_i)\, f_{X_j}(x_j) \left[\int_{\mathrm{D}_Y} \left| f_{Y|X_i}(y) - f_{Y|X_i,X_j}(y) \right| \mathrm{d}y \right] \mathrm{d}x_i\, \mathrm{d}x_j \tag{2.78}$$

$\delta_{j|i}$ represents the sensitivity measure of the output Y to the input X_j when X_i is known.

The properties of the δ_i sensitivity measure derivied from Eqs. (2.75), (2.77) and (2.78) are now described:

1. $0 \leqslant \delta_i \leqslant 1$, the δ_i sensitivity measure is bounded. When $f_Y = f_{Y|X_i}$, $s(X_i)$ is zero. When f_Y and $f_{Y|X_i}$ have no intersection, $s(X_i) = 2$ and $\delta_i = 1$.
2. If Y does not depend of X_i, then $\delta_i = 0$. X_i has no effects on f_Y, $f_Y = f_{Y|X_i}$ and $s(X_i) = 0$.
3. The importance measure of all parameters equals unity : $\delta_{1,2,\ldots,n} = 1$. In this case, $f_{Y|\boldsymbol{X}}$ is a *Dirac function*, $s(\boldsymbol{X}) = 2$ and $\delta_i = 1$.
4. If Y depends on X_i but does not depend on X_j, then $\delta_{i,j} = \delta_i$.
5. The bounds of a bidimensional index $\delta_{i,j}$ are defined by $\delta_i \leqslant \delta_{i,j} \leqslant \delta_i + \delta_{i|j}$.

The δ_i importance measure represents the normalized expected shift in the model output PDF f_Y due to the input parameter X_i. Therefore, evaluating δ_i consists in first approximating the shift, that is the area between the two curves, and then in estimating its expected value. These two steps evaluation scheme might be numerically expensive, especially if the distributions are not known and need to be approximated.

2.5.2 Improvements in the definitions

The δ importance measure introduced in Borgonovo (2007) has been improved in Borgonovo et al. (2011). Considering an open interval $\Omega \in \mathbb{R}$ and two PDFs f and g defined on Ω, the distance between f and g reads:

$$\|f - g\| = \int_{\Omega} |f(x) - g(x)|\, \mathrm{d}x \tag{2.79}$$

Then Eqs. (2.79) and (2.73) are equivalent. A piecewise study of Eq. (2.79) leads to:

$$u(x) = \begin{cases} f(x) - g(x) & x \in \Omega^+ \\ f(x) = g(x) & x \in \Sigma \\ g(x) - f(x) & x \in \Omega^- \end{cases} \tag{2.80}$$

where Σ is the set of points where $f(x) = g(x)$ while Ω^+ and Ω^- are respectively the domains where $f(x) > g(x)$ and $f(x) < g(x)$. Let us now consider the CDFs F and G respectively defined by:

$$F(x) = \int_{-\infty}^{x} f(\xi)\,\mathrm{d}\xi \quad \text{and} \quad G(x) = \int_{-\infty}^{x} g(\xi)\,\mathrm{d}\xi \tag{2.81}$$

The distance defined in Eq. (2.79) can be rewritten:

$$\begin{aligned} \|f - g\| &= 2F\left(\Omega^+\right) - 2G\left(\Omega^+\right) \\ &= 2G\left(\Omega^-\right) - 2F\left(\Omega^-\right) \end{aligned} \tag{2.82}$$

The distance $\|f - g\|$ is equal to twice the probability that $x \in \Omega^+$ under F and the same probability under G. By symmetry, this quantity is also the probability that $x \in \Omega^-$ under G and under F.

Let us now get back to GSA by denoting $\Omega = \mathrm{D}_Y$, $F = F_Y$ and $G = F_{Y|X_i}$. The support of Y can be decomposed in $\mathrm{D}_Y = \left\{\mathrm{D}_Y^+, \mathrm{D}_Y^-\right\}$ with $\mathrm{D}_Y^+ = \left\{y : f_Y(y) > f_{Y|Xi}(y)\right\}$ and $\mathrm{D}_Y^- = \left\{y : f_Y(y) < f_{Y|Xi}(y)\right\}$. As the decomposition $\mathrm{D}_Y = \left\{\mathrm{D}_Y^+, \mathrm{D}_Y^-\right\}$ depends on X_i, the notations D_{Y,X_i}^+ and D_{Y,X_i}^- will be used for convenience. Thus, the shift defined in terms of PDFs in Eq. (2.73) can be written in terms of CDFs:

$$s(X_i = x_i^*) = 2F_Y(\mathrm{D}_{Y,X_i}^+) - 2F_{Y|X_i=x_i^*}(\mathrm{D}_{Y,X_i}^+) \tag{2.83}$$

Consequently, the importance measure δ_i defined previously in Eq. (2.75) also reads:

$$\delta_i = \mathrm{E}_{X_i}\left[F_Y(\mathrm{D}_{Y,X_i}^+) - F_{Y|X_i}(\mathrm{D}_{Y,X_i}^+)\right] \tag{2.84}$$

The difficulty now consists in identifying the set of points $\Sigma = \{y_1, y_2; y_1 < y_2\}$, that is the points at which the two PDFs are crossing themselves as illustrated in Figure 2.3. This will be addressed in details in section 4.3 together with the numerical evaluation of the integrals associated with the expectation.

2.5.3 Conclusion

The δ_i moment-free importance measure represents the normalized expected shift in the model output distribution due to an input parameter X_i. It is defined so that no hypothesis are made neither on the model (linearity for instance) nor on the dependence structure of the input random vector. This makes δ a very general sensitivity index that not only focus on the output variance (the "width" of output PDF) but also on its global shape. This definition has one main drawback, namely the index does not represent a share of variance but is more like a *bounded distance* between the unconditional and conditional distributions. On top of a different scale of values for the indices, the practitioner will also have to deal with a ranking of the most important variables that might differ from the one given by the variance-based methods.

In order to find a link with the ANOVA decomposition, a generalization of the Sobol' decomposition to models with correlated inputs is presented in the next section.

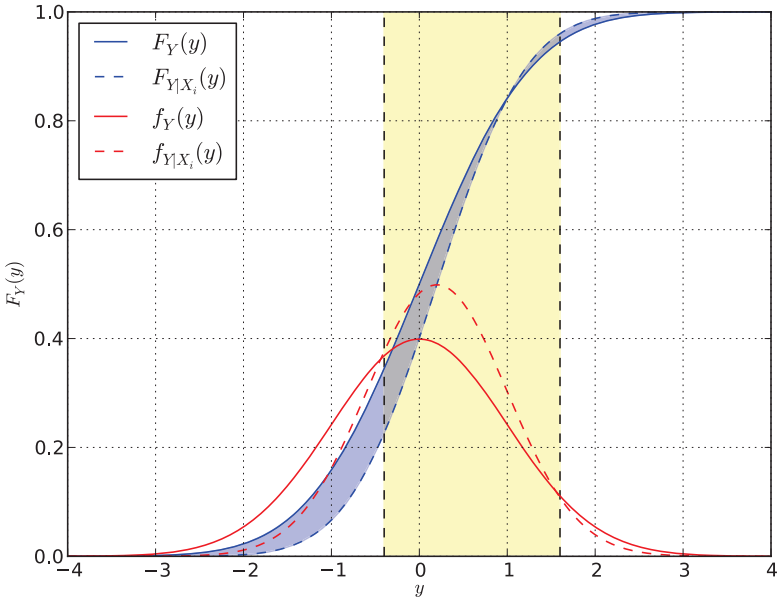


Figure 2.3: *The* shift *between the PDFs f_Y and $f_{Y|X_i}$ (red) is now computed using the CDFs F_Y and $F_{Y|X_i}$ (blue). The domain D_Y^- where $f_Y < f_{Y|X_i}$ is colored in yellow. The values y_1 and y_2 are the abscissa of the vertical black dashed lines.*

2.6 The ANCOVA decomposition

One issue of global sensitivity analysis in engineering for models with correlated inputs (GSA-MCI) is to define quantities that are easily interpretable by the practitioners. This is the case for the variance-based methods because the indices represent the shares of the output variance due to the input parameters. In Jacques (2005) (page 99), the author examines the decomposition of the variance decomposition of the model output when the input variables are no longer independent. In this case, the Sobol' functional decomposition in Eq. (2.15) still holds since the dependence has no influence on the expected values $\mathrm{E}[.]$. The Sobol' decomposition of the variance in Eq. (2.19) is also verified since its last term $V_{1\ldots n}$ is defined as the difference between the variance of Y and the sum of all the lower order variances.

However, the partial variances $V_{i_1\ldots i_s}$ are no longer the variances of the function components $\tilde{V}_{i_1\ldots i_s} = \mathrm{Var}\left[\mathcal{M}_{i_1\ldots i_s}(X_{i_1},\ldots,X_{i_s})\right]$ since the separation of the effects is not transmitted to the decomposition of the variance of Y. Indeed, $\mathcal{M}_{i_1\ldots i_s}(X_{i_1},\ldots,X_{i_s})$ represents the effects of the $s-$dimensional subvector $\boldsymbol{X}_{i_1\ldots i_s} \subseteq \boldsymbol{X}$ that are not taken into account by the effects of the strict subsets $\{X_{i_1},\ldots,X_{i_s}\}$. Thus, $\tilde{V}_{i_1\ldots i_s}$ represents the the share of variance of Y that is due to the interaction of the s variables $X_{i_1},\ldots,X_{i_s}$ but $V_{i_1\ldots i_s}$ longer does. Although the author establishes a link between $\tilde{V}_{i_1\ldots i_s}$ and $V_{i_1\ldots i_s}$, this work keeps investigating the notion of generalized ANOVA for models with correlated input parameters. In this section, a generalization of the ANOVA decomposition for models

with correlated inputs is presented.

2.6.1 Principle

This method has been first introduced in Li and Rabitz (2010) and then rewritten in Chastaing et al. (2012). The objective is to generalize the variance decomposition of the model output for models with correlated input parameters. Let us first consider a model $Y = \mathcal{M}(\boldsymbol{X})$ where $\boldsymbol{X}$ is a $n-$dimensional random vector. No hypothesis are made on its dependence structure. In the same manner as in the independent case, the model $\mathcal{M}$ can be expanded as a sum of functions of increasing dimension:

$$\begin{aligned}\mathcal{M}(\boldsymbol{X}) &= \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(X_i) + \sum_{1\leqslant i<j\leqslant n} \mathcal{M}i,j(X_i, X_j) + \ldots + \mathcal{M}_{1,\ldots,n}(X_1, \ldots, X_n) \\ &= \mathcal{M}_0 + \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\end{aligned} \tag{2.85}$$

where $\mathcal{M}_0$ is the mean $\mathrm{E}[Y]$ and where each functions $\mathcal{M}_{\boldsymbol{u}}$, $\boldsymbol{u} \subseteq \{1, \ldots, n\}$ represents, for any non empty set $\boldsymbol{u}$, the combined contribution of the variables $\boldsymbol{X}_{\boldsymbol{u}}$ to Y.

Let us now rewrite the unconditional variance of the model output Y as follows:

$$\mathrm{Var}[Y] = \mathrm{E}\left[(Y - \mathrm{E}[Y])^2\right] \tag{2.86}$$

$$= \mathrm{E}\left[(Y - \mathcal{M}_0)\left(\sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right)\right] \tag{2.87}$$

$$= \mathrm{Cov}\left[Y, \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right] \tag{2.88}$$

The variance of Y can be written as the covariance of Y and its *functional decompostion* in Eq. (2.85) minus the zero-order term $\mathcal{M}_0$. Thanks to the properties of the covariance, Eq. (2.88) also reads:

$$\mathrm{Var}[Y] = \mathrm{Cov}\left[Y, \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right] \tag{2.89}$$

$$= \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathrm{Cov}\left[Y, \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right] \tag{2.90}$$

Because the left member Y of the covariance also contains the functions $\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})$, Eq. (2.90) also reads:

$$\mathrm{Var}[Y] = \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \mathrm{Cov}\left[Y, \mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right] \tag{2.91}$$

$$= \sum_{\boldsymbol{u}\subseteq\{1,\ldots,n\}} \left[\mathrm{Var}\left[\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right] + \mathrm{Cov}\left[\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}), \sum_{\boldsymbol{v}\subseteq\{1,\ldots,n\}, \boldsymbol{v}\cap\boldsymbol{u}=\emptyset} \mathcal{M}_{\boldsymbol{v}}(\boldsymbol{X}_{\boldsymbol{v}})\right]\right] \tag{2.92}$$

The summands in Eq. (2.91) can be decomposed into the variance of the term $\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})$ which is also contained in Y and the covariance of $\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})$ and the $\mathcal{M}_{\boldsymbol{v}}(\boldsymbol{X}_{\boldsymbol{v}})$, that is Y where $\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})$ has been removed.

This technique is referred to as the *ANCOVA decomposition* (ANalysis of COVAriance) by Li and Rabitz (2010). The partial variances are decomposed into a variance part and a covariance part. When the input variables of the model are independent, the functions $\mathcal{M}_{\boldsymbol{u}}$ and $\mathcal{M}_{\boldsymbol{v}}$ are orthogonal. Consequently, the covariance part is zero and only the variance part remains. Under these conditions, the ANCOVA is equivalent to the ANOVA, which is then a particular case (independence) of the ANCOVA.

2.6.2 ANCOVA-based sensitivity indices

2.6.2.1 Definition

For models with independent input parameters, a single index S_i is used to measure the contribution of the input variable X_i to the variance of Y. The challenge of GSA-MCI is to distinguish which part of the contribution is due to X_i itself and which one is due to its correlation with the other parameters. The ANCOVA decomposition and its variance-covariance separation of the partial variances allows one to separate the uncorrelated and the correlated effects. Let us define the following indices:

$$S_{\boldsymbol{u}} = \frac{\mathrm{Cov}\left[\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}), Y\right]}{\mathrm{Var}\left[Y\right]} \tag{2.93}$$

$$S_{\boldsymbol{u}}^{U} = \frac{\mathrm{Var}\left[\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}})\right]}{\mathrm{Var}\left[Y\right]} \tag{2.94}$$

$$S_{\boldsymbol{u}}^{C} = \frac{\mathrm{Cov}\left[\mathcal{M}_{\boldsymbol{u}}(\boldsymbol{X}_{\boldsymbol{u}}), \sum\limits_{\boldsymbol{v}\subseteq\{1,\ldots,n\},\boldsymbol{v}\cap\boldsymbol{u}=\emptyset} \mathcal{M}_{\boldsymbol{v}}(\boldsymbol{X}_{\boldsymbol{v}})\right]}{\mathrm{Var}\left[Y\right]} \tag{2.95}$$

The first index $S_{\boldsymbol{u}}$ represents the total share of variance of Y due to $X_{\boldsymbol{u}}$. The second index $S_{\boldsymbol{u}}^{U}$ represents the uncorrelated share of variance of Y due to $X_{\boldsymbol{u}}$, that is the *physical* contribution of $X_{\boldsymbol{u}}$. Finally, the third index $S_{\boldsymbol{u}}^{C}$ represents the correlated share of variance of Y due to $X_{\boldsymbol{u}}$, that is the contribution of the correlation of $X_{\boldsymbol{u}}$ with the other input parameters. Due to Eq. (2.92), one gets the following relationship:

$$S_{\boldsymbol{u}} = S_{\boldsymbol{u}}^{U} + S_{\boldsymbol{u}}^{C} \tag{2.96}$$

Due to its definition, $S_{\boldsymbol{u}}^{U}$ is always positive. On the contrary, $S_{\boldsymbol{u}}^{C}$ can be either positive or negative: it depends on the nature of the correlation between $X_{\boldsymbol{u}}$ and the $X_{\boldsymbol{v}}$'s. Therefore, $S_{\boldsymbol{u}}$ can be either positive if the physical contribution is higher than the correlated contribution $S_{\boldsymbol{u}}^{U} > S_{\boldsymbol{u}}^{C}$, or negative in the opposite case $S_{\boldsymbol{u}}^{U} < S_{\boldsymbol{u}}^{C}$, or zero if $S_{\boldsymbol{u}}^{U} = -S_{\boldsymbol{u}}^{C}$.

2.6.2.2 Interpretation: a negative sensitivity index

Sensitivity indices are supposed to represent a share of the variance of the model output. Thus, they are supposed to be positive, or zero. The issue of correlation among the input parameters is that it might couples the effects of the variables. Does one variable have a high contribution because of its physical role in the model $\mathcal{M}$ or because it is strongly correlated to variables with higher contributions? If the correlation might raise the contribution of a variable, it might also take it down : this is how the index $S_{\boldsymbol{u}}^C$ has to be interpreted. It is a corrective term that indicates if the total contribution is overestimated or underestimated because of the correlation between input parameters. If $|S_{\boldsymbol{u}}^C|$ is low, $S_{\boldsymbol{u}}$ is close to $S_{\boldsymbol{u}}^U$, that is the correlation has a weak influence on the contribution of $X_{\boldsymbol{u}}$. On the contrary, if $|S_{\boldsymbol{u}}^C|$ is high, $S_{\boldsymbol{u}}$ is close to $S_{\boldsymbol{u}}^C$, that is the correlation has a strong influence on the contribution of $X_{\boldsymbol{u}}$.

2.6.2.3 Higher-order indices

In the case of an additive model, namely:

$$Y = \sum_{i=1}^{n} a_i X_i, \quad a_i \in \mathbb{R} \tag{2.97}$$

the variance of the model response can be entirely decomposed into the first order contributions since there is no interaction between the variables. Thus the sum of the indices S_i is equal to 1. When the model is no longer additive but multiplicative or event more complex, interaction arises among the variables. Then, the sum of the indices S_i is lower than 1. Similarly to Sobol' indices, one defines the second order indices, namely:

$$S_{ij} = \frac{\mathrm{Cov}\left[\mathcal{M}_{ij}(X_i, X_j), Y\right]}{\mathrm{Var}\left[Y\right]} \tag{2.98}$$

The same sort of index can be defined for subsets of $\boldsymbol{X}_{\boldsymbol{u}} \subset \boldsymbol{X}$. Adding the higher-order indices to the first order index leads to an index S_i^T that is consistent with the Sobol' total index in the case of independent variables. Then the sum of the S_i^T indices may be higher than one because coupling effects are stored in several total indices.

2.6.3 Conclusion

The issue of uncoupling the uncorrelated and correlated effects of correlated input parameters on the variance of the model output has been circumvented by the ANCOVA decomposition provided a functional decomposition as in Eq. (2.85) is available. Issues on the existence and uniqueness of such decompositions are hot topics in the current literature and not yet solved (Kucherenko et al., 2012; Chastaing et al., 2012). The contribution of an input variable $X_{\boldsymbol{u}}$ is now described by a triplet of indices $(X_{\boldsymbol{u}}, X_{\boldsymbol{u}}^U, X_{\boldsymbol{u}}^C)$ that respectively represent the total, physical and correlated contribution of $X_{\boldsymbol{u}}$ to the variance of Y. The interpretation of these three quantities leads to important observations on the coupling effects of the physical model $\mathcal{M}(\boldsymbol{X})$ and the probabilistic model $F_{\boldsymbol{X}}$, that is the joint distribution of the input random vector, on the dispersion of the model response Y.

2.7 Conclusion

In this chapter, several methods to carry out global sensitivity analysis have been presented. The methods for models with independent inputs are today well-established and their improvements only focus on the optimization of their numerical computing efficiency. In the presence of correlation among the input parameters of the model, classical methods cannot be applied directly, therefore, new techniques have been developped along two directions.

The first direction consists in quantifying the influence of a parameter with an *importance measure* which is often kind of a *normalized distance* observed on the output distribution (Iman and Hora, 1990; Chun et al., 2000; Borgonovo, 2007). Importances measures are convenient because no hypothesis are made on the input vector dependence structure, but their interpretation is less attractive because they do not represent a share of variance.

The second direction is an attempt to generalize the Sobol' decomposition for models with correlated inputs. The so-called ANCOVA represents the most advanced work in this domain. Like Xu and Gertner (2008), it allows one to separate the uncorrelated and correlated effects of an input parameter (Li and Rabitz, 2010) using the covariance decomposition of the variance of the model output.

The generalization of the Sobol' decomposition is also revendicated by Kucherenko et al. (2012). His estimation technique for the sensitivity indices for models with correlated input parameters provides results different from Li and Rabitz (2010) but are completely consistent. In other words, several generalizations of the well-established Sobol' indices are available for models involving dependent inputs and the field is still open. The approach proposed in the present thesis, which relies upon polynomial chaos expansions in order to derive a functional decomposition as in Eq. (2.85) contributes to this goal. It will be presented in details in chapter 4.

In any cases, the number of calls to the numerical model $\mathcal{M}$ that are necessary to estimate the variances, covariances or (un)conditional PDFs with accuracy, remains very high, approximately 10^4 to 10^6 calls per index. This issue can be overcome by substituting a surrogate model $\hat{\mathcal{M}}$ to the real model $\mathcal{M}$. This aspect is developped in the next chapter.

3

Surrogate modelling

Contents

3.1 Introduction

Numerical studies such as reliability analysis or sensitivity analysis on physical models may require a high number of model runs in order to catch the modifications in the model response due to the values of the input parameters. In the field of mechanical engineering, the physical models can be analytical when the phenomenon can be modelled by simple equations, a finite-element analysis when the structure is a more complex or even real experiments in extreme cases. For time and costs limitations, the model (or the experiment) cannot reasonably be run, say more than a few hundreds of times, that is far from the number of calls required for the abovementioned analyses. The solution consists in substituting the physical model $\mathcal{M}$ with a mathematical approximation $\hat{\mathcal{M}}$ built from a set of data samples. Such an approximation is referred to as a *surrogate model* (or *metamodel*).

In this chapter an overview on existing methods is first addressed. Three of them, namely the Support Vector Regression, a geostatistics method known as Gaussian process modelling and the high-dimensional model representation, are presented. Then, a particular attention is given to the technique referred to as polynomial chaos expansions. This last technique provides an efficient tool for uncertainty propagation with respect to distribution of the variables.

3.2 Overview on existing methods

This first section proposes an overview of three existing metamodelling techniques, namely, the support vector regression, the Gaussian processes and the high-dimensional model representation.

3.2.1 Support Vector Regression

The Support Vector (SV) algorithm is a tool from the statistical learning theory, or Vapnik-Chervonenkis (VC) theory. It describes learning machines that allow one to observe properties from given data and generalize them to unseen data. Industrial applications are in the field of optical pattern recognition using SV classifiers but also regression and time series prediction. In this section, priority is given to regression applications.

3.2.1.1 Linear case

Let us consider a 2-dimensional sample of observations $\boldsymbol{\mathcal{X}} = \left\{(\boldsymbol{x}^{(k)}, y^{(k)}),\ k = 1, \ldots, N\right\}$, also referred to as *training sample*. In $\epsilon-$regression (Vapnik, 1995), the aim is to find a *flat* function $f(\boldsymbol{x})$ so that $f(\boldsymbol{x}^{(k)})$ has at most ϵ deviation from the corresponding observation $y^{(k)}$, for all pairs of observations in the sample. Indeed, in one hand the regression error is not considered as important as long as it is less than ϵ, but on the other hand no deviation

larger than ϵ will be accepted. Let us begin with linear cases in which the function f reads:

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \text{ with } \boldsymbol{w} \in \mathbb{X} \text{ and } b \in \mathbb{R} \tag{3.1}$$

where $\langle ., . \rangle$ denotes the dot product in the space of the input variables $\mathbb{X} \subset \mathbb{R}^n$. *Flatness* corresponds to small values of $\boldsymbol{w}$. It is obtained by minimizing the norm $\|\boldsymbol{w}\|^2 = \langle \boldsymbol{w}, \boldsymbol{w} \rangle$. This can be written as a convex optimization problem:

$$\text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{3.2}$$

$$\text{subject to } \begin{cases} y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)} \rangle - b & \leqslant \quad \epsilon \\ \langle \boldsymbol{w}, \boldsymbol{x}^{(k)} \rangle + b - y^{(k)} & \leqslant \quad \epsilon \end{cases} \tag{3.3}$$

Eq. (3.3) implies that the function f approximating the given pairs with maximal deviation ϵ exists and that the optimization problem is *feasible*. However, it might sometimes not be the case, or one may simply want the regression to accept larger errors. Therefore, similarly to the concept of *soft margin* loss function for SV machines, one can introduced slack variables ξ, ξ^* to circumvent the issue of *infeasible* optimization problems. The new formulation then reads:

$$\text{minimize } \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{k=1}^{N} (\xi_k + \xi_k^*) \tag{3.4}$$

$$\text{subject to } \begin{cases} y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)} \rangle - b & \leqslant \quad \epsilon \\ \langle \boldsymbol{w}, \boldsymbol{x}^{(k)} \rangle + b - y^{(k)} & \leqslant \quad \epsilon \\ \xi_k, \xi_k^* & \geqslant \quad 0 \end{cases} \tag{3.5}$$

The constant $C > 0$ manages the point up to which deviations larger than ϵ are accepted at the expense of flatness. The algorithm then admits a $\epsilon-$insensitive error, namely:

$$|\xi|_\epsilon := \begin{cases} 0 & \text{if } |\xi| \leqslant \epsilon \\ |\xi| - \epsilon & \text{otherwise} \end{cases} \tag{3.6}$$

As it is depicted in Figure 3.1, only the points outside of the shaded area are linearly penalized.

In order to ease the convergence of the optimization problem in Eq. (3.5), one usually recasts it in its dual formulation that moreover allows one do deal with nonlinear models.

3.2.1.2 Dual formulation

The principle relies in the construction of a Lagrange function of the objective function (also denoted *primal* objective function) and the constraint functions by introducing a dual set of variables. The primal objective function exhibits a *saddle point* at the solution

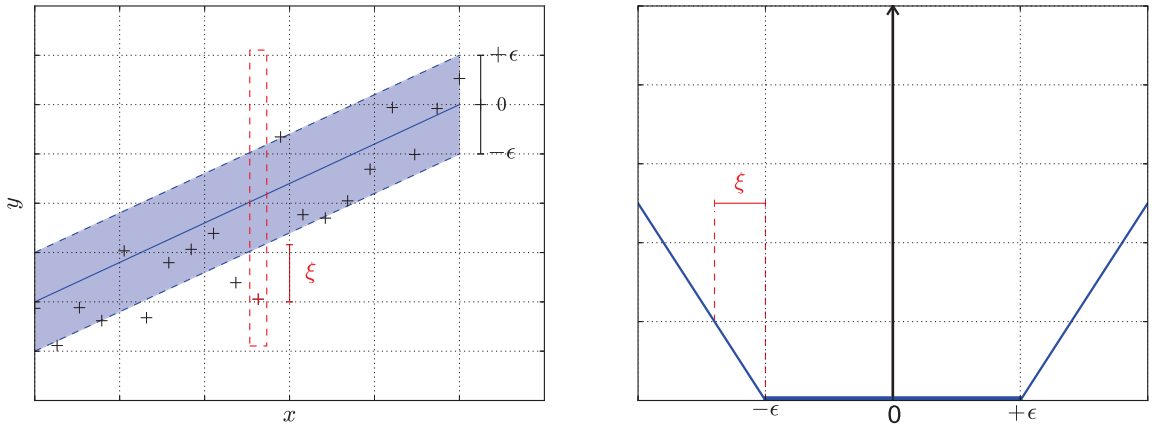


Figure 3.1: *Concept of margin and loss function for the SV machines.*

with respect to primal and dual variables. This reads:

$$\begin{aligned} L :=& \frac{1}{2}\|w\|^2 + C\sum_{k=1}^{N}(\xi+\xi^*) - \sum_{k=1}^{N}(\eta_k\xi + \eta_k^*\xi^*) \\ &- \sum_{k=1}^{N}\alpha_k\left(\epsilon + \xi_k - y^{(k)} + \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle + b\right) \\ &- \sum_{k=1}^{N}\alpha_k^*\left(\epsilon + \xi_k + y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle - b\right) \end{aligned} \tag{3.7}$$

where L is the Lagrangian and η_k, η_k^*, α_k, α_k^* are the Lagrange multipliers that must satisfy the positivity of the constraints functions:

$$\alpha_k, \alpha_k^*, \eta_k, \eta_k^* \geqslant 0 \tag{3.8}$$

The saddle point conditions imply the nullity of the derivatives of L with respect to the primal variables, namely:

$$\frac{\partial L}{\partial b} = \sum_{k=1}^{N}(\alpha_k^* - \alpha_k) = 0 \tag{3.9}$$

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{k=1}^{N}(\alpha_k^* - \alpha_k)\,\boldsymbol{x}^{(k)} = 0 \tag{3.10}$$

$$\frac{\partial L}{\partial \xi_k} = C - \alpha_k - \eta_k = 0 \tag{3.11}$$

$$\frac{\partial L}{\partial \xi_k^*} = C - \alpha_k^* - \eta_k^* = 0 \tag{3.12}$$

By substituting Eqs 3.9 - 3.12 into 3.7, the dual optimization problem reads:

$$\text{maximize} \begin{cases} -\frac{1}{2}\sum_{k=1}^{N}\sum_{l=1}^{N}(\alpha_k - \alpha_k^*)(\alpha_l - \alpha_l^*)\langle \boldsymbol{x}^{(k)}, \boldsymbol{x}^{(l)}\rangle \\ -\epsilon\sum_{k=1}^{N}(\alpha_k + \alpha_k^*) + y^{(k)}\sum_{k=1}^{N}(\alpha_k - \alpha_k^*) \end{cases} \tag{3.13}$$

$$\text{subject to } \sum_{k=1}^{N}(\alpha_k + \alpha_k^*) = 0 \text{ and } \alpha_k, \alpha_k^* \in [0, C] \tag{3.14}$$

Finally, one rewrites Eq. (3.10) as follows:

$$\boldsymbol{w} = \sum_{k=1}^{N}(\alpha_k - \alpha_k^*)\boldsymbol{x}^{(k)} \tag{3.15}$$

and consequently:

$$f(\boldsymbol{x}) = \sum_{k=1}^{N}(\alpha_k - \alpha_k^*)\langle \boldsymbol{x}^{(k)}, \boldsymbol{x}\rangle + b \tag{3.16}$$

Eq. (3.16) is often referred to as *Support Vector* expansion. $\boldsymbol{w}$ can be fully described a linear combination of the training data $\boldsymbol{x}^{(k)}$, $k = 1, \ldots, N$. The constant b can be computed thanks to the so-called Karush-Kuhn-Tucker (KKT) conditions. They stipulate that the product between the dual variables and constraints is zero at the optimal point:

$$\begin{array}{rcl} \alpha_i\left(\epsilon + \xi_k - y^{(k)} + \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle\right) + b & = & 0 \\ \alpha_i^*\left(\epsilon + \xi_k^* + y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle\right) - b & = & 0 \end{array} \tag{3.17}$$

and:

$$\begin{array}{rcl} (C - \alpha_k)\,\xi_k & = & 0 \\ (C - \alpha_k^*)\,\xi_k^* & = & 0 \end{array} \tag{3.18}$$

As a consequence, only the sample points $(\boldsymbol{x}^{(k)}, y^{(k)})$ with corresponding $\alpha_k, \alpha_k^* = C$ lie outside of the $\epsilon-$insensitive area. In addition, $\alpha_k\,\alpha_k^* = 0$, *i.e.* the dual variables in a set (α_k, α_k^*) cannot be both nonzero. Therefore:

$$\begin{array}{lll} \epsilon - y^{(k)} + \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle + b \geqslant 0 \quad \text{and} \quad \xi_k = 0 & \text{if} & \alpha_k < C \\ \epsilon - y^{(k)} + \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle + b \leqslant 0 & \text{if} & \alpha_k > 0 \end{array} \tag{3.19}$$

As the same relationship also holds for α_k^*, b can be bounded by:

$$\begin{array}{r} \max\{-\epsilon + y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle,\ \alpha_k < C \text{ or } \alpha_k^* > 0\} \\ \leqslant b \leqslant \\ \min\{-\epsilon + y^{(k)} - \langle \boldsymbol{w}, \boldsymbol{x}^{(k)}\rangle,\ \alpha_k > 0 \text{ or } \alpha_k^* < C\} \end{array} \tag{3.20}$$

If any $\alpha_k, \alpha_k^* \in [0, C]$, then the inequalities become equalities.

Let us now discuss Eq. (3.17). When $|f(\boldsymbol{x}^{(k)}) - y^{(k)}| \geqslant \epsilon$, *i.e.* the points outside the shaded area in Figure 3.1, the Lagrange multipliers may be nonzero, then $\alpha_k, \alpha_k^* = 0$. When $|f(\boldsymbol{x}^{(k)}) - y^{(k)}| < \epsilon$, the second term in the product in Eq. (3.17) is nonzero, therefore α_k and α_k^* have to be zero so that the KKT conditions are satisfied. The SV provide a sparse expansion $\boldsymbol{w}$ with respect to the $\boldsymbol{x}^{(k)}$, *i.e.* one do not need the N sampling points $(\boldsymbol{x}^{(k)}, y^{(k)})$ to to describe $\boldsymbol{w}$.

3.2.1.3 Conclusion

The SVR metamodelling technique allows one to build a robust approximation of a model $\mathcal{M}$ with a reasonable number of data points $\boldsymbol{\mathcal{X}} = \{(\boldsymbol{x}^{(k)}, \mathcal{M}(\boldsymbol{x}^{(k)})),\ k = 1, \ldots, N\}$. The approach consists in solving a quadratic optimization problem. Finally, the metamodel only depends on a small number of observations referred to as *Support Vectors*.

3.2.2 Gaussian processes

3.2.2.1 Introduction

Kriging is a well-established prediction method in the field of geostatistics named after the South-African mining engineer Daniel Krige (Krige, 1951) and formalized by Matheron (1962). The aim of Kriging is to consider the values of a physical model $\mathcal{M}(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{X}$ as a *sample path* of a *random field* $M(\boldsymbol{x}) = \mathcal{M}(\boldsymbol{x}, \omega)$, $\omega \in \Omega$, $\boldsymbol{x} \in \mathbb{X}$ at any unobserved point $\boldsymbol{x} \in \mathbb{X}$ from a set of observed realizations $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$. This method has lately been extended to the field of mechanical engineering as a powerful tool for approximating the response $y = \mathcal{M}(\boldsymbol{x})$ of a deterministic numerical model when assuming that y is a sample path $m(\boldsymbol{x})$ of a Gaussian process, $\boldsymbol{x} \in \mathbb{X}$. Kriging is also referred to as *Gaussian Process* (GP) predictor. The statistical assumptions in GP are useful for modelling the most probable metamodel $\hat{\mathcal{M}}(\boldsymbol{x})$ given the set of observations $\boldsymbol{\mathcal{X}}$. The GP metamodelling technique is now developed. For a more detailed review of the topic, the reading of the books by Santner et al. (2003) and Rasmussen and Williams (2006) is advised. See also Dubourg (2011), Chapter 1.

3.2.2.2 Stochastic modelling of the model function

Let us consider a physical model $\mathcal{M}$ whose response has only been evaluated at a given set of points so that only the couples $\{\boldsymbol{x}^{(k)}, y^{(k)},\ k = 1, \ldots, N\}$ are known. The true model response $\mathcal{M}$ is assumed to be a sample path of an underlying Gaussian process M. Let us denote by μ the mean of the random field M and by Z a zero-mean stationary GP. Then M reads:

$$M(\boldsymbol{x}) = \mu(\boldsymbol{x}) + Z(\boldsymbol{x}),\ \boldsymbol{x} \in \mathbb{X} \tag{3.21}$$

The mean μ is usually defined as a linear combination of deterministic functions $\{f_j(\boldsymbol{x}),\ j = 1, \ldots, P\}$, namely:

$$\mu(\boldsymbol{x}) = \sum_{j=1}^{P} \beta_j\, f_j(\boldsymbol{x}) \equiv \boldsymbol{f}(\boldsymbol{x})^{\mathsf{T}}\, \boldsymbol{\beta},\ \boldsymbol{x} \in \mathbb{X} \tag{3.22}$$

In Eq. (3.21), the first summand μ represents the global trend of the model response that can be constant, linear or polynomial for instance, whereas the second summand Z corresponds to local perturbations of the model response. The limitation of this assumption lies the choice of decoupling large-scale (the mean μ) and small-scale (the supposed stationary random field Z) effects. Assumptions have to be made on the nature of μ.

Constant or linear trend is usually retained in practice, while Z is assumed to be a second-order stationary random field, *i.e.* Z has a constant finite variance $\sigma^2(\boldsymbol{x}) = \sigma^2$ and its *autocorrelation function* $R(\boldsymbol{x}, \boldsymbol{x}')$ is a function of the shift $\boldsymbol{x} - \boldsymbol{x}'$ only.

3.2.2.3 Conditional distribution of the model response

Let us first recall a few notations. $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(N)}\}$ is $n-$dimensional $N-$sample of the input random vector $\boldsymbol{x}$ also referred to as a *design of experiments* (DOE) denoted by $\mathcal{D}$. The corresponding evaluations of $\boldsymbol{\mathcal{X}}$ are stored in a unidimensional sample $\mathcal{Y} = \{y^{(k)} = \mathcal{M}(x^{(k)}),\ k = 1, \ldots, N\}$. The Gaussian Process approximation consists in building the distribution of M conditioned on to the N observations in $\mathcal{Y}$. Finding the conditional distribution of M is not an easy problem. Hence, it can be recast as a constrained optimization problem using the *fundamental theorem of prediction* (Santner et al., 2003). The following matrix notation is used in the sequel:

$$\boldsymbol{r}(\boldsymbol{x}) = \left\{R\left(\boldsymbol{x}, \boldsymbol{x}^{(1)}\right), \ldots, R\left(\boldsymbol{x}, \boldsymbol{x}^{(N)}\right)\right\}, \boldsymbol{x} \in \mathbb{X} \tag{3.23}$$

$$\mathbf{F} = \left(f_i\left(x^{(k)}\right)\right)_{1 \leqslant i \leqslant P,\, 1 \leqslant k \leqslant N} \tag{3.24}$$

$$\mathbf{R} = \left(R\left(x^{(k)}, x^{(l)}\right)\right)_{1 \leqslant k \leqslant N,\, 1 \leqslant l \leqslant N} \tag{3.25}$$

$\mathbf{F}$ and $\mathbf{R}$ are referred to as the *experiments* and the *autocorrelation matrices*. Then it is shown that the conditional model response at an unevaluated point $\boldsymbol{x}$ has a Gaussian distribution $\mathcal{N}(\mu_M(\boldsymbol{x}), \sigma_M(\boldsymbol{x}))$ with:

$$\mu_M(\boldsymbol{x}) = \boldsymbol{f}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{\beta} + \boldsymbol{r}^{\mathsf{T}}(\boldsymbol{x})\mathbf{R}^{-1}(\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}) \tag{3.26}$$

$$\sigma_M^2(\boldsymbol{x}) = \sigma^2 - \boldsymbol{r}^{\mathsf{T}}(\boldsymbol{x})\mathbf{R}\boldsymbol{r}(\boldsymbol{x}) \tag{3.27}$$

The previous equations involve the properties of M though (mean, autocovariance, etc.) that are typically unknown in practice. Indeed, the autocorrelation function is selected *a priori*. It is usually defined in the form of tensorized stationary functions of the form:

$$R(\boldsymbol{x}, \boldsymbol{x}') = \prod_{i=1}^{n} R_i(x_i - x_i'),\ \boldsymbol{x}, \boldsymbol{x}' \in \mathbb{X} \times \mathbb{X} \tag{3.28}$$

As shown in Eq. (3.28), the correlation between two realizations only depends on the distance between them. This functions can be of several types, namely:

- *linear*:
$$R(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{l}) = \prod_{i=1}^{n} \max\left(0, 1 - \frac{x_i - x_i'}{l_i}\right) \tag{3.29}$$
where $\{l_i,\ i = 1, \ldots, n\}$ are the so-called *scale parameters*.

- *exponential*:
$$R(\boldsymbol{x} - \boldsymbol{x}', \boldsymbol{l}) = \exp\left(-\sum_{i=1}^{n} \frac{|x_i - x_i'|}{l_i}\right) \tag{3.30}$$

- *square exponential* or *Gaussian*:

$$R(\boldsymbol{x}-\boldsymbol{x}',\boldsymbol{l}) = \exp\left(-\sum_{i=1}^{n}\left(\frac{x_i - x_i'}{l_i}\right)^2\right) \tag{3.31}$$

The different autocorrelation functions mentioned above are pictured in Figure 3.2.

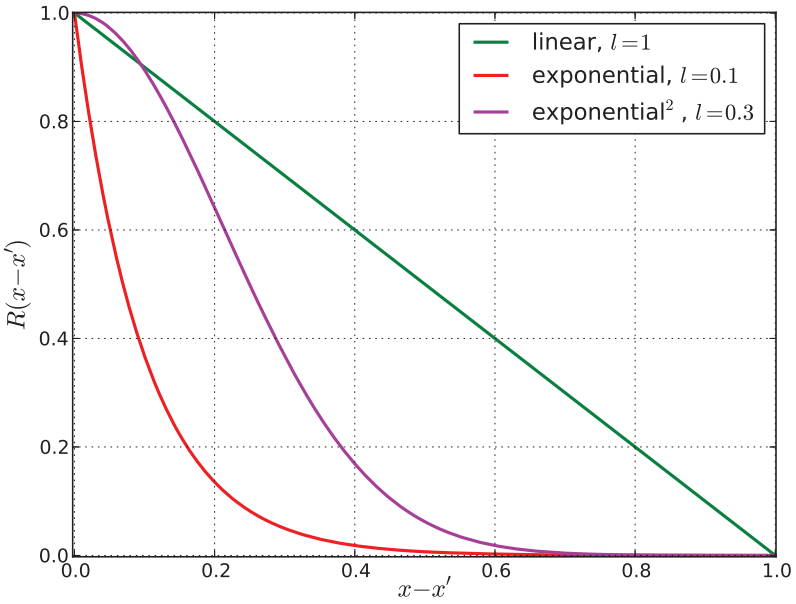


Figure 3.2: *Different families of autocorrelation functions for GP metamodelling.*

Thereby, identifying a Gaussian process involve estimating its parameters :

- the parameters of the mean function μ, *i.e.* the regression coefficients in the $\boldsymbol{\beta}$ vector,
- the variance σ^2,
- the autocorrelation functions parameters $\boldsymbol{l} = \{l_i,\, i = 1, \ldots, n\}$.

3.2.2.4 Estimation of the GP parameters

The GP parameters are then estimated from the data $(\boldsymbol{\mathcal{X}}, \mathcal{Y})$ using *maximum likelihood estimation* (MLE). The ML estimate $\hat{\boldsymbol{\beta}}$ of the regression parameters $\boldsymbol{\beta}$ is a function of the autocorrelation function parameters $\boldsymbol{l}$ and reads:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{F}^{\mathsf{T}}\mathbf{R}_l^{-1}\mathbf{F}\right)^{-1}\mathbf{F}^{\mathsf{T}}\mathbf{R}_l^{-1}\mathcal{Y} \tag{3.32}$$

where the subscript $\boldsymbol{l}$ denotes the dependency of $\mathbf{R}$ on the autocorrelation functions parameters. The ML estimate $\hat{\sigma}_M^2$ of the variance σ^2 is also a function of $\boldsymbol{l}$, namely:

$$\hat{\sigma}_M^2 = \frac{1}{N}\left(\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}^\mathsf{T}\right)^\mathsf{T} \mathbf{R}_l^{-1}\left(\mathcal{Y} - \mathbf{F}\boldsymbol{\beta}^\mathsf{T}\right) \tag{3.33}$$

Finally, the ML estimates $\hat{\boldsymbol{l}}$ of the autocorrelation parameters are obtained by resolving the following optimization problem:

$$\hat{\boldsymbol{l}} = \underset{\boldsymbol{l}}{\operatorname{argmin}} \left(\det \mathbf{R}_{\boldsymbol{l}}\right)^{\frac{1}{N}} \hat{\sigma}_M^2 \tag{3.34}$$

As it might be expected, the computational cost for the estimation of the GP parameters increases with the dimension n of the problem. The choice of an isotropic autocorrelation functions, *i.e.* $l_i = l_0$, $i = 1, \ldots, n$ appears not to be a good choice since the variables may have different autocovariance properties. Nevertheless, it can be a good initial design for an optimization loop. In practice, the optimization of all the hyperparameters can be carried out in a single optimization procedure (O'Hagan, 2006), which is not good. The work by Welch et al. (1992) and more recently by Marrel et al. (2008) based on a variable selection procedure allows one to chose anisotropic autocorrelation functions for problems of dimension up to 10.

3.2.2.5 GP metamodel

The last quantity to be estimate is the mean μ of the GP in Eq. (3.26). Its calculation is based on the estimates of the other GP parameters $\hat{\boldsymbol{\beta}}$, $\hat{\sigma}_M^2$ and $\hat{\boldsymbol{l}}$, namely:

$$\begin{aligned} \hat{\mu}_M(\boldsymbol{x}) &= \hat{\mathcal{M}}(\boldsymbol{x}) \\ &= \boldsymbol{f}^\mathsf{T}(\boldsymbol{x})\hat{\boldsymbol{\beta}} + \boldsymbol{r}_{\hat{\beta}}(\boldsymbol{x})^\mathsf{T}\mathbf{R}_{\hat{l}}^{-1}\left(\mathcal{Y} - \mathbf{F}\hat{\boldsymbol{\beta}}\right) \end{aligned} \tag{3.35}$$

where $\hat{\boldsymbol{l}}$ is the set of the ML estimates of the autocorrelation function hyperparameters. $\hat{\mathcal{M}}(\boldsymbol{x})$ interpolates the model response observations in $\boldsymbol{\mathcal{X}}$ and thus can be used as a metamodel. In addition, the estimation variance of the metamodel, also referred to as *Kriging variance*, in Eq. (3.27) reads:

$$\hat{\sigma}_M^2(\boldsymbol{x}) = \hat{\sigma}_M^2 - \boldsymbol{r}_{\hat{\beta}}(\boldsymbol{x})^\mathsf{T}\mathbf{R}_{\hat{l}}\boldsymbol{r}_{\hat{\beta}}(\boldsymbol{x}) \tag{3.36}$$

which is useful for deriving confidence intervals on the prediction $\hat{\mathcal{M}}(\boldsymbol{x})$, $\boldsymbol{x} \notin \boldsymbol{\mathcal{X}}$.

As an example, the method is applied to the so-called *Runge function* defined by $y = (1 + 25\,x^2)^{-1}$. The approximation is build from a set of observations $\mathcal{X} = \{-0.8, -0.4, -0.1, 0.5, 0.6, 0.9\}$ with the help of the Python package *Scikit-lean* presented in Pedregosa et al. (2011). An illustration is pictured in Figure 3.3. The Kriging variance provides a precious information for an *adaptive* design of experiments: the physical model $\mathcal{M}$ is estimated where the $\hat{\sigma}_M^2(\boldsymbol{x})$ reaches its highest value in order to minimize the uncertainty on the global estimation wherever it matters most, see also Picheny et al. (2010) and Dubourg (2011) for a more detailed description.

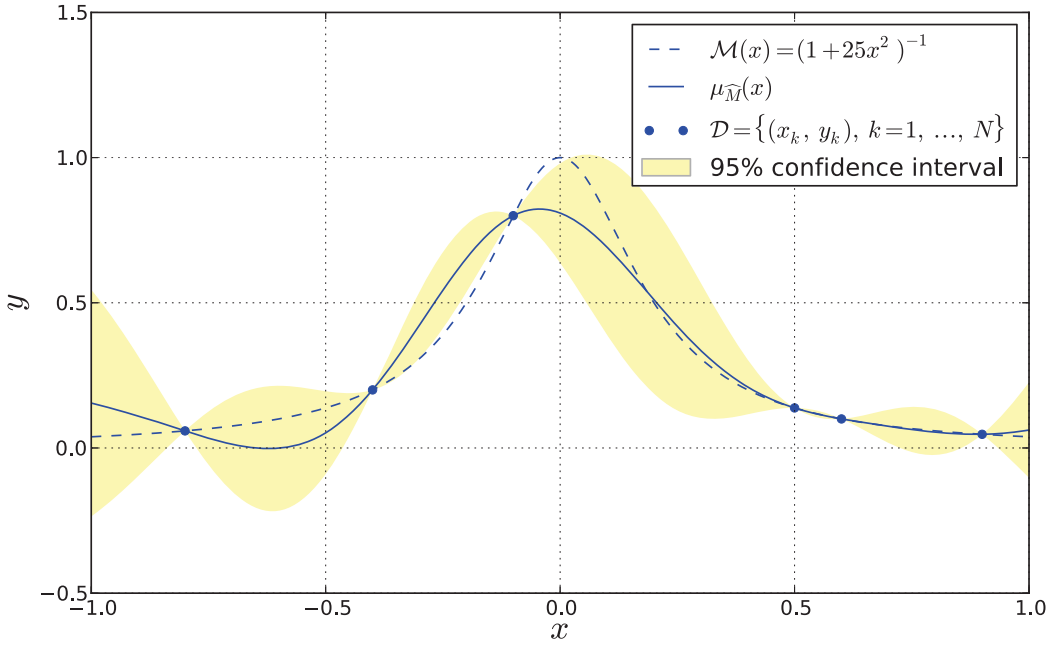


Figure 3.3: *GP metamodelling technique applied to the Runge function.*

3.2.2.6 Conclusion

This subsection has given a short introduction to the Gaussian process metamodelling technique. Its main concept is to consider the model response as a sample path of a Gaussian random field whose parameters are estimated by maximum likelihood. The derived metamodel interpolates the design of experiments, *i.e.* there is no random error when considering a deterministic model function. Then a confidence interval on the prediction is provided where the model has not been evaluated. The limitations of GP metamodels lies in its numerical cost for the estimation of the GP parameters that increases with the problem dimension. An application on a mechanical example of the computation of the Sobol' indices using Gaussian processes has been proposed in Caniou et al. (2011).

3.2.3 High-dimensional model representation

3.2.3.1 Introduction

High-dimensional model representation (HDMR) is a set of tools to build input-output relationships (Li et al., 2002). In this type of representation, each term represents the independent or cooperative contribution of each input parameter upon the model response. A HDMR build from a set of random observations is usually referred to as RS-HDMR, where the RS stands for *random sampling*. The components functions that have to be

determined are selected among well-established basis functions such as orthonormal polynomials, cubic B-splines or simple polynomials. This section gives a short overview on the underlying theory and the computational issues of the HDMR method.

3.2.3.2 Functional decomposition of the model response

Let us consider a physical model $\mathcal{M}$ whose response Y depends on n random input parameters $\boldsymbol{X} = \{X_i,\, i = 1, \ldots, n\}$, $\boldsymbol{X} \in \mathbb{X}$. The principle of HDMR consists in decomposing the model response as a sum of functions of increasing dimension, namely:

$$\mathcal{M}(\boldsymbol{X}) = \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(X_i) + \sum_{1 \leqslant i < j \leqslant n} \mathcal{M}_{ij}(X_i, X_j) + \ldots + \mathcal{M}_{1\ldots n}(\boldsymbol{X}) \tag{3.37}$$

where the zero-th order term is a constant representing the mean of the model response, *i.e.* $\mathcal{M}_0 = \mu_Y$. Each unidimensional term $\mathcal{M}_i(X_i)$ represents the independent contribution of the input X_i in the response Y. The second-order terms $\mathcal{M}_{ij}(X_i, X_j)$ shows the crossed contribution of the couples $(X_i, X_j)_{i \neq j}$. Finally, the highest-order term $\mathcal{M}_{1\ldots n}(\boldsymbol{X})$ corresponds to the cooperative contribution of the n inputs.

As for many families of metamodels, the numerical cost, which corresponds here to the number of functions to be constructed, rapidly increases with the dimension of the problem. As it has been noticed that the high-order of interaction terms are often negligible with respect to the low-order terms, HDMR decomposition are usually truncated to the second order of interaction, namely:

$$\mathcal{M}(\boldsymbol{X}) \approx \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(X_i) + \sum_{1 \leqslant i < j \leqslant n} \mathcal{M}_{ij}(X_i, X_j) \tag{3.38}$$

For the sake of simplicity, the input random variables X_i are first normalized so that the input random vector $\boldsymbol{X}$ is defined on the unit hypercube $\mathbb{K}^n$. Thus, the component functions of the decomposition can be estimated by:

$$\mathcal{M}_0 = \int_{\mathbb{K}^n} \mathcal{M}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.39}$$

$$\mathcal{M}_i(x_i) = \int_{\mathbb{K}^{n-1}} \mathcal{M}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}_{\sim i} - \mathcal{M}_0 \tag{3.40}$$

$$\mathcal{M}_{ij}(x_i, x_j) = \int_{\mathbb{K}^{n-2}} \mathcal{M}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}_{\sim ij} - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j) - \mathcal{M}_0 \tag{3.41}$$

where $\mathrm{d}\boldsymbol{x}_{\sim i}$ (resp. $\mathrm{d}\boldsymbol{x}_{\sim ij}$) is the product of all the $\mathrm{d}x_i$ except the i^{th} one (resp. the i^{th} and j^{th} ones). The highest-order term is finally estimated by the subtraction of all the other component functions to the physical model $\mathcal{M}(\boldsymbol{X})$.

Let us now introduce the multi-index notation $X_{\mathcal{I}}$, $\mathcal{I} = \{i_1, \ldots, i_k\}$ defining a $k-$dimensional subvector of $\boldsymbol{X}$. The HDMR component functions have to satisfy the following property:

$$\int_0^1 \mathcal{M}_{i_1,\ldots,i_k}(x_{i_1}, \ldots, x_{i_k})\, \mathrm{d}x_{i_s}, \quad i_s \in \{i_1, \ldots, i_k\} \tag{3.42}$$

The integral of a component function over the domain $[0,1]$ with respect to any of the variables $X_{i_1},\dots,X_{i_k}$ is zero. From Eq. (3.42) the following orthogonality property between two component functions not sharing variables holds:

$$\int_{\mathbb{K}^n} \mathcal{M}_{i_1,\dots,i_k}(x_{i_1},\dots,x_{i_k})\,\mathcal{M}_{j_1,\dots,j_k}(x_{j_1},\dots,x_{j_l})\,\mathrm{d}\boldsymbol{x} \tag{3.43}$$

with $\{i_1,\dots,i_k\} \neq \{j_1,\dots,j_l\}$.

3.2.3.3 Estimation of the component functions by Monte Carlo simulations

In Li et al. (2002), the component functions are first classically obtained by Monte Carlo simulations. $N-$samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(1)},\dots,\boldsymbol{x}^{(N)}\}$ are randomly generated uniformly in the unit hypercube $\mathbb{K}^n$. Components functions values are estimated by evaluating the functions on $\boldsymbol{\mathcal{X}}$, namely:

$$\begin{aligned}\mathcal{M}_0 &= \int_{\mathbb{K}^n} \mathcal{M}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\\ &\approx \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\end{aligned} \tag{3.44}$$

$$\begin{aligned}\mathcal{M}_i(x_i) &= \int_{\mathbb{K}^{n-1}} \mathcal{M}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{\sim i} - \mathcal{M}_0\\ &\approx \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left((x_i,\boldsymbol{x}_{\sim i})^{(m)}\right) - \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\end{aligned} \tag{3.45}$$

$$\begin{aligned}\mathcal{M}_{ij}(x_i,x_j) &= \int_{\mathbb{K}^{n-2}} \mathcal{M}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}_{\sim ij} - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j) - \mathcal{M}_0\\ &\approx \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left((x_i,x_j,\boldsymbol{x}_{\sim ij})^{(m)}\right)\\ &\quad - \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left((x_i,\boldsymbol{x}_{\sim i})^{(m)}\right) - \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left((x_j,\boldsymbol{x}_{\sim j})^{(m)}\right)\\ &\quad - \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\end{aligned} \tag{3.46}$$

Monte Carlo simulations are certainly the most simple and robust method for identifying the component functions of the decomposition but the massive amount of simulations that are necessary to estimate them with accuracy due to the low convergence rate makes it hardly achievable for a standard computing device.

3.2.3.4 Approximation of the component functions using basis functions

Due to the numerical burden involved by Monte Carlo simulations (MCS) to estimate each component functions values, the authors propose to approximate the component functions

by expansions onto well-established basis functions such as polynomials or splines. The expression of the component functions can be recast as follows:

$$\mathcal{M}_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) \tag{3.47}$$

$$\mathcal{M}_{ij}(x_i, x_j) \approx \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i, x_j) \tag{3.48}$$

where the φ_r and φ_{pq} are one- and two-dimensional basis functions and α_r and β_{pq} are coefficients to be determined. k, l and l' are integers denoting the size of the basis. The complete model approximation then reads:

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}_0 + \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) + \sum_{1 \leqslant i < j \leqslant n} \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i, x_j) \tag{3.49}$$

The coefficients of the expansions can be determined by minimizing the following functional $\int_{\mathbb{K}^n} [\mathcal{M}(\boldsymbol{x}) - \hat{\mathcal{M}}(\boldsymbol{x})]^2 \, \mathrm{d}\boldsymbol{x}$, namely:

$$\min_{\alpha_r^i, \beta_{pq}^{ij}} \int_{\mathbb{K}^n} \left[\mathcal{M}(\boldsymbol{x}) - \mathcal{M}_0 - \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) - \sum_{1 \leqslant i < j \leqslant n} \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i, x_j) \right]^2 \mathrm{d}\boldsymbol{x} \tag{3.50}$$

Knowing that basis functions of different dimensions are orthogonal, that is:

$$\int_{\mathbb{K}^n} \varphi_r(x_i) \, \varphi_{pq}(x_i, x_j) \, \mathrm{d}\boldsymbol{x} = 0 \tag{3.51}$$

one can decouple the one- and two-dimensional effects in the minimization in Eq. (3.50). This is due to the preservation by the approximations in Eqs. (3.47) and (3.48) of the mutual orthogonality defined in Eq. (3.43). The global minimization problem is transformed into as many local least-squares minimization problems as component functions. They read:

$$\min_{\alpha_r^i} \int_0^1 \left[\mathcal{M}_i(x_i) - \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) \right]^2 \mathrm{d}x_i \tag{3.52}$$

$$\min_{\beta_{pq}^{ij}} \int_{[0,1]^2} \left[\mathcal{M}(x_i, x_j) - \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i, x_j) \right]^2 \mathrm{d}x_i \, \mathrm{d}x_j \tag{3.53}$$

For each variable X_i (resp. couple of variables (X_i, X_j)), the set of coefficients $\{\alpha_r^i,\ r = 1, \ldots, k\}$ (resp. $\{\beta_{pq}^{ij},\ p = 1, \ldots, l, q = 1, \ldots, l'\}$) can be obtained by solving a linear equation. As an example, variable X_i, the linear problem is of the form:

$$\mathbf{A}\boldsymbol{y} = \boldsymbol{b} \tag{3.54}$$

where $\mathbf{A}$ is a constant non singular matrix whose terms $A_{rr'}$ are defined by:

$$A_{rr'} = \int_0^1 \varphi_r(x_i) \, \varphi_{r'}(x_i) \, \mathrm{d}x_i, \quad r = 1, \ldots, k \tag{3.55}$$

and where $\boldsymbol{b}$ and $\boldsymbol{y}$ are the following vectors:

$$\boldsymbol{b} = \begin{bmatrix} \int_0^1 \mathcal{M}_i(x_i)\,\varphi_1(x_i)\,\mathrm{d}x_i \\ \vdots \\ \int_0^1 \mathcal{M}_i(x_i)\,\varphi_r(x_i)\,\mathrm{d}x_i \\ \vdots \\ \int_0^1 \mathcal{M}_i(x_i)\,\varphi_k(x_i)\,\mathrm{d}x_i \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} \alpha_1^i \\ \vdots \\ \alpha_r^i \\ \vdots \\ \alpha_k^i \end{bmatrix} \tag{3.56}$$

In order to solve this problem, vector $\boldsymbol{b}$ can be approximated by Eq. (3.45):

$$\boldsymbol{b} \approx \frac{1}{N} \sum_{m=1}^{N} \begin{bmatrix} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\,\varphi_1\left(x_i^{(m)}\right) \\ \vdots \\ \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\,\varphi_r\left(x_i^{(m)}\right) \\ \vdots \\ \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\,\varphi_k\left(x_i^{(m)}\right) \end{bmatrix} \tag{3.57}$$

Finally, the vector of the k coefficients $\{\alpha_r^i,\ r = 1, \ldots, k\}$ corresponding to the component function $\mathcal{M}_i(x_i)$ are obtained by:

$$\boldsymbol{y} = \mathbf{A}^{-1}\boldsymbol{b} \tag{3.58}$$

A similar procedure can be applied to approximate the second-order terms $\mathcal{M}_{ij}(x_i, x_j)$ by computing the set of $l \times l'$ coefficients $\{\beta_{pq}^{ij},\ p = 1, \ldots, l,\ q = 1, \ldots, l'\}$.

3.2.3.5 Estimation of the component functions

The choice of the basis functions is made among well-established functions such as polynomials or splines. The first family of basis functions are the polynomials. According to Eqs. (3.47) and (3.48), one writes:

$$\mathcal{M}_i(x_i) = \sum_{r=1}^{k} \alpha_r^i\, x_i^r \tag{3.59}$$

$$\mathcal{M}_{ij}(x_i, x_j) = \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij}\, x_i^p x_j^q \tag{3.60}$$

Then, estimating the coefficients α_r^i and β_{pq}^{ij} consists in minimizing the following expressions:

$$\min_{\alpha_r^i} \int_0^1 \left[\mathcal{M}_i(x_i) - \sum_{r=1}^{k} \alpha_r^i x_i^r\right]^2 \mathrm{d}x_i \tag{3.61}$$

$$\min_{\beta_{pq}^{ij}} \int_{[0,1]^2} \left[\mathcal{M}(x_i, x_j) - \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} x_i^l x_j^{l'}\right]^2 \mathrm{d}x_i\, \mathrm{d}x_j \tag{3.62}$$

Using simple polynomials implies dealing with a singular matrix $\mathbf{A}$. Therefore, each coefficient α_r^i or β_{pq}^{ij} has to be determined independently by solving the linear equations in Eqs. (3.61) and (3.62) one-by-one. To overcome this singularity problem, one rather opt for orthonormal polynomials, *i.e.* polynomials φ_k that satisfy the following properties on the domain $\mathcal{D} = [a, b]$:

$$\int_{\mathcal{D}} \varphi_r(x)\,\mathrm{d}x = 0 \tag{3.63}$$

$$\int_{\mathcal{D}} \varphi_r^2(x)\,\mathrm{d}x = 1 \tag{3.64}$$

$$\int_{\mathcal{D}} \varphi_r(x)\,\varphi_{r'}(x)\,\mathrm{d}x = 0,\ r \neq r' \tag{3.65}$$

Orthonormal polynomials have zero mean (Eq. (3.63)), unit norm (Eq. (3.64)) and are mutually orthogonal (Eq. (3.65)). For the domain $\mathcal{D} = [0, 1]$, the author proposes the following orthonormal polynomials, which are rescaled Legendre polynomials (the original Legendre polynomials being defined over $\mathcal{D} = [-1, 1]$):

$$\varphi_1(x) = \sqrt{3}(2x - 1) \tag{3.66}$$

$$\varphi_2(x) = 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \tag{3.67}$$

$$\varphi_3(x) = 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \tag{3.68}$$

Using the orthonormality property in Eq. (3.65), the second order component functions read:

$$\mathcal{M}_{ij}(x_i, x_j) = \beta_{pq}^{ij}\varphi_{pq}(x_i, x_j) = \beta_{pq}^{ij}\varphi_p(x_i)\varphi_q(x_j) \tag{3.69}$$

Consequently, the model $\mathcal{M}$ can be approximated by:

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}_0 + \sum_{i=1}^{n} \alpha_r^i \varphi_r(x_i) + \sum_{1 \leqslant i < j \leqslant n} \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_p(x_i)\varphi_q(x_j) \tag{3.70}$$

The matrices $\mathbf{A}$ in Eq. (3.54) are identity matrices for all α_r^i and β_{pq}^{ij} and the coefficients can be approximated by:

$$\begin{aligned} \alpha_r^i &= \int_{\mathbb{K}^n} \mathcal{M}(\boldsymbol{x})\,\varphi_r(x_i)\,\mathrm{d}\boldsymbol{x} \\ &\approx \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\,\varphi_r\left(x_i^{(m)}\right) \end{aligned} \tag{3.71}$$

$$\begin{aligned} \beta_{pq}^{ij} &= \int_{\mathbb{K}^n} \mathcal{M}(\boldsymbol{x})\,\varphi_p(x_i)\,\varphi_q(x_j)\,\mathrm{d}\boldsymbol{x} \\ &\approx \frac{1}{N}\sum_{m=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(m)}\right)\,\varphi_p\left(x_i^{(m)}\right)\,\varphi_q\left(x_j^{(m)}\right) \end{aligned} \tag{3.72}$$

The accuracy of orthonormal polynomials depends on both the degree of the expansions (k, l and l') and the sampling size N. The higher they are, the more accurate the metamodel is but attention has to be given to the numerical cost that is implied.

A third family of basis function that has been tested in Li et al. (2002) are the so called *cubic $B-$splines* named after French automotive engineer Pierre Bézier in the early 60's. A cubic $B-$spline $B_k(x)$, $x \in \mathcal{D} = [a, b]$, $k = -1, 0, 1, \ldots, m+1$ is defined in Eq. (3.73).

$$B_k(x) = \frac{1}{h^3} \times \left\{ \begin{array}{ll} (y_{k+2} - x)^3 & y_{k+1} < x \leqslant y_{k+2} \\ (y_{k+2} - x)^3 - 4(y_{k+1} - x)^3 & y_k < x \leqslant y_{k+1} \\ (y_{k+2} - x)^3 - 4(y_{k+1} - x)^3 + 6(y_k - x)^3 & y_{k-1} < x \leqslant y_k \\ (y_{k+2} - x)^3 - 4(y_{k+1} - x)^3 + 6(y_k - x)^3 - 4(y_{k-1} - x)^3 & y_{k-2} < x \leqslant y_{k-1} \\ 0 & \text{otherwise} \end{array} \right\} \tag{3.73}$$

where:

$$h = \frac{b-a}{m} \tag{3.74}$$

and:

$$y_k = a + kh \tag{3.75}$$

When $\mathcal{D} = [0, 1]$, $h = \dfrac{1}{m}$ and $y_k = kh$. The first- and second-order component functions then read:

$$\mathcal{M}_i(x_i) = \sum_{r=-1}^{m+1} \alpha_r^i \, B_r(x_i) \tag{3.76}$$

$$\mathcal{M}_{ij}(x_i, x_j) = \sum_{p=-1}^{m+1} \sum_{q=-1}^{m+1} \beta_{pq}^{ij} \, B_p(x_i) B_q(x_j) \tag{3.77}$$

Due to the non orthogonality of two B-splines polynomials of different variables, this family of basis function also provide singular matrices $\mathbf{A}$. Therefore, similarly to simple polynomials, all the coefficients α_r^i and β_{pq}^{ij} cannot be determined in the same numerical operation.

3.2.3.6 Conclusion

The RS-HDMR method has been developed to model the behavior of high-dimensional input-output relationships. The method uses a functional decomposition of the physical model so that the contribution of each input variable or group of variables is described by a distinct component function. The component function can be estimated by Monte Carlo simulations but it is numerically more efficient to expand them using an adapted basis of functions such as simple polynomials, orthonormal polynomials or $B-$splines polynomials. The estimation procedure of the coefficients of the expansion capitalizes on the orthogonality property of orthonormal polynomials to be recast as a linear equation system to be solved. In practice, the expansion is often truncated to the second-order of interaction due to the burden of numerical simulations.

3.3 Polynomial chaos expansion

Spectral expansions are well established methods in the field of numerical analysis. In the more specific context of uncertainty propagation, the method consists in expanding the random response of a model onto a suitable finite-dimensional basis $\{\Psi_j,\, j = 0, \ldots, P-1\}$ so that the model approximation reads:

$$\mathcal{M}(\boldsymbol{X}) \approx \sum_{j=0}^{P-1} y_j \Psi_j(\boldsymbol{X}) \tag{3.78}$$

This work focuses on spectral expansions onto bases composed of orthonormal polynomials also referred to as *polynomials chaos expansions* (PCE). Then, the approximation of the random model response in Eq. (3.78) consists in the estimation of the coefficients y_j's of the expansion. Two families of methods may be applied to solve such a problem. The *intrusive* methods, based on a Galerkin scheme have been introduced in the early 90's for solving mechanical problems with spatially random parameters (Ghanem and Spanos, 1991). Their main drawbacks are a high computational cost due to the size of the matrix systems that have to be solved and the *intrusions* of the method in the computer code. Therefore, intrusive methods will not be detailed in this work.

Alternative methods, referred to as *non intrusive methods*, simply benefits from the evaluations of the physical model, which can be a finite element analysis for instance, at soundly chosen points to compute the coefficients of the expansion, without any modification of the computational code (Sudret, 2007). Of interest is this last family of methods. After introducing the general framework of polynomial chaos expansions, the estimation methods for the coefficients and the associated estimation error are studied. Finally, attention is given to the issue of models with *correlated* input parameters.

3.3.1 Mathematical framework of PC expansions

In the Greek mythology, the word chaos is linked to the origin of the world. Chaos is the son of Chronos, personification of Time, and Ananke, personification of destiny, necessity and fate. Then came Gaia (Earth), Eros (love), Tartarus (Hell), Erebus (Hell's darkness) and Nyx (the night). Ovid, in his Metamophoses, gave to this word its signification in use to date, describing it as a gross, amorphous and unorganized mass. Thus, the Chaos can be described by two main aspects:

- *the bottomless abyss down which there is an endless fall : Earth then appears, providing a stable base, which is radically opposed to chaos;*
- *the space with no possible orientation in which one falls in every direction at the same time.*

Nobody knows who first introduced the word chaos in science models. It may be Poincaré who was opposed to the perfectly ordered mechanics of Newton. In any cases, it is prior

to Wiener (1938). Chaos is not necessarily stochastic. There is chaos if the quantity of information that is necessary to predict a sample path must be very large and if a very small modification in this information generates deterministic sample paths that are completely different. Wiener's chaos then appears as the set of all the possible sample paths issued from random data. It depicts the apparently disordered aspect of the behavior of a system as it is originally associated with apparently disordered nature of the Creation of the Universe in the Greek Mythology.

3.3.1.1 Introduction

Let us consider a physical model $\mathcal{M}$ whose random response Y is a function of an input random vector $\boldsymbol{X} = \{X_i,\, i = 1, \ldots, n\}$ having independent components. For the sake of simplicity, a scalar random response Y is assumed but the sequel also holds for each component of an output random vector. The mathematical framework of the probability theory, presented in Chapter 1, Section 1.2, introduces the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Of interest here are continuous random variables X with finite second moment, namely:

$$\mathrm{E}\left[X^2\right] = \int_{\Omega} X^2(\omega)\mathrm{d}\mathbb{P}(\omega) = \int_{\mathrm{D}_X} x^2\, f_X(x)\, \mathrm{d}x < +\infty \tag{3.79}$$

where $\mathrm{D}_X \subset \mathbb{R}$ and f_X are respectively the support and the probability density function of X.

Such variables are defined in the Hilbert space $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ whose associated inner product reads:

$$< X_1, X_2 >_{\mathcal{L}^2} = \mathrm{E}\left[X_1 X_2\right] = \int_{\mathrm{D}_{X_1} \times \mathrm{D}_{X_2}} x_1 x_2\, f_{X_1}(x_1) f_{X_2}(x_2)\, \mathrm{d}x_1 \mathrm{d}x_2 \tag{3.80}$$

where X_1 and X_2 are two finite variance random variables. The inner product defines the $\mathcal{L}^2-$norm:

$$\|X\|_{\mathcal{L}^2} = \sqrt{\mathrm{E}\left[X^2\right]} \tag{3.81}$$

In the sequel, the model $\mathcal{M}$ under consideration takes an input random vector $\boldsymbol{X}$ defined by its joint PDF $f_{\boldsymbol{X}}$. Consequently, its response Y is also a random variable assumed to be square-integrable, *i.e.* $Y \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. This work now studies the spectral decomposition of the model response on bases of $\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ made of orthogonal polynomials. The input random vector $\boldsymbol{X}$ is first assumed to have independent components. Then, the case of correlated input random variables will be addressed.

3.3.1.2 Independent random variables

Let us consider an independent input random vector $\boldsymbol{X} = \{X_i,\, i = 1, \ldots, n\}$. Soize and Ghanem (2004) show that the model response Y may be expanded onto an orthogonal polynomial basis:

$$Y = \mathcal{M}(\boldsymbol{X}) = \sum_{j=1}^{+\infty} y_j \Psi_j(\boldsymbol{X}) \tag{3.82}$$

Moreover, it is shown that these series converges to the true model response in the sense of the $\mathcal{L}^2-$norm, namely:

$$\lim_{P\to+\infty}\left(\left\|\mathcal{M}(\boldsymbol{X})-\sum_{j=1}^{P-1}y_j\Psi_j(\boldsymbol{X})\right\|_{\mathcal{L}^2}\right)^2=\lim_{P\to+\infty}\mathrm{E}\left[\left(\mathcal{M}(\boldsymbol{X})-\sum_{j=1}^{P-1}y_j\Psi_j(\boldsymbol{X})\right)^2\right]=0 \tag{3.83}$$

where the y_j's are P unknown coefficients to be determined and the Ψ_j are multivariate polynomials. Therefore, the series in Eq. (3.82) is usually referred to as *polynomial chaos expansions* (PCE). Such expansions require the building of a suitable basis which is now detailed.

When the input parameters X_i, $i=1,\ldots,n$ are independent, the joint PDF $f_{\boldsymbol{X}}$ of the input random vector $\boldsymbol{X}$ is simply the product of the n marginal PDF, namely:

$$f_{\boldsymbol{X}}(\boldsymbol{x})=\prod_{i=1}^{n}f_{X_i}(x_i) \tag{3.84}$$

Let us denote by $\left\{\pi_j^{(i)},\, j\in\mathbb{N}\right\}$ a family of *univariate orthonormal polynomials* with respect to f_{X_i}. These polynomials satisfy:

$$<\pi_j^{(i)}(X_i),\pi_k^{(i)}(X_i)>_{\mathcal{L}^2}=\mathrm{E}\left[\pi_j^{(i)}(X_i)\,\pi_k^{(i)}(X_i)\right]=\delta_{j,k} \tag{3.85}$$

where the $\delta_{j,k}$ is the Kronecker symbol, *i.e.* $\delta_{j,k}=1$ if $j=k$ and 0 otherwise. The polynomials $\pi_j^{(i)}$ are assumed of degree j for $j>0$ and $\pi_0^{(i)}=1$ for all $X_i\in\boldsymbol{X}$.

Let us now introduce the multi-index notation $\boldsymbol{\alpha}=\{\alpha_1,\ldots,\alpha_n\}$. This notation allows one to define the set of *multivariate polynomials* $\{\Psi_{\boldsymbol{\alpha}},\boldsymbol{\alpha}\in\mathbb{N}^n\}$ by tensorizing the n univariate polynomials families as follows:

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})=\pi_{\alpha_1}^{(1)}(x_1)\times\ldots\times\pi_{\alpha_n}^{(n)}(x_n) \tag{3.86}$$

In the original work by Wiener (1938), the bases were made of *Hermite polynomials* to model stochastic processes from series of Gaussian variables. The first Hermite polynomials are presented in Figure 3.4. The method was then also called *Wiener chaos expansion*. To meet the expectations of the practitioners using not only Gaussian variables, the method has been extended to the main families of random variables with basis functions from the Askey-scheme of hypergeometric polynomials (Xiu and Karniadakis, 2003). When different families of random variables are mixed in a joint distribution, the expansion is referred to as *generalized polynomial chaos* (gPC) expansion. Table 3.1 provides the polynomial families that are associated with the most popular probabilistic distributions.

For other types of distributions, the problem may be be recast using an isoprobabilistic transformation where the gPC expansion is applied on the transformed variables. Finally, families of orthogonal polynomials that suit uncommon distributions can be generated numerically. For a more detailed overview on generalized polynomial chaos expansions, the reading of Blatman (2009) is advised.

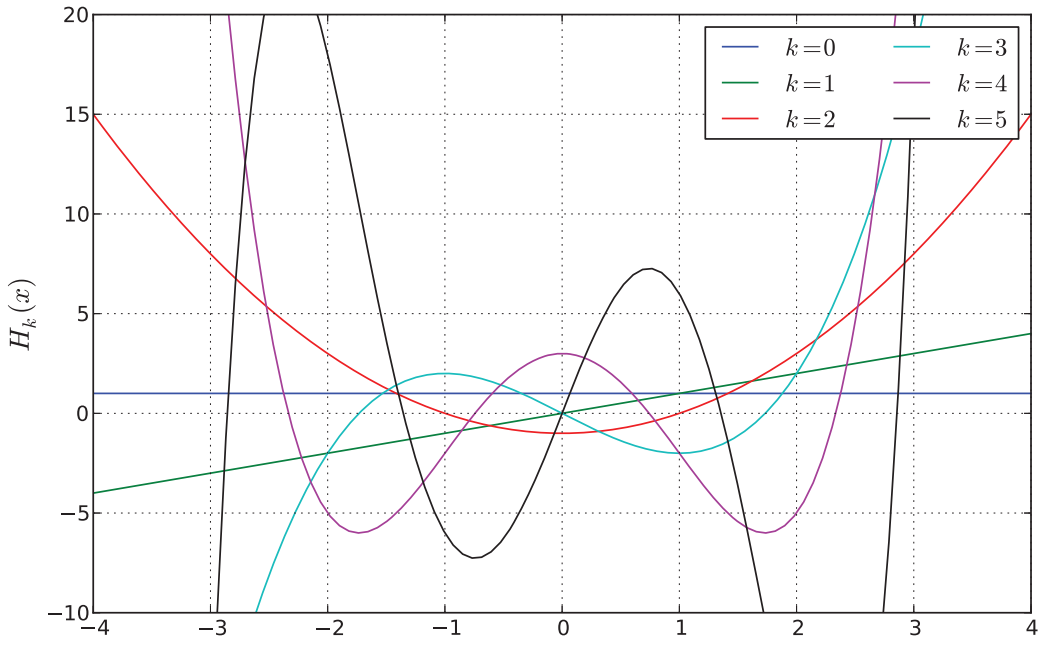


Figure 3.4: *Hermite polynomials H_k, $k = 0, \ldots, 5$.*

Distribution	Support	Polynomial
Gaussian	$\mathbb{R}$	Hermite
Uniform	$[-1, 1]$	Legendre
Gamma	$[0, +\infty]$	Laguerre
Chebyshev	$[-1, 1]$	Chebyshev
Beta	$[-1, 1]$	Jacobi

Table 3.1: *Polynomial families from the Askey-scheme of hypergeometric orthogonal polynomials associated with the main continuous probabilistic distributions.*

3.3.1.3 Truncature of the basis

It has been shown in Eq. (3.83) that the approximation converges in the mean-square sense to the true random response when the size P of the expansion tends to $+\infty$. However, one only retains in practice the multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$ whose total degree $\sum_{i=1}^{n} \alpha_i$ is smaller than a degree p. Then the size of the expansion and therefore the number of unknown coefficients to be estimated reads:

$$P = \binom{n+p}{p} \tag{3.87}$$

The polynomial chaos expansion truncated in such a way that no multivariate polynomials have a total degree greater than p will be denoted by:

$$Y \approx \hat{\mathcal{M}}_p(\boldsymbol{X}) = \sum_{|\boldsymbol{\alpha}| \leqslant p} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{3.88}$$

3.3.2 Advanced truncature strategies

According to Eq. (3.87), the number of coefficients of the expansion rapidly increases with both the dimension n of the problem and the order p of the expansion. When a variable selection procedure has already been performed and the complexity of the problem requires a high order of expansion, the number of coefficients to be estimated might remains to high with respect to the necessary number of calls to the physical model.

It has been observed that in a *complete basis* when p is high, the coefficients corresponding to the highest orders of interaction are often negligible compared with the ones corresponding to the inner contribution of each variable. Thus, there is a need for strategies to reduce the sampling effort, *i.e.* the number of coefficients to be computed, by removing from the set of coefficients those who have the lowest contribution in the expansion. Several strategies have been proposed in Blatman (2009) which are now summarized.

3.3.2.1 Low-rank index sets

Let us recall the notation for the *multi-index set* $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\}$ where α_i is the degree of the univariate polynomial corresponding to the i^{th} input variable. The total degree and the *rank* of $\boldsymbol{\alpha}$ are respectively defined in Eqs. (3.89) and (3.90).

$$|\boldsymbol{\alpha}| = \|\boldsymbol{\alpha}\|_1 = \sum_{i=1}^{n} \alpha_i \tag{3.89}$$

$$\|\boldsymbol{\alpha}\|_0 = \sum_{i=1}^{n} \mathbf{1}_{(\alpha_i > 0)} \tag{3.90}$$

As mentioned already, it is common practice to retain the multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$ whose total degree is not greater than p. The corresponding *index sets* are then defined by:

$$\mathcal{A}^{n,p} = \{\boldsymbol{\alpha} \in \mathbb{N}^n,\ \|\boldsymbol{\alpha}\|_1 \leqslant p\} \tag{3.91}$$

with $\text{card}\,(\mathcal{A}^{n,p}) = P = \mathrm{C}_p^{n+p}$.

The low-rank index strategy consists in only retaining the multi-index sets $\boldsymbol{\alpha}$ whose rank is smaller than an integer $j \leqslant p$, *i.e.* at most j indices $\alpha_i \in \boldsymbol{\alpha}$ are non zero. The corresponding index sets reads:

$$\mathcal{A}^{n,p,j} = \{\boldsymbol{\alpha} \in \mathbb{N}^n,\ \|\boldsymbol{\alpha}\|_1 \leqslant p,\ \|\boldsymbol{\alpha}\|_0 \leqslant j\} \tag{3.92}$$

with the following property:

$$\text{card}\left(\mathcal{A}^{n,p,j}\right) \leqslant \text{card}\,(\mathcal{A}^{n,p}) \tag{3.93}$$

This strategy offers an alternative to high-dimensional models requiring a high order of expansion but the limitation might come from limiting the interaction order to, let us say 2 or 3, and therefore neglecting some interaction terms that might exist in complex physical models.

3.3.2.2 Hyperbolic index sets

Blatman and Sudret (2009) proposes an index selection based on the so-called $q-$norm:

$$\|\boldsymbol{\alpha}\|_q = \left(\sum_{i=1}^{n} \alpha_i^q\right)^{1/q}, \quad 0 < q < 1 \tag{3.94}$$

The corresponding *hyperbolic index set* reads:

$$\mathcal{A}_q^{n,p} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^n,\ \|\boldsymbol{\alpha}\|_q \leqslant p\right\} \tag{3.95}$$

When $q = 1$, the index set verifies $\mathcal{A}_1^{n,p} \equiv \mathcal{A}^{n,p}$. When $q < 1$, the high-order interaction terms are not retained in the multi-index set and the lowest q is, the less high-order interaction terms are retained. The name *hyperbolic index sets* comes from the fact that graphically speaking, the retained indices are located under a hyperbola parametrized by q. The advantage of the hyperbolic sets compared to the low-rank sets is that the PC approximation converges to the true response with respect to the $\mathcal{L}^2-$norm when p increases, whatever the value of q.

In its simplest definition, hyperbolic index sets are *isotropic*, *i.e.* the same importance is given to all the input variables. In order to penalize less the high-order interaction on those variables who have the highest contribution on the model response, an anisotropic definition is also proposed by Blatman (2009). The anisotropic norm is defined by:

$$\|\boldsymbol{\alpha}\|_{q,\boldsymbol{w}} = \left(\sum_{i=1}^{n} |w_i \alpha_i|^q\right)^{1/q}, \quad \omega_i \geqslant 1 \tag{3.96}$$

The corresponding *anisotropic index set* reads

$$\mathcal{A}_{q,\boldsymbol{w}}^{n,p} = \left\{\boldsymbol{\alpha} \in \mathbb{N}^n,\ \|\boldsymbol{\alpha}\|_{q,\boldsymbol{w}} \leqslant p\right\} \tag{3.97}$$

The definition of the weights w_i is related to the total sensitivity index S_i^T that represents the share of variance of Y due to the variable X_i and its interaction with the variables X_j, $j \neq i$ (see Chapter 2, Section 2.3). In other words, S_i^T measures the degree of interaction of X_i in $\mathcal{M}$. A variable with a low total sensitivity index is penalized with a high weight w_i of the form:

$$w_i = 1 + K\frac{S_{max}^T - S_i^T}{\sum_{i=1}^{n} S_i^T} \tag{3.98}$$

where S_{max}^T is the highest total sensitivity index $S_{max}^T = \max\{S_i^T,\ i = 1, \ldots, n\}$ and K is a non negative constant. According to the definition in Eq. (3.98), $K = 0$ leads to the isotropic index sets, *i.e.* $w_i = 1$, $i = 1, \ldots, n$ whereas a high value of K defines a high anisotropy in the index set.

Hyperbolic sets and moreover the anisotropic ones lead to a drastic reduction of the number of expansion coefficients to be computed as demonstrated in Blatman and Sudret (2010). Nevertheless, the total sensitivity indices have to be computed before building the PC expansion, inducing a prior sensitivity analysis of the model.

3.3.2.3 Adaptive sparse polynomial chaos expansion

Adaptive sparse polynomial chaos expansion is due to the recent work of Blatman (2009). The main idea is to replace the *selection rule* defined by the $q-$norm and the anisotropy coefficient K by an algorithm that is able to define which coefficients must be retained and which can be neglected. Several strategies might be used, depending on the initial full basis, the criterion to classify the coefficients, the maximal number of coefficients in the final basis etc.

Of interest is the most advanced adaptive algorithm based on *Least Angle Regression* (LAR) (Efron et al., 2004). Let us consider the data samples $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(k)},\, k = 1, \ldots, N\}$ and $\mathcal{Y} = \{y^{(k)} = \mathcal{M}(\boldsymbol{x}^{(k)}),\, k = 1, \ldots, N\}$ as the design of experiments and the corresponding physical model evaluations. The PC expansion of $Y = \mathcal{M}(\boldsymbol{X})$ reads:

$$\mathcal{M}_{\mathcal{A}}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathcal{A}} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{3.99}$$

where $\mathcal{A}$ is the set of multi-indices. In the case of a $p-$truncated basis, the coefficients $y_{\boldsymbol{\alpha}}$, $\boldsymbol{\alpha} \in \mathcal{A}^{n,p}$ (Eq. (3.91)), are estimated all at the same time by regression. The principle of the variable selection algorithm is to select step by step the polynomial $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$ that is the most correlated with the current residual (in the first step, the current residual is the model response Y). In practice, the *predictors* are evaluated on the training sample $\boldsymbol{\mathcal{X}}$ and the correlation between the corresponding evaluations vector $\{\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(k)}),\, k = 1, \ldots, N\}$, $\boldsymbol{\alpha} \in \mathcal{A}$ and the response sample $\mathcal{Y}$ is studied. The predictor that has the highest correlation with the response is added in the expansion and its coefficient is increased until another predictor has a higher correlation with the residuals. The increase of the coefficients is ruled by fast mathematical derivations instead of slow incremental procedures such as in the *forward stagewise regression* algorithm.

The number of predictors is progressively increased until a stopping criterion (number of predictors, relative error) is reached. This technique allows one to reach high orders of expansion and interaction which lead to an accurate approximation of the model response with less coefficients to compute and consequently less calls to the physical model. The different truncature strategies are illustrated in Figure 3.5.

Once the optimal basis, *i.e.* the set of multivariate polynomials $\Psi_{\boldsymbol{\alpha}}$, has been built, the unknown coefficients of the expansion have to be estimated.

3.3.3 Estimation of the coefficients

The so-called non intrusive methods allows one to estimate the coefficients of the expansions using the evaluations of the physical model at a set of chosen points. Two kinds of methods are studied here, namely the projection and regression methods. These methods are non interpolating. The estimation of the coefficients relies upon the minimization of the mean-square error of the approximation.

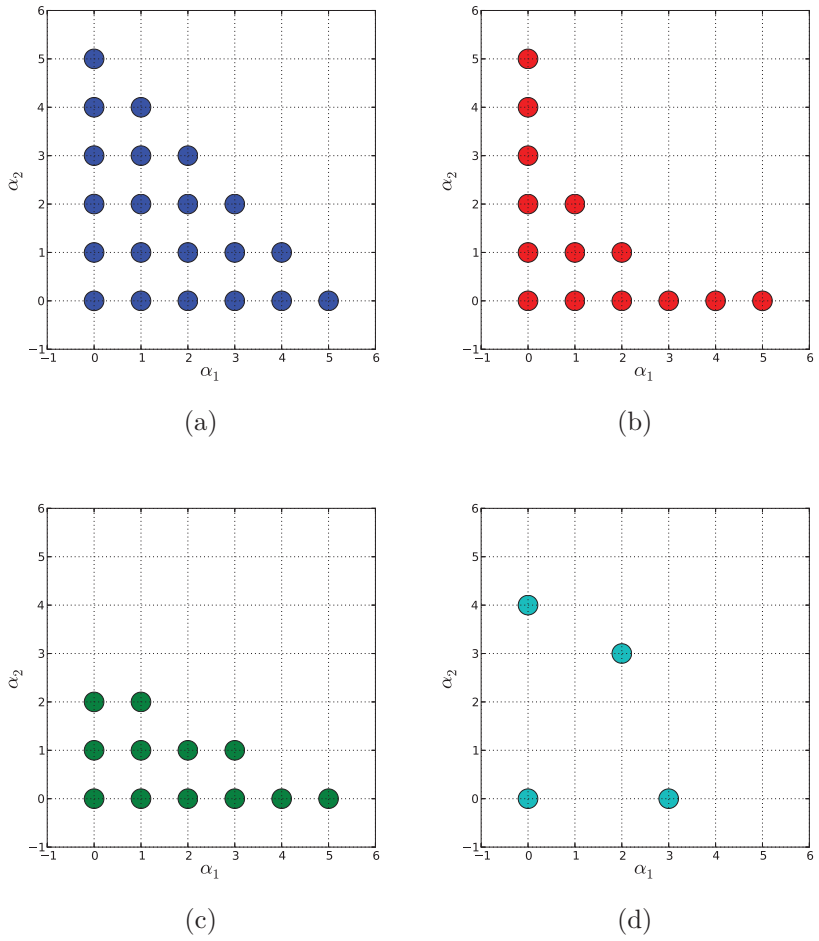


Figure 3.5: *Example of index sets. (a) is a full index set with* $\|\boldsymbol{\alpha}\|_1 \leqslant 5$. *(b) is a hyperbolic index set with* $\|\boldsymbol{\alpha}\|_{0.5} \leqslant 5$. *(c) is an anisotropic index set with* $\|\boldsymbol{\alpha}\|_{1,\boldsymbol{w}} \leqslant 5$, $\boldsymbol{w} = \{1, 2\}$. *(d) is a LAR index set where only the most influent polynomials are retained.*

3.3.3.1 Projection methods

The so-called spectral projection method takes advantage of the orthonormality property of the truncated basis $\{\Psi_{\boldsymbol{\alpha}},\, |\boldsymbol{\alpha}| \leqslant p\}$. The coefficient $y_{\boldsymbol{\alpha}}$ actually equals the expected value of the series times the associated multivariate polynomial $\Psi_{\boldsymbol{\alpha}}$, namely:

$$y_{\boldsymbol{\alpha}} = \mathrm{E}\left[\Psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\, \mathcal{M}(\boldsymbol{X})\right] = \int_{\mathrm{D}_{\boldsymbol{X}}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})\, \mathcal{M}(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} \tag{3.100}$$

Practical implementation of the computation scheme requires the approximation of the expected value in Eq. (3.100). The simplest of the *simulation methods* is the Monte Carlo estimation. Considering a $n-$dimensional $N-$sample $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(k)},\, k = 1, \ldots, N\}$, the

estimate of the coefficient $y_{\boldsymbol{\alpha}}$ reads:

$$\hat{y}_{\boldsymbol{\alpha}} \approx \frac{1}{N} \sum_{k=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(k)}\right) \Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{x}^{(k)}\right) \tag{3.101}$$

Advanced sampling techniques such as *latin hypercube sampling* (LHS) or *quasi-Monte Carlo* (QMC) propose more intelligent space filling strategies compared to the random MC simulations and therefore provide a better accuracy. Nevertheless, having in mind that simulation techniques have a quite slow convergence rate and that each of the P coefficients will require such a computationally heavy procedure, other integration techniques that make the best use of each physical model estimations have to be carried out.

Quadrature methods select a set of specific points $\boldsymbol{x}^{(k)}$ and associated weights $\omega^{(k)}$ so that the integral in Eq. (3.100) reads:

$$\hat{y}_{\boldsymbol{\alpha}} = \sum_{k=1}^{N_q} \omega^{(k)} \mathcal{M}\left(\boldsymbol{x}^{(k)}\right) \Psi_{\boldsymbol{\alpha}}\left(\boldsymbol{x}^{(k)}\right) \tag{3.102}$$

The multivariate nature of the input implies to use multidimensional quadrature scheme such as *sparse grids*. The main drawback of projection methods is that the coefficients have to be estimated one by one.

One may have noticed that the HDMR decomposition presented in Section 3.2.3 is identical to a polynomial chaos expansions of degree $p = 2$ whose component functions, which are in both cases multivariate Legendre polynomials, are obtained by projection, namely:

$$a_j = \mathrm{E}\left[\mathcal{M}(\boldsymbol{X})\Psi_j(\boldsymbol{X})\right] \approx \frac{1}{N} \sum_{k=1}^{N} \mathcal{M}\left(\boldsymbol{x}^{(k)}\right) \Psi_j\left(\boldsymbol{x}^{(k)}\right) \tag{3.103}$$

3.3.3.2 Regression methods

Regression methods allow one to reduce the sampling effort of projection methods by computing all the coefficients at the same time (Berveiller, 2005). Let us introduce the following vector notation:

$$\boldsymbol{y} = \left\{y_{\boldsymbol{\alpha}_0}, \ldots, y_{\boldsymbol{\alpha}_{P-1}}\right\}^{\mathsf{T}} \tag{3.104}$$

$$\boldsymbol{\Psi}(\boldsymbol{X}) = \left\{\Psi_{\boldsymbol{\alpha}_0}(\boldsymbol{X}), \ldots, \Psi_{\boldsymbol{\alpha}_{P-1}(\boldsymbol{X})}\right\}^{\mathsf{T}} \tag{3.105}$$

The regression problem can be recast as an optimization problem:

$$\boldsymbol{y}^* = \underset{\boldsymbol{y}}{\operatorname{argmin}} \underbrace{\mathrm{E}\left[\left(\boldsymbol{y}^{\mathsf{T}}\boldsymbol{\Psi}(\boldsymbol{X}) - \mathcal{M}(\boldsymbol{X})\right)^2\right]}_{\mathcal{I}(\boldsymbol{y})} \tag{3.106}$$

whose optimality condition:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{y}} \mathcal{I}(\boldsymbol{y}^*) = 0 \tag{3.107}$$

leads to:

$$\underbrace{\mathrm{E}\left[\boldsymbol{\Psi}(\boldsymbol{X})\,\boldsymbol{\Psi}^{\mathsf{T}}(\boldsymbol{X})\right]}_{\mathbf{I}}\boldsymbol{y}^{*} = \mathrm{E}\left[\boldsymbol{\Psi}(\boldsymbol{X})\,\mathcal{M}\boldsymbol{X}\right] \tag{3.108}$$

where $\mathbf{I}$ is the $P \times P$ identity matrix. Indeed, $\boldsymbol{\Psi}(\boldsymbol{X})\,\boldsymbol{\Psi}^{\mathsf{T}}(\boldsymbol{X})$ is the covariance matrix of $\boldsymbol{\Psi}(\boldsymbol{X})$ which has orthogonal components by definition. Finally, the vector of the expansion coefficients reads:

$$\boldsymbol{y}^{*} = \mathrm{E}\left[\boldsymbol{\Psi}(\boldsymbol{X})\,\mathcal{M}\boldsymbol{X}\right] \tag{3.109}$$

which is nothing but the expression of the projection-based coefficients in Eq. (3.100). In other words, the theoretical projection-based coefficients minimize the mean-square error of the approximation.

The estimates of the regression-based coefficients are now studied. Let us consider the sample $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(k)},\, k = 1, \ldots, N\}$ and let us define the *information matrix* $\boldsymbol{\Psi}$ where Ψ_{kj} is the estimation of the j^{th} multivariate polynomial $\Psi_{\boldsymbol{\alpha}_j}$ at the k^{th} sampling point $\boldsymbol{x}^{(k)}$, namely:

$$\boldsymbol{\Psi} = \begin{pmatrix} \Psi_{\boldsymbol{\alpha}_0}\left(\boldsymbol{x}^{(1)}\right) & \ldots & \Psi_{\boldsymbol{\alpha}_{P-1}}\left(\boldsymbol{x}^{(1)}\right) \\ \vdots & \ddots & \vdots \\ \Psi_{\boldsymbol{\alpha}_0}\left(\boldsymbol{x}^{(N)}\right) & \ldots & \Psi_{\boldsymbol{\alpha}_{P-1}}\left(\boldsymbol{x}^{(N)}\right) \end{pmatrix} \tag{3.110}$$

The regression-based estimates of the coefficients $\boldsymbol{y}^{*}$ read:

$$\boldsymbol{y}^{*} = \left(\boldsymbol{\Psi}^{\mathsf{T}}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^{\mathsf{T}}\mathcal{Y} \tag{3.111}$$

where $\mathcal{Y} = \{y^{(k)} = \mathcal{M}(\boldsymbol{x}^{(k)}),\, k = 1, \ldots N\}$ is the sample of the model response at the sampling points.

A condition for the estimation of the unknown PC expansion coefficients is that the size N of the sample must be greater than the number of coefficients P. In practice, one may use $N = 3P$ sampling points to increase the accuracy of the estimation. Concerning the sampling techniques, similarly as for projection methods, quasi-random techniques such as LHS or QMC can be used.

3.3.4 Models with correlated inputs

When the input random vector no longer has independent components, the joint cumulative distribution function $F_{\boldsymbol{X}}$ is defined by the marginal CDFs F_{X_i}, $i = 1, \ldots, n$, and the copula C (see Chapter 1, Section 1.5), namely:

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = C\left(F_{X_1}(x_1), \ldots, F_{X_n}(x_n)\right) \tag{3.112}$$

In most cases, the dependence structure of the input random vector is modelled by a $n-$dimensional Gaussian copula $C_{n,\boldsymbol{\Sigma}}$ where $\mathbf{R}$ is the $n \times n$ symmetrical matrix that is linked by Eq. (1.73) to the rank correlation matrix of the components of $\boldsymbol{X}$. Its joint distribution then reads:

$$F_{\boldsymbol{X}}(x_1, \ldots, x_n) = \Phi_{\mathbf{R}}\left(\Phi^{-1}(F_{X_1}(x_1)), \ldots, \Phi^{-1}(F_{X_n}(x_n))\right) \tag{3.113}$$

where Φ is the univariate standard Gaussian CDF. The joint CDF in Eq. (3.113) appears as a function of *correlated* standard Gaussian variables denoted by:

$$\boldsymbol{\xi}^C = \left\{\xi_i^C = \Phi^{-1}(F_{X_i}(x_i)),\, i = 1, \ldots, n\right\} \tag{3.114}$$

As the polynomial chaos expansion require independent input variables, the ξ_i^C have to be transformed into *uncorrelated* Gaussian variables ξ_i^U so that a Hermite polynomial chaos can be processed. The uncorrelated variables read:

$$\boldsymbol{\xi}^U = \boldsymbol{\Gamma}^{-1}\boldsymbol{\xi}^C \tag{3.115}$$

where $\boldsymbol{\Gamma}$ is the Cholesky decomposition of the rank correlation matrix $\mathbf{S}$, namely:

$$\mathbf{S} = \boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\mathsf{T}} \tag{3.116}$$

The model response is recast as a function of independent standard Gaussian variables as follows:

$$Y = \mathcal{M}\left(\boldsymbol{X}\left(\boldsymbol{\xi}^U\right)\right) \tag{3.117}$$

Finally, the model response can be expanded onto a suitable basis made of orthonormal Hermite polynomials in $\boldsymbol{\xi}^U$ as any function of independent Gaussian variables.

When the dependence structure of the input random vector is no longer Gaussian, the Nataf transformation is no longer relevant. The problem has been addressed in Soize and Ghanem (2004). A perfect knowledge of the joint distribution of $\boldsymbol{X}$, that is of the copula C, is required in order to transform the correlated physical variables into independent standard Gaussian variables using the Rosenblatt transformation (see Chapter 1, Section 1.5.7.2).

3.3.5 Accuracy of PC expansions

Polynomial chaos expansions represent an efficient way to substitute a numerically expensive model with a cheaper analytical one. However, the worst cases the practitioner may have a complex response to approximate with a limited number of calls to the physical model at hand and might want to know what is the approximation error in order to build a confidence interval for instance. Therefore, the *generalization error* and the *empirical error* are now introduced.

3.3.5.1 Generalization error and empirical error

Let us consider a sample $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(k)},\, k = 1, \ldots, N\}$ of the input random vector $\boldsymbol{X}$ and the corresponding response sample $\mathcal{Y} = \{y^{(k)} = \mathcal{M}(\boldsymbol{x}^{(k)}),\, k = 1, \ldots, N\}$. This set $(\boldsymbol{\mathcal{X}}, \mathcal{Y})$ of the experimental points and the evaluations of the physical model $\mathcal{M}$ also referred to as the *training sample*, constitutes the design of experiments one may use to build the following $p-$truncated PC expansion:

$$Y \approx \hat{\mathcal{M}}_p(\boldsymbol{X}) = \sum_{|\boldsymbol{\alpha}| \leqslant p} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) \tag{3.118}$$

Let us now denote by Err the approximation error with respect to the $\mathcal{L}^2-$norm, also referred to as *generalization error* (Vapnik, 1995), namely:

$$Err = \mathrm{E}\left[\left(\mathcal{M}(\boldsymbol{X}) - \hat{\mathcal{M}}_p(\boldsymbol{X})\right)^2\right] \tag{3.119}$$

The generalization error may be estimated by computing the mean-squared error at the design points, or *empirical error* Err_{emp} defined by:

$$Err_{emp} = \frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{M}(\boldsymbol{x}^{(k)}) - \hat{\mathcal{M}}_p(\boldsymbol{x}^{(k)})\right)^2 \tag{3.120}$$

Because the empirical error is estimated from the training sample, it is also referred to as the *training error*. At this point, the importance of the training error depends on the values Y can take. Thus, one may norm the training error by the empirical variance of the random response $\hat{\sigma}_Y^2$ to define the *relative* training error, namely:

$$\epsilon_{emp} = \frac{Err_{emp}}{\hat{\sigma}_Y^2} \tag{3.121}$$

where $\hat{\sigma}_Y^2$ reads:

$$\hat{\sigma}_Y^2 = \frac{1}{N-1}\sum_{k=1}^{N}\left(y^{(k)} - \bar{y}\right)^2, \quad \bar{y} = \frac{1}{N}\sum_{k=1}^{N} y^{(k)} \tag{3.122}$$

Finally, one generally refers to the *coefficient of determination* R^2 (Eq. (3.123)) to evaluate the accuracy of the metamodel.

$$R^2 = 1 - \epsilon_{emp} \tag{3.123}$$

A R^2 close to 1 usually indicates a good accuracy of the metamodel whereas a R^2 close to zero characterizes a poor representation of the model response. The accuracy of PC expansions is illustrated in Figure 3.6. The Ishigami function Eq. (3.124) is succesively approximated by PC expansions of order $p = 3, 5, 7, 10$. Legendre polynomials are used to be consistent with the uniformly distributed input variables. The PDF of the model response f_Y is estimated by kernel smoothing based on a sample of size $N_{ks} = 10^6$.

$$Y = \sin(X_1) + 7 \times \sin(X_2)^2 + 0.1 \times X_3^4 \times \sin(X_1) \tag{3.124}$$

with $X_i \sim \mathcal{U}[-\pi, \pi]$, $i = 1, 2, 3$.

Nevertheless, one should not be fooled by the coefficient of determination which is known for underestimating the generalization error. Indeed, due to its definition in Eq. (3.120), Err_{emp} decreases when the maximal degree p (and consequently the size of the DOE) increases while Err may increase. This phenomenon is referred to as *overfitting*. Its most popular example is the so-called *Runge effect* (Blatman, 2009).

3.3.5.2 Leave-One-Out error

The coefficient of determination R^2, which is assumed to overpredict the accuracy, is based on the same experimental design than the one used to build the PC approximation.

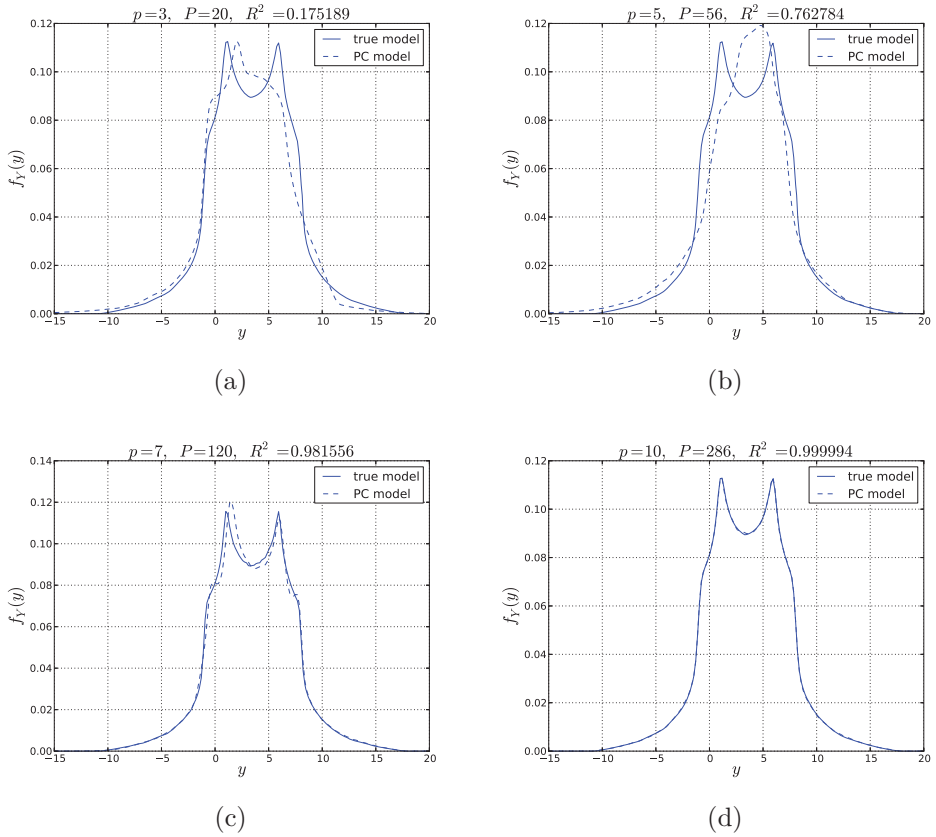


Figure 3.6: *The Ishigami function is approximated by PC expansions of order* $p = 3, 5, 7, 10$ *(from left to right and from top to bottom). The PDF of the model response is estimated by kernel smoothing. The higher the order, the most accurate the response PDF approximation. The coefficient of determination* R^2 *tends to* 1 *when* p *is increased.*

In order to estimate the approximation error more accurately, one might be tempted to compute the generalization error on a different set of points, referred to as *test set* or *validation set*, but that would lead to extra calls to the physical model.

The numerically most expensive step in building a PC expansion often lies in the calls to the model that may be a finite element analysis. Then the time to build the polynomial basis and compute the regression-based estimates of the coefficients might be negligible compared to the DOE computation time. The so-called *leave-one-out* (LOO) technique belongs to the *cross-validation* methods. The principle is to build the PC expansion $\hat{\mathcal{M}}_p^{(\sim k)}$ with the DOE points except the k^{th} one $(\boldsymbol{\mathcal{X}}, \mathcal{Y})^{(\sim k)}$ and to estimate the corresponding *predicted residual* $r^{(k)}$ at the point $\boldsymbol{x}^{(k)}$ that has been removed from the DOE, namely:

$$r^{(k)} = \mathcal{M}\left(\boldsymbol{x}^{(k)}\right) - \hat{\mathcal{M}}_p^{(\sim k)}\left(\boldsymbol{x}^{(k)}\right) \tag{3.125}$$

By successively removing the k^{th} experimental points, $k = 1, \ldots, n$ from the DOE, one

computes the leave-one-out error:

$$Err_{LOO} = \frac{1}{N} \sum_{k=1}^{N} r^{(k)2} \tag{3.126}$$

In the case of linear regression, it is possible to compute the LOO-error analytically. The predicted residuals read:

$$r^{(k)} = \frac{\mathcal{M}\left(\boldsymbol{x}^{(k)}\right) - \hat{\mathcal{M}}_p\left(\boldsymbol{x}^{(k)}\right)}{1 - h_k} \tag{3.127}$$

where h_k is the k^{th} term of $\text{diag}(\boldsymbol{\Psi}\left(\boldsymbol{\Psi}^\mathsf{T}\boldsymbol{\Psi}\right)^{-1}\boldsymbol{\Psi}^\mathsf{T})$, $\boldsymbol{\Psi}$ being the information matrix in Eq. (3.110). The leave-one-out error may be recast as follows:

$$Err_{LOO} = \frac{1}{N} \sum_{k=1}^{N} \frac{\mathcal{M}\left(\boldsymbol{x}^{(k)}\right) - \hat{\mathcal{M}}_p\left(\boldsymbol{x}^{(k)}\right)}{1 - h_k} \tag{3.128}$$

and similarly to the training error, one also defines the relative LOO-error, namely:

$$\epsilon_{LOO} = \frac{Err_{LOO}}{\hat{\sigma}_Y^2} \tag{3.129}$$

and finally, the coefficient of determination R^2 is replaced by Q^2:

$$Q^2 = 1 - \epsilon_{LOO} \tag{3.130}$$

The PC expansions equipped with the abovementioned accuracy indicators provide the practitioner with a powerful tool to build a metamodel from a set of data with an approximation error that is controlled.

3.4 Conclusion

The first part of this chapter gives an overview on several surrogate modelling techniques. The Support Vector Regression, based on the wider branch of Machine Learning, consists in solving a quadratic optimization problem. Its advantage is that the response surface can be built using only a limited number of data points and offers a low sensitivity to *outliers*, *i.e.* observations far from the global trend. Gaussian processes, also known as *Kriging* has been developed for geostatistical applications. It offers the practitioner a stochastic representation of the data since the model is assumed to be a sample path of a Gaussian random field. Then the parameters of the process are estimated by maximum likelihood. Gaussian processes are interesting because they interpolate the data and provide a approximation error through the Kriging variance. Finally the high-dimensional model representation has been developed to model the behavior of chemical models with dozens of input parameters. It consists in a projection of the model on polynomial basis

whose terms are estimated by projection. Note that when the polynomial basis is orthogonal the approach is equivalent to a PC expansions whose coefficients are computed by projection as seen in from the comparison of Eqs. (3.71, 3.72) and Eq. (3.101).

In the second part, the polynomial chaos expansion is presented. The approach relies upon the expansion of the model response onto a suitable polynomial basis. In its simplest form, *i.e.* when the model response is expanded onto the full basis, computing all the expansion coefficients by regression is quite a numerical burden. Advanced solutions based on sparse representations are introduced. The variable selection problem can be solved by neglecting the high-order interaction terms of the expansion or by using adaptive algorithm that successively add and/or suppress polynomials in the basis in order to reach the highest accuracy with the minimum number of terms. Finally, the approximation error is studied through different indicators of accuracy. In this work, the practical implementation for PC applications has been processed using the Python package OpenTURNS (2005).

In Chapter 4, it is shown how surrogate models can be used in order to compute accurately the sensitivity indices that have been introduced in Chapter 2.

4

Computing sensitivity indices using surrogate models

Contents

4.1 Introduction

It has been observed in Chapter 2 that computing global sensitivity indices accurately might be numerically very expensive. The number of calls to the physical model $\mathcal{M}$ approaching 10^3 for a first approximation of an index value makes it almost impossible to perform GSA within a reasonable time when the physical model is a time-demanding process (complex code, finite element simulation). Therefore, several surrogate modelling techniques have been presented in Chapter 3 and more particularly polynomial chaos expansions. This chapter now proposes methodologies to compute Sobol', Borgonovo and ANCOVA indices using polynomial chaos expansions. This work will be shortly published in Caniou and Sudret (2012).

4.2 Postprocessing PC coefficients

Polynomial chaos expansions (PCE) allows one to represent a physical model $\mathcal{M}$ along a polynomial basis which is optimal for the probabilistic description of the input random vector $\boldsymbol{X}$. The PCE of the random response $Y = \mathcal{M}(\boldsymbol{X})$ reads:

$$\hat{Y} = \sum_{j=1}^{P-1} y_j \Psi_j(\boldsymbol{X}) \tag{4.1}$$

The random response Y is fully described by the coefficients of the development y_j, $j = 1, \dots P-1$ which can be post-processed to compute quantities of interest such as its probability density function or its statistical moments (Sudret, 2008). Then, it will be shown how Sobol' sensitivity indices can be derived from the same coefficients without large Monte Carlo simulations. Before presenting the exploitation of the PCE coefficients, a distribution approximation technique referred to as *kernel smoothing approximation* is introduced.

4.2.1 Kernel smoothing approximation

The *kernel smoothing approximation* offers a smoother representation of a random variable probability density function (PDF) than a basic histogram (Wand and Jones, 1995). It consists in a superposition of standard distributions at each observation of the sample as shows Figure 4.1 that is normed in order to verify:

$$\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1 \tag{4.2}$$

Let us consider a set of N observations $\left\{y^{(1)}, \dots, y^{(N)}\right\}$ of a random variable Y with PDF f_Y. The kernel approximation of f_Y reads:

$$\hat{f}_Y(y) = \frac{1}{N h_K} \sum_{i=1}^{N} K\left(\frac{y - y^{(i)}}{h_K}\right) \tag{4.3}$$

where K is a kernel function and h_K its bandwidth parameter. One usually uses the Gaussian kernel function, namely:

$$K(x) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}x^2\right] \tag{4.4}$$

which is nothing but the standard Gaussian PDF expression. The optimal bandwith for the Gaussian kernel is the *Silverman bandwidth* (Silverman, 1986), namely:

$$h_K = \frac{0.9\, m}{N^{\frac{1}{5}}} \quad \text{with} \quad m = \min\left(\hat{\sigma}_X, \frac{IQR(X)}{1.349}\right) \tag{4.5}$$

where N is the size of the sample, $\hat{\sigma}_X$ the sample standard deviation and $IQR(X)$ its interquartile range. The Silverman bandwidth is the value that minimizes the *mean integrated squared error* (MISE) in the case of an underlying Gaussian distribution f:

$$MISE\left[\hat{f}(\cdot;h)\right] = \mathrm{E}\left[\int \left(\hat{f}(x;h_K) - f(x)\right)^2\right] \mathrm{d}x \tag{4.6}$$

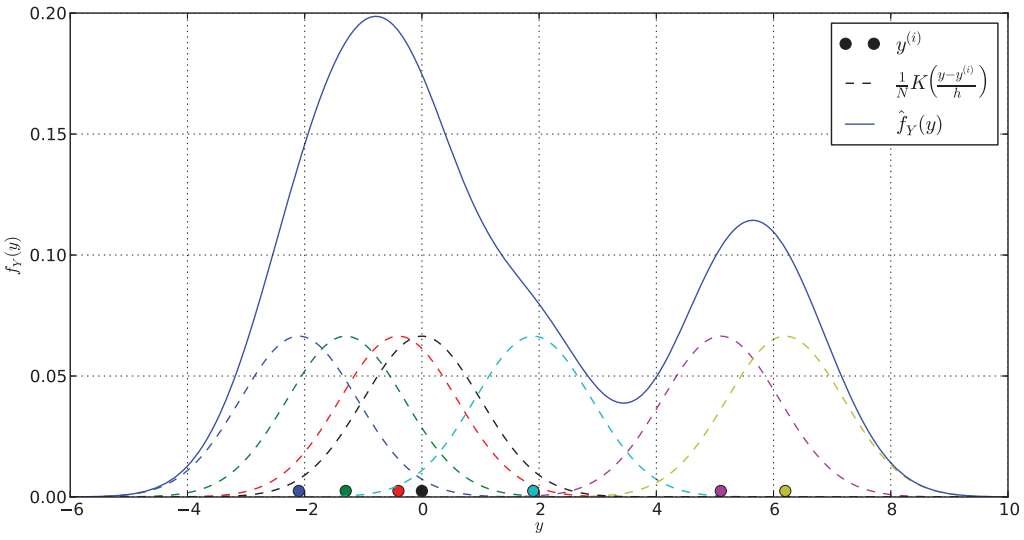


Figure 4.1: *Principle of kernel smoothing estimation using the Gaussian kernel.*

The drawback of the Gaussian kernel is that it is defined on $\mathbb{R}$. Consequently, simulating pseudo-observations from the kernel smoothing estimation of a strictly positive random variable might lead to negative values for at low probabilities which can be inconvenient. In order to circumvent this issue, one usually prefers the *Epanechnikov kernel* Eq. (4.7) which is presented in Figure 4.2.

$$E(x) = \begin{cases} \frac{3}{4}(1-x^2) & \text{if} \quad |x| \leqslant 1 \\ 0 & \text{if} \quad |x| > 1 \end{cases} \tag{4.7}$$

According to Eq. (4.7), the Epanechnikov kernel is bounded and that property prevents one from observing negative realizations from the smoothed PDF. The optimal bandwidth for the Epanechnikov kernel reads:

$$h_K^E = \hat{\sigma}_X \left(\frac{40\sqrt{\pi}}{N} \right)^{\frac{1}{5}} \tag{4.8}$$

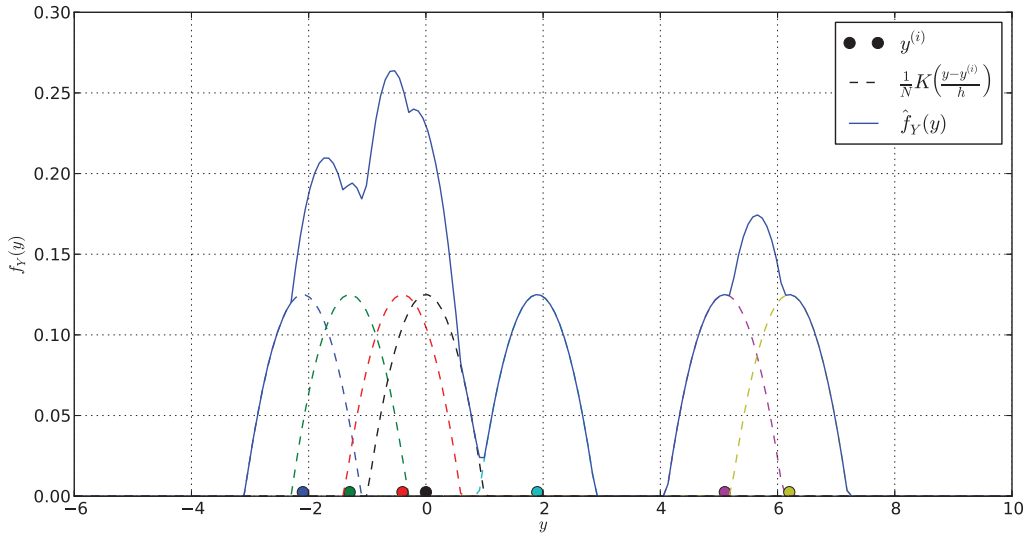


Figure 4.2: *Principle of kernel smoothing estimation using the Epanechnikov kernel.*

Kernel smoothing estimation tends to the true PDF of the random variable when the size of the sample increases. Therefore, the larger the sample, the more accurate the estimation. Moreover, when $N > 10^3 - 10^4$, the approximated density is almost independent of the choice of the kernel function.

4.2.2 Probability density function and statistical moments of the random response

4.2.2.1 Probability density function

The PDF f_Y of the model response Y can be computed from a set of observations $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(N)}\}$ of the random response obtained by performing MC simulations on the surrogate model. The kernel smoothing approximation $\hat{f}_Y$ of the response PDF reads:

$$\hat{f}_Y(y) = \frac{1}{N h_K} \sum_{i=1}^{N} K\left(\frac{y - y^{(i)}}{h_K} \right) \tag{4.9}$$

As the model approximation $\hat{\mathcal{M}}$ is analytical and polynomial, these simulations are far cheaper than those that could have been performed using the real model $\mathcal{M}$. Therefore,

large samples ($N = 10^5 - 10^6$) can be used in order to perfectly describe the PDF of the model response.

4.2.2.2 Statistical moments

The statistical moments of the model response can be estimated by their empirical estimators using the set of pseudo-observations previously described. However, they can also be obtained directly from the coefficients of the development due to the orthogonality properties of the polynomial basis functions. The first and second order moments of the random variable, that is the mean value and the variance, respectively read:

$$\mu_Y^{PC} \equiv \mathrm{E}\left[Y\right] = y_0 \tag{4.10}$$

$$\sigma_Y^{2,PC} \equiv \mathrm{Var}\left[Y\right] = \sum_{i=1}^{P-1} y_i^2 \tag{4.11}$$

The third and fourth order moments of the random variable, that is the skewness and the kurtosis, can also be directly computed by:

$$\delta_Y^{PC} \equiv \frac{1}{\sigma_Y^{3,PC}} \mathrm{E}\left[(Y - y_0)^3\right] = \frac{1}{\sigma_Y^{3,PC}} \sum_{i=1}^{P-1}\sum_{j=1}^{P-1}\sum_{k=1}^{P-1} d_{ijk} y_i\, y_j\, y_k \tag{4.12}$$

$$\kappa_Y^{PC} \equiv \frac{1}{\sigma_Y^{4,PC}} \mathrm{E}\left[(Y - y_0)^4\right] = \frac{1}{\sigma_Y^{4,PC}} \sum_{i=1}^{P-1}\sum_{j=1}^{P-1}\sum_{k=1}^{P-1}\sum_{k=1}^{P-1} d_{ijkl} y_i\, y_j\, y_k\, y_l \tag{4.13}$$

with $d_{ijk} = \mathrm{E}\left[\Psi_i(\boldsymbol{x})\Psi_j(\boldsymbol{x})\Psi_k(\boldsymbol{x})\right]$ and $d_{ijkl} = \mathrm{E}\left[\Psi_i(\boldsymbol{x})\Psi_j(\boldsymbol{x})\Psi_k(\boldsymbol{x})\Psi_l(\boldsymbol{x})\right]$. According to (Sudret, 2008), a second-order PCE allows one to compute the mean and variance accurately, whereas at least a third-order expansion is required for the precise estimation of the skewness and kurtosis.

4.2.3 Sensitivity indices

This subsection highlights the relationship between the functional decomposition used in the ANOVA and the polynomial representation provided by the PCE. This results were originally published in Sudret (2006, 2008).

Let us consider the previously described PC expansion of the model response Y truncated at the degree p:

$$\hat{y} = \sum_{|\boldsymbol{\alpha}| \leqslant p} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \tag{4.14}$$

Eq. (4.14) shows a multi-index notation where $\boldsymbol{\alpha} = \{\alpha_1, \ldots, \alpha_n\} \in \mathbb{N}^n$ denotes all the possible $n-$uplets and:

$$\Psi_\alpha(\boldsymbol{x}) = \prod_{i=1}^{n} \psi_{\alpha_i}^i(x_i) \tag{4.15}$$

In this equation, the polynomials $\psi^i_{\alpha_i}$ are orthogonal with respect to the probability measure associated with the i^{th} random variable X_i. Let us also define the set of multi-indices $\mathcal{I}_{i_1,\ldots,i_s}$ by:

$$\mathcal{I}_{i_1,\ldots,i_s} = \left\{ \boldsymbol{\alpha} : \begin{array}{lll} \alpha_k > 0 & \forall k = 1,\ldots,n & k \in \{i_1,\ldots,i_s\} \\ \alpha_k = 0 & \forall k = 1,\ldots,n & k \notin \{i_1,\ldots,i_s\} \end{array} \right. \tag{4.16}$$

$\mathcal{I}_{i_1,\ldots,i_s}$ corresponds to the set of α's for which only the indices $i_1,\ldots,i_s$ are non zero. For example, $\mathcal{I}_i$ refers to the polynomials depending only on X_i. Using the notation in Eq. (4.16), the summands in Eq. (4.14) are no longer functions of the full input random vector $\boldsymbol{X}$ but are now gathered according to the parameters they only depend on, namely:

$$\begin{aligned} \hat{\mathcal{M}}(\boldsymbol{x}) = y_0 & \\ & + \sum_{i=1}^{n} \sum_{\boldsymbol{\alpha} \in \mathcal{I}_i} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(x_i) \\ & + \sum_{1 \leqslant i_1 < i_2 \leqslant n} \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,i_2}} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(x_{i_1}, x_{i_2}) \\ & \ldots \\ & + \sum_{1 \leqslant i_1 < \ldots < i_s \leqslant n} \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,\ldots,i_s}} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(x_{i_1},\ldots,x_{i_s}) \end{aligned} \tag{4.17}$$

This expression clearly allows one to identify which terms depend on which variables or set of variables and represents in a certain sense a functional decomposition comparable to the one from the ANOVA, that is:

$$\mathcal{M}_{i_1,\ldots,i_s}(x_{i_1},\ldots,x_{i_s}) \equiv \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,\ldots,i_s}} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(x_{i_1},\ldots,x_{i_s}) \tag{4.18}$$

Due to the uniqueness of the Sobol' decomposition, the equivalence relationship in Eq. (4.17) is actually an equality.

Thus, the Sobol' indices can be directly computed from the previous representation. The *PC-based Sobol' indices* are expressed in terms of the PCE coefficients:

$$S^{PC}_{i_1,\ldots,i_s} = \frac{1}{\sigma_Y^{2,PC}} \sum_{\boldsymbol{\alpha} \in \mathcal{I}_{i_1,\ldots,i_s}} y_{\boldsymbol{\alpha}}^2 \tag{4.19}$$

where $\sigma_Y^{2,PC}$ is the *total variance* of Y. The PC-based first order Sobol' index $S^{PC}_{i_1,\ldots,i_s}$, which is the ratio of the partial variance that is due to $(X_{i_1},\ldots,X_{i_s})$ and the total variance, is the ratio of the sum of the squared coefficients of the polynomial depending only on $(X_{i_1},\ldots,X_{i_s})$ and the sum of all the squared coefficients. The author also defines the PC-based total indices $S^{T,PC}_{j_1,\ldots,j_t}$, namely:

$$S^{T,PC}_{j_1,\ldots,j_s} = \sum_{\{i_1,\ldots,i_s\} \subset \{j_1,\ldots,j_t\}} S^{PC}_{i_1,\ldots,i_s} \tag{4.20}$$

Note that when the input random vector $\boldsymbol{X}$ has independent components, the PC-indices are equivalent to the Sobol' indices but if the independence is not verified, the

PC-indices reflects the contribution of the decorrelated variables (after applying the Nataf transformation) which can drastically differ from the input variables.

This section has shown that the PCE provides valuable information on the model response at a numerical cost that is insignificant compared to MCS on the real model once the surrogate model has been built. The cost for all these statistical quantities is limited to the evaluation of the points in the design of experiments.

4.3 Borgonovo importance measure

The *Borgonovo importance measure* δ has been introduced in section 2.5. This distribution-based sensitivity index can be calculated in two ways, either using a PDF-definition or a CDF definition. In this section, computing scheme for both ways are presented.

4.3.1 PDF-based estimation scheme

This work has been originally published in Caniou and Sudret (2011). The importance measure δ_i describing the contribution of the input variable X_i to the distribution of the model response Y reads:

$$\begin{aligned} \delta_i &= \frac{1}{2}\mathrm{E}\left[s(X_i)\right] \\ &= \int_{\mathrm{D}_{X_i}} \left[\int_{\mathrm{D}_Y} \left| f_Y(y) - f_{Y|X_i}(y) \right| \mathrm{d}y \right] \mathrm{d}x_i \end{aligned} \tag{4.21}$$

The estimation of δ_i consists in the evaluation of two integrals, namely the first one for the shift $s(x_i)$ and the second one for its expected value over the support of X_i. Let us assume that the unconditional PDF $f_Y(y)$ and the conditionals PDFs $f_{Y|X_i}(y)$ are not known. In his original paper, Borgonovo (2007) proposes to identify the different PDFs using *maximum likelihood estimation* (MLE) validated by a Kolmogorov-Smirnov goodness-of-fit test. If this method represents a possible solution for this task, it has to be noticed that there are no obvious reason for the model response to follow a common distribution (Gaussian, lognormal, Weibull). For example, the PDF of Y could be bimodal. *Kernel smoothing estimation* (see section 4.2.1) is a more general approach whose only hypothesis requires the PDF to be continuous. In the same paper, the expected value of the shift $\mathrm{E}\left[s(x_i)\right]$ is estimated by performing 10^3 MC simulations. In this subsection, an improved and more robust estimation procedure of δ_i is proposed.

4.3.1.1 Kernel smoothing estimation of the PDFs

Let us consider two $N-$samples of pseudo-observations $\mathcal{Y} = \{y^{(1)}, \ldots, y^{(N)}\}$ and $\mathcal{Y}^{X_i} = \{z^{(1)}, \ldots, z^{(N)}\}$ that have been numerically simulated from a polynomial chaos expansion

with respectively no parameter fixed and X_i fixed at the value x_i. The kernel smoothing estimation of the unconditional and conditional PDFs respectively read:

$$\hat{f}_Y(y) = \frac{1}{Nh_K}\sum_{i=1}^{N} K\left(\frac{y-y^{(i)}}{h_K}\right) \quad \text{and} \quad \hat{f}_{Y|X_i}(y) = \frac{1}{Nh_K}\sum_{i=1}^{N} K\left(\frac{y-z^{(i)}}{h_K}\right) \tag{4.22}$$

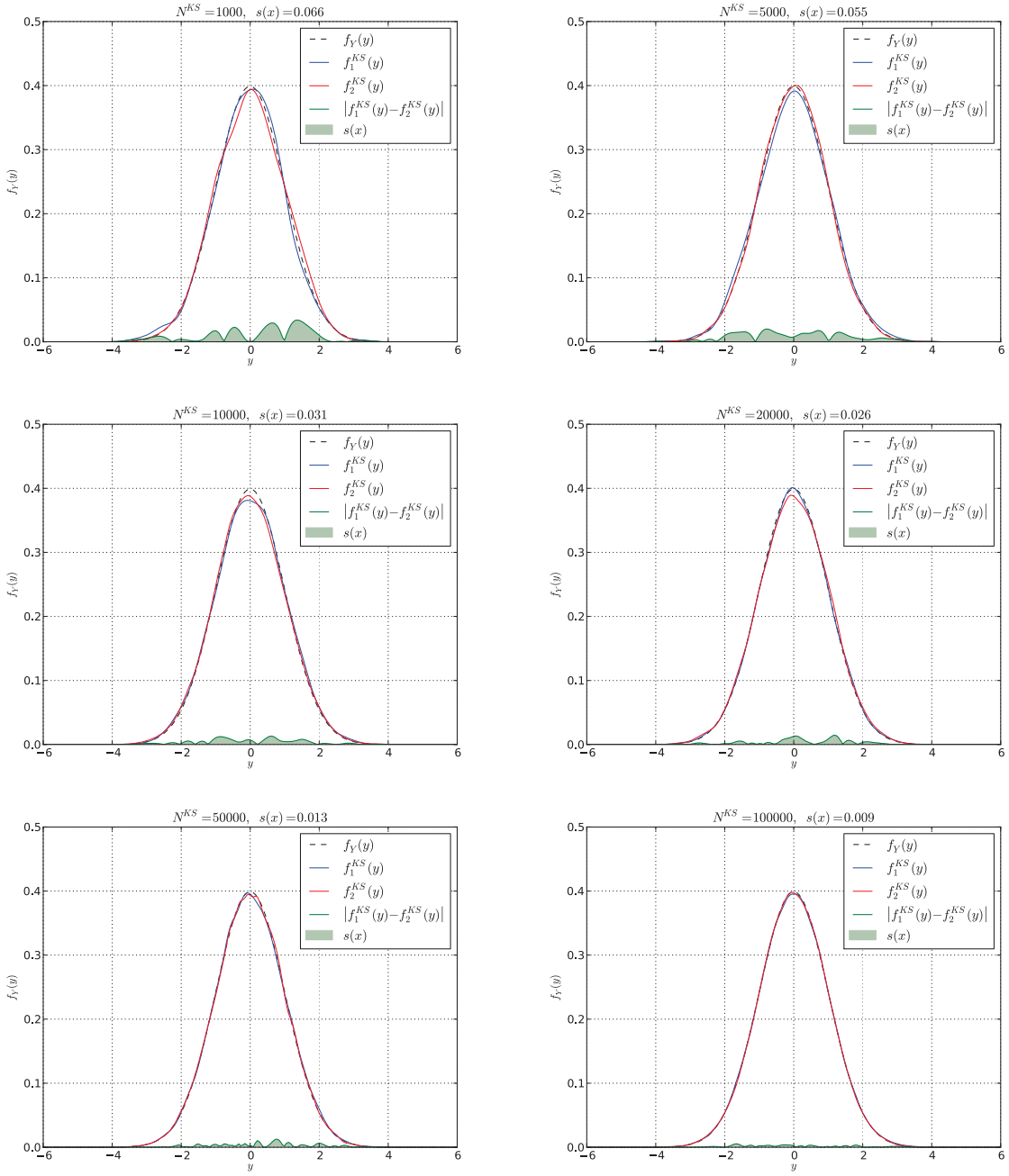
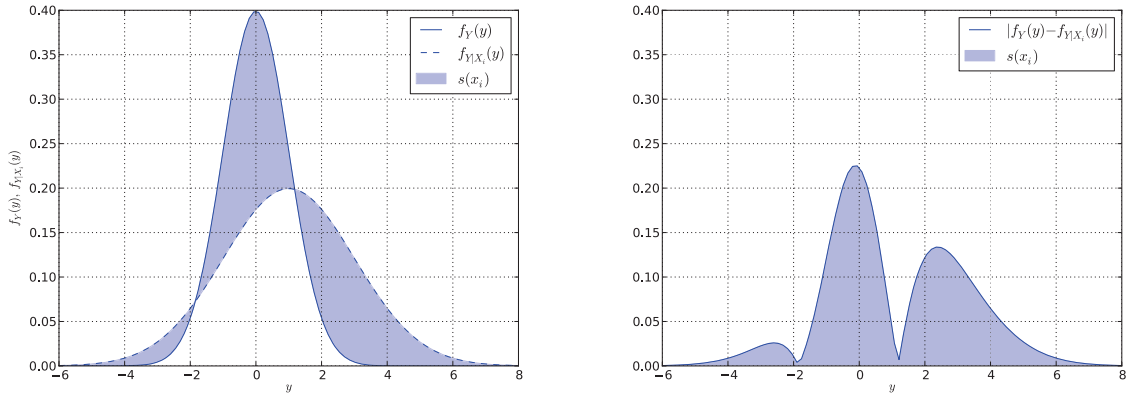


Figure 4.3: *Convergence of the shift between two kernel smoothing estimations of the standard Gaussian PDF $\mathcal{N}(0,1)$ when the size N of the sample is increased.*

It has been shown in Subsection 4.2.1 that the accuracy of the kernel smoothing estimation (KSE) of the PDFs is logically growing with the size N of the sample. The convergence of the KSE is illustrated in Figure 4.3. The shift between two kernel smoothing estimations of the same standard Gaussian PDF $\mathcal{N}(0,1)$, thus supposed to be zero, is studied. If the shift is larger than 0.05 for $N = 1000 - 5000$ (a and b), it becomes smaller than 10^{-2} for $N = 10^5$. In the sequel, $N = 10^5$ will be used as the kernel smoothing sampling size since the corresponding accuracy is acceptable considering that the quantity of interest is a sensitivity index.

4.3.1.2 Estimation of the shift

The inner integral in Eq. (4.21) corresponds to the shift $s(x_i)$, that is the area between $f_Y(y)$ and $f_{Y|X_i}(y)$. An simple illustration for two Gaussian distributions, namely $Y \sim \mathcal{N}(0,1)$ and $Y|X_i \sim \mathcal{N}(1,2)$, is given in Figure 4.4.



(a) (b)

Figure 4.4: *The area between the two PDFs (a) is flattened (b) to ease the integration scheme.*

Classical integration procedures such as trapezoidal rules offer robust solutions but their convergence rate is quite low. In order to evaluate the shift $s(x_i)$ with accuracy, a more efficient integration scheme, namely the Gaussian quadrature rule, is preferred. A quadrature rule allows one to approximate an integral I with a weighted sum of function evaluations at specific points. The Gauss-Legendre quadrature rule reads:

$$I = \int_a^b f(x)\,\mathrm{d}x \approx \sum_{k=1}^{N_q} \omega_k\, f(x_k) \tag{4.23}$$

where the x_k are the integration points and the ω_k the corresponding roots. When the integral in standardized to the domain $[-1,1]$, the x_k's are the roots of the N_q^{th} Legendre polynomial. *Legendre polynomials* are orthogonal polynomials defined by the so-called

Bonnet recurrence formula:

$$P_0(x) = 0 \tag{4.24}$$

$$P_1(x) = x \tag{4.25}$$

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x), \quad \forall n > 0 \tag{4.26}$$

The six first Legendre polynomials are pictured in Figure 4.5. The weights ω_k are defined by (Abbott, 2005):

$$\omega_k = \frac{-2}{(n+1)P_n'(x_k)P_{n+1}(x_k)} = \frac{4}{nP_{n+2}(x_k)P_n'(x_{k+2})} \tag{4.27}$$

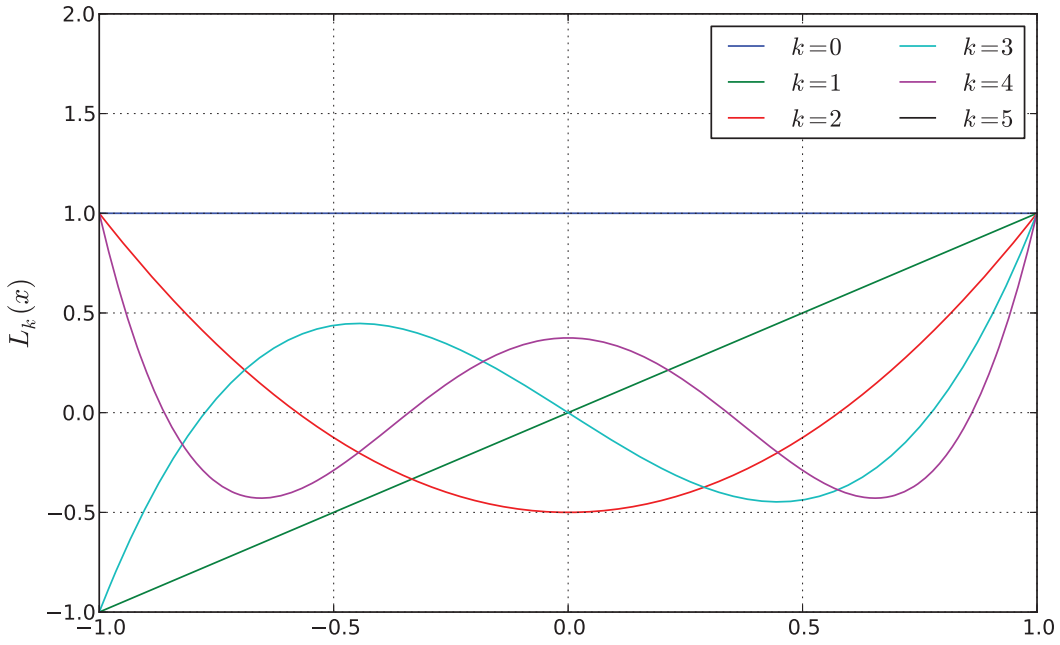


Figure 4.5: *Legendre polynomials over* $[-1, 1]$ *for* $k = 0, \ldots, 5$.

For a Gauss-Legendre quadrature rule over the domain $[-1, 1]$ with $N_q = 3$, the integration points and weights are respectively $\{-\sqrt{3/5}, 0, \sqrt{3/5}\}$ and $\{5/9, 8/9, 5/9\}$. For any interval $[a, b]$, the following linear transformation has to be applied:

$$I = \int_a^b f(x)\,\mathrm{d}x = \frac{b-a}{2}\int_{-1}^{1} f\left(\frac{b-a}{2}x + \frac{a+b}{2}\right)\mathrm{d}x \tag{4.28}$$

Then, the integral I can be approximated by:

$$I \approx \frac{b-a}{2}\sum_{k=1}^{N_q}\omega_k\, f\left(\frac{b-a}{2}x_k + \frac{a+b}{2}\right) \tag{4.29}$$

Thus, the shift $s(x_i)$ can be approximated by:

$$s(x_i) \approx \frac{b-a}{2}\sum_{k=1}^{N_q}\omega_k \left| f_Y\left(\frac{b-a}{2}y_k + \frac{a+b}{2}\right) - f_{Y|X_i}\left(\frac{b-a}{2}y_k + \frac{a+b}{2}\right)\right| \tag{4.30}$$

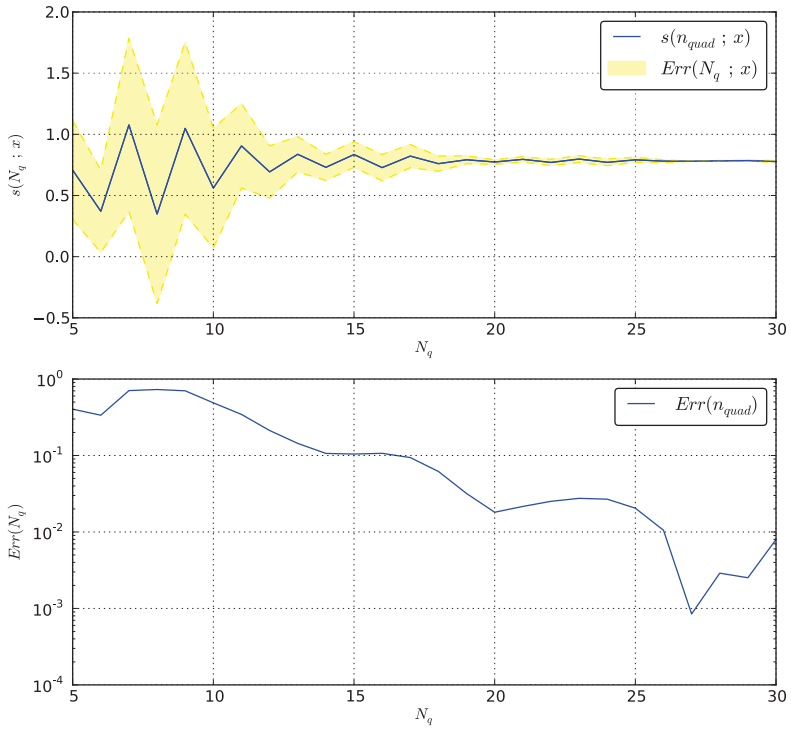


Figure 4.6: *Convergence of the Gauss-Legendre quadrature scheme. The integration error (bottom) at the k^{th} iteration is the difference between the integrals I_{k-1} and I_k.*

where a and b are respectively defined by the q^{th} and $1 - q^{\text{th}}$ quantiles ($q = 10^{-6}$ for instance) of the model response unconditional distribution.

A study of the convergence of the quadrature rule is presented in Figure 4.6. The same test case as in Figure 4.4 is used with $s(x_i) = 0.78$. This study shows that a quadrature order $N_q = 30$ is sufficient to reach a precision of 10^{-2} for the shift where hundreds of KSE evaluations would have been necessary to reach the same accuracy with a trapezoidal rule. In the sequel, $N_q = 30$ quadrature points will be used for the estimation of the shift.

4.3.1.3 Estimation of the expected shift

The outer integral in Eq. (4.21), that is the expected value of the shift $\mathrm{E}\left[s(x_i)\right]$ can also be estimated using another second quadrature rule. The quadrature rule for a the estimation of the expected value of the shift with respect to the probability measure $f_{X_i}(x_i)\,\mathrm{d}x_i$ reads:

$$\mathrm{E}\left[s(X_i)\right] = \int s(x_i)\,\mathbb{P}_{X_i}(\mathrm{d}x_i) \approx \sum_{k=1}^{N_q} \omega_k\, s(x_k) \tag{4.31}$$

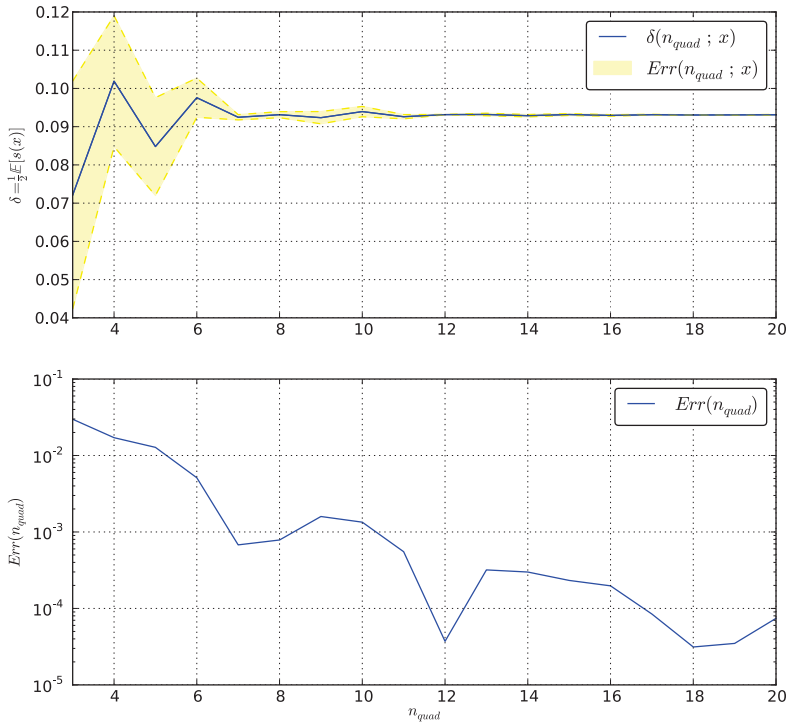


Figure 4.7: *Convergence of the quadrature scheme for the expected shift. The integration error (bottom) at the k^{th} iteration is the difference between the integrals I_{k-1} and I_k.*

where the integration weights and points are now computed from the polynomials that are orthogonal to $\mathbb{P}_{X_i}(\mathrm{d}x_i) = f_{X_i}(x_i)\,\mathrm{d}x_i$. Alternatively, using the Gauss-Legendre integration rule, the estimation reads:

$$\mathrm{E}\left[s(X_i)\right] = \frac{b-a}{2}\sum_{k=1}^{N_q}\omega_k\, s\left(\frac{b-a}{2}x_k + \frac{a+b}{2}\right) f_{X_i}\left(\frac{b-a}{2}x_k + \frac{a+b}{2}\right) \tag{4.32}$$

where a and b are now the q^{th} and $1-q^{\text{th}}$ quantiles of the distribution of X_i. This last method is preferred to the first one for the sake of computer programming simplicity.

Finally, the global expression of the estimator $\hat{\delta}_i$ of δ_i reads:

$$\hat{\delta}_i = \frac{1}{2}\sum_{k=1}^{N_q}\omega_k \sum_{l=1}^{N_q}\omega_l \left|\hat{f}_Y(y_l) - \hat{f}_{Y|X_i=x_k}(y_l)\right| f_{X_i}(x_k) \tag{4.33}$$

The convergence of the quadrature scheme is pictured in Figure 4.7. With $N_q = 30$, the estimation of δ_i requires *only* 31 KSE of the PDFs (1 unconditional and 30 conditional) which is already numerically expensive considering each KSE is built from 10^5 pseudo-observations.

Note that the consistence of the estimator $\hat{\delta}_i$ of δ_i has been shown in (Plischke et al., 2012).

4.3.1.4 Diagram of the procedure

To summarize the PDF-based estimation scheme, the procedure is illustrated by the algorithm in Figure 4.8.

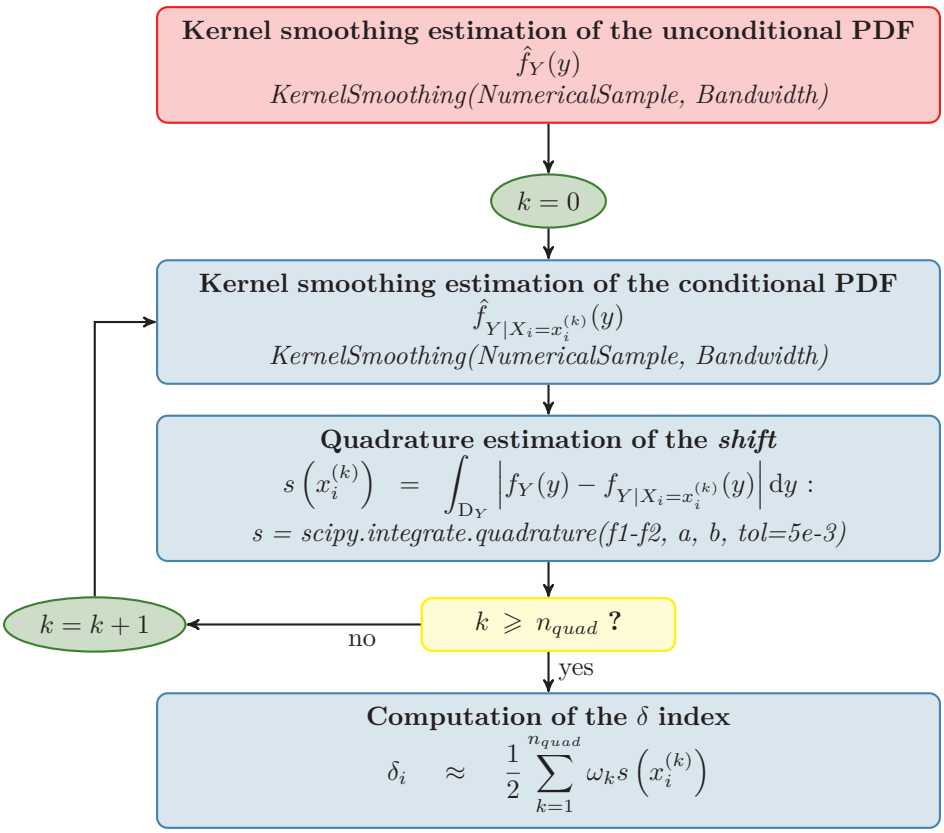


Figure 4.8: *Diagram of the PDF-based estimation of the δ sensitivity measure.*

4.3.2 CDF-based estimation scheme

The estimation of Borgonovo δ indices using kernel smoothing estimation of the PDF is quite expensive due to the large size of the samples involved in the process. A new estimation procedure based on the cumulative distribution functions has been proposed in Borgonovo et al. (2011). In this case, the shift reads (see Chapter 2, Section 2.5):

$$s(X_i = x_i^*) = 2F_Y(\mathrm{D}_{Y,X_i}^+) - 2F_{Y|X_i=x_i^*}(\mathrm{D}_{Y,X_i}^+) \tag{4.34}$$

where D_{Y,X_i}^+ is the domain where $f_Y(y) > f_{Y|X_i}(y)$.

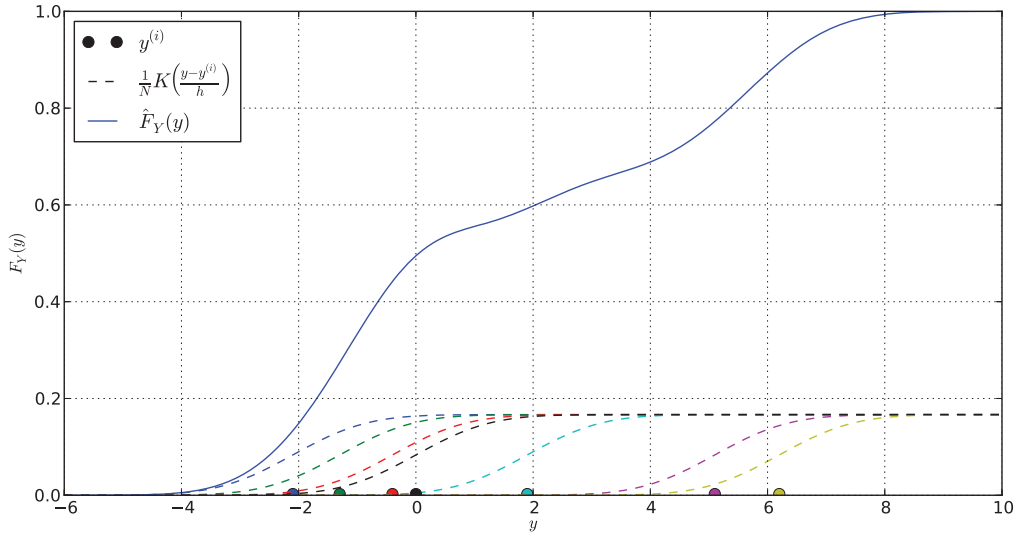


Figure 4.9: *Kernel smoothing estimation of the cumulative distribution function from a sample.*

4.3.2.1 Kernel smoothing estimation of the CDFs

The kernel smoothing estimation of a cumulative distribution function is comparable to a cumulative histogram representation. The same principle is used but now the kernel function is a typically a CDF. An example is pictured in Figure 4.9. The kernel function is the cumulative distribution function of the standard Gaussian distribution:

$$\begin{aligned} K(x) &= \Phi(x) \\ &= \frac{1}{2}\left(1 + \operatorname{erf}\left(\frac{x}{\sqrt{2}}\right)\right) \end{aligned} \tag{4.35}$$

where *erf* denotes the error function.

4.3.2.2 Estimation of the shift

According to Eq. (4.34), estimating the shift consists in determining the bounds of the domains D^+_{Y,X_i} and D^-_{Y,X_i}, that is the points y_1 and y_2 at which $f_Y(y) = f_{Y|X_i}(y)$. According to the relation between the PDF and the CDF of a random variable, one gets for all interval $[a, b]$:

$$\int_0^a f_Y(y)\,\mathrm{d}y = F_Y(a) \tag{4.36}$$

$$\int_a^b f_Y(y)\,\mathrm{d}y = F_Y(b) - F_Y(a) \tag{4.37}$$

These properties are illustrated in Figure 4.10. Thus, the goal is to solve a *root-finding problem* for a continuous function $g(y) = f_Y(y) - f_{Y|X_i}(y)$ on interval $[a, b]$. Although a

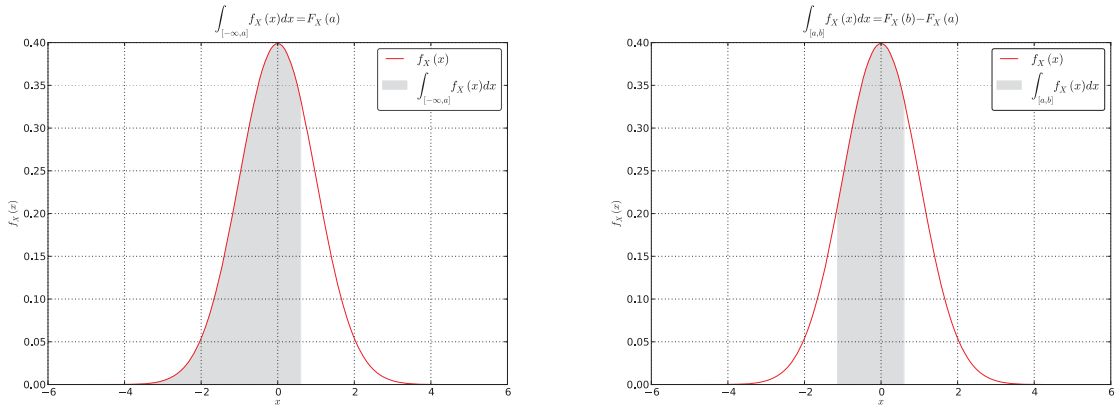


Figure 4.10: *Illustration of the analytical integration of a PDF.*

bisection method may converge slower than a *Newton-Raphson* algorithm, it represents a robust alternative to find both roots with a controlled accuracy. In order to find good initial guesses a and b, the support of Y is first discretized with a grid on which the function g is evaluated. Then, it becomes easy to find two pairs, one for each root, of consecutive terms with opposite signs.

Once the roots y_1, y_2, $y_1 < y_2$ have been found, the shift is computed by evaluating the smoothed CDFs $\hat{F}_Y$ and $\hat{F}_{Y|X_i}$ at the roots:

$$s(X_i) = 2\left[\hat{F}_Y(y_1) + \hat{F}_{Y|X_i}(y_2) - \hat{F}_Y(y_2) - \hat{F}_{Y|X_i}(y_1)\right] \tag{4.38}$$

The principle of the estimation of δ is illustrated in Figure 4.11.

4.3.2.3 Estimation of the expected shift

The last step of the evaluation of δ_i is the computation of the expected value of the shift $s(X_i)$. In the same manner as for the PDF-based estimation scheme, a quadrature rule is preferred to Monte Carlo simulations for the estimation of the expected shift. The estimator of the sensitivity measure δ_i finally reads:

$$\hat{\delta}_i = \sum_{k=1}^{N_q} \omega_q \left[\hat{F}_Y(y_1) + \hat{F}_{Y|X_i=x_q}(y_2) - \hat{F}_Y(y_2) - \hat{F}_{Y|X_i=x_q}(y_1)\right] \tag{4.39}$$

4.3.2.4 Diagram of the procedure

To summarize the CDF-based estimation scheme, the procedure is illustrated by the algorithm in Figure 4.12.

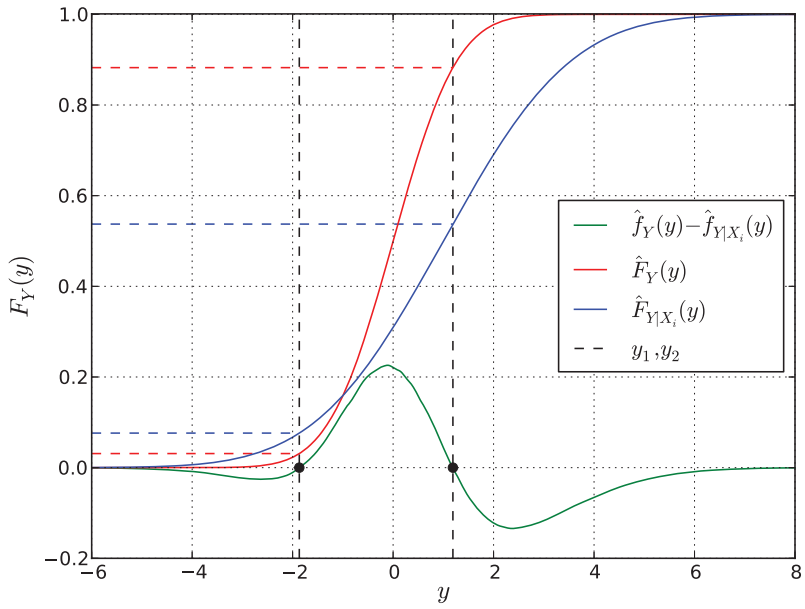


Figure 4.11: *Principle of the estimation of the δ index using CDFs.*

4.3.3 A comparison example

Let us consider a numerical model $\mathcal{M}$ defined by:

$$Y = \mathcal{M}(\boldsymbol{X}) = X_1 + X_2 \tag{4.40}$$

where $X_1 \sim \mathcal{N}(0., 1.)$ and $X_2 \sim \mathcal{N}(1., 2.)$. The δ measures are now computed with both PDF and CDF approaches and for two sizes of sample, *i.e.* 10^3 and 10^4 and $N_q = 15$. The results presented in Table 4.1 are also illustrated in Figure 4.13. The relative error ϵ_{rel} between the estimation based on two different sampling size reads:

$$\epsilon_{rel} = \frac{|\delta^{(10^3)} - \delta^{(10^4)}|}{\delta^{(10^4)}} \tag{4.41}$$

Indices	$N_{ks} = 10^3$	$N_{ks} = 10^4$	ϵ_{rel}
δ_1^{PDF}	0.151	0.161	0.062
δ_1^{CDF}	0.153	0.157	0.025
δ_2^{PDF}	0.521	0.493	0.057
δ_2^{CDF}	0.498	0.506	0.016

Table 4.1: *PDF-based and CDF-based estimators of the δ indices for two sizes of sample.*

The first observation that can be done is that samples of size $N_{ks} = 10^3$ are too small for an accurate estimation of the PDF whereas it seems sufficient for a CDF estimation.

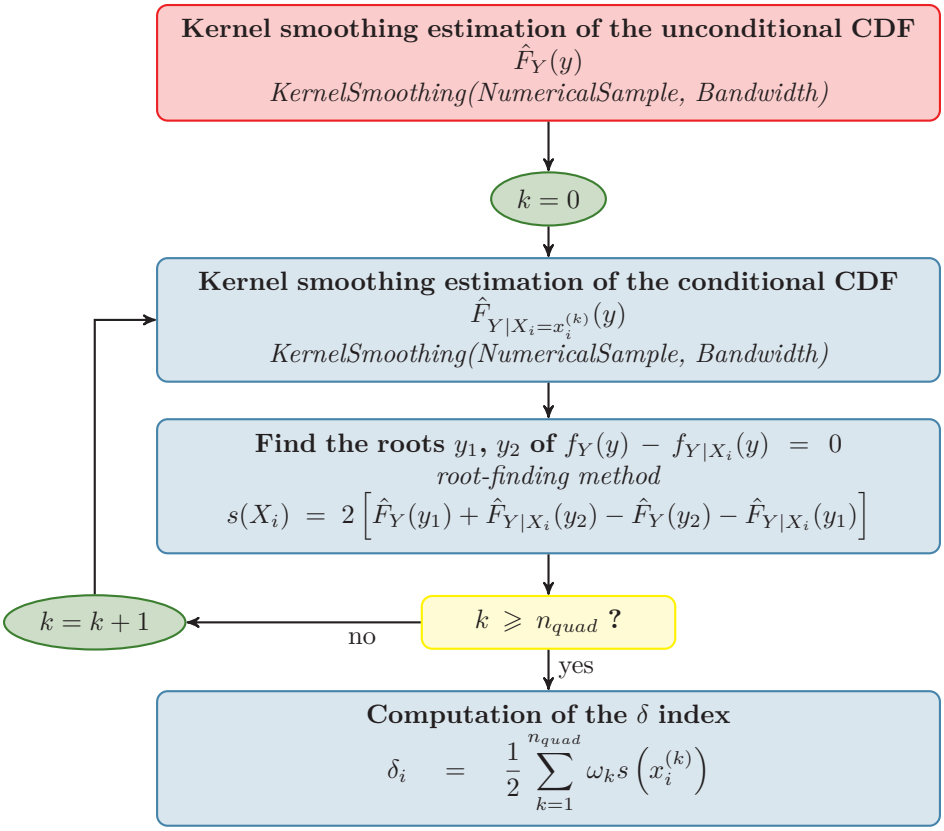


Figure 4.12: *Diagram of the CDF-based estimation of the δ sensitivity measure.*

The second observation is that the δ indices are less sensitive to N_{ks} when they are computed with CDF-based scheme than with the PDF-based scheme. The relative error ϵ_{rel} between the two sampling size is three times lower for the CDF-based estimation than for the PDF-based estimation. Therefore, because the CDF-based estimation scheme of the δ indices requires smaller samples for the same accuracy, it is preferred to the PDF-based estimation scheme in the sequel.

4.4 ANCOVA indices using PC functional decomposition

The so-called ANCOVA indices have been introduced in Li and Rabitz (2010). The principle relies on the covariance decomposition of the model reponse into partial variances and covariances. This covariance decomposition requires a functional decomposition that is a priori unknown for a given model and consequently has to be built. The functional decomposition of a model $\mathcal{M}(\boldsymbol{X})$ reads:

$$\mathcal{M}(\boldsymbol{X}) = \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(X_i) + \sum_{1 \leqslant i,j \leqslant n} \mathcal{M}_{i,j}(X_i, X_j) + \ldots + \mathcal{M}_{1,\ldots,n}(\boldsymbol{X}) \tag{4.42}$$

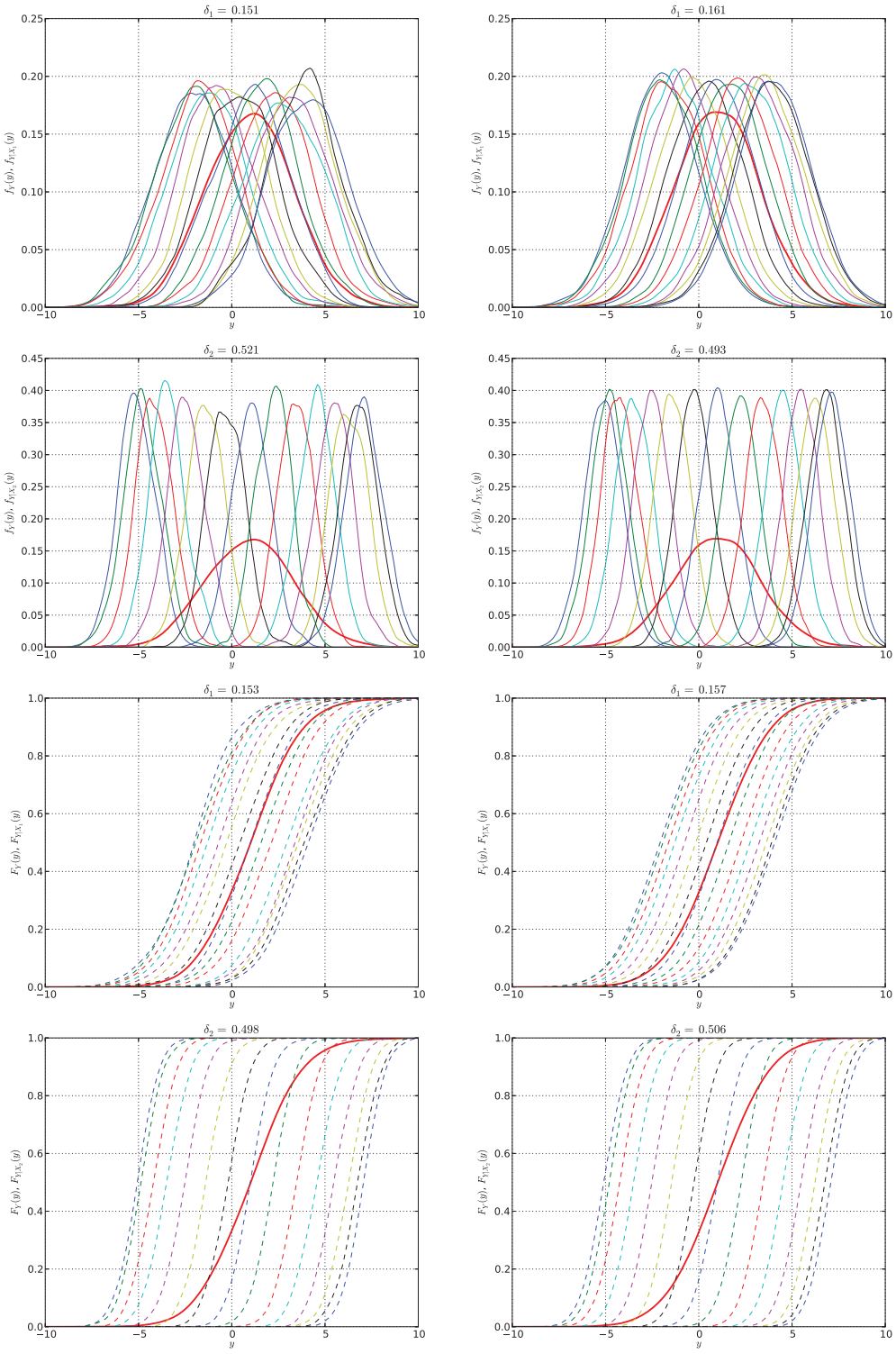


Figure 4.13: *Comparison between PDF-based (top) and CDF-based (bottom) estimators of the δ indices with $N_{ks} = 10^3$ (left) and $N_{ks} = 10^4$ (right).*

Then, the so-called ANCOVA indices described in 2.6 are defined by the following total variance-normed sums:

$$S_i = \frac{\mathrm{Cov}\left[\mathcal{M}_i(x_i), Y\right]}{\mathrm{Var}\left[Y\right]} \tag{4.43}$$

$$S_i^U = \frac{\mathrm{Var}\left[\mathcal{M}_i(x_i)\right]}{\mathrm{Var}\left[Y\right]} \tag{4.44}$$

$$S_i^C = \frac{\mathrm{Cov}\left[\mathcal{M}_i(x_i), Y - \mathcal{M}_i(x_i)\right]}{\mathrm{Var}\left[Y\right]} \tag{4.45}$$

The challenge is consequently to determine properly the subfunctions $\mathcal{M}_i$, $\mathcal{M}_{ij}$ of the decomposition. In this section, two approaches to compute the summands of the functional decomposition are presented.

4.4.1 RS-HDMR decomposition

The Random Sampling High Dimensional Model Representation (RS-HDMR) is the solution that has been proposed in the original paper by Li and Rabitz (2010). The principle is to decompose the model into functions of increasing dimension as in Eq. (4.42). Experience shows (Li et al., 2002) that high-order terms are often negligible and therefore a second-order decomposition can provide a satisfactory description of $\mathcal{M}(\boldsymbol{X})$:

$$\mathcal{M}(\boldsymbol{X}) \approx \mathcal{M}_0 + \sum_{i=1}^{n} \mathcal{M}_i(X_i) + \sum_{1 \leqslant i,j \leqslant n} \mathcal{M}_{i,j}(X_i, X_j) \tag{4.46}$$

The identification of the terms of the decomposition can be made by Monte Carlo integration. For this purpose, $N-$samples of the $n-$dimensional vector $\boldsymbol{X}$ are generated. Then :

$$\mathcal{M}_0 = \int_{\mathrm{D}_{\boldsymbol{X}}} \mathcal{M}(\boldsymbol{x})\mathrm{d}x \approx \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}(\boldsymbol{x}^{(k)}) \tag{4.47}$$

$$\begin{aligned} \mathcal{M}_i(x_i) &= \int_{\mathrm{D}_{\boldsymbol{X}_{\sim i}}} \mathcal{M}(\boldsymbol{x})\mathrm{d}x_{\sim i} - \mathcal{M}_0 \\ &\approx \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}((x_i, \boldsymbol{x}_{\sim i})^{(k)}) - \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}(\boldsymbol{x}^{(k)}) \end{aligned} \tag{4.48}$$

$$\begin{aligned} \mathcal{M}_{i,j}(x_i, x_j) &= \int_{\mathrm{D}_{\boldsymbol{X}_{\sim i,j}}} \mathcal{M}(\boldsymbol{x})\mathrm{d}x_{\sim i,j} - \mathcal{M}_i(x_i) - \mathcal{M}_j(x_j) - \mathcal{M}_0 \\ &\approx \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}((x_i, x_j, \boldsymbol{x}_{\sim i,j})^{(k)}) - \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}((x_i, \boldsymbol{x}_{\sim i})^{(k)}) \\ &\quad - \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}((x_j, \boldsymbol{x}_{\sim j})^{(k)}) - \frac{1}{N}\sum_{k=1}^{N} \mathcal{M}(\boldsymbol{x}^{(k)}) \end{aligned} \tag{4.49}$$

This strategy is numerically very expensive and the authors recommand in the original paper to rather expand the summands on a functional basis that can be orthormal polynomials, splines or simple polynomials, namely:

$$\mathcal{M}_i(x_i) \approx \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) \tag{4.50}$$

$$\mathcal{M}_{i,j}(x_i, x_j) \approx \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^i \varphi_{pq}(x_i, x_j) \tag{4.51}$$

where the α_r^i and β_{pq}^{ij} are constant coefficients to be determined and the φ_r and φ_{pq} are one and two-dimensional basis functions. Thus, the RS-HDMR approximation of $\mathcal{M}(\boldsymbol{X})$ reads:

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}_0 + \sum_{i=1}^{n} \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) + \sum_{1 \leqslant i < j \leqslant n} \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^{ij} \varphi_{pq}(x_i, x_j) \tag{4.52}$$

Provided the basis functions are orthonormal, the coefficients α_r^i and β_{pq}^{ij} can be independently determined by minimization:

$$\min_{\alpha_r^i} \int_{\mathrm{D}_{X_i}} \left[\mathcal{M}_i(x_i) - \sum_{r=1}^{k} \alpha_r^i \varphi_r(x_i) \right] \mathrm{d}x_i \tag{4.53}$$

$$\min_{\beta_{pq}^{ij}} \int_{\mathrm{D}_{X_i}} \int_{\mathrm{D}_{X_j}} \left[\mathcal{M}_{ij}(x_i, x_j) - \sum_{p=1}^{l} \sum_{q=1}^{l'} \beta_{pq}^i \varphi_{pq}(x_i, x_j) \right] \mathrm{d}x_i \mathrm{d}x_j \tag{4.54}$$

Then, the set of coefficients for the decomposition can be obtained by solving a linear equation:

$$\mathbf{A}\boldsymbol{y} = \boldsymbol{b} \tag{4.55}$$

where $\mathbf{A}$ is a constant non singular matrix, $\boldsymbol{b}$ is a vector of integrals over a product of $\mathcal{M}(\boldsymbol{X})$ times the basis functions and $\boldsymbol{y}$ is the vector of the basis functions.

4.4.1.1 Orthonormal polynomial approximation

Polynomials φ_k are referred to as orthonormal on a domain $\mathcal{D}$ if they have zero mean (Eq. (4.56)), unit norm (Eq. (4.57)) and are mutually orthogonal (Eq. (4.58)).

$$\int_{\mathcal{D}} \varphi_k(x)\, \mathrm{d}x = 0, \quad k = 1, \ldots, n \tag{4.56}$$

$$\int_{\mathcal{D}} \varphi_k^2(x)\, \mathrm{d}x = 1, \quad k = 1, \ldots, n \tag{4.57}$$

$$\int_{\mathcal{D}} \varphi_k(x)\, \varphi_l(x)\, \mathrm{d}x = 0, \quad k \neq l \tag{4.58}$$

If $\mathcal{D}$ is $[0,1]$, orthonormal polynomials are derived from the Legendre family:

$$\varphi_1(x) = \sqrt{3}(2x-1) \tag{4.59}$$

$$\varphi_2(x) = 6\sqrt{5}\left(x^2 - x + \frac{1}{6}\right) \tag{4.60}$$

$$\varphi_3(x) = 20\sqrt{7}\left(x^3 - \frac{3}{2}x^2 + \frac{3}{5}x - \frac{1}{20}\right) \tag{4.61}$$

Using the orthonormality properties, the 2-dimensional basis functions in Eq. (4.51) read:

$$\varphi_{pq}(x_i, x_j) = \varphi_p(x_i)\varphi_q(x_j) \tag{4.62}$$

and consequently:

$$\mathcal{M}(\boldsymbol{x}) \approx \mathcal{M}_0 + \sum_{i=1}^{n}\sum_{r=1}^{k}\alpha_r^i\varphi_r(x_i) + \sum_{1\leqslant i<j\leqslant n}\sum_{p=1}^{l}\sum_{q=1}^{l'}\beta_{pq}^{ij}\varphi_p(x_i)\varphi_q(x_j) \tag{4.63}$$

Matrix $\mathbf{A}$ in Eq. (4.55) is equal to identity. Thus, the coefficients α^i and β^{ij} can be evaluated by:

$$\alpha_r^i = \int_{\mathcal{D}} \mathcal{M}(\boldsymbol{x})\,\varphi_r(x_i)\,\mathrm{d}\boldsymbol{x} \approx \frac{1}{N}\sum_{k=1}^{N}\mathcal{M}(\boldsymbol{x}^{(k)})\,\varphi_r(x_i^{(k)}) \tag{4.64}$$

$$\beta_{pq}^{ij} = \int_{\mathcal{D}} \mathcal{M}(\boldsymbol{x})\,\varphi_p(x_i)\,\varphi_q(x_j)\,\mathrm{d}\boldsymbol{x} \approx \frac{1}{N}\sum_{k=1}^{N}\mathcal{M}(\boldsymbol{x}^{(k)})\,\varphi_p(x_i^{(k)})\,\varphi_q(x_j^{(k)}) \tag{4.65}$$

4.4.1.2 Spline function approximation

Spline polynomials can be used instead of orthonormal polynomials for the approximations of the component functions. A cubic $B-$spline $B_k(x)$, $(k=-1,0,\ldots,m+1)$ on an interval $[a,b]$ reads:

$$B_k(x) = \frac{1}{h^3}\times \left\{\begin{array}{ll} (y_{k+2}-x)^3 & y_{k+1} < x \leqslant y_{k+2} \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 & y_k < x \leqslant y_{k+1} \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 + 6(y_k-x)^3 & y_{k-1} < x \leqslant y_k \\ (y_{k+2}-x)^3 - 4(y_{k+1}-x)^3 + 6(y_k-x)^3 - 4(y_{k-1}-x)^3 & y_{k-2} < x \leqslant y_{k-1} \\ 0 & \text{otherwise} \end{array}\right\} \tag{4.66}$$

where:

$$h = \frac{b-a}{m} \tag{4.67}$$

and:

$$y_k = a + kh \tag{4.68}$$

When $a = 0$ and $b = 1$, then $h = 1/m$ and $y_k = kh$. Thus, first and second order basis functions for RS-HDMR representation can be approximated by:

$$\mathcal{M}_i(x_i) \approx \sum_{r=-1}^{m+1} \alpha_r^i B_r(x_i) \tag{4.69}$$

$$\mathcal{M}_{ij}(x_i, x_j) \approx \sum_{p=-1}^{m+1} \sum_{q=-1}^{m+1} \beta_{pq}^{ij} B_p(x_i) B_q(x_j) \tag{4.70}$$

Cubic $B-$splines of different variables are not mutually orthogonal. Consequently, due to the singularity of the matrix $\mathbf{A}$, the coefficients α_r^i and $\beta_{pq}^i j$ cannot be estimated at the same time. The minimization in Eqs. (4.53) and (4.54) are required.

4.4.1.3 Simple polynomial approximation

Simple polynomials can be used as basis functions instead of cubic $B-$splines or orthonormal polynomials:

$$\mathcal{M}_i(x_i) \approx \sum_{r=0}^{k} \alpha_r^i x_i^r \tag{4.71}$$

$$\mathcal{M}_{ij}(x_i, x_j) \approx \sum_{p=0}^{l} \sum_{q=0}^{l'} \beta_{pq}^i x_i^p x_j^q \tag{4.72}$$

Polynomials allow one to describe the model response with simple basis functions but, similarly to spline functions, the main difficulty arises from the singularity of the matrix $\mathbf{A}$ that forces one to minimize the integrals in Eqs. (4.73) and (4.74) in order to evaluate the coefficients.

$$\min_{\alpha_k^i} \int_0^1 \left[\mathcal{M}_i(x_i) - \sum_{r=0}^{k} \alpha_r^i x_i^r \right]^2 \mathrm{d}x_i \tag{4.73}$$

$$\min_{\beta_{ll'}^{ij}} \iint_{[0,1]^2} \left[\mathcal{M}_{ij}(x_i, x_j) - \sum_{p=0}^{l} \sum_{q=0}^{l'} \beta_{pq}^{ij} x_i^p x_j^q \right]^2 \mathrm{d}x_i \mathrm{d}x_j \tag{4.74}$$

4.4.1.4 Conclusion

The formulas to determine the coefficients of the expansion α_r^i and β_{pq}^{ij} are constructed using the orthogonality of the basis functions $\mathcal{M}_i$ and $\mathcal{M}_{ij}$. The evaluation requires large samples for accurate Monte Carlo integration whose error decreases at the rate $1/\sqrt{N}$. Therefore, the performance of a basis for the estimation of RS-HDMR functions strongly depends on the orthogonality of its components and on the size of the samples involved in the approximation of the integrals.

As it appears that orthonormal polynomials are the most efficient basis functions for describing the model response, let us get back to polynomial chaos expansions.

4.4.2 Polynomial chaos decomposition

Polynomial chaos expansions (see Chapter 3, Section 3.3) allow one to represent the response of a model on a suitable polynomial basis. The basis depends on both the marginal distributions of the variables and the maximal degree of the expansion:

$$\hat{Y} = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^n} y_{\boldsymbol{\alpha}} \Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) \tag{4.75}$$

where the $\Psi_{\boldsymbol{\alpha}}$ are multivariate polynomials, namely:

$$\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x}) = \prod_{i=1}^{n} \psi_{\alpha_i}^{i}(x_i) \tag{4.76}$$

4.4.2.1 Identification of the subfunctions

By writing Eq. (4.75) in the form of the functional decomposition appears. Once again, the multi-index notation $\boldsymbol{\alpha} = \{\alpha_i,\, i = 1, \ldots, n\}$ is used.

$$\begin{aligned} \hat{Y} = & \, y_0 \\ & + \sum_{i=1}^{n} \sum_{\alpha_i=1, \alpha_j=0, j \neq i}^{p} y_{\alpha_i} \Psi_{\alpha_i}(\boldsymbol{x}) \\ & + \sum_{1 \leqslant i < j \leqslant n} \sum_{\alpha_i=1}^{p} \sum_{\alpha_j=1}^{p} y_{\boldsymbol{\alpha}_{ij}} \Psi_{\boldsymbol{\alpha}_{ij}}(\boldsymbol{x}) \\ & + \ldots \\ & + y_{1,\ldots,n} \Psi_{i,\ldots,n}(\boldsymbol{x}) \end{aligned} \tag{4.77}$$

The first order terms $\Psi_i(\boldsymbol{x})$ are actually univariate polynomials because when $\boldsymbol{\alpha} = \{\alpha_i\}$, $\psi_{\alpha_j}^{j}(x_j) = 1$, $j \neq i$ and $\Psi_i(\boldsymbol{x})$ only depends on x_i. The second order terms $\Psi_{\boldsymbol{\alpha}_{ij}}(\boldsymbol{x})$ only depend on x_i and x_j because $\psi_{\alpha_k}^{k}(x_k) = 1$, $j \neq i, j$. The same procedure is applied for higher order terms $\Psi_{\boldsymbol{\alpha}}(\boldsymbol{x})$. Note that Eq. (4.77) is identical to Eq. (4.17) with a notation that is closed to the HDMR one. By identification, one obtains:

$$\mathcal{M}_0 = y_0 \tag{4.78}$$

$$\mathcal{M}_i(x_i) = \sum_{\alpha_i=1, \alpha_j=0, j \neq i}^{p} y_{\alpha_i} \psi_{\alpha_i}(x_i) = \sum_{\alpha_i=1, \alpha_j=0, j \neq i}^{p} y_{\alpha_i} \Psi_{\alpha_i}(x_i) \tag{4.79}$$

$$\mathcal{M}_{ij}(x_i, x_j) = \sum_{\alpha_i=1}^{p} \sum_{\alpha_j=1}^{p} y_{\boldsymbol{\alpha}_{ij}} \psi_{\alpha_i}(x_i) \psi_{\alpha_j}(x_j) = \sum_{\alpha_i=1}^{p} \sum_{\alpha_j=1}^{p} y_{\boldsymbol{\alpha}_{ij}} \Psi_{\boldsymbol{\alpha}_{ij}}(x_i, x_j) \tag{4.80}$$

$$\ldots \tag{4.81}$$

where ψ_{α_i} is the basis polynomial of degree α_i corresponding to the marginal distribution of the input variable X_i.

4.4.2.2 Computation of the sensitivity indices

Once the subfunctions have been identified, the partial variances and covariances shall be computed. However, when the input variables of the model are *dependent*, they are decorrelated before the calculation of the coefficients so that the basis stays orthonormal. In the case of a Gaussian copula, the Nataf transformation is used, in other cases, the Rosenblatt transformation, which is non unique, is used instead. Performing an ANCOVA sensitivity analysis on such a basis would on the one hand lead to zero correlative contributions (the covariance between the decorrelated variables are zero) and on the other hand provide sensitivity indices with respect to the decorrelated variables that might broadly differ from the physical variables.

The goal here is to get the terms of the functional decomposition. Whereas the RS-HDMR makes no hypothesis on the probabilistic model to compute subfunctions, the PCE takes both the physical and probabilistic models into account. In the case of correlated input variables, the orthogonality property of the basis functions of the physical variables is lost due to the isoprobabilistic transformation.

To circumvent the issue of decorrelation, it is proposed to build the metamodel with a joint distribution featuring an independent copula to preserve the orthogonality of the basis. The PC expansion thus behaves as a natural response surface that provides a response $y^{(i)}$ for an input vector $\boldsymbol{x}^{(i)}$.

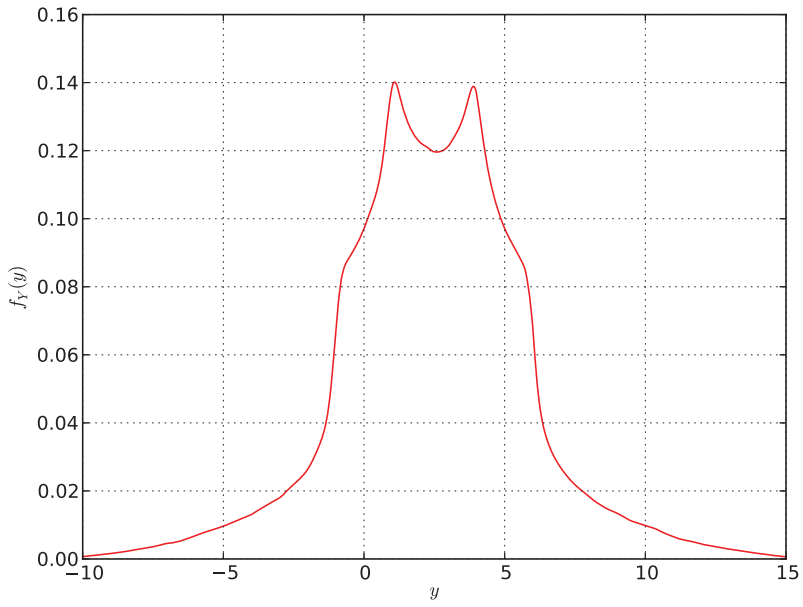


Figure 4.14: *Probability density function of the response of the Ishigami function.*

As an illustration, let us consider the so-called Ishigami function (Ishigami and

Homma, 1990), namely:

$$Y = \sin X_1 + 7 \times \sin^2 X_2 + 0.1 \times X_3^4 \sin X_1 \tag{4.82}$$

with $X_i \sim \mathcal{U}[-\pi, \pi]$, $i = 1, \ldots, 3$. Although in the original work the variables are independent, correlation is added here in order to show that the a PCE built with an independent copula also holds for the same marginal distributions featured with any copula. Due to the multimodal shape of the response PDF (Figure 4.14), the Ishigami requires a high order of expansion (up to $p = 10$). The accuracy of the response approximation is studied for both an independent sample and a correlated one with the rank correlation matrix in Eq. (4.83). Scatterplots of these samples are pictured in Figure 4.15.

$$\mathbf{S} = \begin{bmatrix} 1 & -0.4 & 0.9 \\ -0.4 & 1 & 0 \\ 0.9 & 0 & 1 \end{bmatrix} \tag{4.83}$$

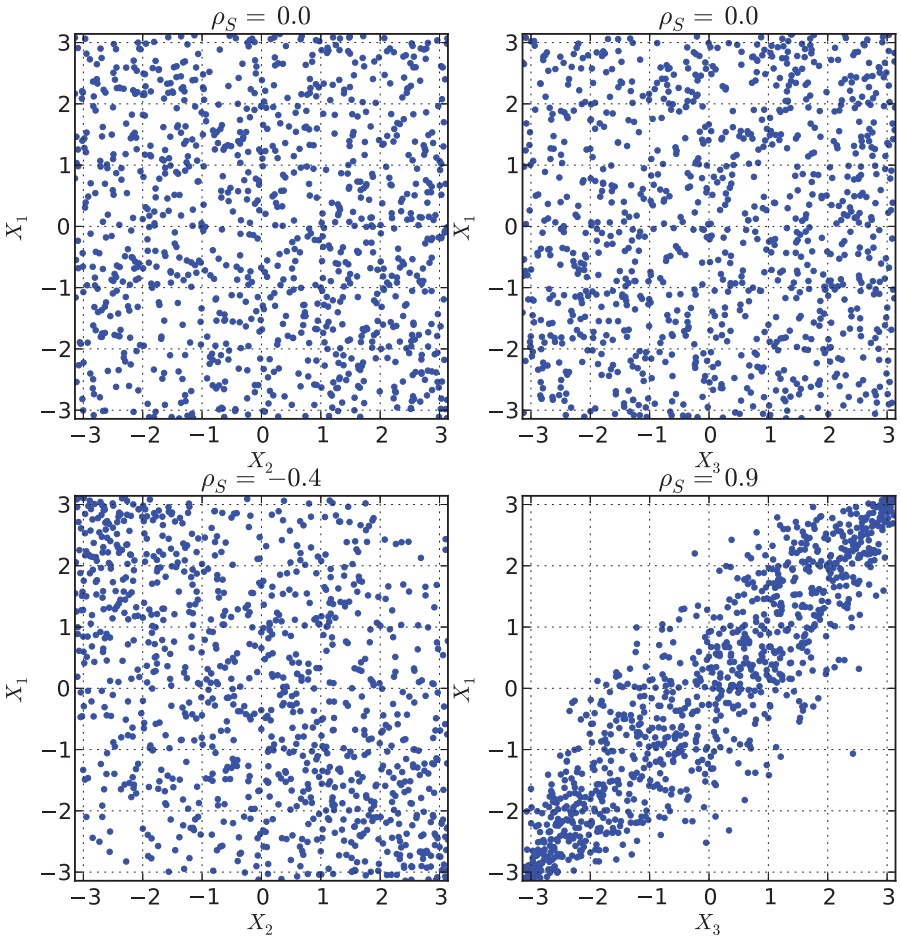


Figure 4.15: *Scatterplots in the independent case (top) and correlated case (bottom) for the PC accuracy comparison.*

For an increasing expansion order p, the *empirical error* (or *training error*) Err_{emp} Eq. (4.84) and the *relative training error* ϵ_{emp} Eq. (4.85) are studied in Figure 4.16.

$$Err_{emp} = \frac{1}{N}\sum_{k=1}^{N}\left(\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \hat{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right)\right)^2 \tag{4.84}$$

$$\begin{aligned} \epsilon_{emp} &= \frac{\sum_{k=1}^{N}\left(\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \hat{\mathcal{M}}\left(\boldsymbol{x}^{(i)}\right)\right)^2}{\hat{\sigma}_{\mathcal{Y}}^2} \\ \hat{\sigma}_{\mathcal{Y}}^2 &= \sum_{k=1}^{N}\left(\mathcal{M}\left(\boldsymbol{x}^{(i)}\right) - \bar{y}\right), \quad \bar{y} = \frac{1}{N}\sum_{k=1}^{N} y^{(i)} \end{aligned} \tag{4.85}$$

Coefficients of determination $R^2 = 1 - \epsilon_{emp}$ are also presented for the *training sample* or design of experiments, *i.e.* the points $\boldsymbol{x}^{(i)} \in \mathcal{X}$ the regression is based on, a independent sample and a correlated sample, all of size $N = 1000$. The coefficients of determination are respectively denoted by $R^2_{\mathcal{Y}_{DOE}}$, $R^2_{\mathcal{Y}_{ind}}$, $R^2_{\mathcal{Y}_{corr}}$. Results show that the accuracy of the metamodel is the same whatever the dependence structure of the sample that is studied. This observation allows one to use the PCE as a classic response surface for the functional decomposition of any model for the computation of the sensitivity indices.

4.4.2.3 First order and total indices

In the original paper by Li and Rabitz (2010), the so-called ANCOVA indices are defined in such a way that:

- the index S_i represents the *total* contribution of the input variable X_i,
- the index S_i^U represents the *uncorrelated* (or *structural*) contribution of the input variable X_i,
- the index S_i^C represents the *correlated* contribution of the input variable X_i,

Therefore, the uncorrelated part of the contribution of X_i is carried by the term of the decomposition $\mathcal{M}_i(X_i)$ that only depends on X_i. The correlated part of the contribution is described by the covariance of $\mathcal{M}_i(X_i)$ and the terms that do not depend on X_i, namely $\mathcal{M}_{\boldsymbol{u}}(X_{\boldsymbol{u}})$, $i \notin \boldsymbol{u}$. Then the total contribution is the sum of the correlative contribution and the uncorrelative contribution $S_i = S_i^U + S_i^C$. This definition is consistent with the Sobol' first order indices in the case of independent variables.

Let us now try to extend this definition to describe the total contribution, in the sense of Sobol' total indices, of a variable. The terms of the functional decomposition depending on X_i are:

$$\mathcal{M}_{i\in\boldsymbol{u}}(\boldsymbol{x}) = \mathcal{M}_i(x_i) + \sum_{j=1, j\neq i}^{n-1} \mathcal{M}_{ij}(x_i, x_j) + \ldots + \mathcal{M}_{1\ldots n}(\boldsymbol{x}) \tag{4.86}$$

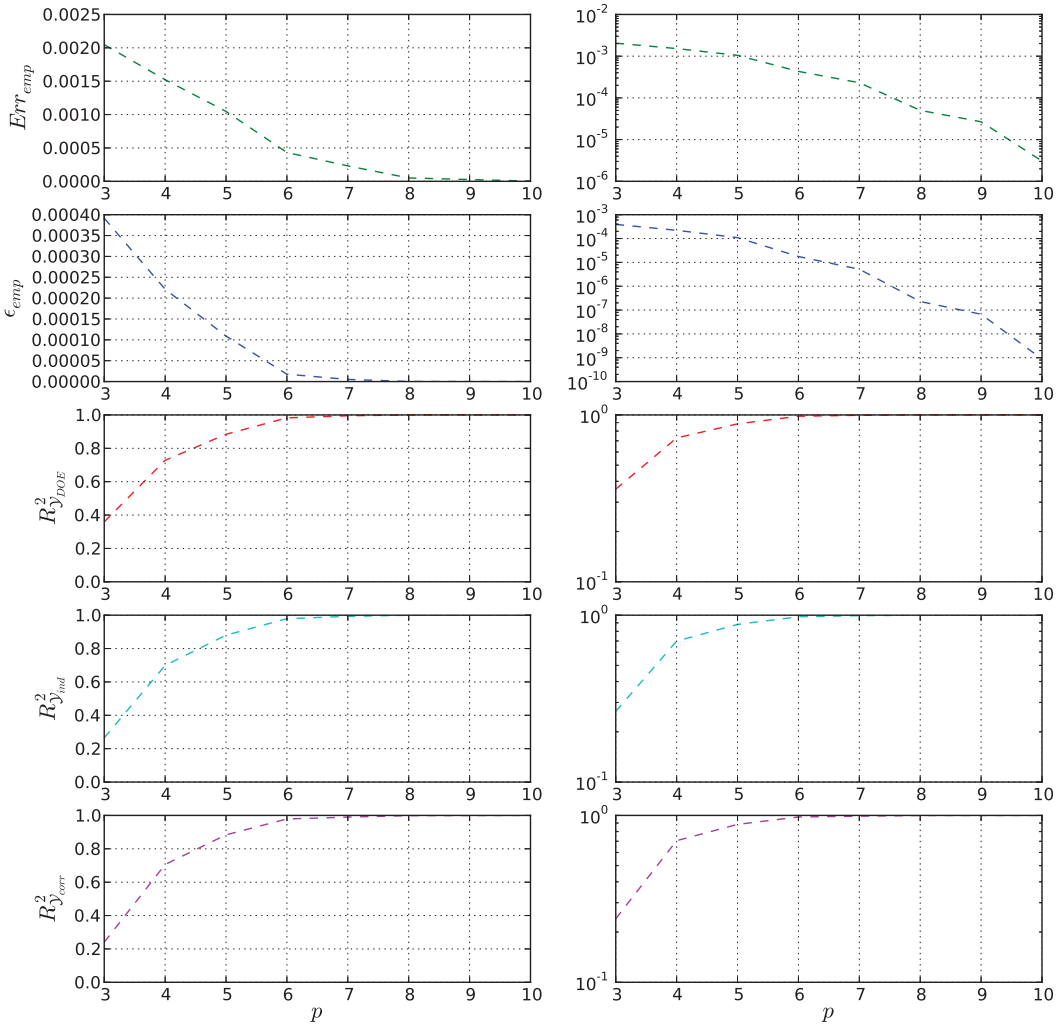


Figure 4.16: *Accuracy of the PC expansion of the Ishigami function for independent and correlated samples.*

One may be tempted to define the total uncorrelative contribution of X_i by:

$$S_i^{U,T} = \frac{\mathrm{Var}\left[\mathcal{M}_{i\in\boldsymbol{u}}(\boldsymbol{x})\right]}{\mathrm{Var}\left[Y\right]} \tag{4.87}$$

and the total correlative contribution by:

$$S_i^{C,T} = \frac{\mathrm{Cov}\left[\mathcal{M}_{i\in\boldsymbol{u}}(\boldsymbol{x}), \mathcal{M}_{i\notin\boldsymbol{v}}(\boldsymbol{x})\right]}{\mathrm{Var}\left[Y\right]} \tag{4.88}$$

On the one hand, in Eq. (4.87), if X_i and X_j are correlated, the term $\mathcal{M}_{ij}(X_i, X_j)$ is stored in the uncorrelated part of the contribution of X_i although the variables are correlated. On the other hand, in Eq. (4.88), the covariance may detect the correlation between $X_{k\in\boldsymbol{u},k\neq i}$ and $X_{l\in\boldsymbol{v}}$ although X_i is not correlated neither with X_k nor X_l. There is a confusion betwwen the interaction and correlation effects.

To circumvent this issue, the interaction and correlation effects may be seperated in the following way:

$$S_i = S_i^U + S_i^I + S_i^C \tag{4.89}$$

where the indices S_i^U, S_i^I and S_i^C respectively represent the uncorrelative contribution, the interactive contribution and the correlative contribution. These three indices read:

$$S_i^U = \frac{\mathrm{Var}\left[\mathcal{M}_i(X_i)\right]}{\mathrm{Var}\left[Y\right]} \tag{4.90}$$

$$S_i^I = \frac{\mathrm{Cov}\left[\mathcal{M}_i(X_i), \mathcal{M}_{i\in\boldsymbol{u}}(\boldsymbol{X})\right]}{\mathrm{Var}\left[Y\right]} \tag{4.91}$$

$$S_i^C = \frac{\mathrm{Cov}\left[\mathcal{M}_i(X_i), \mathcal{M}_{i\notin\boldsymbol{v}}(\boldsymbol{X})\right]}{\mathrm{Var}\left[Y\right]} \tag{4.92}$$

The interactive contribution is described by the covariance of $\mathcal{M}_i(X_i)$ and the component functions $\mathcal{M}_{i\in\boldsymbol{u}}(\boldsymbol{X})$ that also depend on X_i except $\mathcal{M}_i(X_i)$. The correlative contribution is described by the covariance of $\mathcal{M}_i(X_i)$ and the component functions $\mathcal{M}_{i\notin\boldsymbol{v}}(\boldsymbol{X})$ that do not depend on X_i. Thus, S_i^C only represent the correlative part of the contribution of X_i. However, the interactive part may also carry part of the correlation since $X_{i\in\boldsymbol{u}}$ and $X_{j\in\boldsymbol{u}}$ may be correlated.

Defining ANCOVA total indices is not a trivial matter since interactive and correlative contributions may be confused. Therefore, the question of the classification of the interaction terms in one side or the other or both remains open for total indices.

4.4.2.4 Diagram of the procedure

To summarize the PC-based estimation scheme of the ANCOVA indices, the procedure is illustrated by the algorithm in Figure 4.17.

4.5 Validation

In this section, the computation scheme that has been proposed here is confronted to the results from the original papers.

4.5.1 Distribution-based importance measure

The importance measure δ proposed in Borgonovo (2007) has been improved in Borgonovo et al. (2011). In order to compare the rank and scale of the indices to the one in the ANOVA, academic examples are presented, namely an additive and a non additive and non multiplicative models. The results are compared with the computation procedure proposed in this chapter.

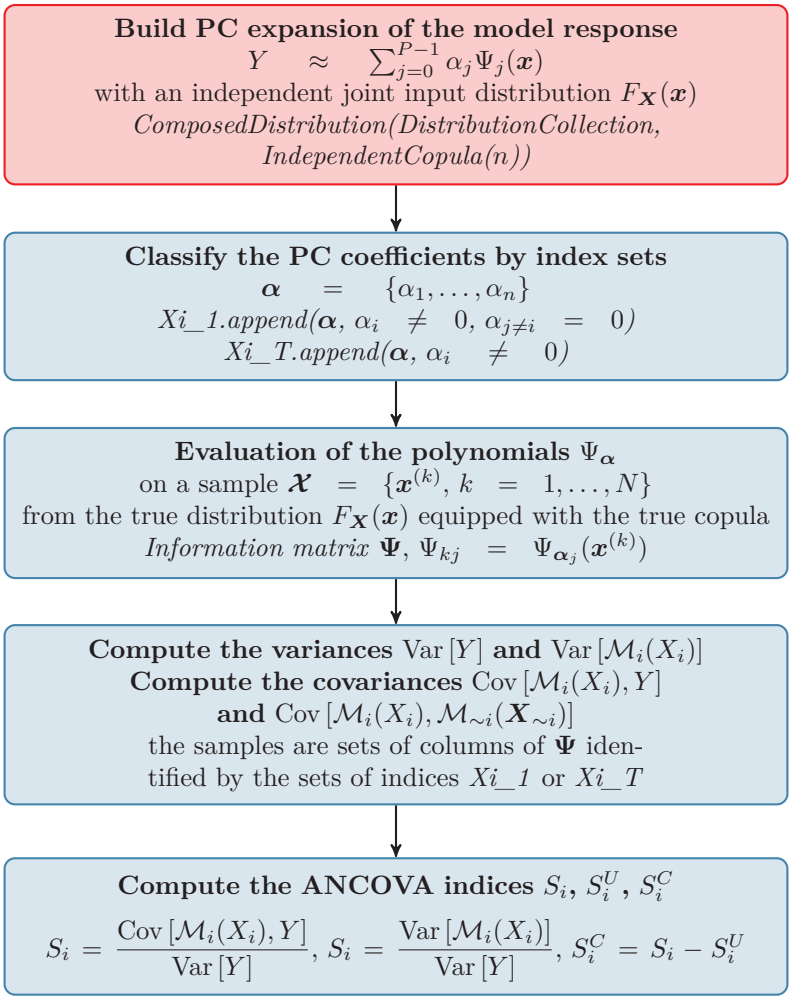


Figure 4.17: *Diagram of the PC-based estimation of the ANCOVA sensitivity measure.*

4.5.1.1 An additive model

The first model consists in a sum of standard Gaussian variables, namely:

$$y = \sum_{i=1}^{n} a_i \, x_i \tag{4.93}$$

with $X_i \sim \mathcal{N}(0,1)$, $i = 1, \ldots n$. The shift can be analytically evaluated and reads:

$$\begin{aligned} s(x_i) = 2\Big[&\Phi(y_1; m_Y, \sigma_Y^2) + \Phi(y_2; m_{Y|X_i}, \sigma_{Y|X_i}^2) \\ &- \Phi(y_2; m_Y, \sigma_Y^2) + \Phi(y_1; m_{Y|X_i}, \sigma_{Y|X_i}^2)\Big] \end{aligned} \tag{4.94}$$

where m_Y and σ_Y^2 are the mean and variance of the model response and (y_1, y_2) are the intersection points of the unconditional and conditional PDFs. The different values of the shift $s(x_i)$ for all the values of x_i are illustrated in Figure 4.18. When X_i takes extreme

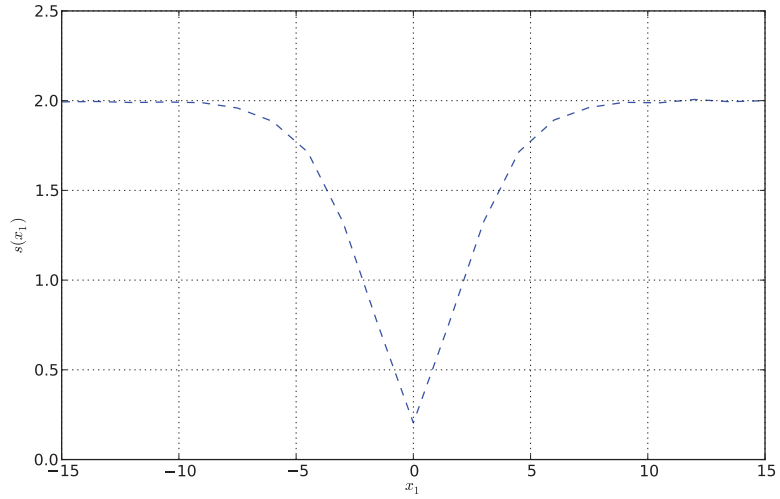


Figure 4.18: *Values of the shift for the additive model ($n = 2$) depending on the value of X_i.*

values, *i.e.* $|x_i| > 10$, the conditional and unconditional PDFs do not cover each other, *i.e.* their respective modes are far away so that they only intersect in their very far tails. Thus the shift is the sum of the area under both distributions is equal to 2.

$$\lim_{x \to \pm\infty} s(x_i) = 2 \tag{4.95}$$

Two cases are studied, $n = 2$ and $n = 3$. For both cases, $a_i = 1$, $i = 1, \ldots, n$ and the input variables are independent. The total response variance σ_Y^2 is equally divided between each input variable. Therefore, the Sobol' first order indices are equal for each variable and are worth $1/n$: $S_i = 0.5$ $(n = 2)$ and $S_i = 0.33$ $(n = 3)$. The corresponding δ measure are also logically equal for each variable because each variable is identically distributed and bring as much variability to the response variance than the other. According to Borgonovo et al. (2011), the analytical values are $\delta_i = 0.306$ for $n = 2$ and $\delta_i = 0.224$ for $n = 3$. The distributions of the δ importance measure for $n = 2, 3$, based on 100 index estimations for each sample size, are illustrated in Figure 4.19 and Figure 4.20. It shows the convergence of the accuracy of the indices estimation when the size of the sample the kernel smoothing is based on increases. The number of quadrature points used for the expected value of the shift is $n_q = 15$.

The figure shows that for small samples, the value of δ is underestimated but the gap between the estimated value and the true value progressively decreases. The final values are $\delta_i = 0.303$ and $\delta_i = 0.222$ for $n = 2$ and $n = 3$ respectively. However, the exact value is not reached even for very large samples ($N > 10^6$) but the precision is approximately 10^{-3}. This is probably due to the residual error in the numerical procedure developed in Section 4.3.

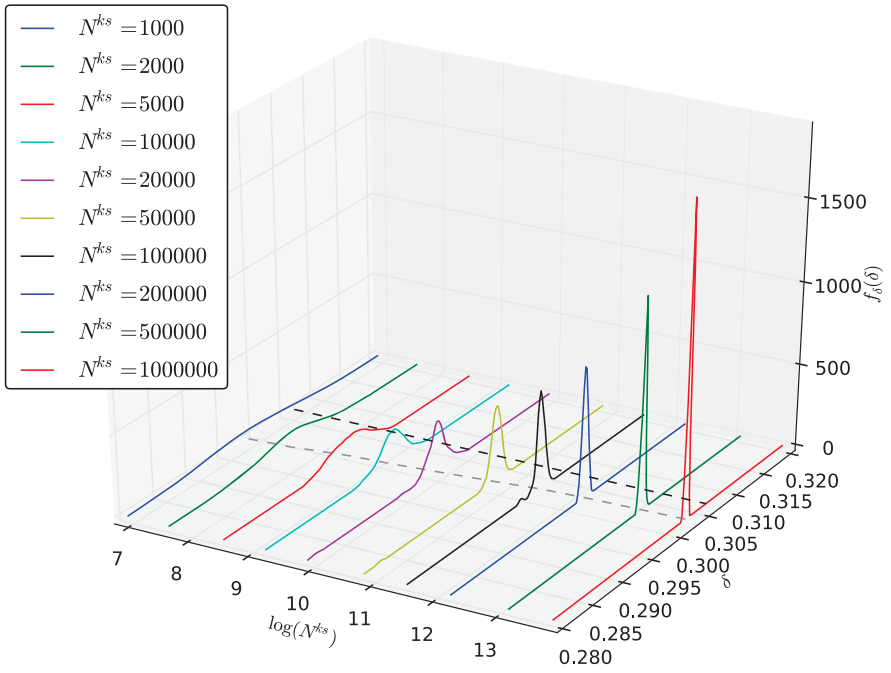


Figure 4.19: *Distributions of the δ importance measure for the additive model with $n = 2$. The estimated mean value (dashed grey line) converges to the analytical value (dashed black line)*

4.5.1.2 A non additive and non multiplicative model

Let us now consider a non additive and non multiplicative model with Gamma distributed input variables $X_1 \sim Gamma(\alpha, \theta)$ and $X_2 \sim Gamma(\beta, \theta)$. The model reads:

$$y = \frac{x_1}{x_1 + x_2} \tag{4.96}$$

The domain of definition of Y is $[0, 1]$ and the output distribution is symmetrical when $\alpha = \beta$, $\theta = 1$. The analytical value of δ_i reads:

$$\begin{aligned} \delta_i = \int_0^1 \Big[& F_Y(y_1) - F_{Y|X_i}(y_2) \\ & + F_{Y|X_i}(y_1) - F_Y(y_2) \Big] \, \mathrm{d}x_i \end{aligned} \tag{4.97}$$

where (y_1, y_2) are the two points at which the conditional PDF $f_{Y|X_i}$ intersect the unconditional PDF f_Y. It is shown in the original paper that when $\alpha = \beta$, $\delta_1 = \delta_2$. The analytical and estimated values of the first order indices when $\alpha = \beta = 1$, 2, 3 and 10 and the relative estimation error such as defined in Eq. (4.41) are presented in Table 4.2. The conditional PDFs are illustrated in Figure 4.21. Results show that for a reasonable sampling effort, *i.e.* $N_{ks} = 10^3$ and $N_q = 30$. the estimated values of the indices are very close to the analytical values with a relative error close to 1%.

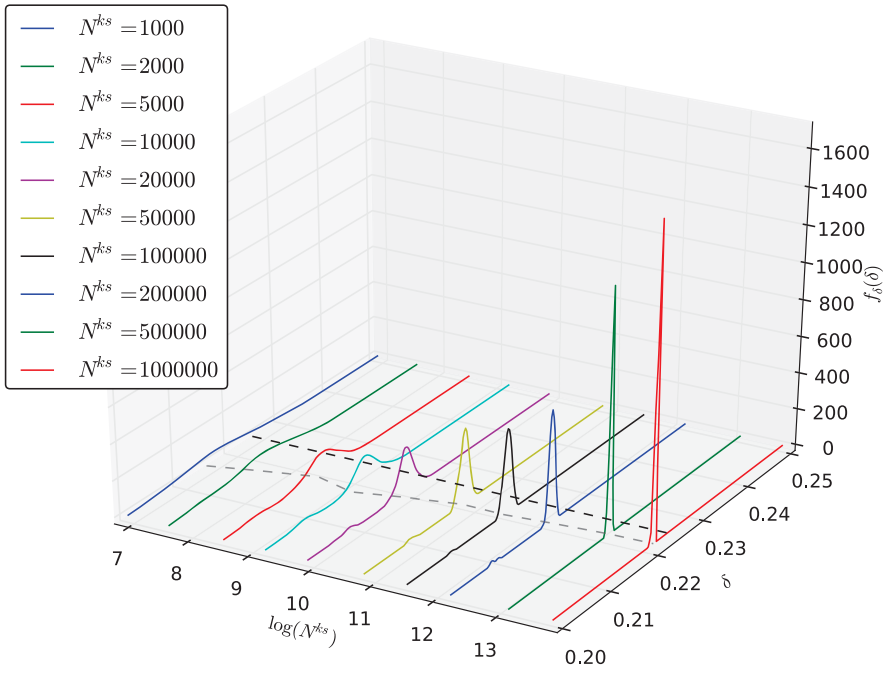


Figure 4.20: *Distributions of the δ importance measures for the additive model with $n = 3$. The estimated mean value (dashed grey line) converges to the analytical value (dashed black line)*

4.5.1.3 Conclusion

Two analytical cases, an additive model and a non additive and non multiplicative model, have been studied. The result show that the estimation scheme that have been proposed in Section 4.3 tends to the analytical value of the indice when the sampling effort, *i.e.* the size N of the sample for the kernel smoothing estimation of the distributions and the number of quadrature points N_q, is increased. However, an estimation error close to 1% is observed.

4.5.2 Ancova Indices

The so-called ANCOVA indices have been proposed in Li and Rabitz (2010). The numerical estimation procedure proposed in Section 4.4 is applied to numerical examples that are now exposed.

4.5.2.1 Equal contribution of independent parameters

The model is the sum of five normally distributed input variables $X_i \sim \mathcal{N}(0.5, 1)$ that are first assumed independent:

$$y = \mathcal{M}(\boldsymbol{x}) = x_1 + x_2 + x_3 + x_4 + x_5 \tag{4.98}$$

α	$\delta_{analytic}$	δ_{CDF}	ϵ_{rel}
1	0.330	0.323	0.021
2	0.319	0.319	0.000
3	0.315	0.317	0.006
10	0.309	0.309	0.000

Table 4.2: *Analytical and estimated value of the δ indices for the non additive and non multiplicative model.*

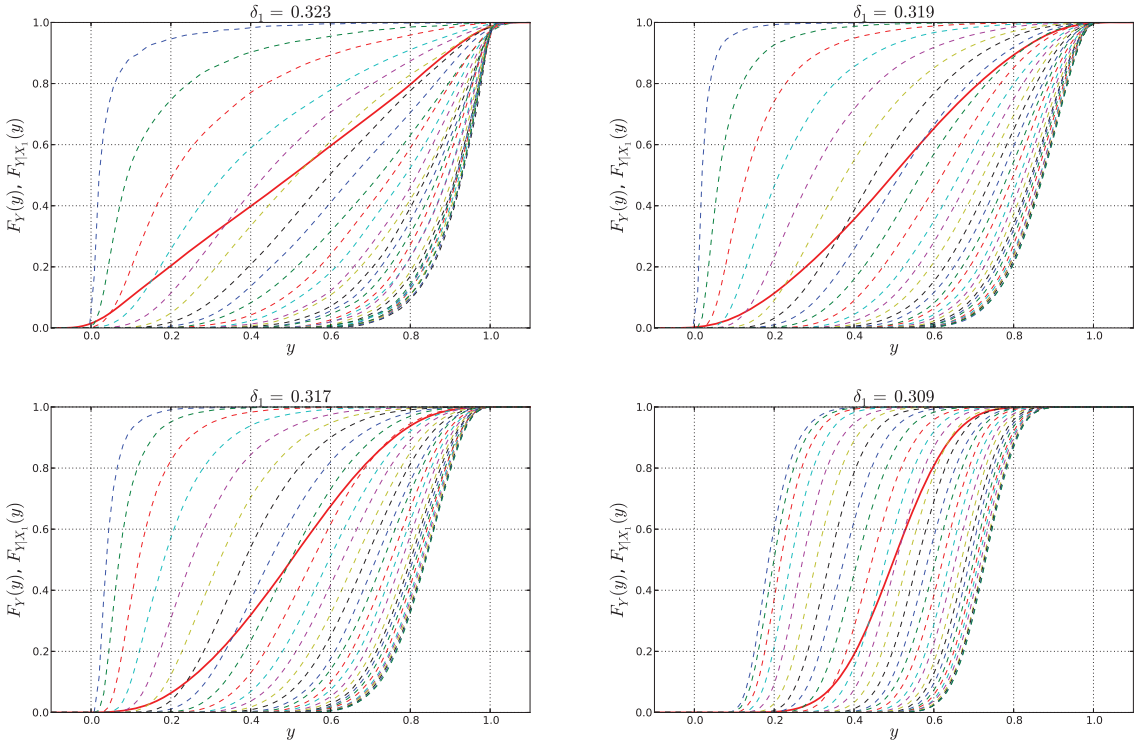


Figure 4.21: *Unconditional and conditionals CDFs for the non additive and non multiplicative model with (from left to right and from top to bottom) $\alpha = 1, 2, 3,$ and 10.*

Due to the polynomial nature of the model, the functional decomposition reads $\mathcal{M}_i(x_i) \equiv x_i$, $i = 1, \ldots, 5$. The sampling size for the estimation of the variances and covariances is $N = 10^4$. The results of the sensitivity analysis are presented in Table 4.3. The total contribution of each parameter is $S_i = 0.20$, that is $1/5$ and this value is consistent with the first order Sobol' indices. The correlative contribution is logically zero due to the independence of the parameters. Therefore, the uncorrelated contribution equals the total contribution.

X_i	S_i	S_i^U	S_i^C
X_1	0.20	0.20	0.00
X_2	0.20	0.20	0.00
X_3	0.20	0.20	0.00
X_4	0.20	0.20	0.00
X_5	0.20	0.20	0.00
$\sum$	1.00	1.00	0.00

Table 4.3: *ANCOVA sensitivity indices with equal structural contribution and independent input parameters.*

4.5.2.2 Equal structural contribution of correlated parameters

The second test case keeps the same physical model $\mathcal{M}$ and marginal distributions $\mathcal{N}(0.5, 1)$ as the first one but now features a Gaussian copula with linear correlation matrix:

$$\mathbf{\Sigma} = \begin{bmatrix} 1 & 0.6 & 0.2 & 0 & 0 \\ 0.6 & 1 & 0 & 0 & 0 \\ 0.2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0.2 & 1 \end{bmatrix} \tag{4.99}$$

The results are given in Table 4.4. The uncorrelative contribution is the same for all the variables $S_i^S = 0.13$ but the correlative contribution depends on the intensity of the correlation. For example, X_1, which is correlated with both X_2 and X_3 is the most correlated input variable, therefore, its correlative contribution is the highest. X_2 also gets a high correlative contribution due the importance of the correlation with X_1, namely $\rho_{1,2} = 0.6$. On the other side, X_4 and X_5 are mutually correlated but because of their low correlation coefficient $\rho_{4,5} = 0.2$, the correlative contribution remains lower than for the other parameters.

X_i	S_i	S_i^U	S_i^C
X_1	0.24	0.13	0.11
X_2	0.24	0.13	0.11
X_3	0.19	0.13	0.06
X_4	0.16	0.13	0.03
X_5	0.16	0.13	0.03
$\sum$	1.00	0.65	0.35

Table 4.4: *ANCOVA sensitivity indices with equal structural contribution and correlated input parameters.*

4.5.2.3 Distinct structural contribution of correlated parameters

Let us now add to the previous example distinct structural contributions. The model now reads:

$$y = \mathcal{M}(\boldsymbol{x}) = 5x_1 + 4x_2 + 3x_3 + 2x_4 + x_5 \tag{4.100}$$

The results of the sensitivity analysis are given in Table 4.5. The ranking and the importance of the contribution is influenced by both the model structure and the correlation between input parameters. The highest total contribution is associated with X_1 which has both the highest uncorrelative and correlative contributions. At the bottom of the ranking, X_4 and X_5 which had equal total contributions in the previous test case are now tied by distinct uncorrelative contributions.

X_i	S_i	S_i^U	S_i^C
X_1	0.44	0.28	0.16
X_2	0.33	0.17	0.16
X_3	0.17	0.11	0.06
X_4	0.04	0.04	0.00
X_5	0.02	0.02	0.00
$\sum$	1.00	0.62	0.38

Table 4.5: *ANCOVA sensitivity indices with distinct structural contribution and correlated input parameters.*

For comparison, the corresponding Sobol' indices in the case of distinct correlative contributions but independent input variables are given in Table 4.6. The first order and total sensitivity indices are equal. Because the model is additive, there is no interaction between the variables. The ranking of the parameters follows the importance of the correlative contributions but the importance of the indices differs from the total contribution of the ANCOVA indices. The ANCOVA indices S_1 and S_2 (0.44 and 0.33) are slightly smoothed compared to the corresponding Sobol' indices (0.46 and 0.29) because of the mutual correlation between the two variables. The same observation can be done for X_4 and X_5 (0.04 and 0.02 versus 0.07 and 0.02).

4.6 Conclusion

This chapter proposes metamodel-based evaluation procedures for the estimation of sensitivity indices for models with correlated input parameters. A particular attention is given to polynomial chaos expansions as a mean to reduce the numerical cost. This metamodelling technique allows one both to substitute efficiently the real model with a polynomial representation and to provide a functional decomposition for the estimation of the ANCOVA indices.

X_i	S_1	S_T
X_1	0.46	0.46
X_2	0.29	0.29
X_3	0.16	0.16
X_4	0.07	0.07
X_5	0.02	0.02
$\sum$	1.00	1.00

Table 4.6: *Sobol' first order and total sensitivity indices with distinct structural contribution and independent input parameters.*

The estimation of the δ importance measure requires large samples for the kernel smoothing estimation of the distributions. An improved technique based on the CDF instead of the PDF has shown better performance with respect to the sampling size. The estimation of the expected shift can also be enhanced by using a quadrature scheme instead of Monte Carlo simulations. The δ importance measure offers a different approach than the ANOVA. The rankings might differ because δ not only focuses on the reponse variance but on its whole distribution.

The ANCOVA decomposition allows one to distinguish the uncorrelative (or structural) and correlative contribution of the input parameters. One important feature is that ANCOVA indices are consistent with ANOVA indices in the case of independence and represent a generalization of the ANOVA. A functional decomposition is required for the estimation. Such a decomposition can be directly derived from the PC expansion avoiding the numerically expensive identification of each terms of the decomposition.

Once sensitivity indices for models with correlated input parameters have been studied, they will be applied to industrial cases. Before that, the particular case of nested models and the associated formalism is introduced in the next chapter.

5

Nested and multiscale modelling

Contents

5.1 Introduction

Model-Based System Engineering (or MBSE for short) is defined by the INCOSE (INternational Council On System Engineering) as "*the formalized application of modeling to support system requirements, design, analysis, verification and validation activities beginning in the conceptual design phase and continuing throughout development and later life cycle phases*".

The SysML (for Systems Modelling Language) is an extension of the more general UML (Unified Modelling Language) for model-based engineering. SysML is richer and more flexible than UML. Contrary to UML software-focused specifications, SysML allows one to model a wider range of systems such as software, hardware, personnel, information, processes or facilities.

The issues behind nested and multiscale modelling lie in the communication between different programs or softwares sharing variables and parameters. The tools that have to be used differ from a case to another because of the multitude of applications (mechanics, accoustics, dynamics, etc.). The structure, which is related to the complexity of the modelling, also requires flexibility. How many scale of description for the modelling? Are there loops?

MBSE and SysML provide a general theory on what can or has to be done and a tool for making programs talk to each other but there is a lack of mathematical and physical framework to treat such problems. In the scope of uncertainty propagation for mechanical engineering applications for instance, uncertain parameters must be considered and classified in such a way that the results of the computational workflow are fully workable.

This chapter first introduces the nested and multiscale modelling aspects. Then, a framework for nested models, based on the graph theory is introduced. Finally, a methodology to assess global sensitivity analysis problems for nested and multiscale problems is proposed.

5.2 Nested and multiscale modelling

5.2.1 System robust engineering

Uncertainty propagation for reliability analysis is presented in a comprehensive way in Lemaire (2009). The approach consists in modelling the input parameters $\boldsymbol{X}$ of a model $\mathcal{M}$ by random variables and in studying the dispersion of the model output Y. The robustness of the system described by the model can be expressed in terms of a probability of failure, *i.e.* the probability that Y exceeds a threshold value that is deemed acceptable. If the methodology to address this kind of problems is nowadays well-established for a single model $\mathcal{M}$, it has not been much extended to system approaches.

5.2.2 System fault trees

A first step has been reached by developping *fault tree analysis* (or FTA). This top down deductive safety analysis combines series of lower-level events to characterize the upper-level state of a system using Boolean logic. In Hähnel (2007), the author proposes a general framework for system reliability engineering. The workflow is modelled by a chain of models with independent inputs. Therefore, the chain can be represented by a tree, a special case of graph (see subsection 5.3.2). This type of representation has been proposed in Sudret et al. (2009).

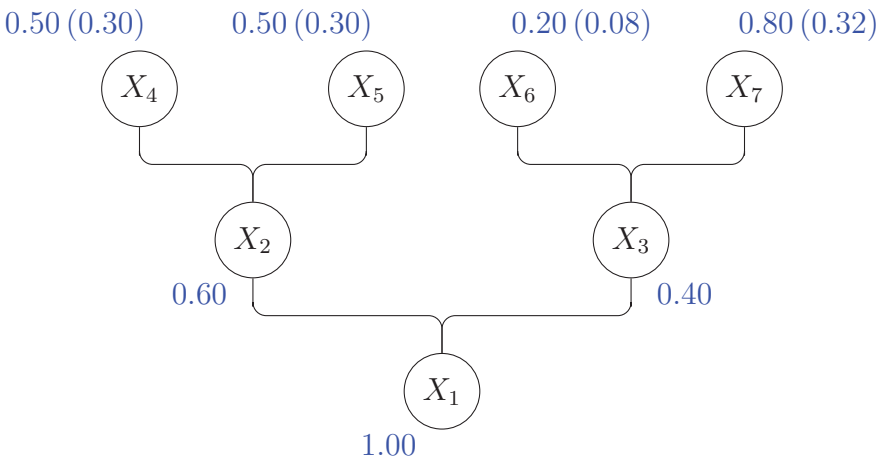


Figure 5.1: *A tree-graph composed of 7 vertices and 6 edges. Vertex X_1 is the* root*, vertices X_4, X_5, X_6 and X_7 are the* leaves *of the tree. The contributions of the leaves to the intermediate nodes (to the root) are mentionned in blue.*

An example of tree is pictured in Figure 5.1. In the system-based approach, the variable X_2 and X_3 are functions of (X_4, X_5) and (X_6, X_7) respectively, and the output of interest X_1 is a function of (X_2, X_3). The uncertainties are propagated from the lowest levels (the *leaves*) to the highest one (the *root*). From a sensitivity analysis point of view, the total variance at one node can only be expressed by the shares of the variables at the previous level and so on. Then, the share of variance of the root variable due to a leaf variable is the product of the shares at each node on the path connecting the leaf to the root. For example, the sensitivity of the variable X_3 to the variable X_7 is 0.80 whereas the sensitivity of X_1 to X_7 is $0.80 \times 0.40 = 0.32$.

This approach is consistent as long as the variables of each level are independent, that is to say that two nodes from the a level k do not share any entry parameter from the level $k-1$. That would not be the case if X_5 and X_6 were one single parameter in the modelling, *i.e.* a common input for computing X_2 and X_3. If it is tempting to omit the correlation between the parameters, it might also lead to substantial errors in the results. Therefore a model representation accompanied by global sensitivity analysis techniques for models with correlated parameters is proposed in the next sections.

5.3 Model representation

In contrast to the algorithm representations, no convention exists in the mechanical engineering field for describing a complex structure. Most of the classical representations are usually composed of a collection of shapes, colors and arrows of different types for input or output parameters, the model itself or the links between the elements.

5.3.1 State-of-the-art

A survey of the litterature has shown that there is no well-established convention for describing nested models as those appearing in mechanical engineering. In the sequel, a general framework based on the graph theory is proposed for this purpose.

5.3.2 The graph theory

The *graph theory* is a scientific discipline that lies at the boundaries between mathematics and computer science. Graphs are mathematical structures intended to model relationships between objects by introducing a collection of *vertices* and a collection of *edges* representing respectively the objects and their relationships. The origin of graph theory can be found in the paper of Leonard Euler addressing *the seven bridges of Königsberg problem* (Euler, 1741). Euler's theorem was actually proven later by Hierholzer (1873). A similar problem referred to as the *Knight's tour*, introduced by Arabic chess theorician Al-Adli in *Kitab ash-shatranj* around 840 (and lost since) has been studied later in Vandermonde (1771). A knight, moving according to the rules of chess, must visit only once every square of a chessboard. Another famous theorem that has been demonstrated using the graph theory is the *four color problem* introduced in 1852 by the English cartographer Francis Guthrie. He noticed that coloring a complex map so that no contiguous have the same color only require four colors and initiated to verify the validity of this property for any map. After several attempts (Cayley, 1879; Kempe, 1879; Petersen, 1891), the theorem was finally demonstrated using a computer in Gonthier (2000). Today, graphs are broadly used in science to model networks, processes in biology, chemistry or physics or social networks. For a complete review of the graph theory, the reader is referred to Bergé (1958). In this section, it is shown how graphs can be employed for the nested modelling of complex systems.

5.3.2.1 Definitions

A graph is an abstract representation of a collection of objects among which pairs are connected by links. The objects are called *vertices* (or *nodes*, or *points*) and the links are called *edges* (or *lines*). A general mathematical designation of a graph is:

$$G = (V, E) \tag{5.1}$$

where V is a collection of vertices and E a collection of edges. Edges of a graph can be either unoriented or oriented. For example, in a graph representing a group of person, if a person A knows a person B, but B does not know A, the link between A and B is oriented from A to B. On the contrary, if A and B both know each other, then the link between A and B is unoriented. In the first case it means that $(A, B) \neq (B, A)$ whereas in the second case $(A, B) = (B, A)$. Oriented edges have an arrow at one end and they are rather referred to as *arches* and denoted A. These different cases are pictured on Figure 5.2.

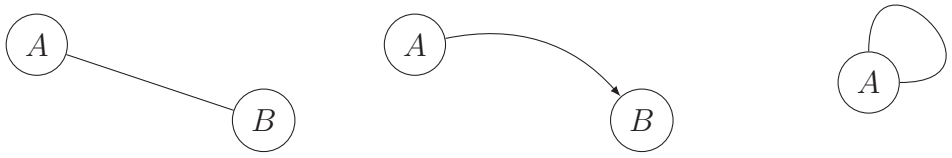


Figure 5.2: *An edge between 2 vertices (A, B) can be unoriented (left) or oriented (center), in this case the edge is also called* arch*. An edge that link a vertex to itself is called a* loop *(right).*

Loop A loop in a graph is an edge E_i that links a vertex V_i to itself, namely:

$$E_i = (V_i, V_i) \tag{5.2}$$

Simple graph An oriented simple graph is a graph with no internal loop and no double edges (or arches), that is :

$$A = \{(V_i, V_j) \in V \times V,\, i \neq j\} \tag{5.3}$$

In other words, no arch links a vertex V_i to itself and only one arch links V_i to V_j but another arch may link V_j to V_i. The edges of an unoriented simple graph read $E \subseteq \mathcal{P}_2(V)$, where $\mathcal{P}_2(V)$ is the powerset of cardinality 2 of V (a set of pairs of vertices $(V_i, V_j) \in E$, $i \neq j$).

Complete graph A complete graph K_n of size n is a graph where each vertex V_i is linked to all the vertices V_j, $j \neq i$. The total number of edges, *i.e.* the cardinality of E is given by:

$$|E| = \frac{n(n-1)}{2} \tag{5.4}$$

Clique A clique C of an unoriented graph G is a subset of vertices of V such that every pair of vertices of C are connected by an edge. It is equivalent to say that the subgraph induced by C is a complete graph. Finding the largest clique of a graph is a NP-complete

problem, *i.e.* a problem with computational complexity that exponentially increases with the size of the graph. The clique number, *i.e.* the number of vertices in the clique, is denoted by the $\omega(C)$.

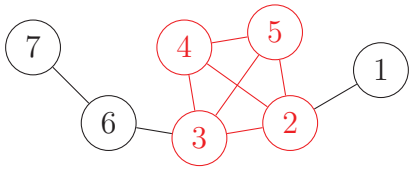


Figure 5.3: *A graph with 7 vertices. The set* $\{V_2, V_3, V_4, V_5\}$ *(in red) is a clique* C *with* $\omega(C) = 4$ *because every vertex* V_i *of* C *is linked to all the other vertices* $V_j \in C$, $i \neq j$.

Coloring A coloring of a loopless graph G is a way to label its vertices such that no two vertices forming an edge has the same color. The terminology refers to the problem of map coloring and the so called *four color problem*. The chromatic number χ_G is the smallest number of colors needed to color a graph G.

Perfect graph A perfect graph (or Bergé graph) is a graph for which the chromatic number of every induced subgraph equals the size of the largest clique of the same subgraph.

5.3.2.2 Graphs and linear algebra

The collection of edges E of an *unoriented* graph G induces symmetrical relationships between the vertices of the collection V. If the edge (A, B) exists, A and B are said to be *adjacent*. It is denoted $A - B$. The adjacency relation of $G = (V, E)$ with a collection of vertices V of size n can be described by its *adjacency matrix* $\mathbf{A}_G$ of size $n \times n$. $\mathbf{A}_G$ is defined by:

$$A_{ij} = \begin{cases} 1 & \text{if} \quad V_i - V_j \\ 0 & \text{either} \end{cases} \tag{5.5}$$

It is clear that $\mathbf{A}_G$ has zero entries on its diagonal and is a symmetrical matrix. Let us now consider the graph $G = (V, E)$ pictured in Figure 5.4 with:

- $V = \{1, 2, 3, 4, 5, 6\}$
- $E = \{(6, 4), (4, 5), (4, 3), (3, 2), (5, 2), (2, 1), (1, 5)\}$.

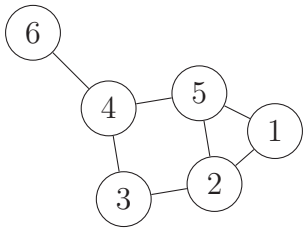


Figure 5.4: *A typical graph G with 6 vertices is taken as an example for the whole section.*

Its 6×6 adjacency matrix $\mathbf{A}_G$ is given by:

$$\mathbf{A}_G = \begin{pmatrix} 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \\ 1 & 1 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \tag{5.6}$$

More informations about a graph can be given by various matrices. The *degree matrix* $\mathbf{D}_G$ of a graph G is $n \times n$ diagonal matrix where each diagonal term D_{ii} corresponds to the number of connexions of the vertex V_i. This value is also referred to as its *degree* p_i.

$$\mathbf{D}_G = \begin{pmatrix} 2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 3 & 0 & 0 & 0 & 0 \\ 0 & 0 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 3 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix} \tag{5.7}$$

The *Laplacian matrix* $\mathbf{L}_G$ of a graph G is given by $\mathbf{L}_G = \mathbf{D}_G - \mathbf{A}_G$. The Laplacian matrix for the graph pictured on Figure 5.4 is given by:

$$\mathbf{L}_G = \begin{pmatrix} 2 & -1 & 0 & 0 & -1 & 0 \\ -1 & 3 & -1 & 0 & -1 & 0 \\ 0 & -1 & 2 & -1 & 0 & 0 \\ 0 & 0 & -1 & 3 & -1 & -1 \\ -1 & -1 & 0 & -1 & 3 & 0 \\ 0 & 0 & 0 & -1 & 0 & 1 \end{pmatrix} \tag{5.8}$$

Note that it gathers all the informations contained in $\mathbf{A}_G$ and $\mathbf{D}_G$. By construction, the sum of each row and each column of $\mathbf{L}_G$ is zero.

5.3.2.3 Graphs for illustrating sensitivity analysis results

The Ishigami function (Ishigami and Homma, 1990) is defined by:

$$Y = \sin(X_1) + 7\sin^2(X_2) + 0.1\, X_3^4 \sin(X_1) \tag{5.9}$$

where the X_i, $i = 1, 2, 3$, are uniformly distributed on $[-\pi, \pi]$. The numerical results of variance-based sensitivity analysis methods (see section 2.3) are given in Table 5.1.

Parameters	S^1	S^T
X_1	0.32	0.56
X_2	0.44	0.44
X_3	0.00	0.24
$\sum$	0.76	1.24

Table 5.1: *Results of variance-based sensitivity analysis for the Ishigami function.*

Sobol' first order indices reveal that the variables with the highest contribution to the output variance are by order of importance X_2 (0.44), X_1 (0.32) and X_3 who has no influence when taken alone. The difference with the Sobol' total indices allow one to deduce the numerical value of the interaction between X_1 and X_3, *i.e.* $S_{13} = 0.24$, which is due to the coupling term $0.1\, X_3^4 \sin(X_1)$. As it represents a collection of figures, it can be complicated to analyze, especially in higher dimension. A graphical representation using the graph theory, referred to as FANOVA-graph, has been proposed in Muehlenstaedt et al. (2012). Each variable X_i is represented by a circle-shaped vertex V_i whose thickness is proportional to the first order Sobol' indice S_i^1. Second-order indices S_{ij} are represented by straight lines between the couple of vertices (V_i, V_j). An illustration is given in Figure 5.5.

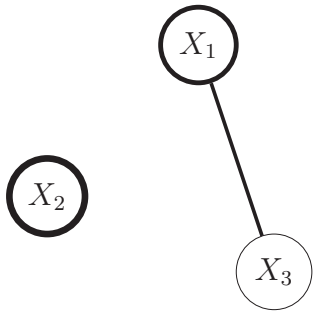


Figure 5.5: *Representation of the results of global sensitivity analysis for the Ishigami function using a graph.*

These conventions ease the comprehension of the results of sensitivity for high-dimensional problems. The (at least) third order, or total index of a collection of variables $X_I = (X_{i_1}, \ldots, X_{i_n})$ could also be represented by filling the area defined by the clique C_I of X_I with a scale of gray.

5.4 Sensitivity analysis for nested and multiscale modelling

In this section, a framework is proposed to represent nested or multiscale modelling of complex structures and to address sensitivity analysis problems.

5.4.1 Proposed methodology

Ths aim of global sensitivity analysis for nested and multiscale modelling is often to compute the sensitivity of the global model response, *i.e.* the output at the end of the chain, to variables from the lower levels. For such an analysis, it is necessary to map the positions of the variables in the workflow. The idea behind the *mapping* of the variables is to use some tools of the graph linear algebra. Let us consider the oriented graph G with 8 vertices and 8 edges described in Figure 5.6 representing a nested model. The variables are graphically organized in 3 levels. The *depth* D of a graph characterizes the maximum number of edges an input parameter has to take to reach the final output. In this case, variable X_8 depends on variables X_6 and X_7 who respectively depend on variables (X_1, X_2, X_3) and (X_3, X_4, X_5). Consequently, $D(G) = 2$.

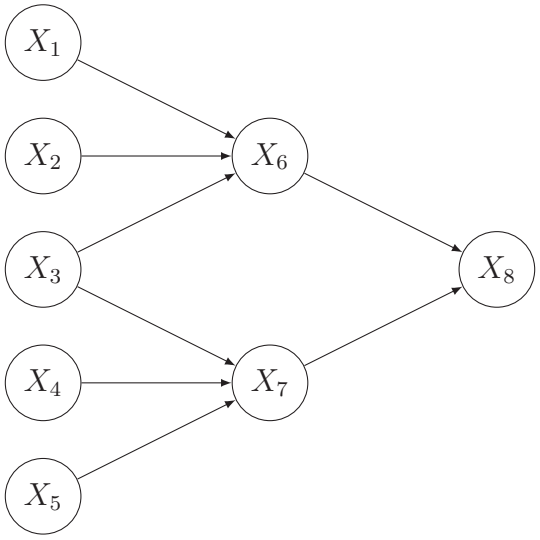


Figure 5.6: *A nested modelling of a structure represented by an oriented graph.*

Let us now define the matrix $\mathbf{I}_G$ by:

$$I_{ij} = \begin{cases} 1 & \text{if} \quad V_j \to V_i \\ 0 & \text{otherwise} \end{cases} \tag{5.10}$$

where $V_j \to V_i$ is the arch from V_i to V_j. In other words, $I_{ij} = 1$ if and only if $V_j \to V_i$ exists. The matrix $\mathbf{I}_G$ is referred to as the *incidence matrix* of the oriented graph G. The

incidence matrix of the graph presented in Figure 5.6 reads:

$$\mathbf{I}_G = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \\ X_7 \\ X_8 \end{array} \begin{array}{c} \begin{array}{cccccccc} X_1 & X_2 & X_3 & X_4 & X_5 & X_6 & X_7 & X_8 \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \tag{5.11}$$

When reading a given line l of $\mathbf{I}$, the ones correspond to the variables X_l depends on. The incidence matrix is derived from the graph linear adjacency and Laplacian matrices $\mathbf{A}_G$ and $\mathbf{L}_G$. The matrix has been rearranged so that the maximum information lies in the minimum number of terms. This way, the 8^{th} line of $\boldsymbol{I}_G$ indicates that the variable X_8 depends on X_6 and X_7. For deepest paternity, a simple algorithm can read the matrix and indicate the analyst that X_8 also depends on $(X_1, X_2, X_3, X_4, X_5)$ at a second level of modelling.

The steps of the general algorithm for an oriented graph G with input parameter X_i and output of interest Y are:

1. Read the line that corresponds to Y and store in a list $\mathcal{L}_1$ the indices of the columns that corresponds to the input X_i where $I_{Y,X_i} = 1$.
2. For each index in $\mathcal{L}_1$, read the line that corresponds to X_i and store in a list $\mathcal{L}_2$ the index of the columns that corresponds to the input X_j where $I_{X_i,X_j} = 1$.
3. Repeat operation until the all the lines to be read contains only zeros, meaning that the lowest level of the modelling has been reached. The last non-empty list is $\mathcal{L}_{D(G)}$.

When reading the matrix $\mathbf{I}_G$ columnwise, the analyst obtains information on the dependence between variables. Indeed, two (or more) variables with non zero terms on the same i^{th} column are correlated because they share the same input variable X_i.

As a conclusion, the first reading allows one to identify quickly which variables are the inputs of the considered output while the second reading provides an overview of the dependence, *e.g.* the copulas to be constructed. The abstract representation described above becomes tractable when linked with specific software for uncertainty modelling. This softwares are now described.

5.4.2 Software development

The aim of this work is to develop a methodology for the propagation of uncertainties through nested models. Thus, one component consists in taking into account the uncertainties by modelling the input parameters by a random vector while a second component

is in charge of modelling the workflow. In this subsection, two softwares corresponding to the previously described components are introduced. Then a coupling technique for the propagation of uncertainty through a nested model is proposed.

5.4.2.1 Handling uncertainty using OpenTURNS

OpenTURNS (for Open source initiative to Treat Uncertainties, Risks'N Statistics) is defined as an *uncertainty engineering software* (OpenTURNS, 2005). OpenTURNS allows the practitioner to propagate uncertainty in physical models. The computing functionalities are compatible with the Python language. Therefore, a wide range of uses (coupling, post-processing) are feasible.

More precisely, OpenTURNS provides efficient tools for the joint probabilistic modelling of variables (marginal distributions and copula), kernel smoothing estimation of model response probability density functions and the computation of response surfaces such as polynomial chaos expansions. In this particular case, one can decompose the model approximation into a polynomial basis and associated coefficients.

5.4.2.2 Drawing nested models with YACS

Salome is an open source software platform for numerical simulations. Among various modules (pre-processing, post-processing), YACS enables one to build and execute a chain of calculations where each link represents a coupling of computer codes (Python scripts, run of a software, *etc.*). YACS is thus a powerful tool to build nested models in a graphical way. The main tasks are:

1. creating the submodels,
2. defining their respectives input / output parameters,
3. associating a script with each model, which can include a call to any external code,
4. specifying the relationships between the models.

A graphical user interface makes it easy to represent graphically the various submodels and their links.

The calculation scheme represents the architecture of the nested model. Each model is represented by a node as shown in Figure 5.7. Then, the relationships between the models are pictured by arrows connecting the output parameter from a first model to the corresponding input parameter of a second one (see Figure 5.8). In addition to the nodes, input and output data modules (see Figure 5.9) can be added for setting the value of the input parameters and post-processing the response value.

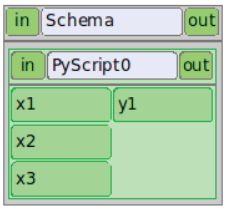


Figure 5.7: *A node figuring a model with 3 input and 1 output parameter in a YACS calculation scheme.*

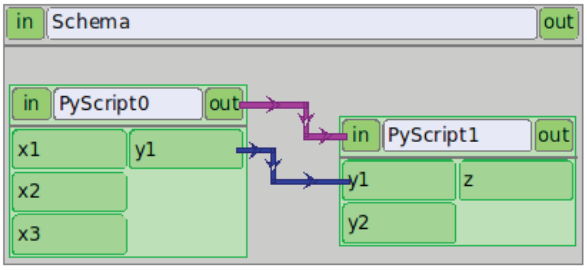


Figure 5.8: *A link between two nodes in YACS.*

The graph G of the YACS scheme featured pictured in Figure 5.9 (featuring the input and output modules) is drawn in Figure 5.10. Its incidence matrix then reads:

$$\boldsymbol{A}_G = \begin{array}{c} \\ X_1 \\ X_2 \\ X_3 \\ Y_1 \\ Y_2 \\ Z \end{array} \begin{array}{c} \begin{array}{cccccc} X_1 & X_2 & X_3 & Y_1 & Y_2 & Z \end{array} \\ \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 0 \end{pmatrix} \end{array} \tag{5.12}$$

The scheme of calculation can be exported in the XML (eXtensible Markup Language)

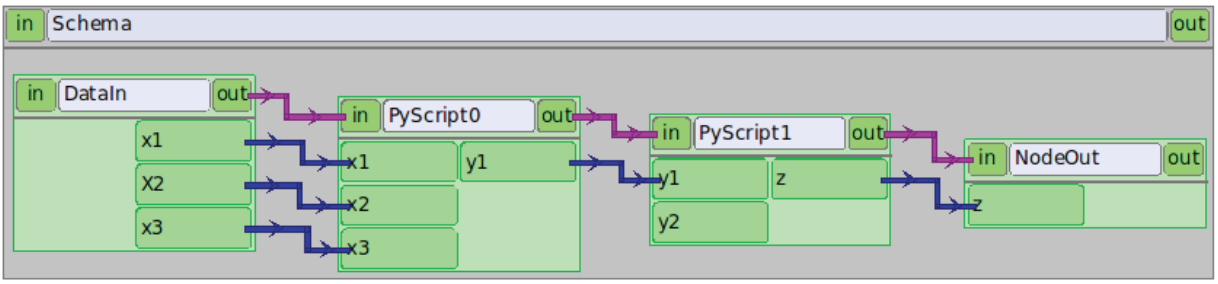


Figure 5.9: *Input and output data modules in YACS.*

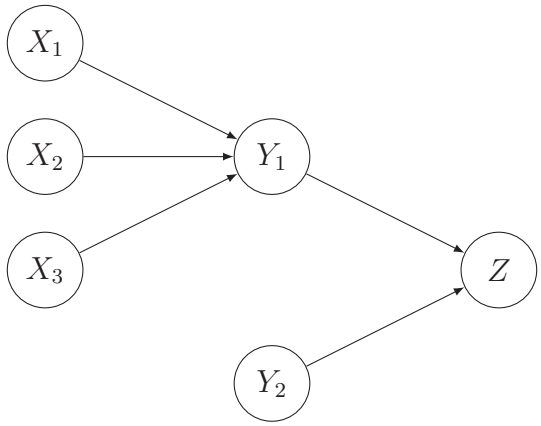


Figure 5.10: *Graph of the example scheme.*

format. The XML file is executable from a simple script. This allows one to run the chain of calculation for any numerical values of the input parameters. Then, providing the input data and collecting the response values can be carried out using input/output (I/O) files. Although this solution is not the most efficient one, the reading/writing durations are often negligible compared to the execution time of the scheme.

5.4.2.3 Run the workflow

On the computational side of multiscale modelling, one may require different softwares. In this work, the probabilistic modelling of the input random vector is addressed using OpenTURNS. Each marginal distribution is defined by its moments when they are available (given or infered) or by kernel smoothing estimation for non usual distributions. The dependence structure is modelled by a $n-$dimensional copula which can be independent, Gaussian or a composition of several copulas (Gaussian, Gumbel, Clayton, etc.).

```
myDistributionCollection = DistributionCollection(n)
myDistributionCollection[0] = Distribution(L0)
myDistributionCollection[1] = Distribution(L1)
...
myDistributionCollection[n] = Distribution(Ln)

myCopula = IndependentCopula(n)
myCopula = NormalCopula(R)
myCopula = ot.ComposedCopula(myCopulaCollection)

myInputDistribution = ComposedDistribution(myDistributionCollection,
    myCopula)
```

The physical modelling is addressed by the YACS platform. Each model of the nested scheme is executed by a Python script (or *Pyscript*) that reads the values of the input variables $\boldsymbol{x}^{(k)}$ in a text file *input.txt*. Once all the Pyscripts have been executed, the scheme writes the corresponding values of all the outputs (intermediary and final) $\boldsymbol{y}^{(k)}$ in

a second text file *output.txt* that is read by the OpenTURNSPythonFunction.

```
class myFunction(OpenTURNSPythonFunction):

    def __init__(self):
        OpenTURNSPythonFunction.__init__(self, 14, 1)

    def f(self, X):
        savetxt('input.txt', array(X))
        os.system('runSession driver schema_concrete.xml')
        Y = loadtxt('output_homog_concrete.txt')
        return Y
```

Finally, the PC approximation is build from the data $\boldsymbol{\mathcal{X}} = \{\boldsymbol{x}^{(k)}\, k = 1, \ldots, N\}$ and $\boldsymbol{\mathcal{Y}} = \{\boldsymbol{y}^{(k)}\, k = 1, \ldots, N\}$.

```
myModel = NumericalMathFunction(myFunction())

polynomialChaosResult = polynomialChaosAlgorithm(myInputDistribution,
    myModel)
```

The full coupling strategy between OpenTURNS and YACS is summarized in the form of an algorithm pictured in Figure 5.11.

5.5 Conclusion

In this chapter, a global framework for global sensitivity analysis in nested modelling is proposed. The methodology is inspired by the graph theory. This branch of mathematics provides concepts and tools which can be adapted to our concerns in this thesis. First, the graph representation helps the practitioner build a graphical model of the structure. Each node is assumed to be a model parameter and the edges between the nodes correspond to the relationships connecting the parameters, *i.e.* the submodels. Then, an adapted version of the adjacency matrix, namely the incidence matrix, is used to map the variables. From a software development point of view, this matrix helps identifying which variables are involved in the computation of an output of interest and at which level of modelling, as well as identifying the dependence between some intermediate variables : this is a useful information in order to determine which copula function has to be identified in the context of uncertainty propagation.

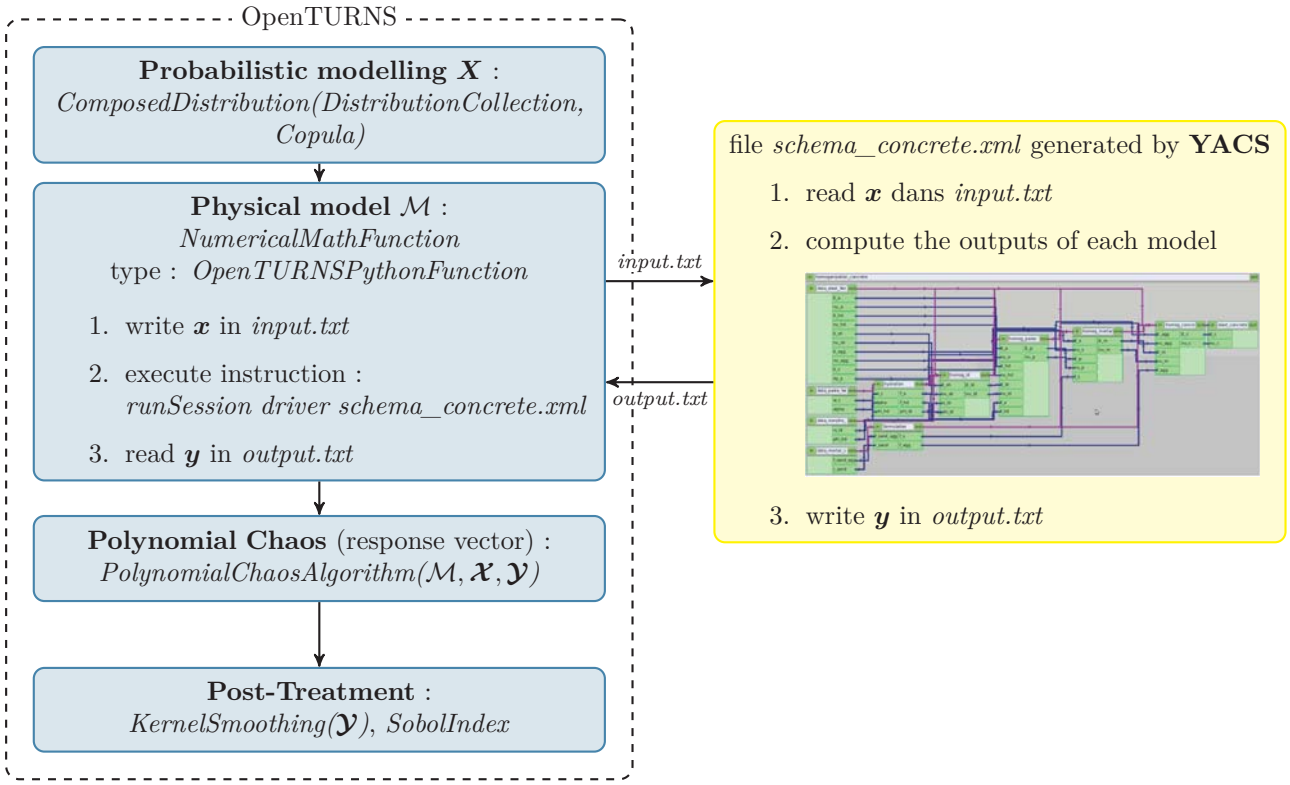


Figure 5.11: *The OpenTURNS / YACS coupling strategy for computing nested models with random input variables.*

6

Industrial applications

Contents

6.1 Introduction

In the first part of this chapter, the sensitivity analysis methods presented in Chapter 2, *i.e.* the δ sensitivity measure and the decomposition of covariance, are applied using analytical test functions in order to compare the results obtained with the computational schemes proposed in Chapter 4 with the reference results in original papers. They are then applied to academical problems with correlated inputs.

In the second part, the global methodology for addressing global sensitivity analysis for nested and multiscale modelling is applied to two mechanical problems, namely the homogenization of concrete material properties and the performance of an automobile diesel engine. The first problem is a multiscale model. Of interest are the sensitivities of the mechanical properties of concrete to the constituents from lower scales such as the mortar or the cement paste and their proportions. In the second problem, the engine is modelled by a multiphysics approach. The sensitivity of the damage in the cylinder head to the mechanical and thermal loads is studied.

6.2 The Ishigami function

6.2.1 An analytical model

The Ishigami function (Ishigami and Homma, 1990) is a numerical test case that has been broadly used to illustrate many global sensitivity analysis methods. The function reads:

$$Y = \sin(X_1) + a\,\sin(X_2)^2 + b\,X_3^4 \sin(X_1) \tag{6.1}$$

with $a = 7$, $b = 0.1$ and $X_i \sim \mathcal{U}[-\pi, \pi]$, $i = 1, 2, 3$. The PDF of the model response illustrated in Figure 6.1 shows a bimodal shape which is quite hard to approximate by kernel smoothing estimation. Indeed, up to $N_{ks} = 10^6$ sampling points are necessary to model the modes accurately. Therefore, trying to compute the δ sensitivity indices by the PDF-based estimation scheme proposed in Chapter 4, section 4.3.1, with accuracy is hardly achievable. The CDF-based estimation scheme proposed in Chapter 4, section 4.3.2, is used instead.

6.2.2 Computation and analysis of the indices

For the three input parameters, $N_q = 30$ conditional CDFs are simulated (see Figure 6.2). Each of them is estimated by kernel smoothing using $N_{ks} = 10^4$ sampling points and a Gaussian kernel whose bandwidth is the optimal Silverman's one. The support of the integration is arbitrary fixed to $[-15, 20]$. The results are given in Table 6.1.

Let us start by comparing the first two columns. The indices δ^{ref} have been calculated in Borgonovo (2007). The original author of this sensitivity measure first proposed a method using maximum likelihood estimation (MLE) for the estimation of the unconditional and conditional PDFs, the shift (here the absolute area between the PDFs) is

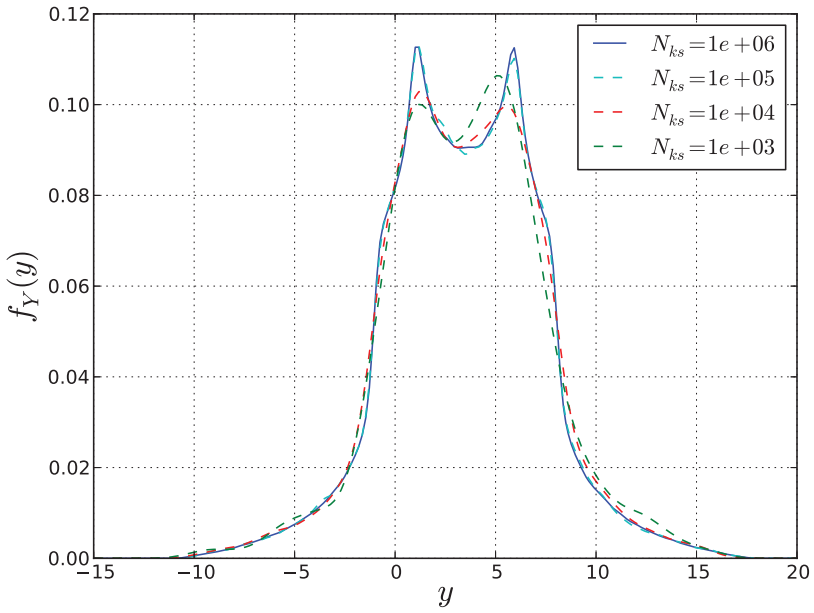


Figure 6.1: *The Ishigami function response PDF approximated by kernel smoothing with* $N_{ks} = 10^3$, 10^4, 10^5 *and* 10^6 *sampling points.*

computed from a simple numerical integration and the expected value of the shift is calculated with 10^3 Monte Carlo simulations. In addition to a probably long computational time, this procedure, and more precisely the estimation of the PDFs might bring inaccuracies to the computation of the shift and especially on this test case where the PDFs cannot be modelled by usual distributions. In other words, the values of δ^{ref} for X_1 and X_3 are likely overestimated.

Parameter	δ^{ref}	δ^{CDF}	S^1	S^T
X_1	0.33	0.224	0.314	0.558
X_2	0.39	0.394	0.442	0.442
X_3	0.28	0.103	0.000	0.244
Σ	1.00	0.721	0.756	1.244

Table 6.1: *Results of the distribution-based sensitivity analysis of the Ishigami function. The* δ^{ref} *are taken from the original paper by Borgonovo (2007).*

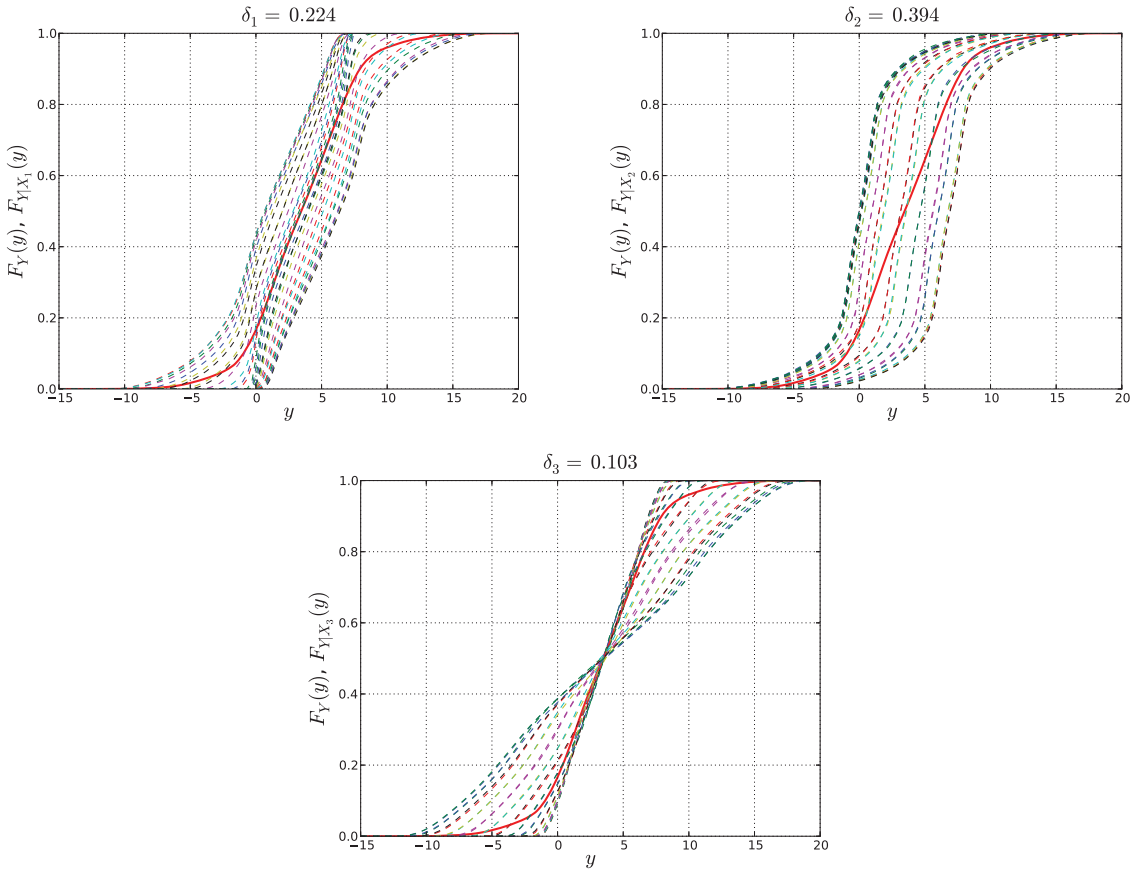


Figure 6.2: *Distribution-based sensitivity analysis of the Ishigami function. Simulations of $N_q = 30$ conditional CDFs for the estimation of the δ sensitivity measures.*

6.2.3 Discussion

The comparison between the δ^{CDF} and the Sobol' indices brings a more global information on the sensitivity analysis. The hierarchy for the importance measures δ^{CDF} and the Sobol' first order indices S^1 is the same: X_2 comes first, followed by X_1 and X_3. The latter has no influence when taken alone. The hierarchy differs when comparing δ^{CDF} and the Sobol' total indices S^T due to the value of $S_{13} = 0.244$. The range of the values is also different. X_3 alone does not contribute to the output variance (it does when coupled with X_1) but it modifies the shape of the response distribution as shown in Figure 6.2. On the one hand, the distribution-based sensitivity analysis appears as a more global sensitivity measure: the absolute area between the unconditional and conditional output distribution not only measures the reduction in the variance (the *width* of the PDF) but also detects the changes in the whole output distribution such as a *mode offset* or a *shape modification*. On the other hand, it is also a less discriminant method for model reduction applications.

6.3 Academical mechanical problems

In this section, the sensitivity analysis methods proposed in this thesis are applied to models with correlated inputs.

6.3.1 A rugby scrum

Rugby is a popular sport mainly played in Europe and in the Southern Hemisphere in which one key stage is the *scrum*. The pack (players from #1 to 8) of one team arranged in three rows tries to push the opposing pack to get the ball that is introduced between them by the scrum half (player #9) as illustrated in Figure 6.3. Of interest are the thrust of the first row (players from #1 to 3) at the shoulders of the *loosehead prop* #1, the *hooker* #2 and the *tighthead* prop #3.

Figure 6.3: *A rugby scrum is composed of three rows of players. The first row is composed of players #1, 2 and 3. The second row is composed of players #4 and 5. The third row is composed of players #6, 7 and 8.*

The difference in the thrust at the different players of the front row might be used to rotate the scrum and win the ball. The orientation of the thrust is governed by the player #8 who indicates to the *flankers* (players #6 and 7) and the *locks* (players #4 and 5) in the second row how to push. As the locks both push the hooker and a prop, the thrust of the latter is correlated. This correlation is reinforced if the thrust of the third row on the second row is considered. Finally, as the players in the front and seconds rows are linked by there shoulders, the thrusts Y_1, Y_2, Y_3 and Y_4 , Y_5 are also correlated. The graph of the model is given in Figure 6.4. The inner thrust of the player $\#i$, $i = 1, \ldots, 8$ is modelled by the input parameter X_i. The resulting thrust at player $\#j$, $j = 1, \ldots, 5$ of the front

row is modelled by a the output variable Y_j.

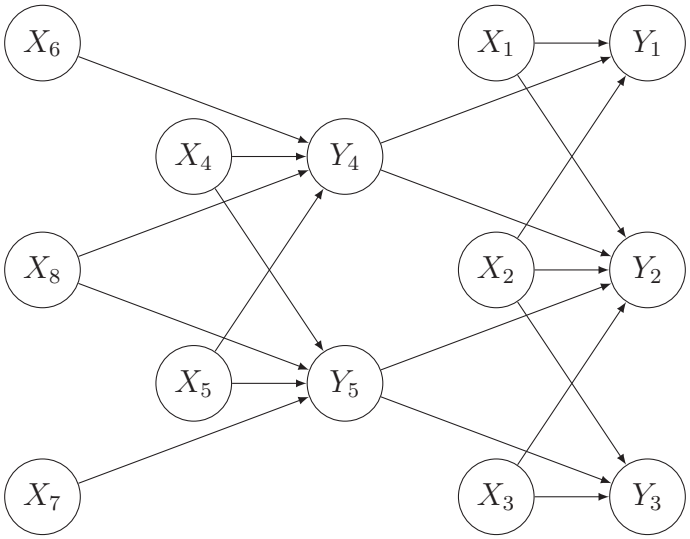


Figure 6.4: *Nested modelling of a rugby scrum.*

From a mathematical point of view, the model reads:

$$Y_4 = X_4 + a_{46}X_6 + a_{48}X_8 + a_{45}X_5 \tag{6.2}$$
$$Y_5 = X_5 + a_{57}X_7 + a_{58}X_8 + a_{54}X_4 \tag{6.3}$$
$$Y_1 = X_1 + a_{14}Y_4 + a_{12}X_2 \tag{6.4}$$
$$Y_2 = X_2 + a_{24}Y_4 + a_{25}Y_5 + a_{21}X_1 + a_{23}X_3 \tag{6.5}$$
$$Y_3 = X_3 + a_{35}Y_5 + a_{32}X_2 \tag{6.6}$$

where the coefficients a_{ji} indicates the percentage of the inner thrust of the player $\#i$ that is transmitted to the resulting thrust of the player $\#j$. That test case being purely of fiction, arbitrary values are taken, namely $a_{46} = a_{48} = a_{58} = a_{57} = 0.6$, $a_{14} = a_{24} = a_{25} = a_{35} = 0.8$ and $a_{12} = a_{21} = a_{32} = a_{23} = a_{45} = a_{54} = 0.2$. The input variables are modelled by normally distributed random variables as mentioned in Table 6.2.

The linear correlation matrices of the resulting thrust in the front row $\mathbf{\Sigma}_1$ and in the second row $\mathbf{\Sigma}_2$ are given in Eq. (6.7).

$$\mathbf{\Sigma}_1 = \begin{bmatrix} 1 & 0.58 & 0.12 \\ 0.58 & 1 & 0.58 \\ 0.12 & 0.58 & 1 \end{bmatrix}, \quad \mathbf{\Sigma}_2 = \begin{bmatrix} 1 & 0.42 \\ 0.42 & 1 \end{bmatrix} \tag{6.7}$$

In order to differentiate the uncorrelated (inner contribution of the player) and correlated (interaction with other players) effects, an ANCOVA sensitivity analysis is processed. The results are given in Table 6.3.

First of all, the uncorrelated contributions are logical due to the additive nature of the model. One may also notice that the contributions to the resulting thrust Y_1 and

Parameter	Distribution	Mean	CV
X_1	Normal	1200 N	10%
X_2	Normal	1300 N	10%
X_3	Normal	1200 N	10%
X_4	Normal	1100 N	10%
X_5	Normal	1100 N	10%
X_6	Normal	900 N	10%
X_7	Normal	900 N	10%
X_8	Normal	1000 N	10%

Table 6.2: *Probabilistic modelling of the inner thrust of the players.*

X_i	S	S^U	S^C
X_1	0.52	0.51	0.01
X_2	0.03	0.03	0.00
X_3	–	–	–
Y_4	0.45	0.44	0.01
Y_5	–	–	–
Σ	1.00	0.98	0.02

(a) Y_1

X_i	S	S^U	S^C
X_1	0.02	0.01	0.01
X_2	0.32	0.32	0.00
X_3	0.02	0.01	0.01
Y_4	0.32	0.22	0.10
Y_5	0.32	0.22	0.10
Σ	1.00	0.78	0.22

(b) Y_2

X_i	S	S^U	S^C
X_1	–	–	–
X_2	0.03	0.03	0.00
X_3	0.52	0.51	0.01
Y_4	–	–	–
Y_5	0.45	0.44	0.01
Σ	1.00	0.98	0.02

(c) Y_3

Table 6.3: *Results of the ANCOVA sensitivity analysis for the rugby scrum. From left to right is the sensitivity of the resulting thrust in the first row Y_1, Y_2, Y_3 to the inner thrust of the first row X_1, X_2, X_3 and the resulting thrust in the second row Y_4, Y_5.*

Y_3 are symmetrical in this case. For the resulting thrusts of the props Y_1 and Y_3, the total contributions are almost entirely carried by the uncorrelated contributions ($S_1^C = S_3^C = 0.01$), *i.e.* the inner thrust of players #1 and 4 and #3 and 5. Concerning the thrust at the hooker #2, the correlated effects of the locks #4 and 5 is more sensitive $S_4^C = S_5^C = 0.10$. As a conclusion, if the resulting thrust on the side of the scrum is mainly governed by the side prop and lock, the resulting thrust in the middle benefits from the interaction between all the players. If in this example the model is symmetrical, it is most of the time not the case, two players having different size and strength. Wittingly, the coaching is able to control and optimize the resulting thrusts in the first row by choosing scrum configuration and strategy.

6.3.2 A composite beam

This problem has been originally assessed in Caniou and Sudret (2010). It deals with a beam on two supports as illustrated in Figure 6.5. The material is composed of a fraction f of fibers and a fraction $1-f$ of matrix. The respective Young's modulus and density of the constituents are E_f, ρ_f and E_m, ρ_m. The beam is of length L and has a rectangular section $b \times h$. Of interest is the maximal mid-span deflection v, namely:

$$v = \frac{5}{384} \frac{qL^4}{E_{\text{hom}} I} \tag{6.8}$$

where q is the distributed load:

$$q = \rho_{\text{hom}} g b h \tag{6.9}$$

and where E_{hom} and ρ_{hom} are the homogeneous Young's modulus and density of the composite material, namely:

$$E_{\text{hom}} = f E_f + (1-f) E_m \tag{6.10}$$

$$\rho_{\text{hom}} = f \rho_f + (1-f) \rho_m \tag{6.11}$$

The module of bending reads:

$$I = \frac{bh^3}{12} \tag{6.12}$$

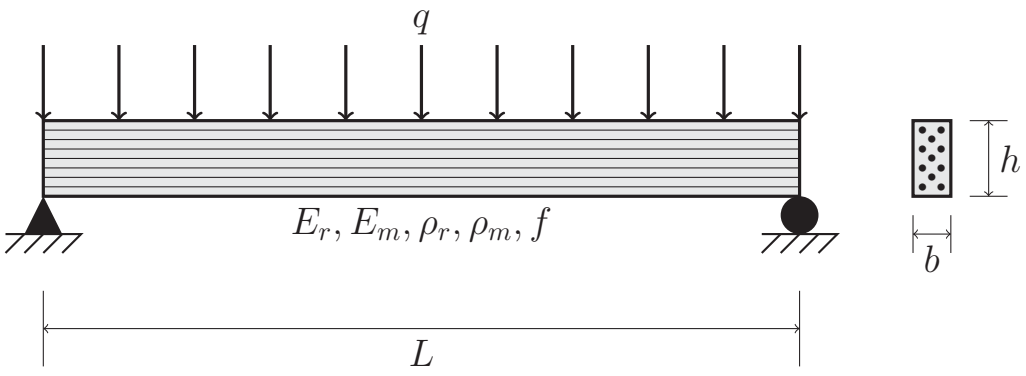


Figure 6.5: *A composite beam on two supports.*

The probabilistic modelling of the input parameters is presented in Table 6.4.

The composite can be seen as a multiscale model since its structure is decomposed up to the constituting parameters of the material. A graph representation of the composite beam is proposed in Figure 6.6. If L has no input variable at a lower level of modelling, note that q and E_{hom} have a common input of level 0 f and that q and I have common inputs of level 0 b and h. Therefore, the parameter of level 1 q is correlated with both E_{hom} and I. Their dependograms are presented in Figure 6.7. The dependence structure of the variables of level 1 $(L, I, q, E_{\text{hom}})$ is modelled by a 4−dimensional Gaussian copula whose parameter matrix is derived from the following Spearman's rank correlation matrix:

Parameter	Distribution	Mean	CV
L	Lognormal	2 m	1%
b	Lognormal	10 cm	3%
h	Lognormal	1 cm	3%
E_f	Lognormal	300 GPa	15%
E_m	Lognormal	10 GPa	15%
ρ_f	Lognormal	1800 kgm^{-3}	3%
ρ_m	Lognormal	1200 kgm^{-3}	3%
f	Beta$[0, 1]$	0.5	10%

Table 6.4: *Probabilistic modelling of the input parameters for the composite beam.*

$$\boldsymbol{\rho}_S = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0.71 & 0 \\ 0 & 0.71 & 1 & 0.20 \\ 0 & 0 & 0.20 & 1 \end{pmatrix} \tag{6.13}$$

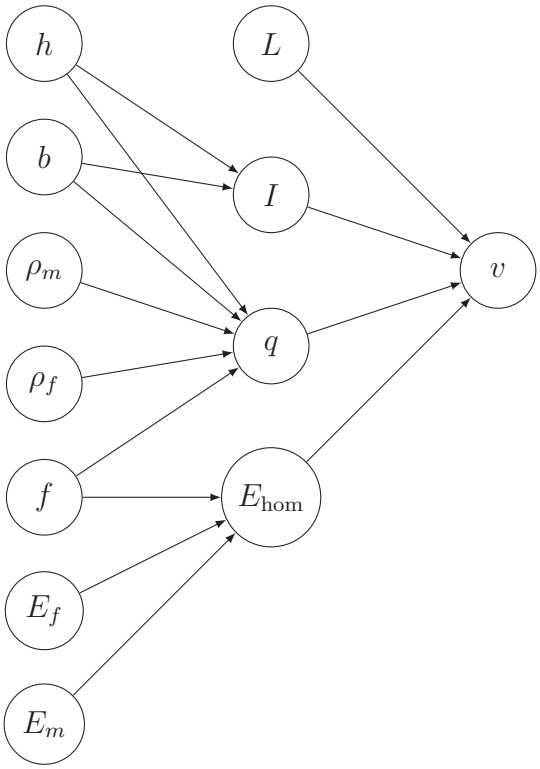


Figure 6.6: *Nested modelling of the composite beam.*

The δ sensitivity measures are computed with $N_{ks} = 10^4$ sampling points for the kernel

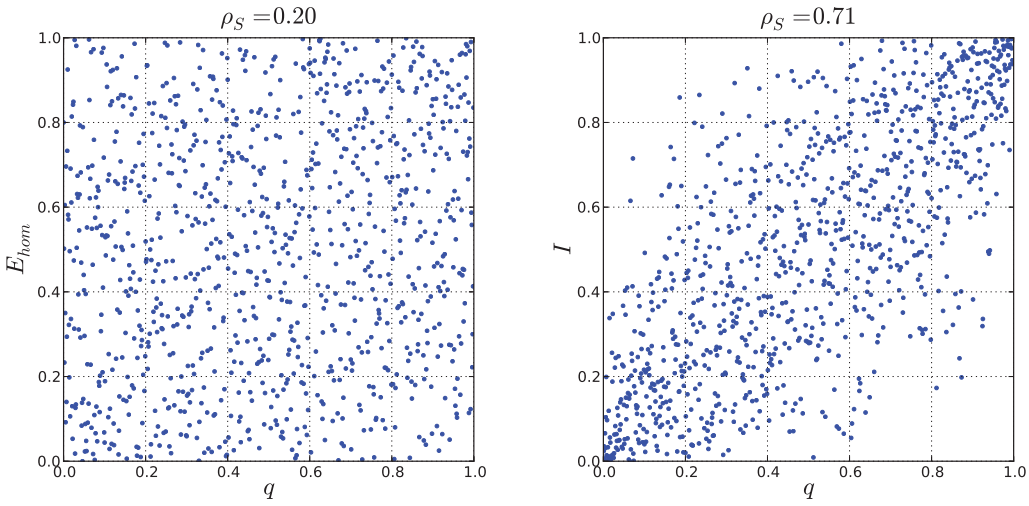


Figure 6.7: *Dependograms of the couples* (q, E_{hom}) *and* (q, I).

smoothing estimation and $N_q = 30$ quadrature points for the computation of the expected shift. The conditional CDFs are presented in Figure 6.8. An ANCOVA sensitivity analysis is also performed. The full results are presented in Table 6.5.

Parameter	δ^{CDF}	S	S^U	S^C
q	0.015	-0.08	0.09	-0.17
L	0.003	0.01	0.01	0.00
E_{hom}	0.546	0.89	0.94	-0.05
I	0.129	0.18	0.30	-0.12
Σ	0.693	1.00	1.34	-0.34

Table 6.5: *Results of the sensitivity indices for the composite beam.*

The analysis of the results clearly shows the domination of the contribution of E_{hom} followed by I, q and L. The low contribution of L is more due to its low coefficient of variation (1%) than to its role in the physical model. In other words, the mechanical properties of the composite material and the section dimensions have so high dispersions that they hide the structural contribution of L. When looking at the ANCOVA indices, the same hierarchy is observed. The correlated contributions are non zero and negative. This can be explained by the definition of the midspan deflection: q, which is located at the numerator, is correlated to both E_{hom} and I which are located at the denominator. Thus, the positive correlation between them is limited by their relative position in the fraction, *i.e.* the correlation tends to reduce the total contribution of the variables. It has a positive effect on the variability of v.

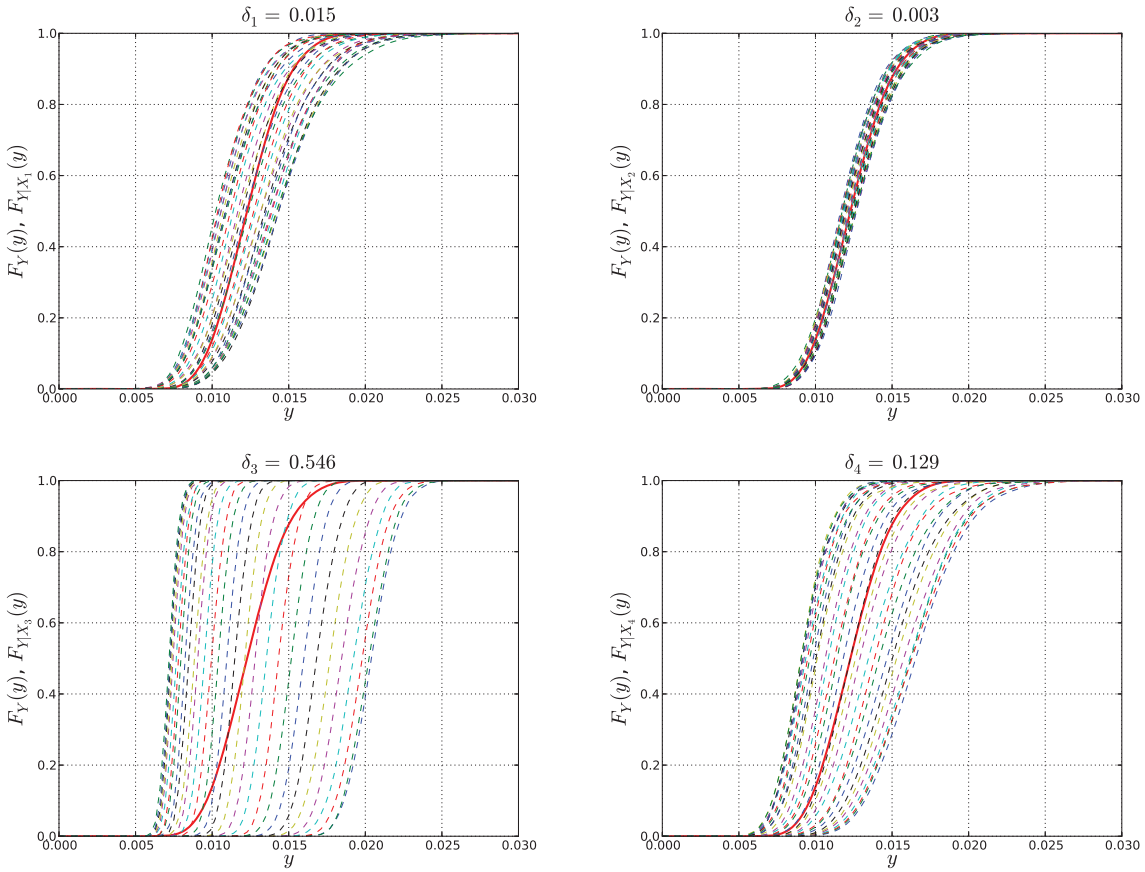


Figure 6.8: *Distribution-based sensitivity analysis of the composite beam. Simulations of* $N_q = 30$ *conditional CDFs for the estimation of the* δ *sensitivity measures.*

6.3.3 A bracket structure

This section proposes a second mechanical example that has already been addressed in Chateauneuf and Aoues (2008) and Dubourg (2011). The results have been originally presented in Caniou et al. (2012b).

The bracket structure is composed of two beams of section $w \times t$. On top of the dead load, a vertical load is applied at the right tip of the upper beam as sketched in Figure 6.9. The probabilistic modelling of the inputs is given in Table 6.6. Of interest is the *bending stress* σ_B reading:

$$\sigma_B = \frac{6M_B}{w_{CD}t^2} \tag{6.14}$$

where:

$$M_B = \frac{PL}{3} + \frac{\rho g w_{CD} t L^2}{18} \tag{6.15}$$

is the bending moment of the structure.

The parameters w_{CD} and t, namely the width and height of the beam CD are supposed

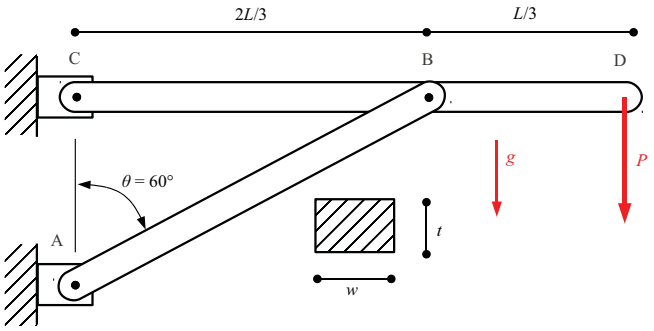


Figure 6.9: *Definition drawing of the bracket structure.*

Parameter	Unity	Distribution	Mean μ	σ/μ
P	kN	Gumbel	100	10%
ρ	kg.m^{-3}	Weibull	7860	10%
L	m	Normal	5	5%
w_{CD}	mm	Normal	125	10%
t	mm	Normal	250	10%

Table 6.6: *Probabilistic modelling of the variables.*

correlated ($\rho(\omega_{CD}, t) = 0.8$) due to the manufacturing process. An ANCOVA sensitivity analysis is carried out using a polynomial chaos expansion of degree $p = 7$ ($1-Q^2 \approx 10^{-8}$). The variances and covariances are computed using samples of size $N = 10^4$. The results are presented in Table 6.7.

The sensitivity analysis reveals that the input parameters with the most important contribution to the variability of the bending stress σ_B are the length L of the upper beam AB and the height t of the beams. Although the correlation between w_{CD} and t is quite strong, the correlative part of their contribution remains small, *i.e.* $S_t^C = S_{w_{CD}}^C = 0.02$. This is mainly due to the low total contribution of w_{CD} to the variability of σ_B.

Let us now study the evolution of the uncorrelated and correlated parts of the contributions as a function of the correlation. The computation of the ANCOVA indices is done for $\rho(\omega_{CD}, t)$ varying from -0.9 to 0.9. The results are presented in Figure 6.10. As a con-

Parameter	δ^{CDF}	S	S^U	S^C
P	0.02	0.00	0.00	0.00
ρ	0.01	0.00	0.00	0.00
L	0.27	0.42	0.42	0.00
w_{CD}	0.02	0.02	0.00	0.02
t	0.22	0.56	0.54	0.02
Σ	0.65	1.00	0.96	0.04

Table 6.7: *Results of the ANCOVA sensitivity analysis for the bracket structure.*

clusion, the correlative part of the contribution of one input parameter depends on both the strength of the correlation with other parameters and inner (structural) contributions of the correlated variables.

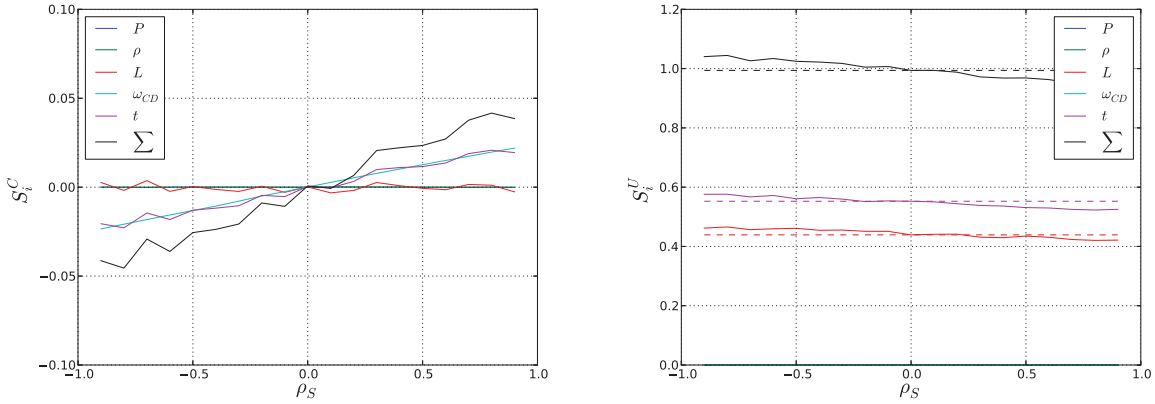


Figure 6.10: *Evolution of the correlated and uncorrelated parts of the contribution when* $\rho(\omega_{CD}, t) = 0.8$ *varies.*

6.3.4 Electrical connectors

This problem proposed by the RADIALL Company in Gayton et al. (2011) is a contribution to the field of manufacturing tolerancing. The issue is to determine the probability that two electronic connectors, one male and one female cannot assemble. Indeed, the quality of the connection depends on the backlash between the functional surfaces. On the one hand this backlash shall not be too large in order to preserve the connection and on the other hand it shall not be to small or the male connector will not fit with the female one. The connectors are sketched in Figure 6.11.

Of interest is the variability of the contact length Y of the pin (light gray) in the socket (light blue) (see Figure 6.11). As a functional requirement, Y shall not exceed a threshold

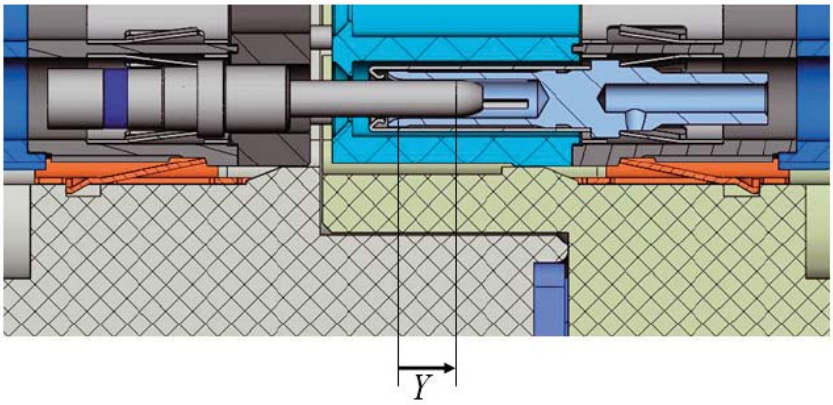


Figure 6.11: *CAE drawing of the connectors problem.*

value $t = 1.75$ mm otherwise the minimal contact length to enable a good conductivity in an electrical circuit will not be fulfilled. More precisely, using geometrical considerations, the contact length Y reads:

$$c = \frac{cm_4 + cm_8 + cm_7}{2} \tag{6.16}$$

$$i = \frac{ima_{17} + ima_{19} + ima_{20}}{2} \tag{6.17}$$

$$h = ima_2 + ima_3 \tag{6.18}$$

$$\alpha = \arccos\left(\frac{c}{\sqrt{i^2 + h^2}}\right) - \arccos\left(\frac{i}{\sqrt{i^2 + h^2}}\right) \tag{6.19}$$

$$r_2 = \frac{\dfrac{ima_{19} + ima_{20}/2}{2} - \dfrac{cm_8 + cm_7/2}{2\cos(\alpha)}}{\tan(\alpha)} \tag{6.20}$$

$$z = \frac{r_2}{\cos(\alpha)} + \left(\frac{cm_9 + cm_{10}/2}{2} + \frac{cm_7}{4}\right)\tan(\alpha) \tag{6.21}$$

$$J_1 = (cm_1 - cm_3 - z)\sin(\alpha) \tag{6.22}$$

$$J_2 = \frac{cm_7}{4}\cos(\alpha) \tag{6.23}$$

$$J_3 = \frac{cm_5 + cm_6}{2}\cos(\alpha) \tag{6.24}$$

$$Y = J_1 + J_2 + J_3 \tag{6.25}$$

All the dimensions are sketched in Figure 6.12. Because the different surfaces are machined during the same manufacturing operation, their dimension are highly correlated.

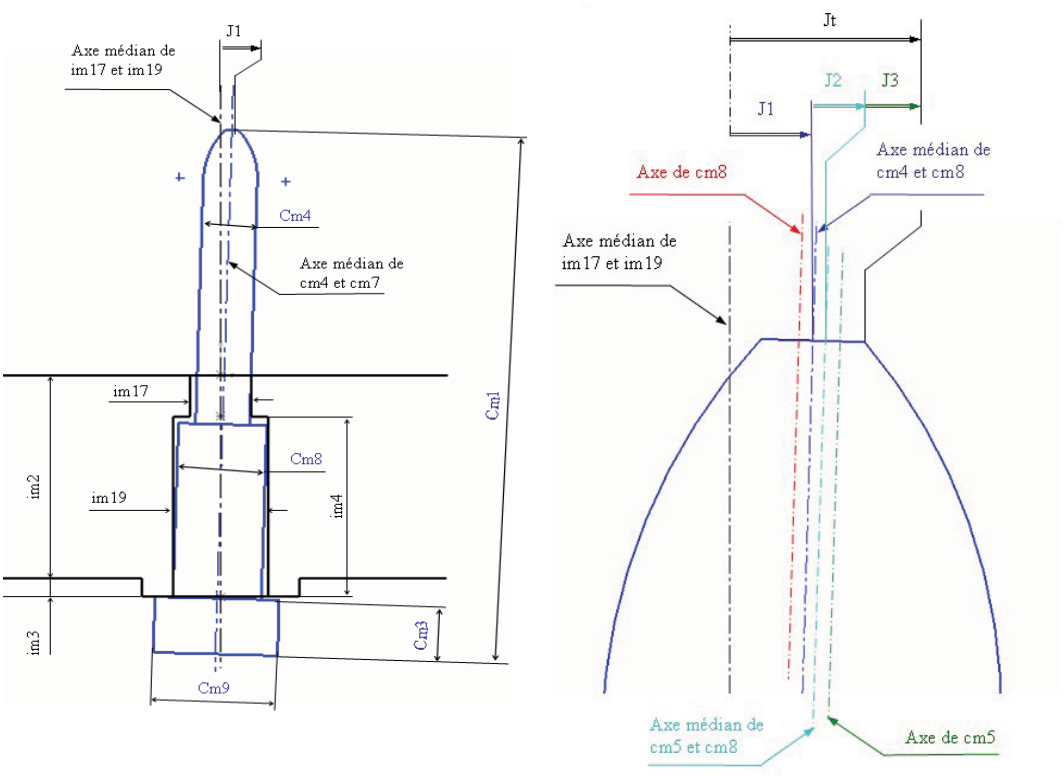


Figure 6.12: *Definition drawing of the connectors problem according to Gayton et al. (2011).*

The different Spearman correlation coefficients read:

$$\rho(cm_1, cm_3) = 0.8 \tag{6.26}$$

$$\rho(cm_4, cm_8) = \rho(cm_4, cm_9) = \rho(cm_8, cm_9) = 0.8 \tag{6.27}$$

$$\rho(ima_{17}, ima_{19}) = \rho(ima_{17}, ima_{20}) = \rho(ima_{19}, ima_{20}) = 0.8 \tag{6.28}$$

The derived Gaussian copula is parameterized by the transformed matrix in Eq. (1.73).

The δ sensitivity measures are computing using $N_{ks} = 10^4$ sampling points for the kernel smoothing of the CDFs and $N_q = 30$ quadrature points for the estimation of the expected shift. An ANCOVA sensitivity analysis is also carried out on the model. A polynomial chaos expansion of degree $p = 2$ is used. Due to the large dimension of the problem, *i.e.* $n = 14$, the number of coefficients to compute is high, *i.e.* $P = 120$.

According to the results presented in Table 6.9, the main contributors to the variability of Y are by order of importance the dimensions cm_7, cm_5, ima_{17}, ima_{19} and ima_{20}. Only the three last ones have significant correlative contributions. This can be explained by both the high rank correlation between them ($\rho = 0.8$) and the substantial uncorrelative correlation ($S^U \geqslant 5\%$) of each of them. Note that once again, the δ sensitivity measures are hierarchically consistent with the Sobol' indices although their range sensibly differs.

Parameter	Unity	Distribution	Mean μ	σ
cm_1	mm	Normal	10.530	0.200/6
cm_3	mm	Normal	0.750	0.040/6
cm_4	mm	Normal	0.643	0.015/6
cm_5	mm	Normal	0.100	0.200/6
cm_6	mm	Normal	0.000	0.060/6
cm_7	mm	Normal	0.000	0.200/6
cm_8	mm	Normal	0.720	0.040/6
cm_9	mm	Normal	1.325	0.050/6
cm_{10}	mm	Normal	0.000	0.040/6
ima_2	mm	Normal	3.020	0.040/6
ima_3	mm	Normal	0.400	0.040/6
ima_{17}	mm	Normal	0.720	0.040/6
ima_{19}	mm	Normal	0.970	0.040/6
ima_{20}	mm	Normal	0.000	0.040/6

Table 6.8: *Probabilistic modelling of the dimensions and their tolerances.*

6.4 Homogenization of concrete

6.4.1 Introduction

This main case study aims at understanding the behavior of concrete through the so-called homogenization modelling (Sanahuja et al., 2007). It has been proposed by EDF R&D as a multiscale problem, *i.e.* the material is studied at different scales of modelling, from the molecules at the nanoscale to the largest aggregates. This work has been originally presented in Caniou et al. (2012a).

Concrete is a composite material broadly used in civil engineering (earthwork, walls, columns, foundations). Its is composed of cement paste, water and aggregates such as sand and gravel. Its fabrication consists in mixing the cement powder with water to form a paste that is left to dry to obtain a hard material. During the drying, the powder particles dissolve in the water, ions precipitate to form crystals of gypsum. The hydration mechanism is of type dissolution-precipitation.

6.4.2 Homogenization of concrete

The hardening mechanism of cement can be decomposed into several stages. The first of them is the setting phase.

Parameter	δ^{CDF}	S	S^U	S^C
cm_1	0.00	0.00	0.00	0.00
cm_3	0.00	0.00	0.00	0.00
cm_4	0.02	0.02	0.01	0.01
cm_5	0.11	0.15	0.15	0.00
cm_6	0.01	0.02	0.02	0.00
cm_7	0.30	0.52	0.51	0.01
cm_8	0.03	0.03	0.02	0.01
cm_9	0.00	0.00	0.00	0.00
cm_{10}	0.00	0.00	0.00	0.00
ima_2	0.01	0.00	0.00	0.00
ima_3	0.01	0.00	0.00	0.00
ima_{17}	0.07	0.11	0.05	0.06
ima_{19}	0.05	0.07	0.02	0.05
ima_{20}	0.07	0.09	0.03	0.06
Σ	0.68	1.00	0.81	0.19

Table 6.9: *Results of the sensitivity analysis for the electrical connectors.*

6.4.2.1 Setting phase

Hydration starts when anhydrous cement powder is mixed with water. The properties of the obtained cement paste evolve over time. At the beginning of the hydration mechanism, the paste is sufficiently malleable to conform to the shape of a mold or a formwork. Advancing in time, the crystals of hydrate are growing in space, the viscosity increases and the paste hardens. The setting phase is followed by the hardening phase.

6.4.2.2 Hardening phase

Hydration can last up to several months during which the mechanical properties of the cement paste such as its Young's modulus, yield strength and ultimate tensile strength increase. The hardening mechanism is now described. First the constituents of the cement paste (or *clinker*) are presented.

Common cement or *Portland cement* (Tennis and Jennings, 2000) is composed of several anhydrous constituents:

- tricalcium silicate (or *alite*) $3CaO.SiO_2$ (C_3S)
- bicalcium silicate (or *belite*) $2CaO.SiO_2$ (C_2S)
- tricalcium aluminate (or *celite*) $3CaO.Al_2O_3$ (C_3A)
- tetracalcium aluminoferrite $4CaO.Al_2O_3.Fe_2O_3$ (C_4AF)

For the sake of a simpler writing, one prefers to use the following equivalences:

$$CaO = \boldsymbol{C}, \quad SiO_2 = \boldsymbol{S}, \quad Al_2O_3 = \boldsymbol{A}, \quad Fe_2O_3 = \boldsymbol{F}, \quad SO_3 = \boldsymbol{s}, \quad H_2O = \boldsymbol{H}$$

The calcium silicates C_2S and C_3S (70% of the mass) are the constituents that bring its strength to the cement paste. C_3A and C_4AF (20% of the mass) are required as melting agents, *i.e.* they help the setting of the calcium silicates.

The first constituent to react during hydration is the Celite C_3A. It dissolves and recrystallizes. It is followed by an hydrolysis reaction of the alite C_3S which forms a film of tobermolite gel around the particles and bring Ca^+ ions in the solution. Hydrolysis and recrystallization of C_3A are so fast that they must be slowed down to prevent the cement to become unstable, and consequently unusable, by a too stiff setting.

When mixing anhydrous cement paste and water, a paste in which the molecules of water surround each cement particle is obtained. A capillary network is formed. Water molecules react with the surface of the anhydrous particles to form hydrated compounds. The calcium silicates C_2S and C_3S (hydrated lime) progressively dissolve and hexagonal-shaped $Ca(OH)_2$ crystals settle. The hydrated calcium silicates form a gel composed of fine needles on the surface of cement particles. These needles grow in size and number filling the capillary network between the particles. Thus the paste becomes harder.

The more the gel fills the capillary network, the stronger the cement paste becomes. This strength increases gradually as the gel becomes compact, firstly because the cohesion between the needles strengthens and secondly because they are welded to the cement grains. Unhydrated cement grains still remain in the hardened cement paste. The hydration continues for months or even years provided the tobermolite gel is surrounded by water.

6.4.2.3 Principal chemical reactions

Due to the complexity of the hydration reaction of Portland cement, only the principal chemical reactions are presented in this section for a better comprehension of the hardening mechanism.

$$2(3CaO + SiO_2) + 6H_2O = 3CaO.3SiO_2.3H_2O + 3Ca(OH)_2 \tag{6.29}$$
$$2(2CaO + SiO_2) + 3H_2O = 3CaO.3SiO_2.3H_2O + 2Ca(OH)_2 \tag{6.30}$$
$$3CaO.Al_2O_3 + 6H_2O = 3CaO.Al_2O_3.6H_2O \tag{6.31}$$
$$4CaO.Al_2O_3.Fe_2O_3 + mH_2O = 3CaO.Al_2O_3.6H_2O + 3CaO.Fe_2O_3.nH_2O \tag{6.32}$$
$$\tag{6.33}$$

$$\underbrace{4CaO.Al_2O_3 + 3(\overbrace{CaSO_4.2H_20}^{\text{gypsum}})}_{\text{Portland cement}} + 26H_2O = \underbrace{3CaO.Al_2O_3.3Ca.CaSO_4.32H_2O}_{\text{calcium hydrosulfate aluminate}} \tag{6.34}$$

The two first reactions describe the hydration of calcium silicates C_3S (Eq. (6.29)) and C_2S (Eq. (6.30)). The third and fourth equations respectively refer to the tricalcium aluminate C_3A (Eq. (6.31)) and C_4AF (Eq. (6.32)).

6.4.2.4 Cement, mortar, concrete

Cement paste can be mixed with aggregates to reinforce its structural properties. Mortar refers to cement paste with sand. When bigger aggregates such as gravel are added to mortar, the obtain composite material is referred to as concrete. The mechanical properties of the final product, whether mortar or concrete, strongly depends on the properties of the constituents and their respective proportions. Similarly, the constitution of the cement powder plays an important role in the performance of concrete.

The so-called homogenization method, that allows one to compute the mechanical properties at the different scales of modelling, establishes the framework of this study.

6.4.3 Multiscale modelling of the mechanical properties of concrete

The submodels corresponding to the different scales of modelling are introduced. For a more detailed overview on multiscale modelling of concrete the reading of the book by Bernard et al. (2003) is advised.

6.4.3.1 Hydration model

The hydration model involves two types of models:

- The Powers model (Powers and Brownyard, 1947) establishes a link between on the one hand the water / cement ratio and on the other hand the volume fractions of the constituents that are the anhydrous, the hydrates, the capillary pores, water and the total pores *i.e.* including the volume occupied by the gel.
- The Tennis and Jennings model (Tennis and Jennings, 2000) allows one to compute the volume fractions of the low- and high-density hydrates in terms of water / cement ratio and hydration degree.

The input parameters of the hydration model are:

- w/c, the water / cement ratio,
- α, the hydration degree,
- ϕ_{hd}, the porosity of the high-density hydrates.

whereas its output parameters are:

- f_{a}, the volume fraction of the anhydrous,
- f_{hd}, the volume fraction of the hydrates,
- ϕ_{ld}, the porosity of the low-density hydrates.

6.4.3.2 Dosing in sand and aggregates model

A formula allows one to determine the volume fraction of sand in mortar and the volume fraction of aggregates in concrete from the volume fraction of sand and aggregates in concrete and the ratio between the volume of sand and the volume of sand and aggregates.

The input parameters of the dosing model are:

- $f_{\mathrm{sand/agg}}$, the volume fraction of sand and aggregates in concrete,
- r_{sand}, the ratio between the volume of sand and the volume of sand and aggregates.

whereas its output parameters are:

- f_{sand}, the volume fraction of sand in mortar,
- f_{agg}, the volume fraction of aggregates in concrete.

The output parameters respectively read:

$$f_{\mathrm{sand}} = f_{\mathrm{sand/agg}} \times \frac{r_{\mathrm{sand}}}{1 - f_{\mathrm{sand/agg}}(1 - r_{\mathrm{sand}})} \tag{6.35}$$

$$f_{\mathrm{agg}} = f_{\mathrm{sand/agg}} \times (1 - r_{\mathrm{sand}}) \tag{6.36}$$

6.4.3.3 Low-density hydrates homogenization model

The low-density hydrates homogenization model allows one to compute the mechanical properties of the low-density hydrates E_{ld} and ν_{ld}. The equations are those from the micro-mechanics. They involve the mechanical properties of the solid hydrates E_{sh} and ν_{sh}, the shape parameter of the flat particles rs_{ld} and the porosity of the low-density hydrates ϕ_{ld}. The low-density hydrates homogenization model follows an *self-consistent* scheme (Hill, 1965; Bornert et al., 2010).

The input parameters of the low-density hydrates homogenization model are:

- E_{sh}, the Young's modulus of the solid hydrates,
- ν_{sh}, the Poisson's ratio of the solid hydrates,
- rs_{ld}, the shape parameter of the flat particles,
- ϕ_{ld}, the porosity of the low-density hydrates.

whereas its output parameters are:

- E_{ld}, the Young's modulus of the low-density hydrates,
- ν_{ld}, the Poisson's ratio of the low-density hydrates.

6.4.3.4 Cement paste homogenization model

The cement paste homogenization model calls complex equations from the micro-mechanics. The latter link the mechanical properties of the cement paste E_{p} and ν_{p} to the mechanical properties of its constituents, *i.e.* the anhydrous E_{a} and ν_{a}, the high-density hydrates E_{hd} and ν_{hd}, the low-density hydrates E_{ld} and ν_{ld} and their respective volume fractions f_{a} and f_{hd}.

The input parameters of the cement paste homogenization model are:

- E_{a}, the Young's modulus of the anhydrous,
- ν_{a}, the Poisson's ratio of the anhydrous,
- E_{hd}, the Young's modulus of the high-density hydrates,
- ν_{hd}, the Poisson's ratio of the high-density hydrates,
- E_{ld}, the Young's modulus of the low-density hydrates,
- ν_{ld}, the Poisson's ratio of the low-density hydrates,
- f_{a}, the volume fraction of anhydrous in the cement paste,
- f_{hd}, the volume fraction of hydrates in the cement paste.

whereas its output parameters are:

- E_{p}, the Young's modulus of the cement paste,
- ν_{p}, the Poisson's ratio of the cement paste.

The cement paste homogenization model follows a Mori-Tanaka (MT) scheme (Mori and Tanaka, 1973). More precisely, spherical composite inclusions are considered with an intermediate density between the kernel of the inclusions and the matrix.

6.4.3.5 Mortar homogenization model

The mortar homogenization model corresponds to the stage where sand is added to the cement paste. The mechanical properties of the mortar E_{m} and ν_{m} are derived using the micro-mechanics equations from the mechanical properties of the sand E_{s} and ν_{s}, the mechanical properties of the cement paste E_{p} and ν_{p} and the volume fraction of sand f_{s} in the mortar.

The input parameters of the mortar homogenization model are:

- E_{s}, the Young's modulus of the sand,
- ν_{s}, the Poisson's ratio of the sand,
- E_{p}, the Young's modulus of the cement paste,
- ν_{p}, the Poisson's ratio of the cement paste,
- f_{s}, the volume fraction of sand in the mortar.

whereas its output parameters are:

- E_{m}, the Young's modulus of the mortar,
- ν_{m}, the Poisson's ratio of the mortar.

The mortar homogenization model follows a Mori-Tanaka (MT) scheme. More precisely, spherical inclusions in the matrix are considered.

6.4.3.6 Concrete homogenization model

The concrete homogenization model corresponds to the stage where aggregates are added to the mortar. The mechanical properties of the concrete E_{c} and ν_{c} are derived using the micro-mechanics equations from the mechanical properties of the aggregates E_{agg} and ν_{a}, the mechanical properties of the mortar E_{m} and ν_{m} and the volume fraction of aggregates f_{agg} in the concrete.

The input parameters of the concrete homogenization model are:

- E_{agg}, the Young's modulus of the aggregates,
- ν_{agg}, the Poisson's ratio of the aggregates,
- E_{m}, the Young's modulus of the mortar,
- ν_{m}, the Poisson's ratio of the mortar,
- f_{agg}, the volume fraction of aggregates in the concrete.

whereas its output parameters are:

- E_{c}, the Young's modulus of the concrete,
- ν_{c}, the Poisson's ratio of the concrete.

The concrete homogenization model follows a Mori-Tanaka (MT) scheme. More precisely, spherical inclusions in the matrix are considered.

6.4.4 Multiscale modelling of the homogenization

The computation of the intermediate and final output variables is carried out using a coupling between OpenTURNS and YACS. The YACS scheme from which the executable XML file is obtained is presented in Figure 6.13. The level 0 (lowest scale of modelling) contains all the input parameters of the model, *i.e.* mechanical properties of the constituents of the cement paste, volume fractions or porosities. At the level 1 are the models of hydration and dosing. They provide the intermediate variables for the low-density homogenization model at the level 2. The latter computes the mechanical properties of the low-density hydrates which are input parameters for the cement paste homogenization model at level 3. At levels 4 and 5 are respectively the mortar homogenization model and the concrete homogenization model.

One may have noticed that for a level n, the model not only uses output variables from the level $n-1$ but also input variables from the level 0. For instance, the mortar homogenization level at the level 4 uses as input variables the output variables from level 3 (E_{p} and ν_{p}) but also variables from level 0 (E_{s} and ν_{s}) and level 1 (f_{s}).

As YACS only works with deterministic values, one has to build the probabilistic model using OpenTURNS and couple both softwares.

6.4.5 Probabilistic modelling of the parameters

6.4.5.1 Marginal distribution of the variables

The mechanical properties of the constituents of concrete are described by two parameters, namely the Young's modulus and the Poisson's ratio. The Young's modules are modelled by lognormal distributions while the Poisson coefficients are modelled by Beta distributions over the domain $[0, 0.5]$. The other ratios denoted by the letter f are modelled by Beta distributions over the domain $[0, 1]$. The complete probabilistic modelling of the parameters is given in Table 6.10. The degree of hydration at which the mechanical properties are calculated is fixed at $\alpha = 0.5$, *i.e.* half the water has been consumed by the reaction. The ratio water / cement powder is assumed to be perfectly known and is fixed at 0.3, *i.e.* 30 liters of water for 100 kg of cement paste.

Parameter	Distribution	Mean μ	σ/μ
E_{a}	Lognormal	135 GPa	10 %
ν_{a}	Beta [0,0.5]	0.3	10 %
E_{hd}	Lognormal	31 GPa	10 %
ν_{hd}	Beta [0,0.5]	0.24	10 %
E_{sh}	Lognormal	72 GPa	10 %
ν_{sh}	Beta [0,0.5]	0.27	10 %
E_{agg}	Lognormal	70 GPa	10 %
ν_{agg}	Beta [0,0.5]	0.23	10 %
E_{s}	Lognormal	70 GPa	10 %
ν_{s}	Beta [0,0.5]	0.23	10 %
α	Deterministic	0.5	-
w/c	Deterministic	0.3	-
rs_{ld}	Beta[0,1]	0.033	10 %
ϕ_{hd}	Lognormal	0.3	15 %
$f_{\mathrm{sand/agg}}$	Beta[0,1]	0.7	7,5 %
r_{sand}	Beta[0,1]	0.4	5 %

Table 6.10: *Marginal distributions of the parameters.*

6.4.5.2 Dependence structure of the variables

The couples formed by the Young's modulus and the Poisson's ratio (E, ν) are usually assumed independent. Their joint PDF is simply defined by the product of the marginal PDFs, namely:

$$f_{E,\nu}(E,\nu) = f_E(E) \times f_\nu(\nu) \tag{6.37}$$

In theory, the bulk modulus K and the shear modulus G could also be considered as independent. They can be related to E and ν by the following relationships:

$$\begin{aligned} E &= \frac{9KG}{3K+G} \\ \nu &= \frac{3K-2G}{2(3K+G)} \end{aligned} \tag{6.38}$$

According to Eq. (6.38), if K and G are assumed independent, then$E$ and ν are correlated because they share input variables. Because the distributions of K and G are often unknown, their densities are estimated by kernel smoothing using the distributions of E and ν given in Table 6.10 and the relations in Eq. (6.39).

$$\begin{aligned} K &= \frac{E}{3(1-2\nu)} \\ G &= \frac{E}{2(1+\nu)} \end{aligned} \tag{6.39}$$

Once the marginal distributions of K and G are known, realizations of E and ν can be sampled using Eq. (6.38). The Spearman's rank correlation of the pairs (E, ν) in the constituents are finally calculated at the various levels. They are presented in Table 6.11.

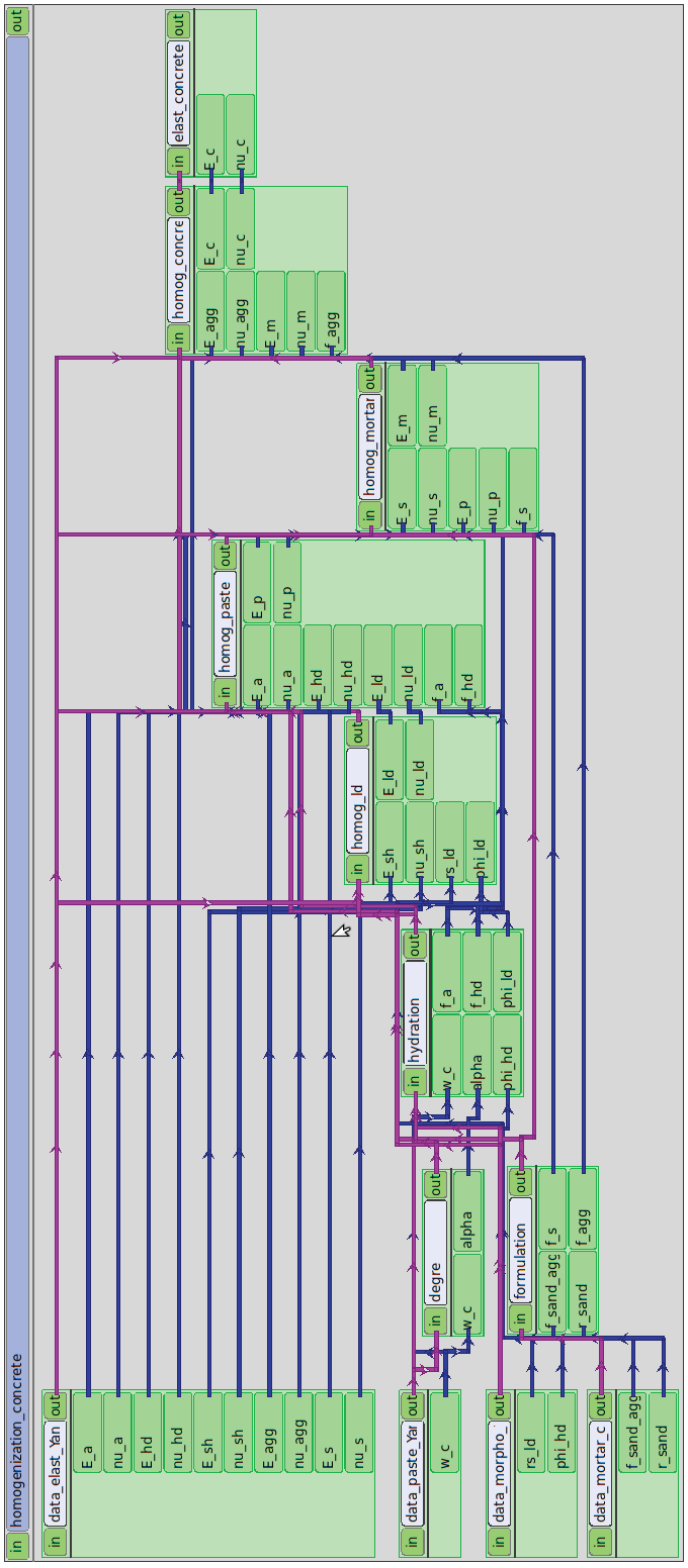


Figure 6.13: *Multiscale modelling for homogenization of concrete featuring 14 input parameters and 13 output parameters.*

Constituent	Rank correlation coefficient ρ_S
anhydrous	-0.57
high-density hydrates	-0.48
low-density hydrates	-0.56
aggregates	-0.54
sand	-0.54

Table 6.11: *Spearman's rank correlation coefficients of the pairs (E, ν) when assuming that the bulk modulus K and the shear modulus G are statistically independent.*

The average rank correlation coefficient ρ_S between E and ν equals -0.55. The dependogram of the mechanical properties of the sand is pictured in Figure 6.14. Due to the shape of the scatterplot, a Gaussian copula is tested for modelling of the 2-dimensional dependence structure. Its parameter θ is derived from the rank correlation coefficient ρ_S using the following relationship:

$$\theta = 2 \sin\left(\frac{\pi}{6}\rho_S\right) \tag{6.40}$$

The corresponding Kendall Plot is pictured in Figure 6.14. The copula of the full set of parameters is defined by a global copula, as presented in Chapter 1, Section 1.5.3.6, composed of five $2-$dimensional Gaussian copulas and a $4-$dimensional independent copula, namely:

$$\begin{bmatrix} \mathbf{S}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{2\times2} & \mathbf{S}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{S}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{S}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{0}_{2\times2} & \mathbf{S}_{2\times2} & \mathbf{0}_{2\times4} \\ \mathbf{0}_{4\times2} & \mathbf{0}_{4\times2} & \mathbf{0}_{4\times2} & \mathbf{0}_{4\times2} & \mathbf{0}_{4\times2} & \mathbf{I}_{4\times4} \end{bmatrix} \tag{6.41}$$

6.4.6 Multiscale sensitivity analysis

Due to the dependence structure of the input variables, classical ANOVA cannot be carried out. In this work, a sensitivity analysis using the δ indices is carried out. The goal is to identify which parameters are the main contributors to the variability of the mechanical properties of concrete E_c and ν_c in order to reduce their variabilities. Therefore, of interest are the sensitivities of E_c and ν_c to the mechanical properties of mortar and cement paste and more generally to any input parameter in the multiscale modelling. The distributions of the intermediate and final output parameters are estimated by kernel smoothing from samples of size $N_{ks} = 10^4$. Their PDFs and moments are respectively reported in Figure 6.15 and Table 6.12.

The sensitivity analysis is decomposed in two subanalyses:

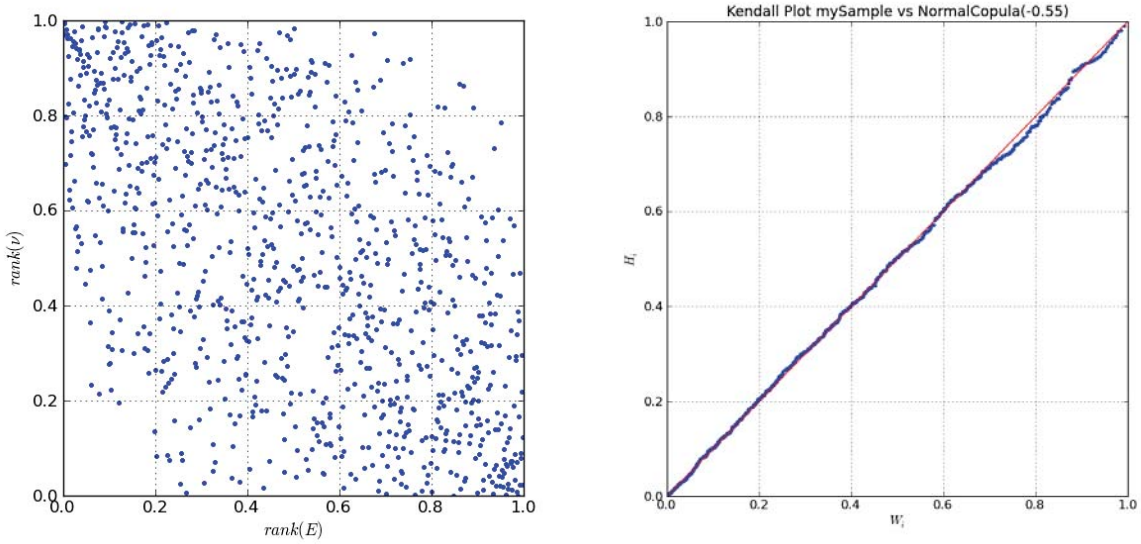


Figure 6.14: *Dependogram of the couple (E_s, ν_s) (left) and Kendall Plot of the sample versus a Gaussian copula parametrized by deriving $\rho_S = -0.55$.*

- the couples (E, ν) are assumed independent and the sensitivity of the mechanical properties of the cement paste, the mortar and the concrete to the parameters of level 0 is studied;

- the couples (K, G) are assumed independent, *i.e.* (E, ν) are assumed correlated and the sensitivity of the mechanical properties of the cement paste, the mortar and the concrete to the parameters of level $n - 1$ is studied.

Variable	Mean μ	Standard deviation σ	σ/μ
E_{p}	20.73	3.56	17 %
ν_{p}	0.23	0.01	3 %
E_{m}	35.44	4.73	13 %
ν_{m}	0.25	0.01	4 %
E_{c}	46.51	4.78	10 %
ν_{c}	0.27	0.01	4 %

Table 6.12: *Mean, standard deviation and coefficient of variation of the ciment paste, mortar and concrete.*

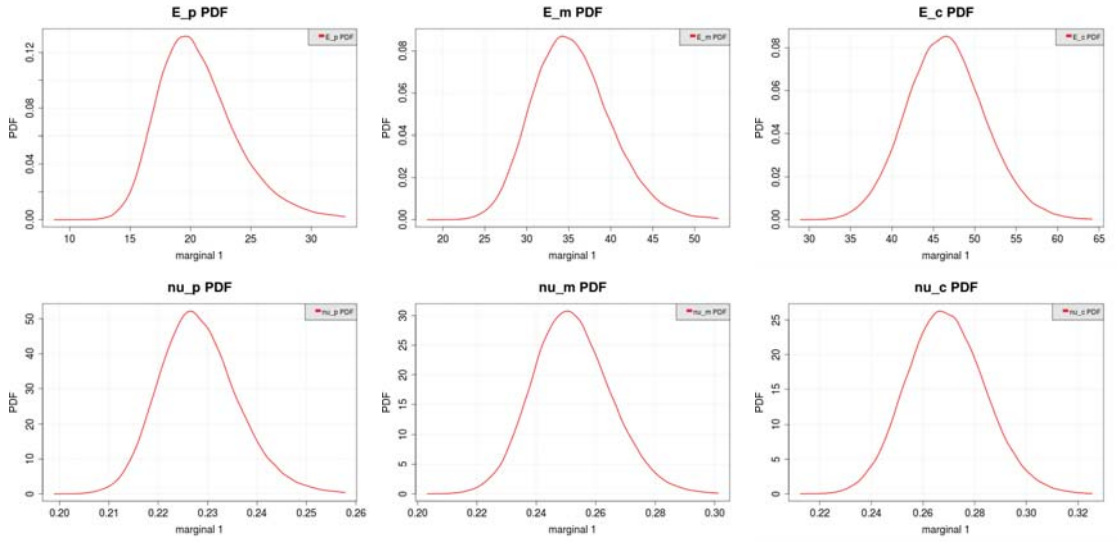


Figure 6.15: *Kernel smoothing estimation of the PDFs of the mechanical properties (E, ν) of the model output parameters (cement paste, mortar and concrete).*

6.4.6.1 Sensitivity to the level 0 parameters

In this section, the sensitivity of the mechanical properties of the cement paste, the mortar and the concrete to the parameters of level 0 is studied. The couples (E, ν) are assumed independent, thus a ANOVA is carried out. In addition, the δ sensitivity measures are also computed.

The results of the sensitivity analysis of the cement paste mechanical properties E_{p} and ν_{p} to the parameters of level 0 are presented in Table 6.13. The results of the sensitivity analysis of the mortar mechanical properties E_{m} and ν_{m} to the parameters of level 0 are presented in Table 6.14. The results of the sensitivity analysis of the concrete mechanical properties E_{c} and ν_{c} to the parameters of level 0 are presented in Table 6.15.

Parameter	S_1	S_T	δ
E_a	0.00	0.00	0.00
ν_a	0.00	0.00	0.00
E_{hd}	0.03	0.03	0.07
ν_{hd}	0.00	0.00	0.00
E_{sh}	0.15	0.15	0.17
ν_{sh}	0.00	0.00	0.00
rs_{ld}	0.02	0.02	0.00
ϕ_{hd}	0.80	0.81	0.40

Parameter	S_1	S_T	δ
E_a	0.07	0.08	0.00
ν_a	0.07	0.07	0.10
E_{hd}	0.24	0.25	0.00
ν_{hd}	0.30	0.32	0.28
E_{sh}	0.01	0.01	0.00
ν_{sh}	0.08	0.08	0.11
rs_{ld}	0.01	0.01	0.00
ϕ_{hd}	0.19	0.20	0.11

Table 6.13: *Sensitivity indices of E_p (left) and ν_p (right) to the level 0 parameters.*

Parameter	S_1	S_T	δ
E_a	0.00	0.00	0.00
ν_a	0.00	0.00	0.00
E_{hd}	0.02	0.02	0.00
ν_{hd}	0.00	0.00	0.00
E_{sh}	0.09	0.09	0.06
ν_{sh}	0.00	0.00	0.00
E_s	0.08	0.09	0.06
ν_s	0.00	0.00	0.00
rs_{ld}	0.00	0.00	0.00
ϕ_{hd}	0.47	0.47	0.21
$f_{sand/agg}$	0.31	0.31	0.17
r_{sand}	0.01	0.01	0.00

Parameter	S_1	S_T	δ
E_a	0.01	0.01	0.00
ν_a	0.01	0.01	0.00
E_{hd}	0.01	0.02	0.00
ν_{hd}	0.03	0.03	0.06
E_{sh}	0.00	0.00	0.00
ν_{sh}	0.01	0.01	0.00
E_s	0.01	0.02	0.00
ν_s	0.71	0.73	0.31
rs_{ld}	0.00	0.00	0.00
ϕ_{hd}	0.09	0.10	0.08
$f_{sand/agg}$	0.08	0.10	0.08
r_{sand}	0.00	0.00	0.00

Table 6.14: *Sensitivity indices of E_m (left) and ν_m (right) to the level 0 parameters.*

The first important result is that the variability of a Young's modulus (resp. a Poisson coefficient) mainly depends on the Young's modulus (resp. the Poisson coefficient) of the constituents at the previous scale. The second one is that the deeper (in the sense of the graph theory) in the modelling a parameter, the lower its contribution to the variability of the output of interest.

More generally, the level 0 main contributors to the variability of the mechanical properties of the cement paste, mortar and concrete are the porosity of the high-density hydrates ϕ_{hd} and the volume ratio $f_{\text{sand/agg}}$ between the sand and the sand and gravel. The porosity ϕ_{hd} is a hardly measurable parameter. Therefore, a high coefficient of variation ($\sigma/\mu = 15\%$) is set. The ratio $f_{\text{sand/agg}}$ behaves as a reinforcement of the cement paste and consequently determines the strength of mortar and concrete. In order to reduce the variability of the mechanical properties of concrete, one must seriously pay attention to the proportion of sand and aggregates in the mixing process. A better knowledge of the porosity ϕ_{hd} may also provide a more accurate modelling of the homogenization of concrete.

6.4.6.2 Sensitivity to the level $n-1$ parameters

The mechanical properties (E, ν) of the constituents are now assumed correlated, *i.e.* K and G are assumed independent. Of interest are the sensitivity of the mechanical properties of the cement paste, mortar and concrete to the lower level parameters, *i.e.* the sensitivity of the mortar mechanical properties to the cement paste and sand mechanical properties and their relative volume fraction. As the hypothesis of independence of the model input parameters is not verified anymore, the Sobol' indices are not applicable and only the δ sensitivity measures are computed.

Parameter	S_1	S_T	δ
E_a	0.00	0.00	0.00
ν_a	0.00	0.00	0.00
E_{hd}	0.01	0.02	0.00
ν_{hd}	0.00	0.00	0.00
E_{sh}	0.06	0.06	0.06
ν_{sh}	0.00	0.00	0.00
E_{agg}	0.14	0.15	0.09
ν_{agg}	0.00	0.00	0.00
E_s	0.05	0.05	0.05
ν_s	0.00	0.00	0.00
rs_{ld}	0.01	0.01	0.00
ϕ_{hd}	0.30	0.30	0.15
$f_{sand/agg}$	0.42	0.43	0.23
r_{sand}	0.00	0.00	0.00

Parameter	S_1	S_T	δ
E_a	0.00	0.00	0.00
ν_a	0.00	0.00	0.00
E_{hd}	0.00	0.00	0.00
ν_{hd}	0.01	0.01	0.00
E_{sh}	0.00	0.00	0.00
ν_{sh}	0.00	0.00	0.00
E_{agg}	0.29	0.30	0.00
ν_{agg}	0.37	0.37	0.31
E_s	0.00	0.00	0.00
ν_s	0.18	0.18	0.14
rs_{ld}	0.00	0.00	0.00
ϕ_{hd}	0.04	0.05	0.05
$f_{sand/agg}$	0.08	0.09	0.07
r_{sand}	0.00	0.00	0.00

Table 6.15: *Sensitivity indices of E_c (left) and ν_c (right) to the level 0 parameters.*

The results of the sensitivity analysis of the cement paste mechanical properties E_{p} and ν_{p} to the lower level parameters are presented in Table 6.16. The results of the sensitivity analysis of the mortar mechanical properties E_{m} and ν_{m} to the lower level parameters are presented in Table 6.17. The results of the sensitivity analysis of the concrete mechanical properties E_{c} and ν_{c} to the lower level parameters are presented in Table 6.18.

Parameter	δ_{E_p}	δ_{ν_p}
E_a	0.06	0.04
ν_a	0.06	0.11
E_{hd}	0.07	0.04
ν_{hd}	0.06	0.03
f_a	0.06	0.04
f_{hd}	0.13	0.04
E_{ld}	0.55	0.10
ν_{ld}	0.07	0.16

Table 6.16: *Sensitivity indices of E_p and ν_p to the $(n-1)$ level parameters.*

The sensitivities of the mechanical properties of the cement paste, the mortar and the concrete to the correlated parameters at the $(n-1)$ level have been measured by the δ sensitivity measures. At the scale of the cement paste, the mechanical properties E_{p} and ν_{p} are mainly sensitive to the properties of the low-density hydrates E_{ld} and ν_{ld}. At the

Parameter	δ_{E_m}	δ_{ν_m}
E_s	0.10	0.05
ν_s	0.08	0.04
f_s	0.24	0.13
E_p	0.37	0.07
ν_p	0.08	0.10

Table 6.17: *Sensitivity indices of $E_m$and ν_m to the $(n-1)$ level parameters.*

Parameter	δ_{E_c}	δ_{ν_c}
E_{agg}	0.11	0.04
ν_{agg}	0.04	0.37
f_{agg}	0.11	0.05
E_{m}	0.49	0.04
ν_{m}	0.04	0.22

Table 6.18: *Sensitivity indices of E_c and ν_c to the $(n-1)$ level parameters.*

upper scale, the mechanical properties of the mortar E_{m} and ν_{m} are more sensitive to the mechanical properties of the cement paste E_{p} and ν_{p} than to those of the sand. Finally, at the highest scale of modelling, the mechanical properties of the concrete E_{c} and ν_{c} are mainly sensitive to the properties of the mortar E_{m} and ν_{m}.

From a global point of view, the mechanical properties of the product, mortar or concrete, is more sensitive to the lower scale product, *i.e.* cement paste or mortar, than to the aggregates that are supposed to increase their strength. Finally, as noticed for the previous sensitivity analysis in Section 6.4.6.1, the Young's modules are more sensitive to Young's modules and the Poisson's ratios are more sensitive to Poisson's ratios. The fraction of each constituent also plays an important role in the variability of the mechanical properties but it is almost twice as important for the Young's modules than for the Poisson's ratios. This result has also been observed in the previous sensitivity analysis where the input $f_{\text{sand/agg}}$ exhibits high contributions to the variability of the Young's modules.

6.4.7 Conclusion

The problem of finding where the variability of the mechanical properties of concrete comes from has been addressed by performing two sensitivity analyses. First the method to compute the quantities of interest referred to as homogenization method has been decomposed into several stages corresponding to the different phases of the production of concrete. A specific model with suitable input and output parameters is associated to each

of these phases. They are linked together to form a multiscale model represented by the graph in Figure 6.13. Secondly, computing the intermediate and final output distributions is carried out through a coupling between OpenTURNS (probabilistic modelling) and YACS (nested modelling). A polynomial chaos expansion with vectorial output, that is built from the obtained data, provides a powerful tool to perform the sensitivity analyses.

The first sensitivity analysis aims at estimating the sensitivities of the intermediate and final output parameters to the input parameters of the lowest scale assumed as independent. The classical variance-based Sobol' sensitivity indices are compared to the distribution-based δ sensitivity measures. As mentioned earlier in this thesis, the δ indices are in accordance with the Sobol' indices concerning the hierarchy of the parameters. Their scale of values remains less discriminant than the one of the Sobol' indices.

The second sensitivity analysis focuses on the sensitivity of the output parameters to the intermediate parameters located at the lower scale of the modelling. Because several models are sharing input parameters, the intermediate parameters are correlated. Therefore, only the δ indices are computed. On the computing side, the lower scale input parameters are identified using the incidence matrix (see Chapter 5, Section 5.3.2.2) of the graph. The results allows the engineer to determine on which constituent or fraction of constituent he may pay more attention in order to limit the variability of the mechanical properties of concrete at the macroscopic scale.

6.5 Damage of a cylinder head

Renault is a French motorist and car manufacturer. Since 75% of the cars are equipped with diesel engines, Renault aims at ensuring the reliability of its products. The M9R is the flagship of the diesel engine range. This engine features a displacement of 2 liters and develops up to 180 horsepower thanks to common rail injection and turbocompression. The Renault M9R diesel engine is presented in Figure 6.16. This example has been proposed in Hähnel (2007) to illustrate system reliability analysis. The results have been originally presented in Caniou and Hähnel (2012).

6.5.1 How do diesel engines work

The diesel engine was invented by Rudolph Diesel in 1858. The internal combustion engine uses the heat of compression to initiate ignition and burn the fuel injected in the combustion chamber. Diesel engines has the highest thermal efficiency due to the very high compression ratio. A diesel engine is constituted of *pistons* sliding inside *cylinders*, closed by a *cylinder head* connecting the cylinders to the *intake* and *exhaust manifolds* by *poppet valves* driven by a *camshaft*.

Air is initially introduced in the combustion chamber and is then compressed with a very high compression ratio (between 14:1 and 25:1) due the injection pressure. The compression heats the air up from 700 to 900°C. When the piston reaches its highest point, fuel is injected into the compressed air in the combustion chamber. The heat of the

compressed air vaporizes the fuel and ignite it. The expansion of the combustion gases pushes the piston downward which induces the rotation of the crankshaft through the rod.

Figure 6.16: *Overview of the M9R Renault diesel engine.*

The four strokes of a diesel engine are:

1. *intake*: the piston descends along the cylinder and air is aspirated inside the cylinder due to the reduction of the internal pressure. The intake valves then close.

2. *compression*: with both exhaust and intake valves closed, the piston rises up to the top of the cylinder due to the rotation of the camshaft. The air is compressed inside the combustion chamber against the cylinder head.

3. *power*: fuel is injected into the highly compressed air and is instantly ignited due to the very high temperature. The expansion of the combustion gases pushes the piston downward.

4. *exhaust*: the piston once again rises up to the top of the cylinder, expelling the brunt gases through the opened exhaust valves.

Common rail engines use a unique very high pressure (up to 2000 bars) line for the injection of fuel into the cylinders whereas standard diesel engines had one for each cylinder.

This technology improves the fuel repartition in the cylinder and consequently increases the efficiency of the engine. On the other hand, the couple pressure and temperature (P, T) in the cylinder is much higher than in the standard diesel engine and causes higher mechanical and thermal damage to the cylinder head.

In the sequel, the damage of the *fire face* (the area above the cylinder) and *bridge* (the zone between two valve holes) of the cylinder head is studied.

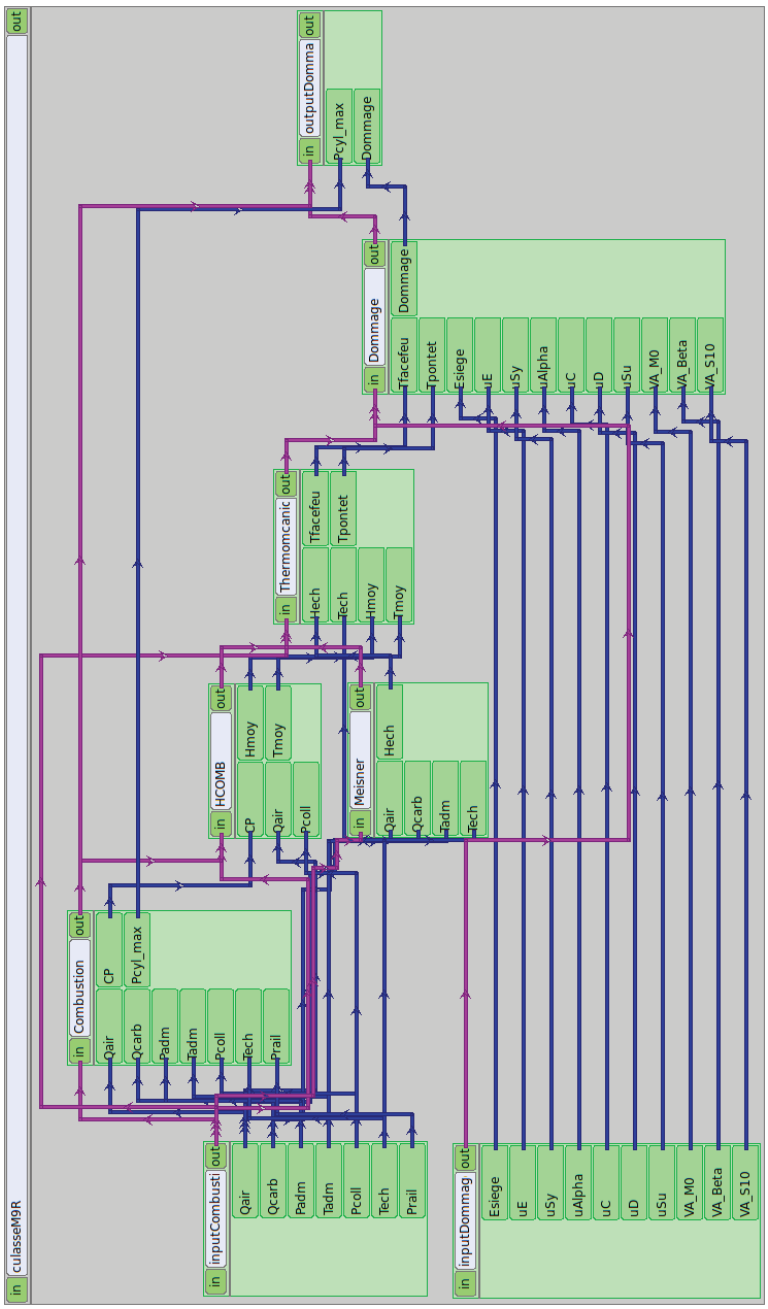


Figure 6.17: *Graph representation of the combustion in a diesel engine from a multiphysics point of view.*

6.5.2 Multiphysics modelling

The modelling of the fuel combustion in a diesel engine is a quite complicated phenomenon involving multiple physics fields. Therefore, it is decomposed into several submodels corresponding to the computation of physical quantities. A first model concerns the combustion of the fuel. It provides input parameters for a second model dealing with the thermal phenomena in the cylinder head. Then, a third model computes the variables of interest thanks to a thermomechanical model. The multiphysics modelling of the engine is described by the graph in Figure 6.17.

The first stage is the mechanical *combustion model*. It calculates the maximal pressure P_{cylmax} in the cylinder from the flows of air and fuel together, respectively denoted by Q_{air} and Q_{fuel} with the pressure P_{adm} and temperature T_{adm} in the intake, the pressures in the manifold P_{coll} and in the common rail P_{rail} and the temperature of the exhaust gases T_{ech}.

The second stage is divided into two thermal models HCOMB and Meisner. The first one computes the mean exchange coefficient H_{moy} and temperature T_{moy} in the cylinder with the pressure curve CP, the flow of air Q_{air} and the intake pressure P_{adm} whereas the second one computes the exchange coefficient H_{ech} in the exhaust manifold with the flows of air Q_{air} and fuel Q_{fuel} and the temperatures of the intake (T_{adm}) and exhaust (T_{ech}) manifolds.

The thermomechanical model at the third stage computes the temperatures at the two weak points on the cylinder head, namely the fire face T_{face} and the bridge T_{bridge} (see Figure 6.18) from the exhaust exchange coefficient H_{ech} and temperature T_{ech} and mean exchange coefficient H_{moy} and temperature T_{moy}.

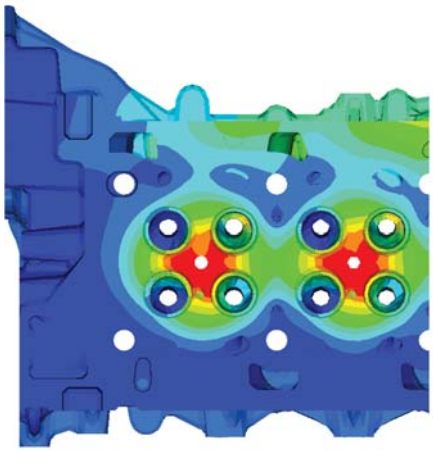


Figure 6.18: *Thermomechanical stresses on the cylinder head. The fire face is located in the middle of the four valves whereas a bridge separates two valves (picture supplied by Renault).*

Finally, the damage model computes the damage D in the cylinder head from the

temperatures T_{ech} and T_{moy} computed by the thermomechanical model, the mechanical properties such the Young's modulus E, the plastic stress S_y or the ultimate tensile stress S_u and a collection of fatigue parameters ($M0$, $Beta$, $S10$).

6.5.3 Sensitivity of the damage on the cylinder head

Of interest are the sensitivities of the damage D and maximal pressure in the cylinder P_{cylmax} to the intermediate parameters in the multiphysics modelling.

6.5.3.1 Computing the responses of the model

A nested model of the fuel combustion is built using YACS. Each model is called from a Pyscript in a node. The global model is driven from OpenTURNS where the realizations of the input parameters are simulated according to the probabilistic modelling described in Table 6.19.

Parameter	Unity	Distribution	Mean μ	σ
Q_{air}	-	Normal	μ_1	σ_1
P_{coll}	bar	Normal	μ_2	σ_2
Q_{carb}	-	Normal	μ_3	σ_3
P_{adm}	bar	Normal	μ_4	σ_4
T_{adm}	°C	Normal	μ_5	σ_5
T_{ech}	°C	Normal	μ_6	σ_6
P_{rail}	bar	Normal	μ_7	σ_7
E_{siege}	MPa	Normal	μ_8	σ_8
E_{head}	-	Normal	0	1
σ_Y	-	Normal	0	1
α_{head}	-	Normal	0	1
C	-	Normal	0	1
D	-	Normal	0	1
R_m	-	Normal	0	1
VA_{M0}	-	Normal	μ_9	σ_9
VA_{Beta}	-	Normal	μ_{10}	σ_{10}
VA_{S10}	-	Normal	μ_{11}	σ_{11}

Table 6.19: *Probabilistic modelling of the engine parameters.*

The distribution of the parameters are defined either by measures or by experimental feedback. A polynomial chaos expansion of degree $p = 4$ with 17 input and 7 output parameters is built from a set of data samples of size $N = 10^4$ as $P = 5985$ coefficients

have to be computed. The accuracy of the metamodel is controlled by the coefficient of determination R^2 as illustrated in Figs. 6.19 and 6.20.

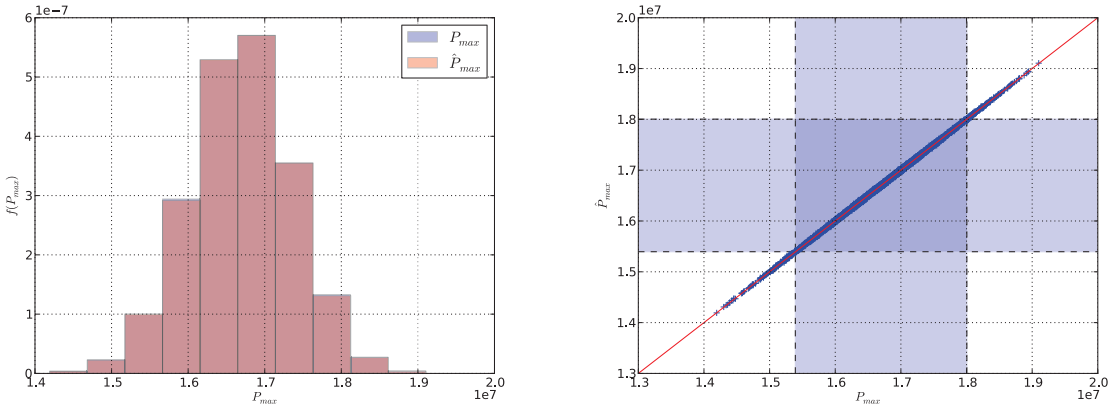


Figure 6.19: *Histogram comparison and coefficient of determination R^2 for the maximum pressure P_{cylmax}.*

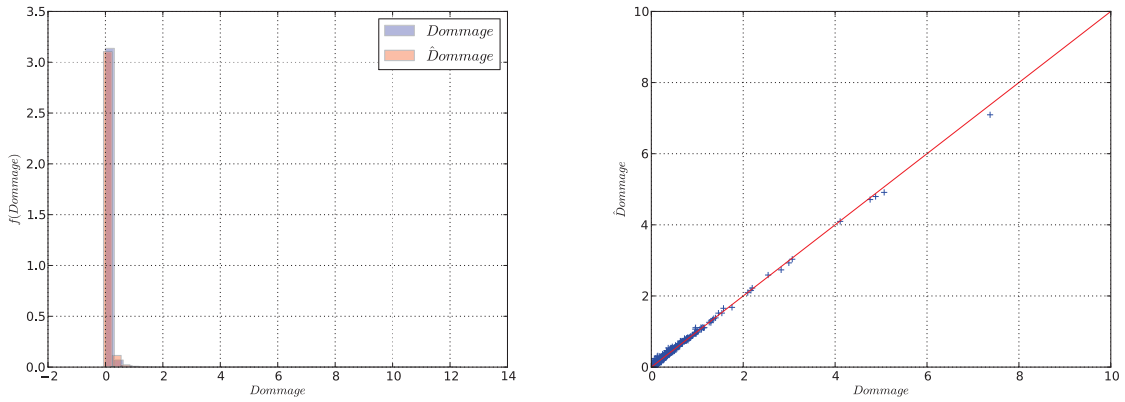


Figure 6.20: *Histogram comparison and coefficient of determination R^2 for the damage D.*

6.5.3.2 Global sensitivity analysis

First the sensitivities of the intermediate and final output parameters, *i.e.* the maximal pressure P_{cylmax} and the damage D, to the input parameters of level 0 are computed. Under the hypothesis of independence of these variables, the Sobol' first order and total indices are computed. The results are presented in Table 6.20.

Parameter	P_{max}	T_{moy}	H_{moy}	H_{ech}
Q_{air}	0.36 (0.36)	0.80 (0.80)	0.73 (0.73)	0.97 (0.97)
P_{coll}	0.00 (0.00)	0.00 (0.00)	0.02 (0.02)	0.00 (0.00)
Q_{carb}	0.12 (0.12)	0.15 (0.15)	0.18 (0.18)	0.02 (0.02)
P_{adm}	0.52 (0.52)	0.05 (0.05)	0.06 (0.06)	0.00 (0.00)
T_{adm}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
T_{ech}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
P_{rail}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
E_{siege}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
E_{head}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
σ_Y	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
α_{head}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
C	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
D	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
R_m	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
VA_{MO}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
VA_{Beta}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
VA_{S10}	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)

Parameter	T_{face}	T_{bridge}	D
Q_{air}	0.26 (0.30)	0.26 (0.30)	0.00 (0.07)
P_{coll}	0.00 (0.00)	0.00 (0.00)	0.00 (0.07)
Q_{carb}	0.44 (0.48)	0.44 (0.48)	0.00 (0.07)
P_{adm}	0.11 (0.12)	0.11 (0.12)	0.00 (0.09)
T_{adm}	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)
T_{ech}	0.14 (0.14)	0.14 (0.14)	0.00 (0.09)
P_{rail}	0.00 (0.00)	0.00 (0.00)	0.00 (0.07)
E_{siege}	0.00 (0.00)	0.00 (0.00)	0.00 (0.05)
E_{head}	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)
σ_Y	0.00 (0.00)	0.00 (0.00)	0.00 (0.07)
α_{head}	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)
C	0.00 (0.00)	0.00 (0.00)	0.00 (0.05)
D	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)
R_m	0.00 (0.00)	0.00 (0.00)	0.02 (0.16)
VA_{MO}	0.00 (0.00)	0.00 (0.00)	0.26 (0.59)
VA_{Beta}	0.00 (0.00)	0.00 (0.00)	0.13 (0.35)
VA_{S10}	0.00 (0.00)	0.00 (0.00)	0.00 (0.06)

Table 6.20: *Global sensitivity analysis of the model outputs to the level 0 parameters S_1 (S_T).*

.

Concerning the output parameters P_{cylmax}, T_{moy}, H_{moy} and H_{ech}, the contributors are clearly identified: they are the flows of air Q_{air} and fuel Q_{carb} and the pressure of the admission gases. In addition comes the temperature of the exhaust gases T_{ech} for the temperatures at the fire face T_{face} and bridge T_{bridge}.

The first order contributors to the damage D are the ultimate tensile stress R_m, and the fatigue parameters VA_{MO} and VA_{Beta} whose distribution are defined by the state of the art. It is to be noticed that the total contributions, taking the interactions between the input variables into account, are non zero for all parameters. It means that complex relationships exist in the multiphysics modelling and consequently that the sensitivity analysis has to investigate deeper.

6.5.3.3 Nested sensitivity analysis

Of interest are now the sensitivities of the output parameters to the input parameters of the lower level. The concerned variables are identified in the graph thanks to the incidence matrix. Let us begin by studying the temperatures T_{face} and T_{bridge}. Due to the nature of the model, these input parameters, namely H_{ech}, T_{ech}, H_{moy} and T_{moy}, are correlated. Their rank correlation matrix reads:

$$\boldsymbol{\rho}_S = \begin{bmatrix} 1 & 0 & -0.54 & -0.81 \\ 0 & 1 & 0 & 0.05 \\ -0.54 & 0 & 1 & 0.92 \\ -0.81 & 0.05 & 0.92 & 1 \end{bmatrix} \tag{6.42}$$

An ANCOVA sensitivity analysis is performed in order to distinguish the uncorrelative and correlative contributions. The results are presented in Table 6.21.

Parameter	S	S^U	S^C
H_{ech}	1.17	1.83	-0.67
T_{ech}	0.09	0.07	0.01
H_{moy}	-0.01	0.01	-0.03
T_{moy}	-0.24	0.29	-0.53
Σ	1.00	2.21	-1.21

Parameter	S	S^U	S^C
H_{ech}	1.17	1.83	-0.67
T_{ech}	0.09	0.07	0.01
H_{moy}	-0.01	0.01	-0.03
T_{moy}	-0.24	0.29	-0.53
Σ	1.00	2.21	-1.21

Table 6.21: *ANCOVA sensitivity indices of T_{face} (left) and T_{bridge} (right) to the lower level parameters. Due to the very close location of the zones these temperatures are measured on, the sensitivities are identical.*

Due to the very close proximity of the zones at which the temperatures are measured, namely the fire face and the bridge, the sensitivities are exactly the same. Based on the uncorrelative contributions, the most influent variables are the exchange coefficient of the exhaust gases H_{ech} and the mean temperature in the cylinder T_{moy}. By adding the correlative contributions, the global contributions are lowered. The high and possibly negative correlations in Eq. (6.42) tends to reduce the inner effects of the main variables.

A second ANCOVA sensitivity analysis is performed for the damage D in the cylinder head. The results are presented in Table 6.22. In this case, the first order ANCOVA indices

Parameter	S	S^U	S^U
T_{face}	0.02	0.01	0.01
T_{bridge}	0.02	0.01	0.01
E_{siege}	0.00	0.00	0.00
u_E	0.00	0.00	0.00
u_{S_y}	0.00	0.00	0.00
u_α	0.00	0.00	0.00
u_C	0.00	0.00	0.00
u_D	0.00	0.00	0.00
u_{S_u}	0.02	0.02	0.00
VA_{MO}	0.20	0.20	0.00
VA_{Beta}	0.12	0.12	0.00
VA_{S10}	0.00	0.00	0.00
Σ	0.38	0.36	0.02

Table 6.22: *ANCOVA sensitivity indices of the damage D in the cylinder head to the lower level parameters.*

cannot explain the full variability of the damage D in the cylinder head. It means that 62% of this variability is due to the interactions between the input parameters as observed in the global sensitivity analysis. Otherwise, the same contributors as for the global sensitivity analysis are identified, namely the fatigue parameters VA_{MO} and VA_{Beta}.

6.5.4 Conclusion

The combustion of the fuel in an engine is described by a multiphysics modelling. The workflow composed of mechanical, thermal and thermomechanical models allows the engineer to study the reliability of the engine by measuring the damage in the cylinder head. The simulations on the model are performed by carrying out a coupling between OpenTURNS and YACS respectively for the probabilistic ans physical modelling. A polynomial chaos expansion is built from the data to perform ANCOVA sensitivity analyses.

Firstly the temperatures on the fire face and bridge of the cylinder head are studied. Their sensitivities to the lower level parameters are computed and the exchange coefficient of the exhaust gases H_{ech} is identified as the main contributor to the variability of these temperatures. Secondly, the damage D of the cylinder head is studied. The most influent variables to the variability of D are the fatigue parameters VA_{MO} and VA_{Beta} which have high coefficient of variation, respectively 20% and 10%.

However, in the last case, the variability of the damage D cannot be entirely explained by the first order effects. This means that the interactions between the variables is relatively strong and thus complex to explain. Total order ANCOVA indices should be computed but the interpretation that can be done might mix up the interactive and correlative contributions.

6.6 Conclusion

A general methodology to address the problems of global sensitivity analysis in nested models has been proposed in Chapters 2, 3, 4 and 5. The approach aims at computing the sensitivity of one output parameter to any input parameters, whatever their position in a complex workflow.

The global sensitivity analysis methods for models with correlated inputs are first trained on analytical functions (Section 6.2) and academical problems (Section 6.3), namely a bracket structure, a composite beam, a electrical connector or a rugby scrum. As shown by the diversity of these examples, the methodology has a wide spectrum of applications.

The first industrial problem that is addressed deals with a model which predicts the mechanical properties of concrete through homogenization methods. A multiscale modelling where each submodel corresponds to a specific phase of the fabrication is proposed. The dependence structure of the input variables and more specifically the couples (E, ν) of the low level constituents is introduced in order to refine the probabilistic modelling. Then, thanks to the graph approach developed in Chapter 5, one is able to identify the input parameters of a quantity of interest throughout the different scales of modelling and compute the corresponding δ sensitivity indices. The results indicate the engineer on which parameters (volume fraction, mechanical property) and at which stages of the fabrication of concrete he must pay attention in order to reduce the variability of the mechanical properties of concrete so as to ensure the robustness of the structure.

The second industrial problem deals with the reliability of a diesel engine. More specifically, the damage in the cylinder head is studied. Due to the complexity of the different thermomechanical phenomena, a multiphysics approach is proposed. As in the first industrial application, an OpenTURNS / YACS coupling is carried out in order to propagate the uncertainty in the input parameters towards the intermediate and final output parameters. The objective is to distinguish if the contribution are due to the structural effects of the variables or to their correlation with other inputs. Therefore, an ANCOVA sensitivity analysis is performed. The latter utilizes the functional decomposition provided by a polynomial chaos expansion to compute the uncorrelative and correlative contributions of the variables on the damage D. The results show that only 40% of the variability of the damage is due to the inner effects and that the correlative contributions are almost zero. The variability may be explained by the complexity of the interactions between the parameters in models that represent different fields of physics.

As a conclusion, the uncertainty in the output parameters of complex structures can

be explained by identifying the most influent input parameters, whether they are independent or not. Two methods are proposed. On the one hand, the distribution-based δ sensitivity measure offers a global information on the contribution but they cannot be interpreted in terms of shares of variance. On the other hand, the indices based on the decomposition of the covariance of the model output allow one to distinguish the uncorrelative (structure) and correlative (interaction) contributions. This triplet of indices require a functional decomposition that can be provided by a surrogate modelling method named polynomial chaos expansion. However, depending on the nature of the models (additive, multiplicative), the entire variability may not be explained by first order indices. Total order indices have been proposed in Chapter 4 but the distinction between interaction and correlation remains to be discussed.

Conclusion

Summary and main contributions

This work was intended to develop a global methodology to address sensitivity analysis problems in nested models. The challenge was to unify different mathematical theories in order to propose a rigorous framework with a potential compatibility with a forthcoming automation of the process and high performance computing. The two major industrial applications that have been treated in Chapter 6, Sections 6.4 and 6.5 exhibit on the one hand the capabilities of such a method but on the other hand its limitations and possible improvements.

Surrogate modelling

Several surrogate modelling techniques, namely the Support Vector regression, the Gaussian processes, the high-dimensional model representation and the polynomial chaos expansions, have been presented in Chapter 3. Among these four, only the last one has been retained for its suitability for sensitivity analysis. No major improvements of this technique have been proposed. The key results presented here are taken from the recent works by Sudret (2007) and Blatman (2009). Nonetheless, the functional decomposition provided by the PC expansions built from an independent joint distribution has been exposed with the purpose of computing the ANCOVA sensitivity indices.

Computing sensitivity indices

Over the last ten years, the problem of defining sensitivity indices for models with correlated inputs has been addressed with different approaches that have been reviewed in Chapter 2. Although no new quantities have been defined in this thesis, practical approaches to compute the so-called δ importance measure and ANCOVA indices respectively introduced in Borgonovo (2007) and Li and Rabitz (2010) have been developed with the sake of computational efficiency.

The distribution-based importance measure δ can be computed in two ways, whether the probability distribution function or the cumulative distribution function of the re-

sponse is considered. In both cases, the distributions are estimated using a kernel smoothing technique and a Gaussian quadrature rule is preferred to Monte Carlo simulations for the integration.

The so-called ANCOVA decomposition is presented as a generalization of the ANOVA for correlated input parameters. Both methods require the computation of many statistical moments, *e.g.* expected value, variance, covariance, but the ANCOVA adds a functional decomposition of the model in its prerequisites. Fortunately (this was not originally the reason why this technique has been retained), one is provided by the PC expansions. However, the PC expansions must be built with a joint distribution featuring an independent copula in order to preserve the orthogonality of component functions with respect to the physical variables. Otherwise the component functions will correspond to the decorrelated variables from the isoprobabilistic transformation. In contrast, the simulations that are performed with the true copula.

Nested modelling of complex structure

Complex systems can be seen as an imbrication of models corresponding to different components, physics or scales of modelling. Yet, no established framework stands out from the literature review. Therefore, an approach based on the *graph theory* is proposed in Chapter 5. Although the modelling that is offered by this method may sound like a simple graphic representation, the linear algebra which is adjoined allows one to map the whole workflow onto an *incidence matrix*.

This matrix appears as a key feature of the global methodology since it defines on the one hand the incidence relationships between the variables from different levels of the modelling and on the other hand the correlation that may exist between the outputs of two models sharing one or more inputs. The incidence matrix can easily be read by an algorithm in order to find all the contributors to an output of interest but so far, it still has to be filled by the analyst.

Contributions to the industry

So far, when dealing with high-dimensional models of complex structure and probabilistic modelling, the engineer often neglect the correlation between the variables for the sake of simplicity or simply due to a lack of adapted methods. As shown in Section 1.4, errors induced by such simplifying assumptions may lead to harmful mistakes in the interpretation of the results. The framework proposed in this thesis provides the practitioner with a suitable methodology for taking the dependence structure into account when designing a complex structure.

The multiscale modelling involves several models that do not necessarily call the same softwares. Therefore, running a complete workflow with its corresponding interactions represents quite a burden for whom is not equipped with a multidisciplinary platform. The YACS module of the simulation software Salome has been developed for this purpose.

Its limitation relies in only considering deterministic numerical values of the parameters. Thus, a coupling with the OpenTURNS Python probabilistic modelling library has been carried out in order to propagate uncertainties in the workflow.

As shown on two major industrial examples, namely, a multiscale approach for computing the mechanical properties of concrete and a multiphysics modelling of the fuel combustion in a Diesel engine, the proposed methodology allows one to identify which parameter or which level of the modelling has the largest influence on the global performance of the structure.

Future work

The work presented in this thesis proposes a global methodology that allows one to compute sensitivity indices when the model has correlated inputs and a framework for nested models at the same time. Its application on academical and industrial problems has highlighted findings, limitations and possible ways of improvements.

(*i*) In Chapter 1, emphasis is put on the joint probabilistic modelling of random variables and more particularly on the corresponding copula functions. There are many families of copulas, *e.g.* elliptical, Archimedean, extreme-value copulas, but in practice, it is clear that in 99% of mechanical applications, a Gaussian copula fits the dependence structure of the data. Nonetheless, it might not be the case in the fields of finance and insurance where extreme values are of interest.

(*ii*) In Chapter 2, the so-called ANCOVA method is presented as a generalization of the well-established Sobol' sensitivity indices when the input parameters of a model are correlated. This technique is also presented from a mathematical point of view in Chastaing et al. (2012). The same concept of generalization is also claimed by Kucherenko et al. (2012) although the techniques provide different results on analytical examples. Both approaches have arguments to oppose to each other. From the experience of the author, the ANCOVA is a more comprehensive way to apprehend the sensitivities since it distinguishes the uncorrelative and correlative parts of the contributions, although the separation between the effects is disputable (see Section 4.4.2.3). On the other side, the negative correlative contributions may repel the practitioner until its signification is assimilated.

(*iii*) In mechanical engineering, and in all scientific fields, more important than the numerical value of a quantity is the confidence that can be attached to it. This is often seen as error bars, confidence intervals or distributions. Sensitivity indices make no exception to the rule. Different approaches for computing analytical confidence intervals on the Sobol' indices are presented in Section 2.3.2.2. In the two major techniques retained in this work, such confidence intervals are not analytically computable and must be estimated using bootstrapping techniques. In addition, a stack of approximations is involved on the one side by the surrogate modelling, and on the other side by the computing scheme of the index (kernel smoothing estimation,

integration using a Gaussian quadrature rule, etc.). The proposed methodology aims at providing an efficient accuracy / computing cost ratio in the sense of the practical implementation. Therefore, the algorithm parameters given throughout this thesis allow one to compute the sensitivity indices with an error less than 10^{-2}, that is at the nearliest percent, which is sufficient for decisions based on sensitivity indices.

(*iv*) In Chapter 5, the linear algebra that is derived from the graph theory is used to build the so-called incidence matrix. The latter maps the relationships between the parameters (contribution, correlation) so that an algorithm is able to find the contributors of an output of interest and build its joint distribution. Nevertheless, the practitioner has to fill the incidence matrix by hand because to date, no disposition has been made for an automatic matrix filling. In the future work, one could imagine that a initial run of the workflow might be able to detect the relationships between the variables and to fill the matrix accordingly. From this point, a graphical representation of the nested model could be drawn for validation at the same time.

(*v*) One objective of the methodology proposed in the introduction was to define sensitivity indices that are suitable to models with correlated input parameters. This objective has been fulfilled by the two methods presented in Chapter 2. The contributions of these methods has been demonstrated with success on analytical test cases. Unfortunately, when dealing with the two proposed industrial applications, taking the correlation into account when computing sensitivity indices does not make a much significant difference when comparing the results with the independent case. The ANCOVA correlative contribution rarely exceed 10% of the total contribution. Indeed the global framework suits any kind of problems but it seems that in some cases the game is not worth the candle. In future research, a *measure of modification* of the model response when the correlation is taken into account or not should determine which sensitivity analysis method is to be carried out.

Bibliography

Abbott, P. (2005). Tricks of the trade: Legendre-gauss quadrature. *Mathematica J.*, 9, 689–691.

Bergé, C. (1958). *Théorie des graphes et ses applications*. Dunod.

Bernard, O., F.-J. Ulm, and E. Lemarchand (2003). A multiscale micromechanics-hydration model for the early-age elastic properties of cement-based materials. *Cement Conc. Res.*, 33(9).

Berveiller, M. (2005). *Eléments finis stochastiques: approches intrusive et non intrusive pour des analyses de fiabilité*. Ph. D. thesis, Université Blaise Pascal - Clermont II.

Blatman, G. (2009). *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. Ph. D. thesis, Université Blaise Pascal - Clermont II.

Blatman, G. and B. Sudret (2009). Anisotropic parcimonious polynomial chaos expansions based on the sparsity-of-effects principle. In *In Proc. ICOSSAR'09, Int Conf. on Structural Safety And Reliability, Osaka, Japan*.

Blatman, G. and B. Sudret (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Prob. Eng. Mech.*, 25(2), 183–197.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliab. Eng. Sys. Safety*, 92, 771–784.

Borgonovo, E., W. Castaings, and S. Tarantola (2011). Moment independent importance measures: New results and analytical test cases. *Risk Analysis*, 31, 404–428.

Bornert, M., T. Bretheau, , and P. Gilormini (2010). *Homogénéisation en mécanique des matériaux 1*. Série : Alliages Métalliques, Hermes Science Publications,.

Caniou, Y., G. Defaux, V. Dubourg, B. Sudret, and G. Petitet (2011). Caractérisation indirecte de défauts géométriques de forme liés au process de fabrication d'un élément d'essuie-glace. In *20ème Congrès Français de Mécanique*.

Caniou, Y. and A. Hähnel (2012). Analyse de sensibilité multi-échelles - application à une culasse de moteur diesel. In *Congrès NAFEMS France 2012*.

Caniou, Y. and B. Sudret (2010). Distribution-based global sensitivity analysis using polynomial chaos expansions with dependent parameters. In *Sixth International Conference on Sensitivity Analysis of Model Output, Milan*.

Caniou, Y. and B. Sudret (2011). Distribution-based global sensitivity analysis in case of correlated input parameters using polynomial chaos expansions. In *11th International Conference on Applications of Statistics and Probability in Civil Engineering*.

Caniou, Y. and B. Sudret (2012). Computational methods for sensitivity indices based on polynomial chaos expansions – case of dependent variables. In preparation.

Caniou, Y., B. Sudret, and A. Micol (2012a). Analyse de sensibilité globale pour des variables corrélées - Application aux modèles imbriqués et multiéchelles. In *Journées de la fiabilité des matériaux et des structures*.

Caniou, Y., B. Sudret, and A. Micol (2012b). Global sensitivity analysis for models with correlated input parameters. In *MASCOT 2012 Meeting*.

Cayley, A. (1879). On the colourings of maps. *Proc. Royal Geographical Society*, 1, 259–261.

Charpentier, A. (2006). *Dependence structures and limiting results, with applications in finance and insurance*. Ph. D. thesis, Katolieke Universiteit Leuven.

Chastaing, G., F. Gamboa, and C. Prieur (2012). Generalized Hœffding-Sobol decomposition for dependent variables - application to sensitivity analysis. Technical Report hal-00649404, GdR MASCOT-NUM (Méthodes d'Analyse Stochastique des Codes et Traitements Numériques).

Chateauneuf, A. and Y. Aoues (2008). *Structural design optimization considering uncertainties*, chapter 9, pp. 217–246. Taylor & Francis.

Choi, S. (1977). Test of equality of dependent correlations. *Biometrika*, 64, 645–647.

Chun, M., S. Han, and N. Tak (2000). An uncertainty importance measure using a distance metric for the change in a cumulative distribution function. *Reliab. Eng. Sys. Safety*, 70, 313–321.

Dubourg, V. (2011). *Adaptive surrogate models for reliability analysis and reliability-based design optimization*. Ph. D. thesis, Université Blaise Pascal - Clermont II.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least Angle Regression. *Annals of Statistics*, 32, 407–499.

Efron, B. and C. Stein (1981). The jackknife estimate of variance. *The Annals of Statistics*, 9, 586–596.

Embrechts, P. (2009). Copulas: A personal view. *Journal of Risk and Insurance*, 76, 639–650.

Embrechts, P., F. Lindskog, and A. McNeil (2001). Modelling dependence with copulas and applications to risk management. Technical report, ETHZ - Department of Mathematics, ETHZ CH-8092 Zürich.

Embrechts, P., A. McNeil, and D. Straumann (1999). Correlation: pittfalls and alternatives. Technical report, Department Mathematik, ETH Zentrum, CH-8092 Zürich.

Embrechts, P., A. McNeil, and D. Straumann (2002). Correlation and dependence in risk management: properties and pittfalls. *Risk Management: Value at Risk and Beyond*, pp. 176–223.

Euler, L. (1741). Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8, 128–140.

Fermanian, J. (2005). Goodness-of-fit tests for copulas. *J. of Multivar. Anal.*, 95, 119–152.

Fieller, E., H. Hartley, and E. Pearson (1957). Tests for rank correlation coefficients. *Biometrika*, 44, 470–481.

Fisher, R. (1915). Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika*, 10, 507–521.

Fisher, R. (1921). On the "probable error" of a coefficient of correlation deduced from a small sample. *Metron*, 1, 3–32.

Fréchet, M. (1951). Sur les tableaux de corrélation dont les marges sont données. *Ann. Univ. Lyon*, Sect. A 9, 53–57.

Gayton, N., P. Beaucaire, J.-M. Bourinet, E. Duc, M. Lemaire, and L. Gauvrit (2011). Apta: advanced probability-based tolerance analysis of products. *Mécanique & Industries*, 12, 71–85.

Genest, C. and J.-C. Boies (2003). Detecting Dependence with Kendall Plots. *Amer. Stat.*, 57, 275–284.

Genest, C. and A.-C. Favre (2007). Everything you always wanted to know about copula modeling but were afraid to ask. *J. Hydro. Eng.*, 12(4), 347–368.

Genest, C. and J. MacKay (1986). The Joy of Copulas: Bivariate Distributions with Uniform Marginals. *Amer. Stat.*, 40(4), 280–283.

Genest, C. and J. Neslehova (2007). A primer on copulas for count data. *Astin Bulletin*, 37(2), 475–515.

Genest, C. and B. Rémillard (2008). Validity of the parametric bootstrap for goodness-of-fit testing in semiparametric models. *Annales de I'Institut Henri Poincaré Probabilités et Statistiques*, 44, 1096–1127.

Genest, C., B. Rémillard, and D. Beaudoin (2007). Goodness-of-fit tests for copulas: A review and a power study. *Insurance: Mathematics & Economics*, 44(2), 199–213.

Ghanem, R. and P. Spanos (1991). *Stochastic Finite Elements : A Spectral Approach*. Springer Verlag. (Reedited by Dover Publications, 2003).

Gonthier, G. (2000). Le théorème des quatre couleurs. Technical report, Ecole polytechnique.

Hähnel, Y. (2007). *Approche mécano-probabiliste système en conception pour la fiabilité*. Ph. D. thesis, Université Blaise Pascal - Clermont II.

Hierholzer, C. (1873). Über die Möglichkeit, einen Linienzug ohne Wiederholung und ohne Unterbrechung zu umfahren. *Mathematische Annalen*, 8, 30–32.

Hill, R. (1965). A self consistent mechanics of composite materials. *Mechanics and Physics of Solids*, 13, 213–222.

Hoeffding, W. (1940). Masstabinvariante Korrelationstheorie. *Schriften des Mathematischen Instituts und des Instituts für Angewandte Mathematik der Universität Berlin*, (5), 179–233.

Homma, T. and A. Saltelli (1996). Importance measures in global sensitivity analysis of model output. *Reliab. Eng. Sys. Safety*, 52 (1), 1–17.

Iman, R. L. and S. C. Hora (1990). Robust measure uncertainty importance for use in fault tree system analysis. *Risk Analysis*, 10, 401–406.

Ishigami, T. and T. Homma (1990). An importance quantification technique in uncertainty analysis for computer models. In *Proc. ISUMA '90, First International Symposium on Uncertainty Modeling and Analysis, University of Maryland*, pp. 398–403.

Jacques, J. (2005). *Contributions à l'analyse de sensibilité et à l'analyse discriminante généralisée*. Ph. D. thesis, Université Joseph Fourier - Grenoble I.

Janon, A., T. Klein, A. Lagnoux, M. Nodet, and C. Prieur (2012). Asymptotic normality and efficiency of two Sobol index estimators. Rapport de recherche hal-00665048, HAL.

Jansen, M. J. W. (1999). Analysis of variance designs for model output. *Comp. Phys. Comm.*, pp. 35–43.

Jansen, M. J. W., W. A. H. Rossing, and R. A. Daamen (1994). *Predictability and Nonlinear Modelling in Natural Sciences and Economics*, chapter Monte Carlo estimation of uncertainty contributions from several independent multivariate sources, pp. 334–343. Kluwer Academic Publishers, Dordrecht.

Kempe, A. B. (1879). On the geographical problem of the four colours. *American Journal of Mathematics*, 2, 193–200.

Kendall, M. (1955). *Rank Correlation Methods*. Hafner Publishing Co.

Kostadinov, K. (2005). Non-parametric estimation of elliptical copulae with application to credit risk.

Krige, D. (1951). A statistical approach to some basic mine valuation problems on the Witwatersrand. *J. of the Chem., Metal. and Mining Soc. of South Africa*, 52(6), 119–139.

Kucherenko, S., S. Tarantola, and P. Annoni (2012). Estimation of global sensitivity indices for models with dependent variables. *Computer Physics Communications*, 183, 937–946.

Kullback, S. and R. Leibler (1951). On information and sufficiency. *Annals of Mathematical Statistics*, 22, 79–86.

Lambert, P. (2007). Archimedean copula estimation using Bayesian splines smoothing techniques. *Comput. Stat. & Data Anal.*, 51(12), 6307–6320.

Laurencelle, L. (2009). Le tau et le tau-b de kendall pour la corrélation de variables ordinales simples ou catégorielles. *Tutorials in Quantitative Methods for Psychology*, 5, 51–58.

Lebrun, R. and A. Dutfoy (2009a). A generalization of the Nataf transformation to distributions with elliptical copula. *Prob. Eng. Mech.*, 24(2), 172 – 178.

Lebrun, R. and A. Dutfoy (2009b). Do Rosenblatt and Nataf isoprobabilistic transformations really differ? *Prob. Eng. Mech.*, 24(4), 577–584.

Lebrun, R. and A. Dutfoy (2009c). An innovating analysis of the Nataf transformation from the copula viewpoint. *Prob. Eng. Mech.*, 24(3), 312 – 320.

Lemaire, M. (2009). *Structural Reliability*. Wiley.

Li, D. X. (1999). On default correlation: a copula function approach. Technical report, RiskMetrics Group.

Li, G. and H. Rabitz (2010). Global Sensitivity Analysis for Systems with Independent and/or Correlated Inputs. *J. Phys. Chem.*, 114, 6022–6032.

Li, G., S. Wang, and H. Rabitz (2002). Practical approaches to construct RS-HDMR component functions. *J. Phys. Chem.*, 106, 8721–8733.

Mara, T. and S. Tarantola (2012). Variance-based sensitivity indices for models with dependent inputs. *Reliab. Eng. Sys. Safety*, 107, 125–131.

Marrel, A., B. Iooss, F. Van Dorpe, and E. Volkova (2008). An efficient methodology for modeling complex computer codes with Gaussian processes. *Comput. Statist. Data Anal.*, 52, 4731–4744.

Martinez, J.-M. (2011). Analyse de sensibilité globale par décomposition de la variance. In *Journée thématique des GdR Ondes & Mascot Num.*

Matheron, G. (1962). *Traité de géostatistique appliquée*. Editions Technip.

Mikosch, T. (2006). Copulas: Tales and facts. *Extremes*, 9(1), 3–62.

Mori, T. and K. Tanaka (1973). Average stress in matrix and average elastic energy of materials with misfitting inclusions. *Acta Metallurgica*, 21, 1605–1609.

Morris, M. (1991). Factorial sampling plans for preliminary copmputational experiments. *Technometrics*, 33(2), 161–174.

Muehlenstaedt, T., O. Roustant, L. Carraro, and S. Kuhnt (2012). Data-driven Kriging models based on FANOVA-decomposition. *Statistics & Computing*, 22, 723–738.

Nataf, A. (1962). Détermination des distributions dont les marges sont données. *C. R. Acad. Sci. Paris*, 225, 42–43.

Nelsen, R. B. (1999). *An Introduction to Copulas*. Lecture Notes in Statistics. Springer.

O'Hagan, A. (2006). Bayesian analysis of computer code outputs: A tutorial. *Reliab. Eng. Sys. Safety*, 91(10-11), 1290–1300. Proc. 4th International Conference on Sensitivity Analysis of Model Output - SAMO 2004.

OpenTURNS (2005). *OpenTURNS, an Open source initiative to Treat Uncertainties, Risks'N Statistics in a structured industrial approach*. EDF R&D, EADS IW, PHIMECA Engineering S.A.

Owen, A. B. (2012a). Better estimation of small Sobol' sensitivity indices. Technical report, Stanford University.

Owen, A. B. (2012b). Variance components and generalized Sobol' indices. Technical report, Stanford University.

Park, C. K. and K. I. Ahn (1994). A new approach for measuring uncertainty importance and distributional sensitivity in probabilistic safety assessment. *Reliab. Eng. Sys. Safety*, 46, 253–261.

Pedregosa, F., G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, 12, 2825–2830.

Petersen, J. (1891). Die Theorie der regulären Graphen. *Acta Mathematica*, 15, 193–220.

Picheny, V., D. Ginsbourger, O. Roustant, and R. Haftka (2010). Adaptive designs of experiments for accurate approximation of a target region. *J. Mech. Des.*, 132(7), 071008.

Plischke, E., E. Borgonovo, and C. L. Smith (2012). Estimating global sensitivity statistics from given data. *European Journal of Operations Research*, submitted, XXXX–XXXX.

Powers, T. C. and T. L. Brownyard (1947). Studies of the physical properties of hardened Portland cement paste. *Journal of the American Concrete Institute*, 18, 101–132.

Rasmussen, C. and C. Williams (2006). *Gaussian processes for machine learning*. Adaptive computation and machine learning. MIT Press, Cambridge, Massachusetts, Internet edition.

Rosenblatt, M. (1952). Remarks on the multivariate transformation. *Annals of Mathematics and Statistics*, 43, 470–472.

Salmon, F. (2009). Recipe for disaster: The formula that killed wall street. *WIRED*, 17.03, 000–000.

Saltelli, A. (2002). Making best use of model valuations to compute sensitivity indices. *Comput. Phys. Comm.*, 145, 280–297.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola (2010). Variance based sensitivity analysis of model output. design and estimator for the total sensitivity index. *Comput. Phys. Comm.*, 181, 259–270.

Saltelli, A., S. Tarantola, and F. Campolongo (2000). Sensitivity analysis as an ingredient of modelling. *Statistical Science*, 15 (4), 39–56.

Saltelli, A., S. Tarantola, F. Campolongo, and M. Ratto (2004). *Sensitivity analysis in practice*. John Wiley & Sons, Ltd.

Sanahuja, J., L. Dormieux, and G. Chanvillard (2007). Modelling elasticity of a hydrating cement paste. *Cement Conc. Res.*, 37, 1427–1439.

Santner, T., B. Williams, and W. Notz (2003). *The design and analysis of computer experiments*. Springer series in Statistics. Springer.

Saporta, G. (2006). *Probabilités, analyse des données et Statistiques*. Technip, 2^e edition.

Silverman, B. (1986). *Density estimation for statistics and data analysis*. Chapman & Hall, London.

Sklar, A. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Sobol', I. (1993). Sensitivity estimates for nonlinear mathematical models. *Mat Model*, 2, 112–8.

Sobol', I. (2007). Global sensitivity analysis indices for the investigation of nonlinear mathematical models. *Matematicheskoe Modelirovanie*, 19, 23–24. (in Russian).

Sobol', I., S. Tarantola, D. Gatelli, S. Kucherenko, and W. Mauntz (2007). Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliab. Eng. Sys. Safety*, 92, 957–960.

Soize, C. and R. Ghanem (2004). Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput.*, 26(2), 395–410.

Spearman, C. (1904). The proof and measurement of association between two things. *Amer. J. Psychol.*, 15, 72–101.

Sudret, B. (2006). Global sensitivity analysis using polynomial chaos expansions. In Spanos, P. and G. Deodatis (Eds.), *Proc. 5th Int. Conf. on Comp. Stoch. Mech (CSM5), Rhodos, Greece*.

Sudret, B. (2007). *Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods*. Ph. D. thesis, Université Blaise Pascal - Clermont II. Habilitation à diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France.

Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Sys. Safety*, 93, 964–979.

Sudret, B., T. Yalamas, E. Noret, and P. Willaume (2009). Sensitivity analysis of nested multiphysics models using polynomial chaos expansions. In Takada, T. and H. Furuta (Eds.), *Proc. 10th Int. Conf. Struct. Safety and Reliability (ICOSSAR'2009), Osaka, Japan*.

Tennis, P. D. and H. M. Jennings (2000). A model for two types of calcium silicate hydrate in the microstructure of portland cement pastes. *Cement Conc. Res.*, 30, 855–863.

Vandermonde, A.-T. (1771). Remarques sur des problèmes de situation. *Mémoires de l'Académie Royale des Sciences*, X, 566–574.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Wand, M. and M. Jones (1995). *Kernel smoothing*. Chapman and Hall.

Welch, W., R. Buck, J. Sacks, H. Wynn, T. Mitchell, and M. Morris (1992). Screening, predicting, and computer experiments. *Technometrics*, 34(1), 15–25.

Wiener, N. (1938). The homogeneous chaos. *American Journal of Mathematics*, 60, 897–936.

Xiu, D. and G. Karniadakis (2003). Modelling uncertainty in steady state diffusion problems via generalized polynomial chaos. *Comput. Methods Appl. Mech. Engrg.*, 191(43), 4927–4948.

Xu, C. and G. Gertner (2008). Uncertainty and sensitivity analysis for models with correlated parameters. *Reliab. Eng. Sys. Safety*, 93, 1563–1573.

Abstract

This thesis is a contribution to the nested modelling of complex systems. A global methodology to quantify uncertainties and their origins in a workflow composed of several models that can be intricately linked is proposed. This work is organized along three axes. First, the dependence structure of the model parameters induced by the nested modelling is rigorously described thanks to the copula theory. Then, two sensitivity analysis methods for models with correlated inputs are presented: one is based on the analysis of the model response distribution and the other one is based on the decomposition of the covariance. Finally, a framework inspired by the graph theory is proposed for the description of the imbrication of the models. The proposed methodology is applied to different industrial applications: a multiscale modelling of the mechanical properties of concrete by homogenization method and a multiphysics approach of the damage on the cylinder head of a diesel engine. The obtained results provide the practitioner with essential informations for a significant improvement of the performance of the structure.

Keywords: Global sensitivity analysis, correlation, copula theory, graph theory, nested modelling, multiscale modelling.

Résumé

Cette thèse est une contribution à la modélisation imbriquée de systèmes complexes. Elle propose une méthodologie globale pour quantifier les incertitudes et leurs origines dans une chaîne de calcul formée par plusieurs modèles pouvant être reliés les uns aux autres de façon complexe. Ce travail est organisé selon trois axes. D'abord, la structure de dépendance des paramètres du modèle, induite par la modélisation imbriquée, est modélisée de façon rigoureuse grâce à la théorie des copules. Puis, deux méthodes d'analyse de sensibilité adaptées aux modèles à paramètres d'entrée corrélés sont présentées : l'une est basée sur l'analyse de la distribution de la réponse du modèle, l'autre sur la décomposition de la covariance. Enfin, un cadre de travail inspiré de la théorie des graphes est proposé pour la description de l'imbrication des modèles. La méthodologie proposée est appliquée à des exemples industriels d'envergure : un modèle multiéchelles de calcul des propriétés mécaniques du béton par une méthode d'homogénéisation et un modèle multiphysique de calcul de dommage sur la culasse d'un moteur diesel. Les résultats obtenus fournissent des indications importantes pour une amélioration significative de la performance d'une structure.

Mots-clés: Analyse de sensibilité globale, corrélation, théorie des copules, théorie des graphes, modélisation imbriquée, modélisation multiéchelles.