Università degli Studi di Milano–Bicocca

DOTTORATO DI RICERCA IN

STATISTICA

CICLO XXIV

Coordinatore Prof.ssa Fulvia Mecatti

TOWARDS MORE ACCURATE MEASURES OF GLOBAL SENSITIVITY ANALYSIS

Investigation of first and total order indices

Dottorando

Paola Gioia

Tutor

Dott. Stefano Tarantola

Acknowledgement

I am grateful to have had the support and guidance of many outstanding people.

First of all I would like to offer my gratitude to my thesis supervisor, Stefano Tarantola. This thesis would not have been possible without his mentorship. I am also grateful to Stefano for teaching me that behind every challenge to statistical models, there is an exciting scientific problem waiting to be solved. I admire him greatly, and I look forward to our continued collaborative work and friendship in the future.

Even though I am not his student, I would like to thank Andrea Saltelli, who enthusiastically gave me an office in his Institute and invited me to his seminars and meetings. I am very grateful for his support and open advice over the last year. I also wish to thank the JRC staff of the Econometrics and applied Statistics Unit.

I would like to express my appreciation to the Institute of Disposal Research of Clausthal University of Technology, for hosting me as a visitor. In particular, I would like to thank Elmar Plischke, who firstly proposed my invitation, for continuously inspiring me to work harder, and for being a good friend. I would also like to thank him for helping me every time I needed to discuss some controversial impasse or to have an advice.

Of course, I am grateful to the Department of Statics of the University of Milan–Bicocca for providing me an excellent learning environment and to all my PhD Professors for being generous with their time and advice whenever I needed them. In particular, I would like to thank my PhD Director Fulvia Mecatti.

Contents

Introduction

Sensitivity Analysis is the study of how a variation in the output of a statistical model can be apportioned, qualitatively and quantitatively, to different sources of variation. It is used to identify which inputs are most influential in inducing the uncertainty in the output. Such information is important for understanding the behavior of the simulator and determining where better input information is needed and where the model might be improved.

Among the different existing schools of thought, can be identified the Global Sensitivity Analysis. This approach explores the whole range of uncertainty of the model inputs by letting them vary simultaneously.

So–called variance–based methods for computing sensitivity indices are based on the decomposition of the variance of the outputs into terms corresponding to the different inputs and their interactions. In this way, they can assess the manner in which the uncertainty in an output is apportioned across the inputs, and across interactions between them. Variance–based measures are attractive because they measure sensitivity across the whole input space (i.e. it is a global method), they can deal with nonlinear responses, and they can measure the effect of interactions in non–additive systems.

The drawback of variance–based measures is their computational cost especially in the presence of a high number of factors. Estimating the sensitivity coefficients takes many model runs, as we shall discuss later, and this is the reason why much recent research aims to find efficient numerical algorithms for their computation.

Our intentions in preparing this thesis were to attempt to overcome this drawback, in order to minimize the computational cost for achieving the required accuracy of sensitivity measures.

We first developed a new technique for estimating variance–based total sensitivity indices from given data and investigated its performance through a new algorithm.

Although the results did not completely meet our expectations, this new technique is worthy of further study and analysis. So, we focused on a new approach for the estimation of the first order effects given a specific sample design. This method adopts the RBD approach published by Tarantola *et al*., (2007) for the computation of first order sensitivity indices in association to Quasi–Random numbers.

From these preconditions, the present thesis is organized in four chapters that are now presented.

In chapter one we give a general introduction and an overview of the Sensitivity Analysis. In chapter two we describe variance–based methods in more details. In chapter three we devise a new methodology for estimating total sensitivity indices from given data which takes into consideration the overall effect of interactions among model inputs. In this chapter four we present a new approach for the estimation of the first order effects given a specific sample design.

Capitolo 1

INTRODUCTION TO SENSITIVITY ANALYSIS

1.1 History of sensitivity analysis

The origin of Sensitivity Analysis can be tracked from the theory of Design of Experiments (DOE), which was first introduced by Fisher (1935) in the context of physical experimentation.

To this purpose the analyst needs to design "a priori" an experiment able to highlight the relationship between the response and the inputs. The most natural way to design an experiment is to use the one–factor at–a–time approach (OAT) (see Daniel, 1958, 1973), which varies the input factors of interest one at a time, with the remaining ones are held constant to their nominal values. However, the result of an OAT depends on the nominal values used for the other factors. Often the behavior of the response function is described

only locally in the input space, i.e. by perturbing the factor of interest of a small amount around its nominal value.

Mathematical models are developed to approximate engineering, physical, environmental, social, and economic phenomena of various complexity. Model development consists of several logical steps, one of which should be the determination of parameters which are most influential on model output. The *National Research Council's Science and Decisions: Advancing Risk Assessment* (2009) defines a model as a "simplification of reality that is constructed to gain insights into select attributes of a particular physical, biologic, economic, or social system. Mathematical models express the simplification in quantitative terms" (page. 96). Model input factors are "terms in a model that determine the specific model form. For computational models, these terms are fixed during a model run or simulation, and they define the model output. They can be changed in different runs as a method of conducting sensitivity analysis or to achieve a calibration goal" (page 97).

Sensitivity analysis can be seen as the modern evolution of DOE. In short:

Sensitivity Analysis is the study of how the uncertainty in the output of a model can be apportioned to different sources of uncertainty in the input. (Saltelli, 2002)

The object of SA is to increase the confidence in the model and its predictions by providing an understanding of how the model output variables respond to changes in the inputs.

Models of varying complexity are developed to approximate or mimic systems and processes of different nature (e.g. physical, environmental, social, or economic). Many processes are so complex that physical experimentation is too time-consuming too expensive or even impossible. As a result, to explore systems and processes, investigators often turn to mathematical or computational models.

A mathematical model is defined by a series of equations, parameters and variables aimed to characterize the process being investigated. Model inputs are subject to many sources of uncertainty including errors of measurement, incomplete information and poor or partial understanding of the driving forces and mechanisms. This imposes a limit

on our confidence in the response of the model. Further, models have to cope with the natural intrinsic variability of the system such as the occurrence of stochastic events.

Good modeling practice requires that the modeler provides an evaluation of the confidence in the model possibly assessing the uncertainties associated with the modeling process and with the outcome of the model itself.

Originally (Tomovic and Vukobratovic, 1972), SA was created to deal simply with uncertainties in the input variables and model parameters. Over the course of time the ideas have been extended to incorporate model conceptual uncertainty, i.e. uncertainty in model structures, assumptions and specifications (Helton and Burmaster, 1996; Draper *et al.*, 1999).

A view of modeling that may help to illustrate the role of sensitivity analysis in the scientific process is taken from the work of the biologist Robert Rosen (1991) (see also Saltelli *et al.*, 2000, page 3–4). According to Rosen, the world and the model are linked via the process of "encoding" from world to model and "decoding" from model to world. While inside "world" and inside "model" causality reigns, encoding and decoding are not themselves entailed by anything, i.e. they are the objects of the modeler craftsmanship. Experience has shown that even when the world is indeed a well–defined and closed system, for instance an artifact, an artificial device or a piece of machinery, different modelers can generate different nonequivalent descriptions of it, that is, models whose outputs are compatible with the same set of observations but whose structures are not reconcilable with one another. The "encoding" and "decoding" activities are the essence and the purpose of the modeling process: one writes a model in the hope that the decoding operation will provide insight on the world. This is only possible if the uncertainty in the information provided by the model (the substance of use for the decoding exercise) is carefully apportioned to the uncertainty associated with the encoding process.

Practitioners of modeling have come to live with the rather unpleasant reality that more than one model may be compatible with the same set of data or evidence. Some have gone so far as to coin a word for this paradox: *equifinality* (Beven, 1993, 2001; see also Saltelli *et al.*, 2004) meaning that different models can lead to the same end. Others refer to the phenomenon as model indeterminacy.

Another general definition of SA took inspiration from the previous Rosen's formalization of the modeling activity:

Sensitivity Analysis studies the relationships between information flowing in and out of the model (Saltelli et al, 2000, pag.4).

SA is hence part of model building. It is used to improve the understanding of the model and to increase the confidence in its predictions. It shows how the model response variables react to changes in the inputs, where by input or factors it is intended data, model structure and model parameters. SA is thus closely linked to uncertainty analysis (UA), which aims to quantify the overall uncertainty associated with the response as a result of uncertainties in the model input.

Our point of departure is a mathematical or computational model $Y = f(X_1, X_2, ..., X_k)$ where some of the input factors X_i are uncertain. We know something about their range of uncertainty. This knowledge might come from a variety of sources: measurements, expert opinion, physical bounds, analogy with factors for similar species, etc. This latter may be seen as a particular case of expert opinion. We may further have information (e.g. via observation) on the joint probability distribution of the factors. The model may be used in a prognostic (forecast) or diagnostic (estimation) mode. In the former, all our knowledge about model input is already coded in the joint probability distribution of the input factors. In the latter, the information on the input constitutes *a priori* knowledge and the analysis might be aimed at updating either the distribution of the input factors or the model formulation based on the evidence. It is customary to propagate uncertainty through different model structures or formulations. In this case some of the input factors are triggers that drive the selection of a structure versus another (Saltelli, 2002).

The input is a quantity which is allowed to vary in order to study its effect on the output. A sensitivity analysis will in turn instruct the modelers as to the relative importance of the inputs in determining the output. An obvious consequence of this is that the modeler will remain ignorant of the importance of those variables which have been kept fixed (not been included in the analysis). This is of course a hazard for the modeler, as a variable deemed non influential and kept fixed could have negative effect the results of the analysis if it is actually influential (Saltelli *et al.*, 2008).

It seems, therefore, that one should be as careful and objective as possible in deciding on the input for uncertainty and sensitivity analyses. Clearly, the more variables we promote to the rank of input, and allow to vary, the greater the variance to be expected in the model prediction. This could lead to a situation in which we discover that, having incorporated all uncertainties, the model prediction varies so wildly as to be of no practical use. This trade-off has been brilliantly summarized by the econometrician Edward E. Leamer (1990):

"I have proposed a form of organized sensitivity analysis that I call 'global sensitivity analysis' in which a neighborhood of alternative assumptions is selected and the corresponding interval of inferences is identified. Conclusions are judged to be sturdy only if the neighborhood of assumptions is wide enough to be credible and the corresponding interval of inferences is narrow enough to be useful".

Note Leamer's emphasis on the need for 'credibility' in the selection of assumptions. The easiest way to invalidate a model is to demonstrate it fragile with respect to shaky assumptions. Note, however, that the trade-off may not be as dramatic as one might expect, and that increasing the number of input factors does not necessarily lead to an increased variance in model output. Practitioners have recorded that in most uncertainty and sensitivity analyses the input factors' importance is distributed similarly to wealth in nations, with a few factors creating almost all the uncertainty and the majority making only a negligible contribution. Hence, if the 'key' factors have been judiciously chosen, adding further variables to the analysis may add to its completeness and defensibility without adversely increasing the variance in the output. (Saltelli *et al*., 2008, page. 10)

As mentioned, the quality of a model is largely a function of its fitness for purpose. If modeling is a craft and models cannot be proven true (because of the pervasive nature of uncertainty and the difficulty of separating observation from observer and facts from values), then the modeler has a moral obligation, and indeed it is in the modelers' own practical interest, to be as rigorous as possible when assessing the robustness of model inference.

Doing so should produce better and more parsimonious models, and will strengthen the analyst's defense of the results in the case of scientific controversy or public policy debate (Saltelli *et al*., 2008).

1.2 Reasons for conducting Sensitivity Analysis

In the context of numerical modelling, SA means very different things to different people. For example, for a chemist SA could be the analysis of the strength of the relation between kinetic or thermodynamic inputs and measurable output of a reaction system. For an economist the task of SA could be to appraise how stable the estimated parameters of a model (customarily derived via regression) are with respect to all factors that were excluded from the regression, thus ascertaining whether parameter estimation is robust or fragile. For a statistician involved in statistical modelling SA is mostly known and practised under the heading of "robustness analysis". Statisticians are mostly interested in distributional robustness intended as insensitivity with respect to small deviations from the assumptions about the underlying distribution assumed for the data (Huber, 1981).

These different types of analyses have in common the aim to investigate how a given computational model responds to variations in its inputs. Modellers conduct SA to determine (Saltelli *et al*., 2008):

a) If a model resembles the system or process under study.
b) The factors that mostly contribute to the output variability and that require additional research to strengthen the knowledge base, thereby reducing output uncertainty.
c) The model parameters (or parts of the model itself) that are insignificant and that can be eliminated from the final model.
d) If there is some region in the space of input factors for which the model variation is maximum.
e) The optimal regions within the space of the factors for use in a subsequent calibration study.
f) If and which (group of) factors interact with each other.

Under (a) the model does not properly reflect the processes involved if it exhibits strong dependence on supposedly non-influential factors or if the range of model predictions is

not a sound one. In this case SA highlights the need to revise the model structure. It often happens that the model turns out to be highly tuned to a specific value of a factor, up to the point that necessary changes, e.g. resulting from new evidence, lead to unacceptable variation in the model predictions. When this happens it is likely that in order to optimize the simulation some parameter values have been chosen incorrectly. This reflects lack of conceptual understanding of the role of the parameters in the system.

Under (b) SA can assist the modeller in deciding whether the parameter estimates are sufficiently precise for the model to give reliable predictions. If not further work can be directed towards improved estimation of those parameters that give rise to the greatest uncertainty in model predictions. If the model sensitivity seems congruent with (i.e. does not contradict) our understanding of the system being modelled, SA will open up the possibility of improving the model by prioritizing measurement of the most influential factors. In this way the impacts of measurement errors on computational results can be minimized.

Under (c) we mean insignificant in the sense of "not affecting the variation of the output"; according with some investigators when the model is used in a case of conflicting stakes (e.g. siting a facility or licensing a practice) the model should not be more complex than needed and factors/processes that are insignificant should be removed.

As far as (e) is concerned we stress the need for "global" optimization. One should investigate the space of the factors in its entirety and not just around some nominal points.

1.2.1 Why one should perform SA?

Sensitivity analysis can serve a number of useful purposes in the economy of modeling. It can surprise the analyst, uncover technical errors in the model, identify critical regions in the space of the inputs, establish priorities for research, simplify models and defend against falsifications of the analysis. In the context of models used for policy assessment,

sensitivity analysis can verify whether policy options can be distinguished from one another given the uncertainties in the system, and so on.

SA can be employed prior to a **calibration** exercise to investigate the tuning importance of each parameter, i.e. to identify a candidate set of important factors (which factor is most deserving further analysis or measurements) for calibration since the difficulty of calibrating models against field or laboratory data increases with the number of processes to be modeled (and hence the number of parameter to be estimated). SA may allow a dimensionality reduction of the parameter space where the calibration is made resulting in some factors of the model to. In this setting, SA can also help to ensure that the problem is not ill-conditioned. Quantitative SA methods (i.e. those which tell how much more important one factor is than another) can be appropriate when both the model inputs and the available data are affected by uncertainties. The question answered is: "what factors can be calibrated and at what confidence, given the data and their uncertainty"?

While SA was originally created to deal with the uncertainties in the input factors, recent developments have seen some of the ideas being extended to incorporate structural uncertainty as mentioned above. In this way SA also touches on the difficult problem of **model quality** and is an important element of judgment for the corroboration or falsification of the scientific hypotheses embedded in a model (is the inference robust? Is the model overly dependent on fragile assumptions?). SA can be used to ensure that the response of the model to its input factors can be accounted for, that the model does not exhibit strong dependence on supposedly non-influential factors and that the range of model predictions is a sensible one.

SA can be an effective tool for **model identification**. By pinpointing experimental conditions in which the ability to discriminate among the various models is a maximum, SA can identify the most appropriate model structures and competing specification that describe available evidence.

This is closely related to mechanism reduction determining a subset of input factors accounting for the output variance. This enables the insignificant factors to be identified and eliminated from the final model. In this way irrelevant parts of the model can be

dropped or a simpler model can be built or extracted form a more complex one (model lumping).

The above points have some epistemological implications concerning the relevance of a model. It has been argued that often the complexity of models largely exceeds the actual requirements. The view of Oreskes et al. (1994) is that models should be heuristic constructs built for a task. They would be relevant when their input factors actually cause variation in the model response that is the object of the analysis. Model irrelevance would flag a bad model, a model used out of context or a model unable to provide the answer being sought.

Another possible goal for SA is to determine if there is some region in the space of inputs for which the model variation is maximum or divergent. This is useful in control theory where one might also be interested in the initiation of chaotic behavior for some combinations of model parameters.

1.2.2 Properties of an ideal sensitivity analysis method

We plan to use methods that are global and model-free, in the sense of being independent from assumptions about the model, such as linearity, additivity and so on. These methods must be capable of testing the robustness and relevance of a model-based analysis in the presence of uncertainties. Whenever possible, we would also like our methods to be quantitative. The desirable properties of sensitivity analysis are as follows:

- *The ability to cope with the influence of scale and shape.* The influence of the input should incorporate the effect of the range of input variation and the form of its probability density function (pdf). It matters whether the pdf of an input factor is uniform or normal, and what the distribution parameters are.
- *To include multidimensional averaging*. In a local approach to SA (*e.g.* $S_i = \partial Y / \partial X_i$), one computes partial derivatives. This is the effect of the variation of a factor when all others are kept constant at the central (nominal) value. A global method should instead evaluate the effect of a factor while all others are also varying.

- *Being model independent*. The method should work regardless of the additivity or linearity of the model. A global sensitivity measure must be able to appreciate the so-called interaction effect, which is especially important for non-linear, non-additive models. These arise when the effect of changing two factors is different from the sum of their individual effects.
- *Being able to treat grouped factors as if they were single factors*. This property of synthesis is essential for the agility of the interpretation of the results. One would not want to be confronted with an SA made of dense tables of sensitivity measures.

Beside the properties above, we would like the setting for the SA itself to be as stringent as possible. It may well happen that using different measures of sensitivity, different experts obtain different relative ranking of the influence of the various input factors (see OECD, 1993 for an example). This happens if the objective of the analysis is left unspecified. Just as there are several definitions of risk (Risk Newsletter, 1987), there may be several definitions of importance.

1.3 Local Vs. Global approaches

Two quite different schools of thought may be identified (Saltelli *et al*., 1997): the local sensitivity analysis school and the global analysis one. In the first school, the local response of the output(s), obtained by varying parameters one at a time and holding the others fixes to a central (nominal) value, is investigated; this involves partial derivatives, possibly normalized by the nominal value of the parameter or by its standard deviation. All the analysis is run at a given central point in the space of the input parameters, and the volume of the region explored is nil. The second school is more ambitious in two respects: firstly, the input parameters space is explored within a finite (or even infinite) region and secondly, the variation of the output induced by a parameter is taken globally, i.e. averaged over the variation over all parameters.

A pragmatic and empirical approach has dominated the global sensitivity school, due to the intrinsic difficulty of building an effective global sensitivity measure over a finite space of variation for the input parameters. In this respect the work of researchers like Iman, Conover, Helton, was fundamental (see Helton, 1993 for a review). These investigators have tested and promoted the use of robust and reliable methods based on Monte Carlo regression and correlation analysis, and on the use of scatterplots. Methods such as Standardized Regression Coefficients (SRC), Correlation measures (Pearson), Partial Correlation Coefficients (PCC), have been used with some success. The aforementioned authors favored the use of rank transformed measures (Standardized Rank Regression Coefficients, SRRC, Spearman correlation, Partial Rank Correlation Coefficients, PRCC) for non-linear models. These methods offer a robust and easy to implement SA, provided that the input-output relationship is monotonic. The ordering of importance of the input factors based on these statistics must be considered with caution, especially when the associated model coefficient of determination is smaller than one. In any case, the analysis should be considered as qualitative rather than quantitative (for instance, SRC gives information on the liar regression model that is used to describe the system model, not on the system model itself; for rank-transformed statistics see Saltelli *et al*, 1993 and Saltelli & Sobol', 1995).

Non–linear, non–monotonic problems are often encountered in everyday model building. These problems call for a non-linear SA which is independent form assumptions about the model structure.

1.3.1 Local sensitivity analysis

Until the 90's, and often today as well, sensitivity analysis was conceived as a **local** measure of the effect of a given input on a given output. In these local sensitivity measures the effect of X_j is observed while assuming all other factors fixed. This approach falls, hence, in the class of the one–factor–at–a–time (OAT) methods.

The simplest and most intuitive way to obtain a local sensitivity index is to compute derivatives (Tomovic & Vukobratovic, 1972; see Varma *et al*., 1999; Grievank, 2000

for recent review). The sensitivity of the output Y to a perturbation of an input factor X_j is estimated at a given value, X_j^* as

$$Y'_{X_i} = \left.\frac{\partial Y}{\partial X_i}\right|_{X_i = X_i^*} \tag{1.1}$$

In situations where Y and X_j have different ranges of uncertainty, a more balanced measure can be obtained normalizing the derivatives by the factors' standard deviations:

$$S_{X_i}^{\sigma} = \left.\frac{\sigma X_i}{\sigma Y}\frac{\partial Y}{\partial X_i}\right|_{X_i = X_i^*} \tag{1.2}$$

The estimation of these local measures can be easily by solving systems of derivatives or taking incremental ratios.

Local sensitivities are useful for a variety of applications, such as the solution of inverse problems, e.g., relating macroscopic observables of a system, such as kinetic constants, to the quantum mechanics properties of the system,6 or the analysis of runaway and parametric sensitivity of various types of chemical reactors. (Turanyi, 1990; Rabitz 1989;, Saltelli *et al*., 2012).

Local sensitivities provide the slope of the calculated model output in the parameter space at a given set of values. In many applications, this is exactly the information needed. In other areas, such as uncertainty analysis, local SA is a computationally efficient technique that allows a rapid preliminary exploration of the model. The calculation of local sensitivities is much faster than that of global sensitivities.

One shortcoming of the linear sensitivity approach is that it is not possible to assess effectively the impact of possible differences in the scale of variation of the input factors

(unless the model itself is linear). When significant uncertainty exists in the input factors, the linear sensitivities alone are not likely to provide a reliable estimator of the output uncertainty in the model. When the model is non-linear and various input variables are affected by uncertainties of different orders of magnitude, a global sensitivity method should be used.

1.3.2 Global sensitivity analysis

In global SA, the probability density functions for each factor provide the input for the analysis. These distributions are valuable since they represent our knowledge (or lack of it) with respect to the model and its parameterization. A SA experiment is usually considered to be global when all the parameters are varied simultaneously and the sensitivity is measured over the entire range of each input parameter.

Global SA techniques have been discussed by Cukier *et al*. (1978), Iman & Helton (1988), Sobol' (1990b), Helton *et al*. (1991) and Saltelli & Homma (1992), among others.

One advantage of these methods is that they explore the entire interval of definition of each factor. Another advantage is that each 'effect' for a factor is in fact an average over the possible values of all the other factors. Global methods have the following two properties (Satelli *et al*., 2000):

1. the *inclusion of influence of scale and shape*. The sensitivity estimates of individual factors incorporate the effect of the range and the shape of their probability density functions.
2. *multidimensional averaging*. The sensitivity estimates of individual actors are evaluated varying all other factors as well.

A global SA technique thus incorporates the influence of the whole range of variation and the form of the probability density function of the input.

Another general consideration with respect to the global, explorative nonparametric methods for the sensitivity analysis just described is that these have a better chance of

being resilient towards type II errors than local (derivative–based) methods. The possibility of important factors being overlooked or dangerous or critical combinations of input factors neglected decreases with the level of exploration of the space of the input factors (Farrell, 2007). The attention paid in global methods to interaction effects is also a protection against type II errors. In Saltelli *et al.* (2012) we show that, for even a relatively simple and well-studied chemical reactor system, global sensitivity analysis can lead to the identification of a larger 'runaway' portion in the space of the input factors, than could previously be identified.

Several global methods have been developed since the 90's: screening methods by Morris (1991), non-parametric or regression-based in Saltelli & Marivoet (1990), Helton (1993), variance-based methods (Sobol' 1993, Oakley & O'Hagan 2004), density-based (Park & Ahn 1994, Chun *et al*. 2000, Borgonovo 2007, Liu & Homma, 2009) and expected–value–of–information (EVI) based (Oakley *et al*. 2010). The common feature of the last three classes of methods is that they are on the one hand the most informative in terms of uncertainty appraisal and on the other hand the most computationally intensive (Plischke et al., 2013).

Operatively, global sensitivity analysis is performed according to the following steps:

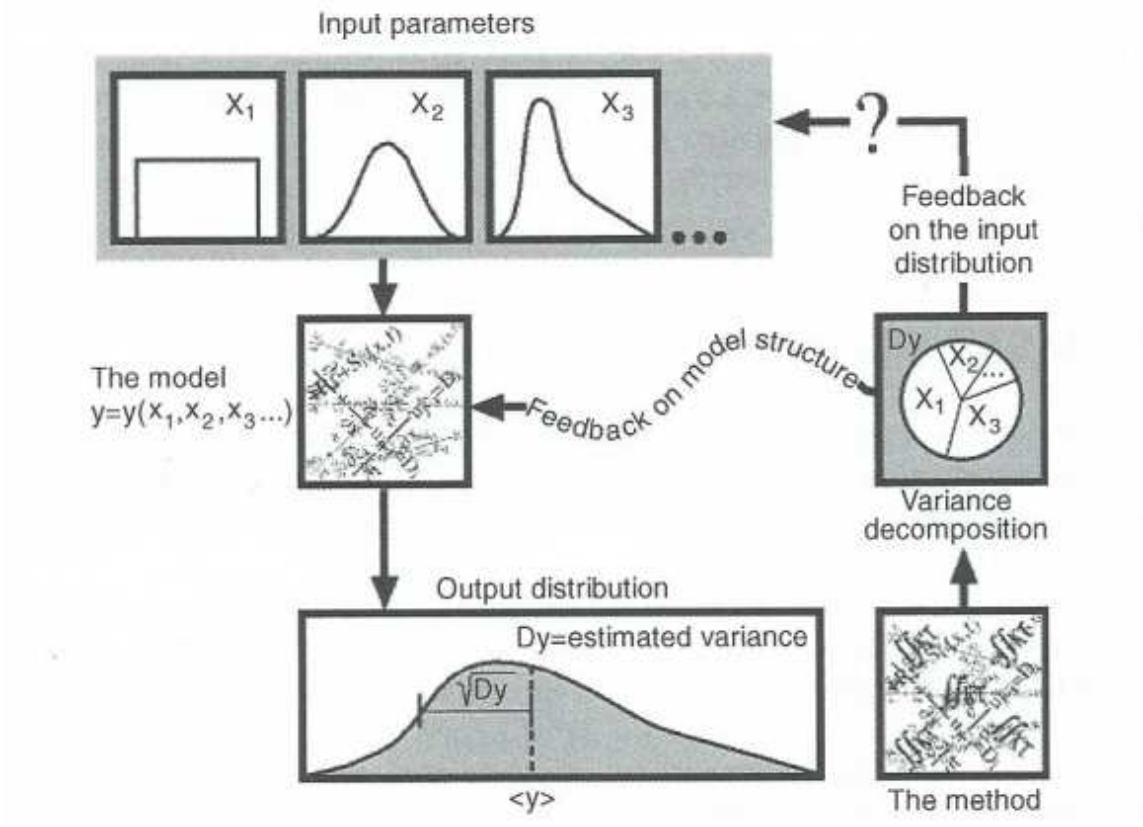
1. specify the target function and select the inputs of interest
2. assign a distribution function to the selected factors, form available data, expert opinion or physical bounding considerations or via an estimation process
3. generate an input set (sample) of size N from the factors distribution according to an appropriate design
4. evaluate the model at each sample point obtaining N values for the target function
5. estimate the influence or relative importance of each input factor on the target function

In point 3, the input set can be generated by using a number of sampling designs: purely random generation from the assigned distributions (simple Monte Carlo), Latin hypercube sampling (LHS) (McKay *et al*, 1979), $LP\tau$ sequences (Sobol', 1967), winding stairs (Jansen, 1994), or other more or less sophisticated techniques. We have

found considerable advantages in using quasi-random $LP\tau$ sequences (Homma & Saltelli, 1994, 1996).

The first four points constitute uncertainty analysis. The N values computed for the target function can be used to display the empirical distribution of the model output, thus quantifying the variation in the model response. The fifth point is sensitivity analysis: different methods can be used to apportion the uncertainty in the target function to the inputs. In variance-based methods, for instance, a typical representation of the results is in the form of a pie chart (Fig. 1.1) that partitions the variance of the output according to the contribution of each input factor. In this thesis we focus exclusively on variance-based methods.

Fig. 1.1



A schematic view of sampling–based sensitivity analysis

Capitolo 2

VARIANCE–BASED SENSITIVITY ANALYSIS

2.1 Introduction

Variance–based methods are based on the decomposition of the variance of the model output into terms of increasing dimensionality (Sobol', 1993).

The idea of using variance in SA dates back to the early 1970s (Cukier *et al*., 1973). Cukier and colleagues not only proposed conditional variances for a SA based on first–order effects, but were already aware of the need to treat higher–order terms and of the underlying variance decomposition theorems. Their method, known as FAST (Fourier Amplitude Sensitivity Test) although quite effective, enjoyed limited success among practitioners not least because of the difficulty in encoding it. The method did not allow the computation of higher–order indices, although this was much later made possible by an extension of the method, EFAST, developed by Saltelli *et al*., (1999).

Hora and Iman (1986), introduced the uncertainty importance of a factor x_i defined as:

$$I_i = \sqrt{Var(y) - E[Var(y|x_i)]} \tag{2.1}$$

Later, the same authors (Iman & Hora, 1990) proposed a new statistic based on estimating the following quantity:

$$\frac{Var_{x_i}[E(\log y|x_i)]}{Var[\log y]} \tag{2.2}$$

where Var_{x_i} stands for variance over all the possible values of x_i and $E[\log y|x_i]$ is estimated using linear regression. This solution has the advantage of robustness but, as observed by the authors, the conclusions drawn on $\log y$ are not easily converted back to y.

In 1993, Sobol' developed an original extension of Design of Experiments (DOE) to the world of numerical experiments in which the total variance of the model output is assumed to be made up of terms of increasing dimensionality (Sobol', 1993). Sobol' indices are superior to the original FAST in that the computation of the higher interaction terms is very natural and is similar to the computation of the main effects. Each effect (main or otherwise) is computed by evaluating a multidimensional integral via a Monte Carlo (MC) method. Saltelli, (2002) and Saltelli *et al*. (2010) further improved this method.

2.4 Properties of Variance–based methods

Variance–based methods are model independent: they work regardless to the additivity or linearity of the model. We can thus study the presence of interactions among the input factors also for non–linear, non–additive models.

They can also capture the influence of the full range of variation of each factor and they are capable of dealing with groups of factors: uncertain factors might pertain to different logical types, and it might be desirable to decompose the uncertainty according to these types.

The drawback of variance–based measures is their computational cost in terms of number of model runs and the fact that the information on the uncertainty of the model output is captured by the second order moment; in this case, we lose the full information of the uncertainty of the output which is given by its distribution.

2.5 Variance decomposition and sensitivity indices

The assessment and presentation of the effects of uncertainty are now widely recognized as important parts of analyses for complex systems. At the simplest level, such analyses can be viewed as the study of functions of the form

$$\mathbf{y} = \mathbf{f}(\mathbf{x}) \tag{2.3}$$

where the function $\mathbf{f}$ represents the model under study, $\mathbf{x} = [x_1, x_2, ...]$ is a vector of model inputs defined over a multi-dimensional space Ω, and $\mathbf{y} = [y_1, y_2, ...]$ is a vector of model predictions. The goal of an uncertainty analysis is to determine the uncertainty

in the elements of $\mathbf{y}$ that results from uncertainty in the elements of $\mathbf{x}$. A typical adjunct to an uncertainty analysis is a sensitivity analysis, which attempts to determine how the uncertainty in individual elements of $\mathbf{x}$ affects the uncertainty in the elements of $\mathbf{y}$. In practice, $\mathbf{f}$ can be quite complex (e.g., one or more computer programs involving complex algorithms and many thousands of lines of programming); further, $\mathbf{x}$ and $\mathbf{y}$ are often of high dimension.

To carry out uncertainty and sensitivity analyses, the uncertainty in the elements of $\mathbf{x}$ must be characterized.

For simplicity we treat each input x_i as random variable which is independent and uniformly distributed over the unit interval

$$x_i \in U[0,1] \tag{2.4}$$

So

$$p(x_i) = 1 \tag{2.5}$$

and all the integrals can be written without integration limits

$$\int_{\Omega} x_i \, p(x_i) \, dx_i \equiv \int x_i \, dx_i \tag{2.6}$$

For notational convenience the output $\mathbf{y}$ in *Eq.* (1) will be assumed to be scalar. With this assumption, the representation in Eq. (1) becomes

$$y = f(\mathbf{x}) \tag{2.7}$$

In a variance–based method we are interested in the variance of the model output and its decomposition. For this purpose we start from a decomposition of our model $y = f(\mathbf{x})$ into a High Dimensional Model Representation (HDMR), Rabitz et al. (1999)

$$y = f(\mathbf{x}) = f_0 + \sum_i f_i(x_i) + \sum_i \sum_{j>i} f_{ij}(x_i, x_j) + \ldots + f_{1,2,\ldots,k}(x_1, x_2, \ldots, x_k) \tag{2.8}$$

$$f(\mathbf{x}) \in L^2[0,1] \qquad \mathbf{x} \in [0,1]^k \tag{2.9}$$

where $f(\mathbf{x})$ is a square integrable function f over Ω, the $k-$dimensional unit hypercube.

This decomposition is not unique (infinite ways exist to build an HDMR) as the lower order terms can be selected arbitrarily and the highest order terms can be written as the difference between $f(\mathbf{x})$ and the terms of lower order.

2.5.1 ANOVA-HDMR decompositions

It can be proven that if each term in the HDMR is chosen such that:

$$\int_0^1 f_{i_1 \ldots i_s}(x_{i_1}, \ldots, x_{i_s})\, dx_{i_k} = 0 \qquad \text{for} \quad i_k = i_1, \ldots, i_s \tag{2.10}$$

then all the terms in (2.8) are orthogonal and can be expressed as integrals of $f(\mathbf{x})$. Indeed,

$$f_0 = E(Y) \tag{2.11}$$

$$f_i(X_i) = E(Y \mid X_i) - E(Y) \tag{2.12}$$

$$f_{ij}(X_i, X_j) = E(Y \mid X_i, X_j) - f_i(X_i) - f_j(X_j) \tag{2.13}$$

$$f_{ijk}(X_i, X_j, X_k) = E(Y \mid X_i, X_j, X_k) - f_{ij}(X_i, X_j) - f_{ik}(X_i, X_k) - f_{jk}(X_j, X_k) \tag{2.14}$$

The HDMR decomposition is called ANOVA-HDMR and is unique. The $f_i(X_i)$ are referred to as main effects of X_i, the $f_{ij}(X_i, X_j)$ are two–way interactions between the pairs (X_i, X_j), etc.

Squaring (2) and integrating over Ω we get the ANOVA decomposition:

$$V(Y) = \sum_i V_i + \sum_{i<j} V_{ij} + \sum_{i<j<k} V_{ijk} + \ldots + V_{12\ldots k} \tag{2.15}$$

where:

$$V_i = V[f_i(X_i)] = V[E(Y \mid X_i)] \tag{2.16}$$

$$V_{ij} = V[f_{ij}(X_i, X_j)] = V[E(Y \mid X_i, X_j) - f_i(X_i) - f_j(X_j)] = V[E(Y \mid X_i, X_j)\} - V_i - V_j \tag{2.17}$$

$$V_{ijk} = V[f_{ijk}(X_i, X_j, X_k)] = V[E(Y \mid X_i, X_j, X_k)] - V_{ij} - V_{ik} - V_{jk} \tag{2.18}$$

The single terms $V_i, V_{ij}, V_{ijk}, \ldots$, are called partial variances and they are orthogonal. No covariances are involved in the decomposition.

We can define the following sensitivity indices:

$S_i = \frac{V_i}{V(Y)}$ are first order sensitivity indices (also called main effects).

$S_{ij} = V_{ij} / V(Y)$ are second order sensitivity indices (highlighting two–way interactions)

$S_{ijk} = V_{ijk} / V(Y)$ are third order sensitivity indices (highlighting three–way interactions),

and so on. The equality $S_{i_1 \ldots i_s} = 0$ means that $f_{i_1 \ldots i_s}(x_{i_1}, \ldots, x_{i_s}) \equiv 0$. Thus, the functional structure of f(**x**) can be investigated by estimating the indices $S_{i_1 \ldots i_s}$.

2.5.2 Main effects

The term V_i is the expected amount of variance that would be removed from the total output variance, if we were able to learn the true value of X_i within its uncertainty range.

S_i indicates the relative importance of an individual input X_i in driving output variance and can be seen as indicating where to direct effort in the future in order to reduce that uncertainty.

If it were possible to observe one of the X_i, learn its true value exactly, then we would choose that with the largest main effect. Of course, it is very rare that the true value of a given input can be learnt exactly.

Nevertheless, the analysis shows where there is the greatest potential for output uncertainty reduction through new research.

This type of measure is used before conducting a calibration experiment on a given input. A high value for the main effect of a given input, indicates that this input is a good candidate for calibration via observations of the model output.

We can also interpret $1-S_i$ as the minimum value of the expected quadratic loss when we approximate $f(\mathbf{X})$ with the function $E(Y \mid X_i)$. If X_i is important, then the approximating function $E(Y \mid X_i)$ explains much of the variance of $f(\mathbf{X})$ and S_i is high. Simultaneous variation of all the other inputs is acknowledged when estimating these indices.

2.5.3 Joint effects and closed indices

If we approximate $f(\mathbf{X})$ by a two-variable function $E(Y \mid X_i, X_j)$, then the minimum expected quadratic loss is $V(Y)-V[E(Y \mid X_i, X_j)]$, which corresponds to the maximum value of $V[E(Y \mid X_i, X_j)]$. We denote this term as V_{ij}^c (c stands for closed). This term can be interpreted as:

- → the expected reduction of output variance when we have jointly learnt the true value of the pair (X_i, X_j), or
- → the expected fraction of the output variance that is removed when the true value of X_i and X_j is learnt, or
- → the fraction of the output variance that is explained by the approximating function $E(Y \mid X_i, X_j)$.

For orthogonal inputs, and only in this case, we have:

$$V[E(Y \mid X_i, X_j)] = V_i + V_j + V_{ij}. \tag{2.19}$$

Hence, V_{ij} is the fraction of the output variance due solely to the interaction between X_i and X_j. When we learn about both X_i and X_j, then V_{ij} is the extra amount of output variance removed over and above the variances V_i and V_j.

2.5.4 Total effects

The total effect index accounts for the total contribution to the output variation due to X_i, i.e. its first–order effect plus all higher–order effects due to interactions and it is given by the sum of all the sensitivity indices which include the factor in question, not considering the sensitivity indices that do not contain that factor.

For a three–factor model, for example, the total effect of X_1 is:

$$S_{T1} = S_1 + S_{12} + S_{13} + S_{123} \tag{2.20}$$

Total sensitivity indices are useful because they are an overall measure of importance of a given factor. For k factors, it would be very demanding to estimate all indices at any order given that this number is $2^k - 1$, problem known as "*the curse of dimensionality*" (Rabitz, 1999). Total indices can be estimated directly without the need to estimate each term of the decomposition. (Homma and Saltelli, 1996). For this reason we customarily

tend to compute the set of all S_i plus the set of S_{Ti} which gives a fairly good description of the model sensitivities at a more reasonable cost.

2.6 Implications and interpretations of the sensitivity indices

By definition, S_{Ti} is surely greater than S_i (or equal to S_i in the case X_i is not involved in any interaction with other factors).

The difference $S_{Ti} - S_i$ is a measure of how much X_i is involved in interactions with any other factor.

The sum of all the S_i is usually less than 1 (for non–additive models).

The sum of all the S_i is equal to 1 if the model is perfectly additive (no interactions between factors)

The sum of all the S_{Ti} is usually greater than 1 (for non–additive models).

The sum of all the S_{Ti} is equal to 1 if the model is perfectly additive.

An indicator of the presence of interactions in a model is given by $1 - \sum_i S_i$: this value is the fraction of the output variance that is not explained by the single factors.

The difference $\sum_i S_{Ti} - 1$ is another indicator of the presence of interactions, but this indicator weights interactions of higher order much more than interactions of lower order.

The condition $S_{Ti} = 0$ is necessary and sufficient for X_i to be a non influential factor. If $S_{Ti} \cong 0$ then X_i can be fixed at any value within its range of uncertainty without appreciably affecting the value of the output variance $V(y)$.

Another way to define total indices is to decompose the output variance $V(y)$ in terms of main effect and residual conditioning with respect to all the factors but the one of interest, i.e. $\mathbf{x}_{-i}$ (Homma and Saltelli, 1996):

$$V(y) = V(E(y \mid \mathbf{x}_{-i})) + E[V(y \mid \mathbf{x}_{-i})] \tag{2.21}$$

The measure $V(y) - V(E(y \mid \mathbf{x}_{-i})) = E[V(y \mid \mathbf{x}_{-i})]$ is the remaining variance of y that would be left on average if we could determine the true values of $\mathbf{x}_{-i}$, of the $k-1$ remaining factors. The average is calculated over all possible combinations of $\mathbf{x}_{-i}$ since $\mathbf{x}_{-i}$ are uncertain factors and their true values are unknown. Dividing by $V(y)$ we obtain the total effect index for X_i:

$$S_{T_i} = \frac{E[V(y \mid \mathbf{x}_{-i})]}{V(y)} = 1 - \frac{V[E(y \mid \mathbf{x}_{-i})]}{V(y)} \tag{2.22}$$

2.7 The Jansen formula for the computation of sensitivity indices

Variance based methods have assessed themselves as versatile and effective among the various available techniques for sensitivity analysis of model output. Practitioners can in principle describe the sensitivity pattern of a model $Y = f(X_1, X_2, ..., X_k)$ with k uncertain input factors via a full decomposition of the variance V of Y into terms depending on the factors and their interactions. (Saltelli *et al.*, 2009)

In this section we present a measure to compute the main effects and the total effects using the mean–square difference proposed by Jansen *et al.*, (1994).

2.7.1 Assumptions

Jansen *et al.* (1994) assumed a scalar model output Y depending on a number of stochastically independent random input vectors, $Y = f(X_1, X_2, ..., X_k)$ *:*

$$Y = f(X_1, X_2, ..., X_k) \tag{2.23}$$

The function f is deterministic, it is evaluated by simulation and may represent a single output or a combination of outputs. In this case the variability of Y will be characterized by its variance. Jansen *et al.*, (1994) assumed that Y has finite mean and variance; this is guaranteed for instance when f is bounded. The use of the variance as measure of uncertainty has an economic rationale: if the loss caused by a prediction error is proportional to the square of that error, the expected loss is proportional to the variance.

Uncertainty analysis consists of the investigation of the output distribution, given the model and the distribution of the inputs. One may investigate the *full variance*, that is the variance of Y induced by all sources X_i collectively. Let U denote a group of one or more sources of uncertainty X_i*;* then, by assumption, U is independent of all the other sources of uncertainty. With respect to U, two variance components are particularly interesting. Firstly, the *top marginal variance* from U, which is defined as the expected reduction of the variance of Y in case U should become fully known, whereas the other inputs remain as variable as before. Secondly, the *bottom marginal variance* from U, defined as the expected value of the variance of Y in case all inputs except U should become fully known, U remaining as variable as before. Since one does not know in advance at what value the sources will become fixed, one can only determine the distribution of the two variances mentioned; Jansen *et al.*, (1994) content themselves with the mean of these distributions.

2.7.2 Decomposition of the variance caused by two input factors

Jansen *et al.*, (1994) started first from the decomposition of the prediction variance caused by two stochastically independent input factors. The extension to a larger number of independent inputs is shown in the next section.

If an input vector, say U, becomes fully known at value u, whereas the other input vector, say V, independent of U, remains as uncertain as before, the best prediction of Y in the least squares sense will be the mean of $f(u,V)$. The situation is illustrated in Table l for the case that U and V can assume a finite number (4 and 5, respectively) of equiprobable values.

Table 1

	v_1	v_2	v_3	v_4	v_5	
u_1	$\bar{y}_{11}$	$\bar{y}_{12}$	$\bar{y}_{13}$	$\bar{y}_{14}$	$\bar{y}_{15}$	$\bar{y}_{1\cdot}$
u_2	$\bar{y}_{21}$	$\bar{y}_{22}$	y_{23}	$\bar{y}_{24}$	$\bar{y}_{25}$	$\bar{y}_{2\cdot}$
u_3	$\bar{y}_{31}$	$\bar{y}_{32}$	$\bar{y}_{33}$	$\bar{y}_{34}$	$\bar{y}_{35}$	$\bar{y}_{3\cdot}$
u_4	$\bar{y}_{41}$	$\bar{y}_{42}$	$\bar{y}_{43}$	$\bar{y}_{44}$	$\bar{y}_{45}$	$\bar{y}_{4\cdot}$
	$\bar{y}_{\cdot 1}$	$\bar{y}_{\cdot 2}$	$\bar{y}_{\cdot 3}$	$\bar{y}_{\cdot 4}$	$\bar{y}_{\cdot 5}$	$\bar{y}_{\cdot\cdot}$

A dot index indicates that the mean has been taken over the index; $f(u_i,v_i)$ is denoted by y_{ij}. The left column and the upper row contain the values assumed by U and V. The best predictions are conditional means of the model output Y. The bottom right element $y_{\cdot\cdot}$ is the best prediction when neither U nor V are known. The right column and the bottom row contain the best predictions at the given value of U respectively V. The output y_{ij} can be decomposed into general mean, main effects and interactions, as usual in analysis of variance:

$$y_{ij} = y..+(y_i.-y..)+(y._j-y..)+[\,y_{ij}-y..-(y_i.-y..)-(y._j-y..)] \tag{2.24}$$

In the general case, the function $f(u,v)$ can be decomposed as follows. Let f_0 denote the best prediction when U and V are unknown:

$$f_0 = E\ f(U,V) \tag{2.25}$$

The best predictions when U or V have become fixed at u or v respectively, are given by $E\,f(u,V)$ and $E\,f(U,v)$ respectively. Let $f_u(u)$ and $f_v(v)$ denote the corrections to f_0 when U respectively V get fixed:

$$f_u(u) = E\ f(u,V) - f_0 \tag{2.26}$$

$$f_v(v) = E\ f(U,v) - f_0 \tag{2.27}$$

and let $f_{uv}(u,v)$ denote what is left. Then $f(u,v)$ may be decomposed:

$$f(u,v) = f_0 + f_u(u) + f_v(v) + f_{uv}(u,v) \tag{2.28}$$

which is sometimes called the *analysis of variance decomposition of* f. Accordingly f_0 is called the *general mean*, $f_u(u)$ and $f_v(v)$ are called *main effects* of u and v while $f_{uv}(u,v)$ is called the *interaction* of u and v. The *full variance of* f neatly falls apart:

$$\text{Var}\, f(U,V) = \text{Var}\, f_u(U) + \text{Var}\, f_v(V) + \text{Var}\, f_{uv}(U,V) \qquad (2.29)$$

If U were to become fixed at u, the best prediction would be $f_0 + f_u(u)$, leaving $f_v(V) + f_{uv}(u,V)$ as prediction error, with reduced variance $\text{Var}\, f_v(V) + \text{Var}\, f_{uv}(U,V)$, which is a function of u.

It is not known in advance at which value U should become fixed. One might wish to calculate the distribution of this reduced variance, but we will be content with the mean of the reduced variance over U, that is $\text{Var}\, f_v(V) + \text{Var}\, f_{uv}(U,V)$. Accordingly, the *top marginal variance* from U, the expected variance reduction in the output variance due to the fixing of U while V remains as variable as before, is given by $\text{Var}\, f_u(U)$. Similarly, the *bottom marginal variance* from V, the expected variance left over when only V remains uncertain, equals $\text{Var}\, f_v(V) + \text{Var}\, f_{uv}(U,V)$.

The top marginal variance from U is seen to be the variance of the main effect of U, whereas the bottom marginal variance from V is equal to the sum of the variances of the main effect of V and the interaction between U and V.

Let u_1 and u_2 denote two independent realizations of U, and let v denote some fixed value that can be assumed by V. Then $f(u_1,v)$ and $f(u_2,v)$ are independent realizations of $f(U,V)$ given $V = v$. Thus $d \equiv f(u_1,v) - f(u_2,v)$ has zero expectation, while its variance, i.e. its expected square, is twice the variance of $f(U,V)$ given $V = v$. So $\frac{1}{2}d^2$ is an unbiased estimate of this latter variance. It follows that if v is a random realization of V, $\frac{1}{2}d^2$ is an unbiased estimate of what we defined as the bottom marginal variance from source U:

$$\text{Bottom marginal variance} = \frac{1}{2}\int [f(x) - f(u',v)]^2 dx du' \qquad (2.30)$$

Note that the top marginal variance from U is obtained by subtracting the bottom marginal variance from V from the full variance of Y.

$$\text{Var } f_u(U) = \text{Var } f(U,V) - \left(\text{Var } f_v(V) + \text{Var } f_{uv}(U,V)\right) \tag{2.31}$$

The top and bottom marginal variances can be generalized to the case of k input factors as follows (Jansen et al., 1999):

Top marginal variance of X_i: $Var\left[E(Y|X_i)\right]$

Bottom marginal variance of X_i: $E\left[Var\ (Y|X_{-i})\right]$

These are, respectively, the numerators of the first order and total sensitivity indices:

$$S_i = \frac{Var_{X_i}(E_{\mathbf{X}_{-i}}[Y \mid X_i])}{Var(Y)} \tag{2.32}$$

$$S_{T_i} = \frac{E_{\mathbf{X}_{-i}}[Var_{X_i}(Y \mid \mathbf{X}_{-i}])}{Var(Y)} = 1 - \frac{Var_{\mathbf{X}_{-i}}(E_{X_i}[Y \mid \mathbf{X}_{-i}])}{Var(Y)} \tag{2.33}$$

where $\mathbf{X}_{-i}$ denotes the set of all factors but X_i and the mean of Y is taken over all possible values of $\mathbf{X}_{-i}$ while keeping X_i fixed. The outer variance is taken over all possible values of X_i.

In such case, the formula of Jansen for the total indices becomes:

$$S_i^T = \frac{1}{2Var(Y)} \int [f(x) - f(x_i', x_{-i})]^2 dx dx_i' \tag{2.34}$$

This formula has been proven to be more efficient with respect to that proposed by Sobol' (Saltelli et al., 2000, p.177) in the sense that:

$$Var\left(S_{Ti}^{Jansen}\right) \leq Var\left(S_{Ti}^{Sobol'}\right) \tag{2.35}$$

2.7.3 Monte Carlo implementation of the Jansen Formula

Here we describe the Monte–Carlo implementation of Jansen formula for the estimation of the total–effect indices. The procedure, described in Saltelli *et al*. (2010), starts by generating two ($N \times k$) independent sample matrices, $\mathbf{A}$ and $\mathbf{B}$, where N is the base sample size (i.e. the basis for the Monte Carlo computation of the multi–dimensional integral) and k the number of input factors. Usually, values of N is of the order of a few hundreds.

We now introduce another matrix, $\mathbf{A}_{\mathbf{B}}^{(i)}$, where all columns are from $\mathbf{A}$ except the i^{th} column which is from $\mathbf{B}$. Jansen's formula (2.34) for calculating S_{Ti} can be computed from the pair of matrices $\mathbf{A}$ and $\mathbf{A}_{\mathbf{B}}^{(i)}$ as follows:

$$\frac{1}{2N} \sum_{j=1}^{N} \left[f(\mathbf{A})_j - f(\mathbf{A}_{\mathbf{B}}^{(i)})_j \right]^2 \tag{2.36}$$

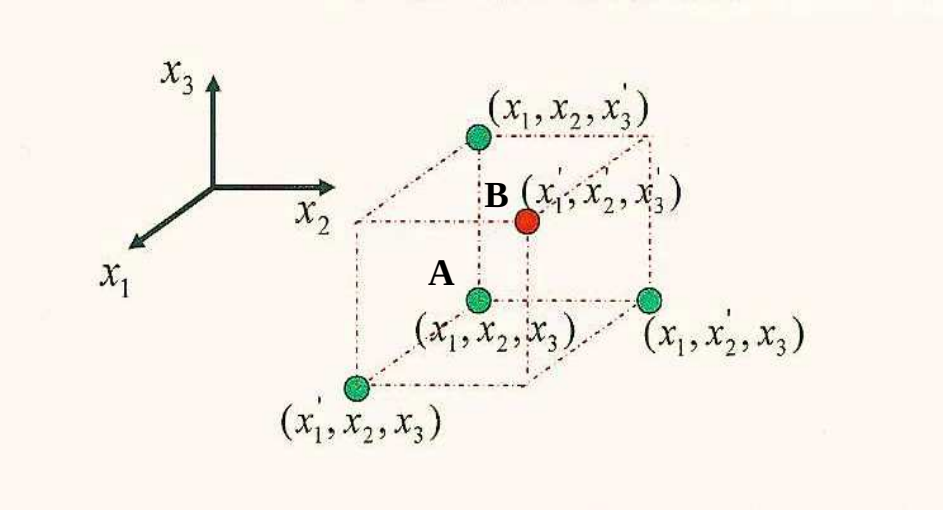
where $(\mathbf{A})_j$ denotes the j^{th} row of matrix $\mathbf{A}$ (Sobol', 1990). The correspondent total sensitivity index is

$$S_{Ti} = \frac{\frac{1}{2N}\sum_{j=1}^{N}\left[f(\mathbf{A})_j - f(\mathbf{A}_{\mathbf{B}}^{(i)})_j\right]^2}{Var(Y)} \tag{2.37}$$

Here the arguments $\mathbf{A}$ and $\mathbf{A}_{\mathbf{B}}^{(i)}$ have in common the coordinates $\mathbf{X}_{-i}$, and can thus be seen as separated by a step in the X_i direction.

Fig. 1 summarize the way in which we can construct the specific design required to calculate the sensitivity indices using the Jansen formula with three input factors. The red point represent matrix $\mathbf{B}$, the vertex of the cube opposite to points represented by matrix $\mathbf{A}$.

Fig.1



Summary of the Jansen formula

The cost of the Monte–Carlo implementation can be summarized as follows.

A number of simulation equal to $(2 \cdot N)$ is needed for computing Y corresponding to matrix $\mathbf{A}$, while $(k \cdot N)$ simulations are needed to compute Y from matrices $\mathbf{A}_{\mathbf{B}}^{(i)}$ for all factors. As a result the cost of the analysis is $N(k+1)$.

CHAPTER 3

TOTAL SENSITIVITY MEASURES FROM GIVEN DATA

3.1 Introduction

Sensitivity analysis often requires a considerable number of model executions, one for each sample point considered. This is especially true when the simulation model is very expensive to run (e.g. one run requires minutes or hours). It is custom practice to try to reduce the number of executions at most, in order to obtain estimates of sensitivity indices of a given accuracy.

The analyst wishing to perform sensitivity analysis can find himself in two different situations: one in which he still has to execute the model runs, and one in which some model runs are already available.

In the first case the analyst can use a suitable design of the input space and adopt, for example, the Sobol' method (or random balance design and other Fourier- based approaches, see chapter 5 for details), to estimate variance-based sensitivity indices. In other words, the analyst decides where to locate the input points, and runs the model on these. Then he uses the model output to compute Sobol' indices.

In the second case, the input points and the corresponding model outputs are "given", i.e. data might come either from measurements or experiments, or from a design that is not specifically intended for sensitivity analysis. Hence, given data are already available and the analyst wishes to use them for the sensitivity analysis in order to save computational time. Therefore, techniques should be available to estimate sensitivity indices for given data. The initial objective of the present thesis is to devise a methodology of this kind.

Some approaches to sensitivity analysis are already available for given data (Paruolo *et al.*, 2011) but they focus on the estimation of first order indices. The objective of the thesis is more general: to devise a "given data" approach for the estimation of total sensitivity indices, which takes into consideration the overall effect of interactions among model inputs. No "given data" methodology is available today for estimating total sensitivity indices.

3.2 Estimating total effects from given data

The Monte Carlo implementation of the Jansen formula, presented in section **2.4.3**, can be used when the analyst can choose the design points and build the matrices $\mathbf{A}$ and $\mathbf{A}_{\mathbf{B}}^{(i)}$. But when data are given (they are indeed represented by matrix $\mathbf{A}$), it is not possible to build $\mathbf{A}_{\mathbf{B}}^{(i)}$ starting from $\mathbf{A}$, as we can not generate another independent matrix, $\mathbf{B}$. Therefore, $\mathbf{A}_{\mathbf{B}}^{(i)}$ needs to be found from the set of given data.

As shown in the previous chapter, each row in $\mathbf{A_B^{(i)}}$ represents a step in the X_i direction with respect to (the rows of) $\mathbf{A}$. We need to find, among the given data, those that are as closest as possible to the X_i direction (i.e. the line passing through $\mathbf{A}$ and $\mathbf{A_B^{(i)}}$).

We can construct the region of search for points $\mathbf{A_B^{(i)}}$ within a cone (see Fig. 4.1) with vertex in $\mathbf{A}$ and axis along the X_i direction.

For each point[1] $\vec{x}_j = (x_{j1}, x_{j2}, ..., x_{jk})$, j=1,…, N of matrix **A,** the best candidate for $\mathbf{A_B^{(i)}}$ is the point of $\mathbf{A}$ that maximizes

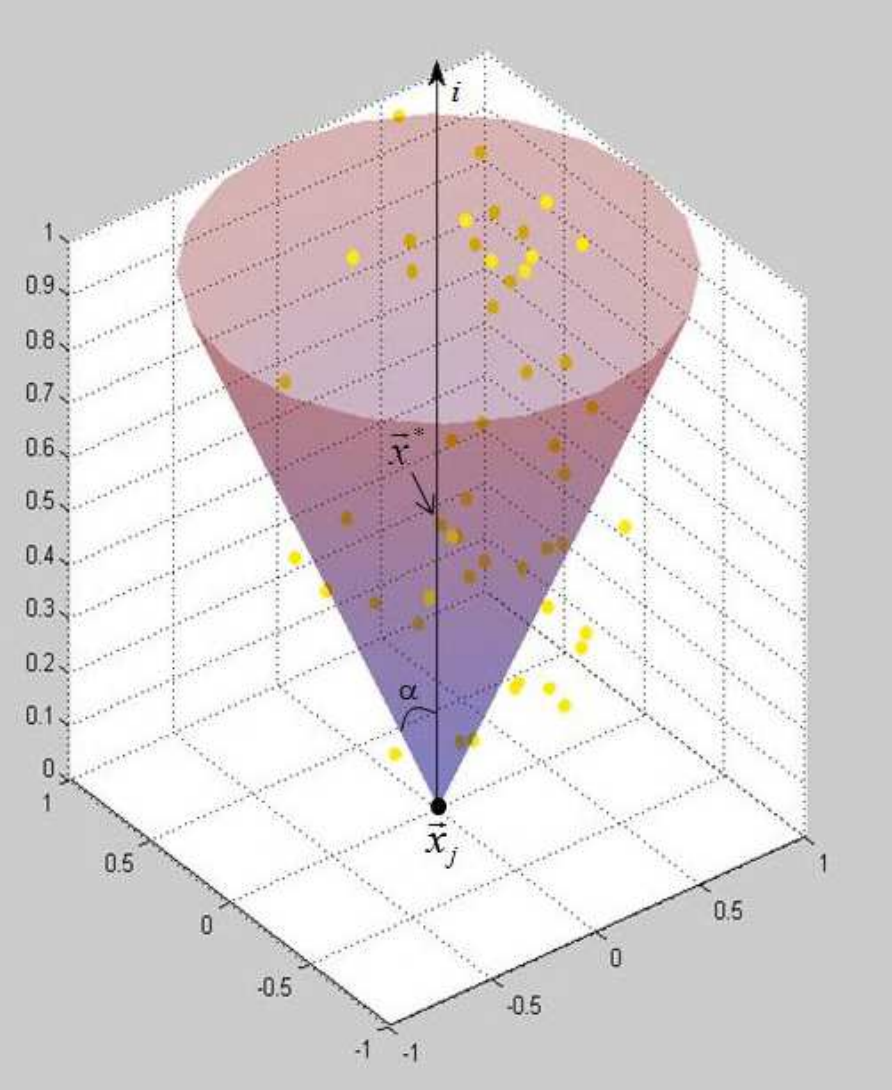
$$cos\,\alpha\,(\vec{x}_{ji}, \vec{x}^{\,*}) = \frac{x_{ji} - x_i^*}{\sqrt{\sum_{i=1}^{k} (\vec{x}_{ji} - \vec{x}^{\,*})^2}} \tag{3.1}$$

with respect to $j = 1, ..., N$.

We can do this by calculating for each point $\vec{x}_j = (x_{j1}, x_{j2}, ..., x_{jk})$ its distance from all the other points of matrix $\mathbf{A}$, and for each direction i (column of $\mathbf{A}$) we compute the angle between each point $\vec{x}_j$ and the projection of the others on the direction i. We want in this way to identify, for each vector $\vec{x}_j$ of matrix $\mathbf{A}$, the vector (point) $\vec{x}^*$ inside the cone, which is as close as possible to the direction of $\vec{x}_j$. For a given cone aperture we expect there are some points around it. We collect all the points with the smallest angle and then we run the model. We proceed then through the Jansen formula for the estimation of the total–effect indices. If there are no points inside the cone, then the point $\vec{x}_j$ is skipped. The aperture of the cone can be increased to improve the likelihood that points fall within it; but, in this case the approximation error would increase, too.

[1] Each point is a row in the matrix. Given the matrix and the output we want to estimate ST_i.

Fig. 4.1



Given the aperture α, for each point $\vec{x}_j$, inside the cone, we choose the one, $\vec{x}^*$, closest as possible to the direction i. The darker dots are the points inside the cone

We coded this approach in a MATLAB script for performing the sensitivity analysis described above. We made several tests of the code to verify the reliability of the results.

In the next section we describe the code in detail and present the results obtained.

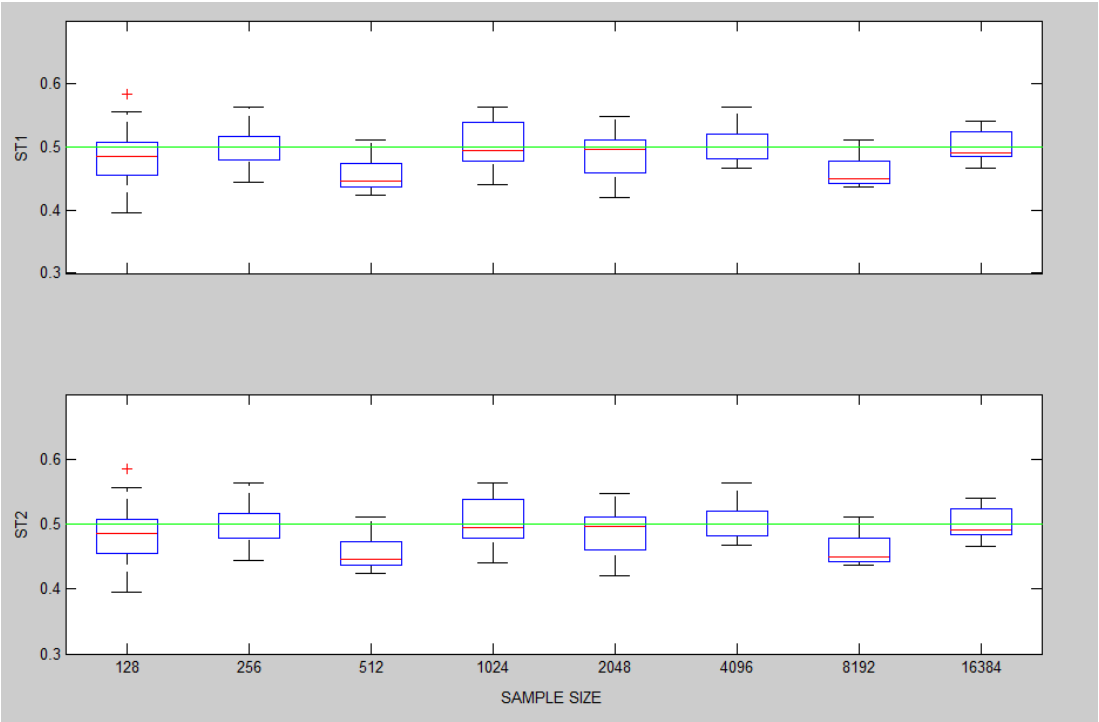
3.3 Total sensitivity indices estimated using the cone approach

We first created a MATLAB code (appendix A.1). Once set the aperture of the cone (alpha), this code is able to capture, for each point of the matrix of interest and for each direction, all the partner points with the "minimal aperture" ($\leq$ alpha). From those points the code selects those which are better than the others, i.e. throwing out the points where the partner is not aligned "straight enough" with the direction of the cone. We then used the Jansen formula to calculate the total sensitivity indices, hence automatically generating a training data index set.

First of all we wanted to test the new code on a simple additive function and so we chose the Corner Peak[2] function. We tested the code at different sample sizes ranging from 128 to 16384 at a step of multiple of two for 100 replicates. We wanted to test its performance at increasing number of input factors: $k = 2$ $k = 3$ $k = 4$.

The results are reported below and plotted in figures for each number of input factors.

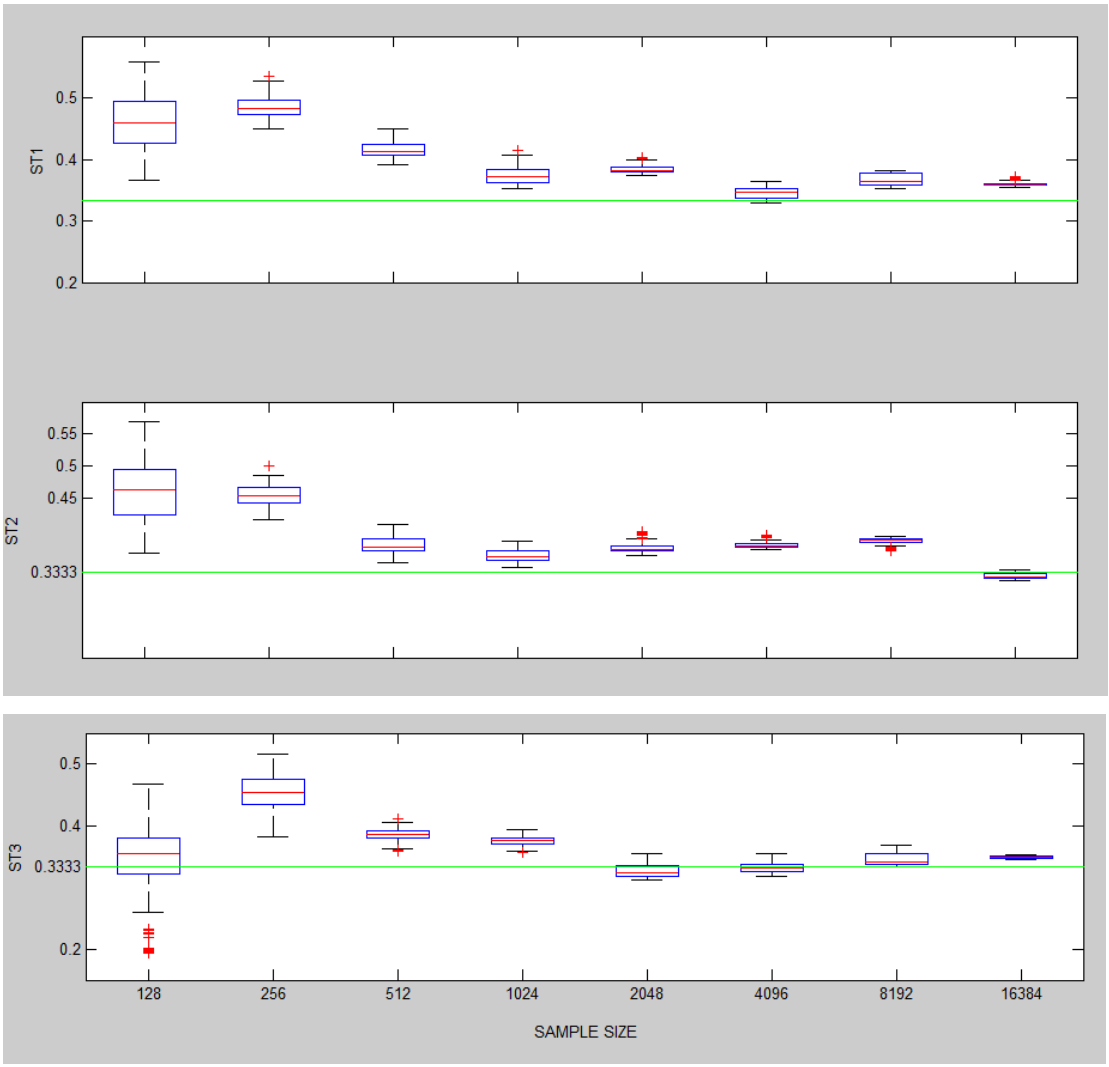
Fig. 4.2



Boxplots of total sensitivity indices against the sample size for the Corner Peak function test case with 2 factors. Cone aperture is set at alpha=5. Analytical value are shown by a green line

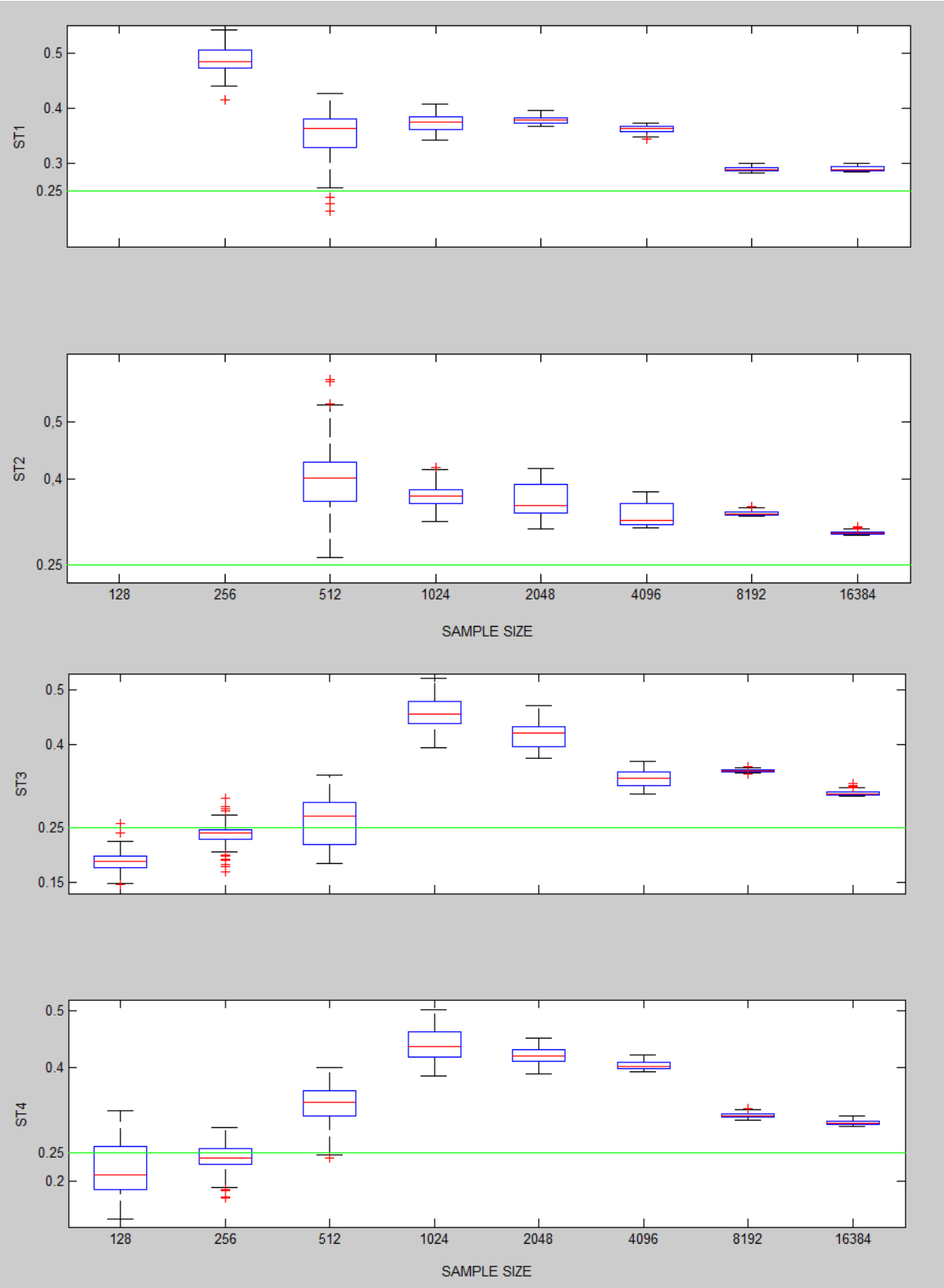
[2] See paragraph 4.5.1 for details on this function

Fig. 4.3



Boxplots of total sensitivity indices against the sample size, for the Corner Peak function test case with 3 factors. Cone aperture is set at alpha=5. Analytical value are shown by a green line

Fig. 4.4



Boxplots of total sensitivity indices against the sample size for the Corner Peak function test case with 4 factors. Cone aperture is set at alpha=5. Analytical value are shown by a green line

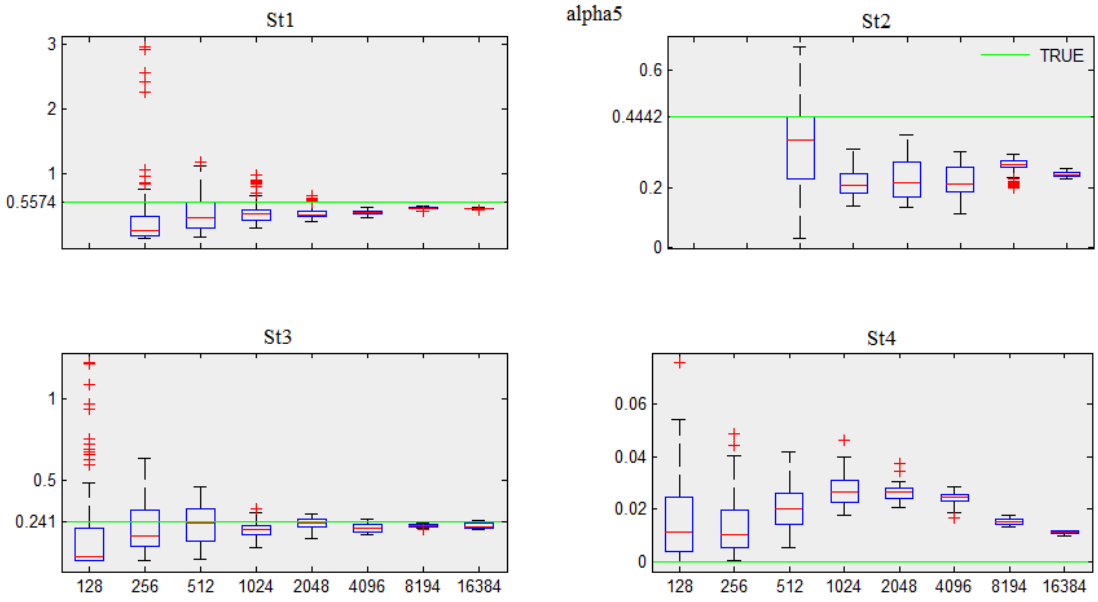
We wanted then to test the performance of the cone approach under different complexity. We chose a non–monotonic test function, the Ishigami[3], using QR numbers. The peculiarity of this function is the absence of additive effect on Y but the presence of interaction between X_3 and X_1.

We tested the code at different apertures of the cone from alpha=5 to alpha=20 at step of 2.5 and at different sample size, from 128 to 16384 at step of multiple of two each of one run for 100 replicates. We added a dummy variable.

The results are reported below and plotted in figures for each cone aperture.

S_{T1} converges to a value under the analytical value and the underestimation grows as the aperture grows and as sample size increases. S_{T3} shows a better convergence than S_{T1} but it is still quite underestimated. S_{T2} shows an evident variability well below the analytical value. S_{T4} has a high variability up to a sample size of 2048 in correspondence of the apertures alpha=5 and alpha=7.5. From alpha=10 it converges to zero as the sample size increases.

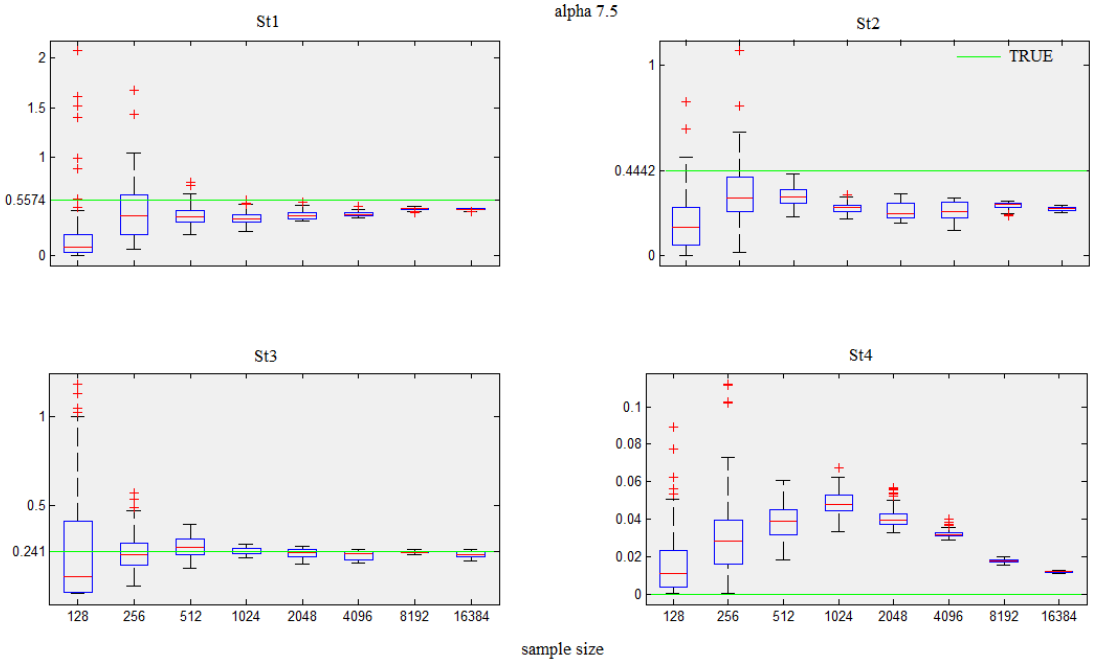
Fig. 4.5



Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=5. Analytical value are shown by a green line

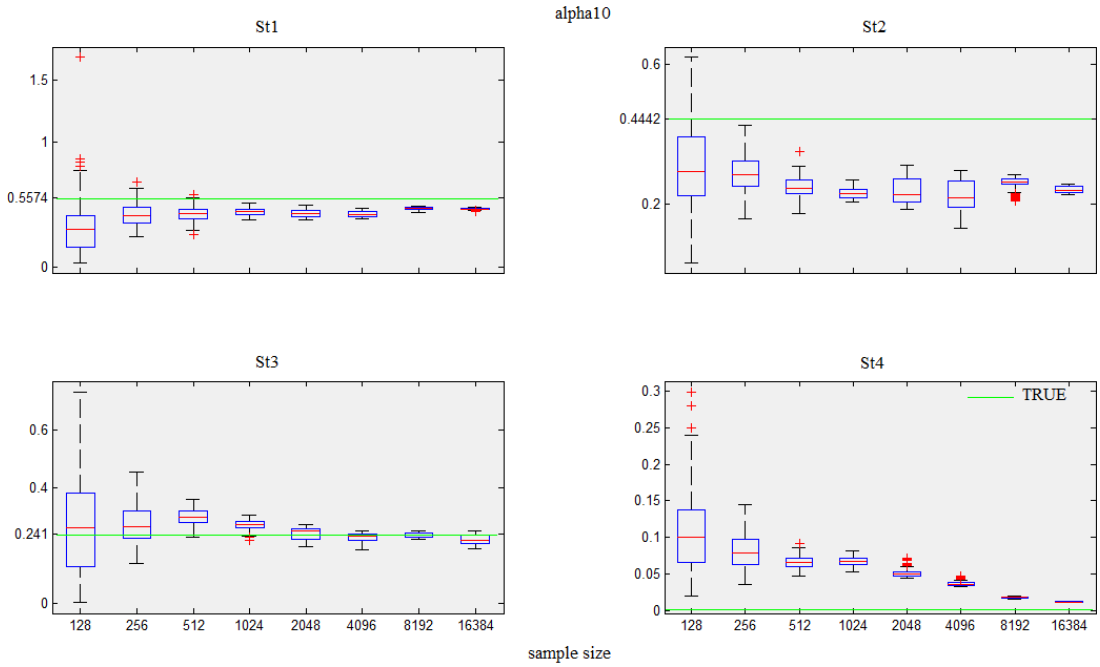
[3] See paragraph 4.5.1 for details on this function

Fig: 4.6



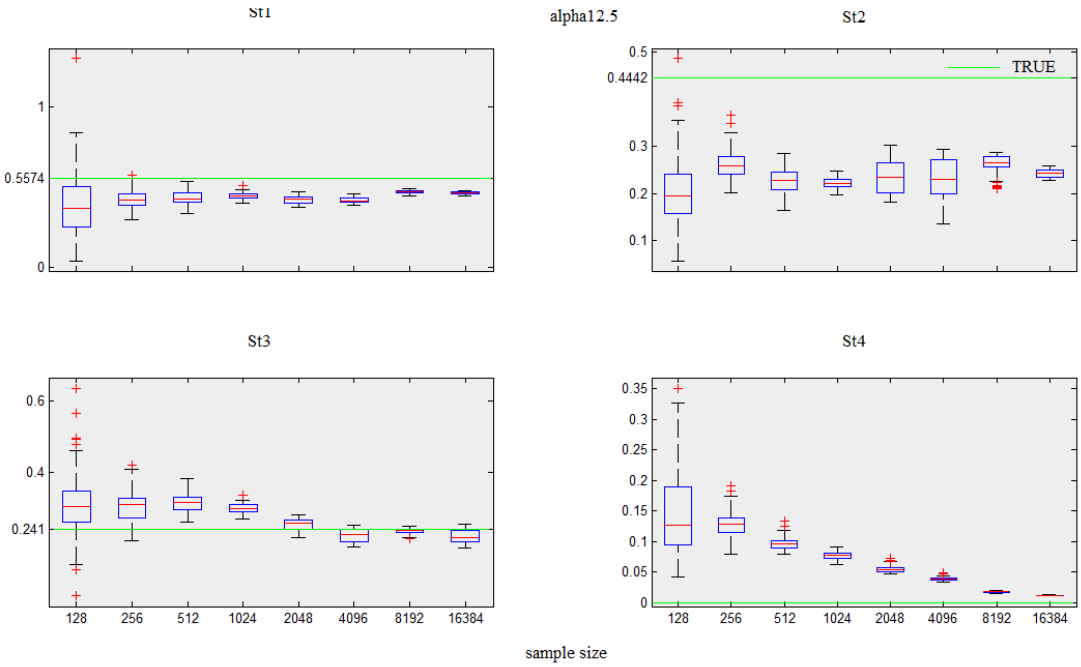
Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=7.5. Analytical value are shown by a green line

Fig. 4.8



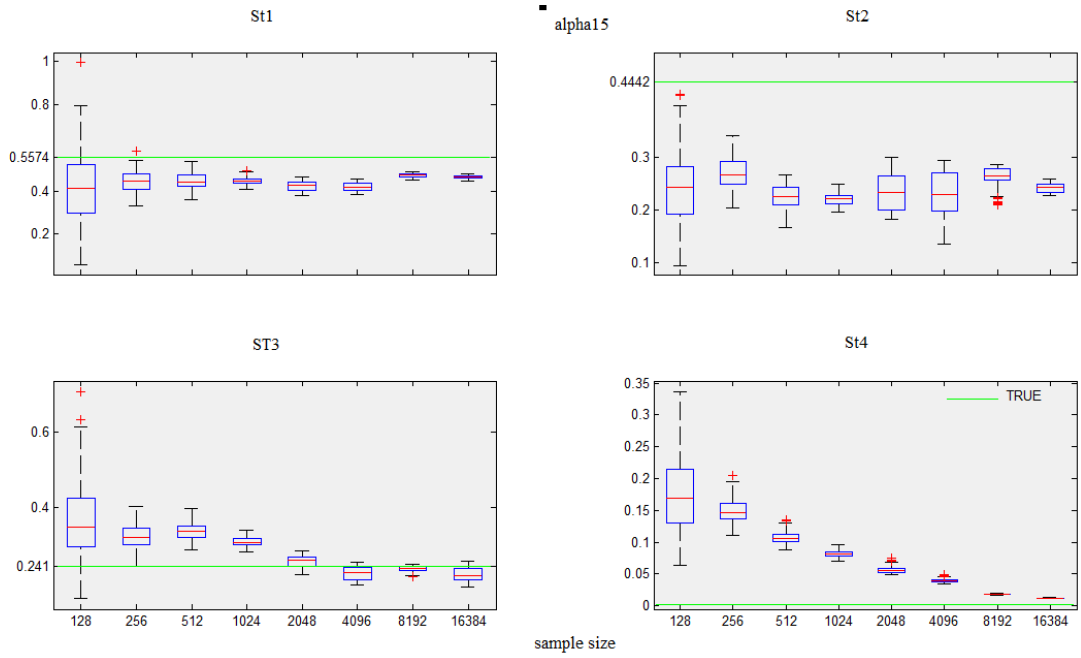
Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=10. Analytical value are shown by a green line

Fig. 4.7



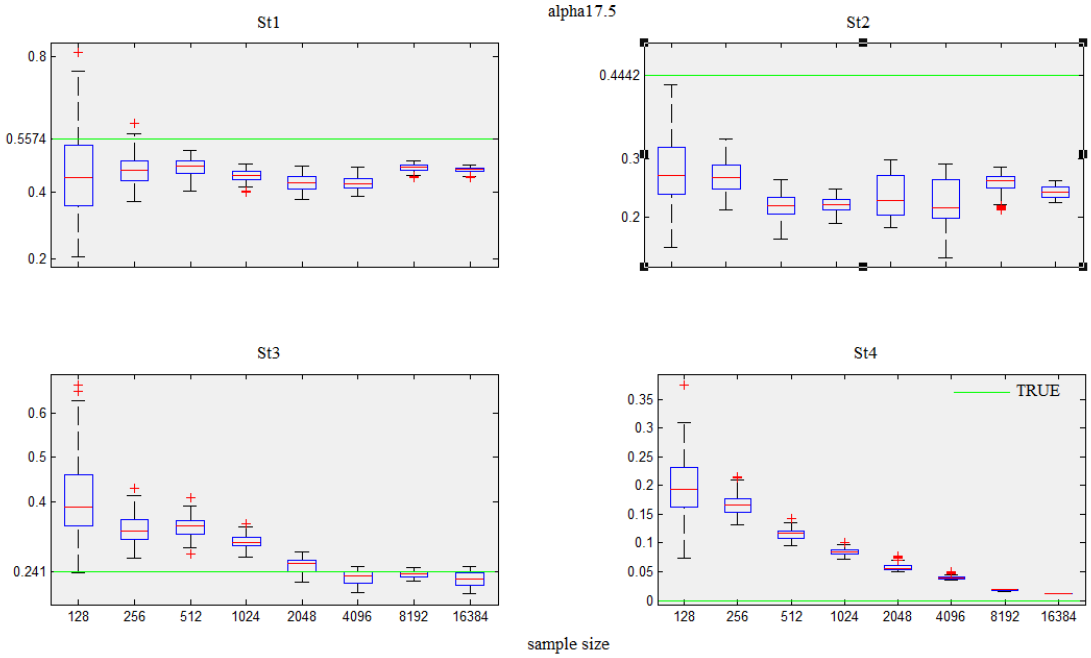
Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=12.5. Analytical value are shown by a green line.

Fig. 4.9



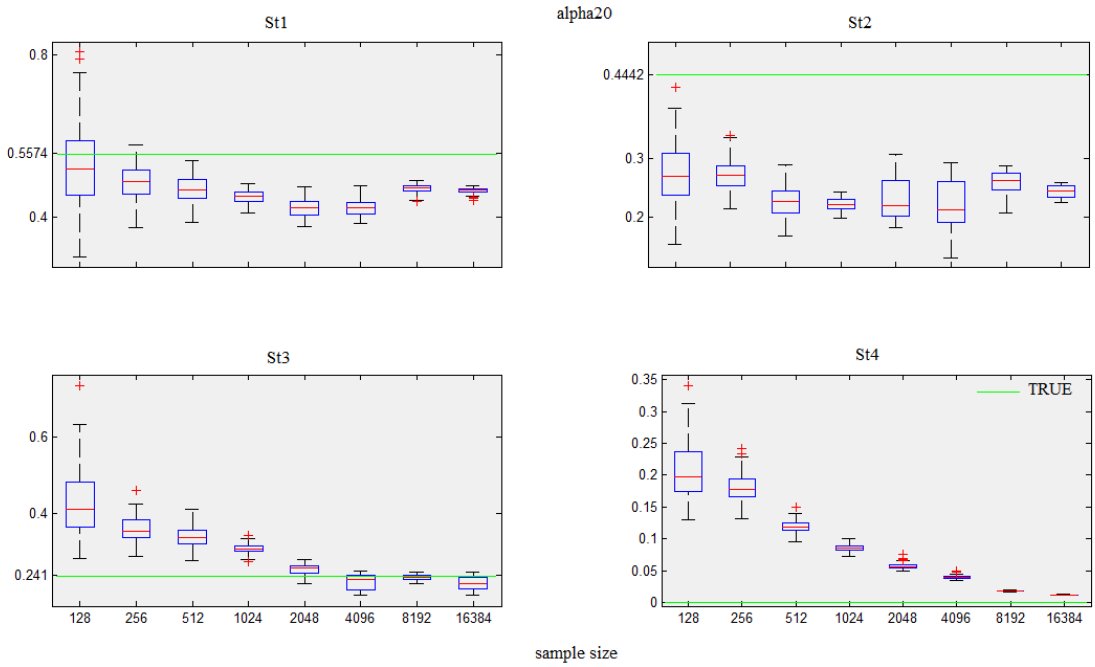
Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=15. Analytical value are shown by a green line

Fig. 4.10



Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=17.5. Analytical value are shown by a green line

Fig. 4.11



Boxplots of total sensitivity indices against the sample size for the Ishigami function test case with 4 factors. Cone aperture is set at alpha=20. Analytical value are shown by a green line.

Since the error bands seem to be underestimanted, we computed the standard deviation (RMSE) for each total index at different apertures and at increasing sample size. We reported the values in the following tables and plotted them in figures (linear and logarithmic scales).

Table 1

S_{T1} RMSE

	SAMPLE SIZE							
ALPHAS	128	256	512	1024	2048	4096	8192	16384
5	NaN	NaN	0,328039	0,237512	0,189313	0,15711	0,09284	0,093044
7.5	NaN	0,313593	0,188105	0,19047	0,157532	0,143551	0,087579	0,09259
10	0,327981	0,166792	0,141566	0,117348	0,130651	0,138187	0,087345	0,092595
12.5	0,255881	0,144303	0,130042	0,11587	0,137096	0,139842	0,08728	0,094583
15	0,209443	0,122059	0,116149	0,112253	0,135702	0,139429	0,08728	0,094583
17.5	0,161419	0,103153	0,088709	0,111127	0,132963	0,133672	0,087373	0,093286
20	0,105726	0,084431	0,094203	0,108926	0,135593	0,136124	0,088111	0,091866

Table 2

RMSE ST2

	SAMPLE SIZE							
ALPHAS	128	256	512	1024	2048	4096	8192	16384
5	NaN	NaN	0,183069	0,231008	0,224038	0,229366	0,169039	0,197598
7.5	NaN	0,198912	0,146638	0,198134	0,215758	0,219881	0,185675	0,201513
10	0,181338	0,164619	0,195832	0,212648	0,208876	0,2274	0,185766	0,201941
12.5	0,247967	0,186332	0,219505	0,222493	0,212235	0,220056	0,183626	0,20114
15	0,211586	0,177738	0,220036	0,223834	0,212329	0,220055	0,183626	0,20114
17.5	0,178119	0,177127	0,22536	0,222236	0,210656	0,225889	0,186949	0,200866
20	0,178556	0,173053	0,219431	0,222162	0,215533	0,230366	0,186483	0,200678

Table 3

RMSE ST3

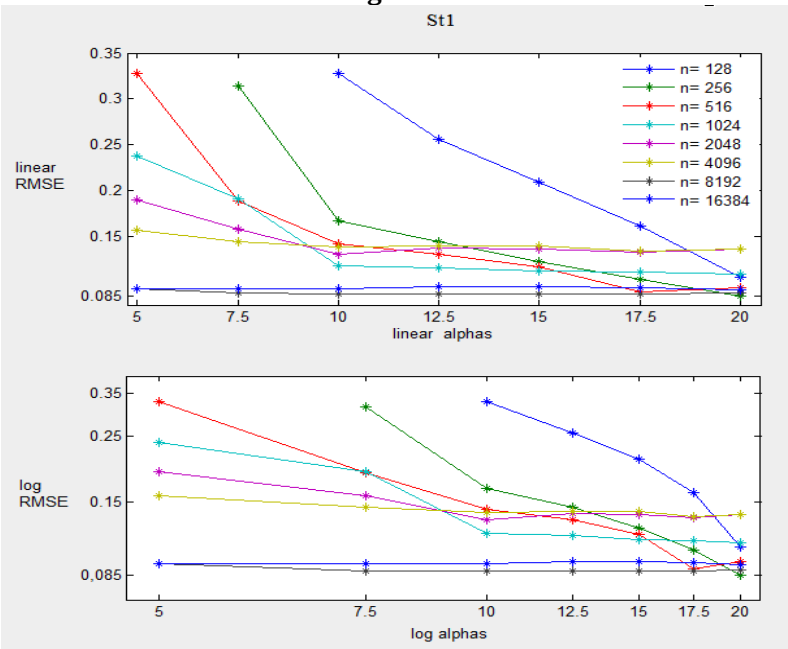
	SAMPLE SIZE							
ALPHAS	128	256	512	1024	2048	4096	8192	16384
5	0,304444	0,167278	0,122508	0,073034	0,037135	0,043697	0,023469	0,028186
7.5	0,330927	0,103788	0,061942	0,017677	0,029258	0,033854	0,0092	0,025832
10	0,167721	0,07493	0,065635	0,034666	0,021834	0,023898	0,008485	0,025807
12.5	0,112686	0,074107	0,078229	0,058457	0,020514	0,025954	0,008686	0,028037
15	0,159332	0,087462	0,098058	0,067879	0,021131	0,025749	0,008687	0,028037
17.5	0,184008	0,101586	0,102094	0,068714	0,021817	0,021568	0,008654	0,023719
20	0,20698	0,123912	0,103203	0,069057	0,019939	0,025448	0,008364	0,02609

Table 4
RMSE ST4

		SAMPLE SIZE							
ALPHAS		128	256	512	1024	2048	4096	8192	16384
5		0,214915	0,180612	0,119395	0,08592	0,057566	0,040115	0,018182	0,011988
7.5		0,024651	0,037717	0,039773	0,049078	0,041317	0,032143	0,017724	0,011855
10		0,123858	0,08529	0,066869	0,068011	0,051673	0,037491	0,018067	0,011993
12.5		0,157889	0,129993	0,09727	0,077454	0,055492	0,039448	0,018055	0,011955
15		0,183238	0,151488	0,107694	0,082334	0,057238	0,040018	0,018056	0,011955
17.5		0,206269	0,166532	0,116383	0,085301	0,058002	0,03991	0,018198	0,01194
20		0,214915	0,180612	0,119395	0,08592	0,057566	0,040115	0,018182	0,011988

When the sample size is small the RMSE decreases as the aperture of the cone grows. This is because the greater the aperture, the better points are captured by the cone. But when the sample size increases, the best points have already been captured by the cone even at small aperture. Increasing the aperture does not change anything. This is well expressed in Fig. 4.12 and Fig. 4.14 at sample size 128 and 16384.

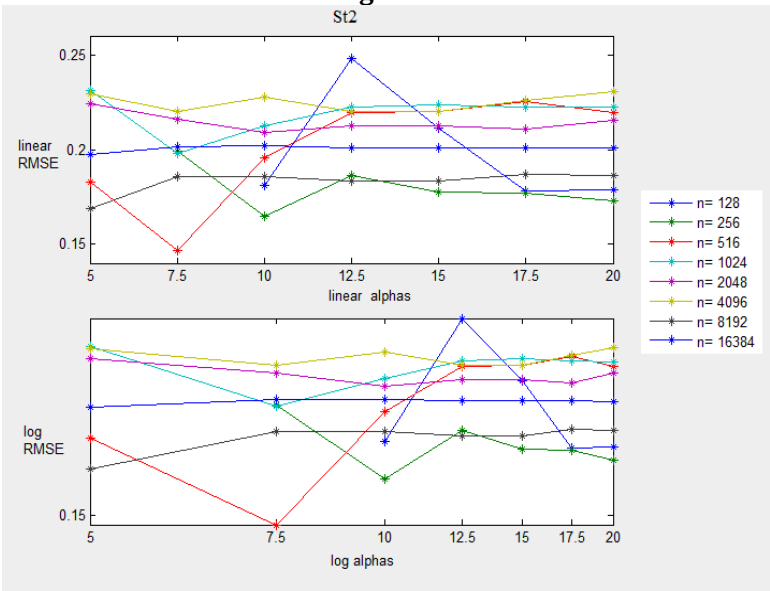
Fig. 4.12



Graphic of the trend of S_{T1} RMSE

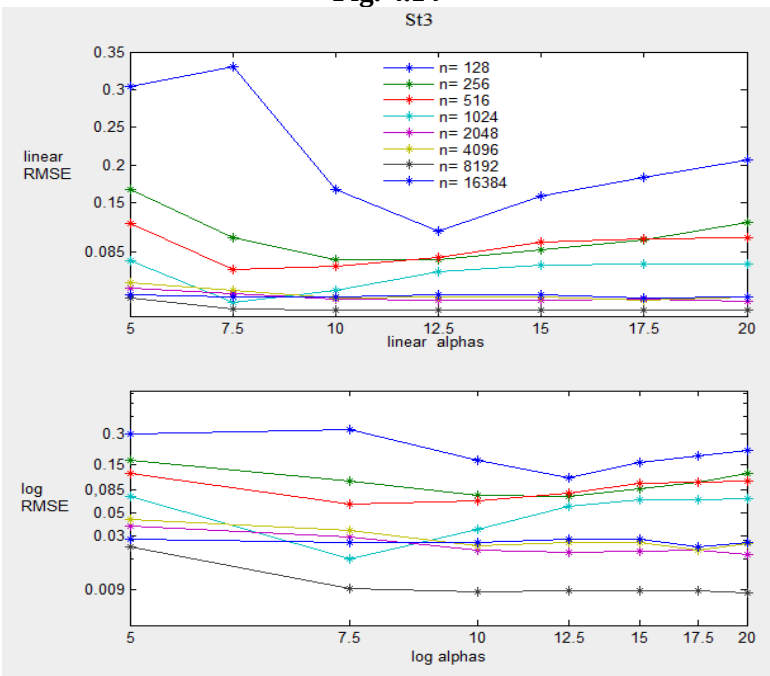
Fig. 4.13 reflects the irregular trend of the index S_{T2}. Apparently, the cone aperture does not influence the estimates except for the smallest sample size (128 and 256).

Fig. 4.13



Graphic of the trend of S_{T2} RMSE

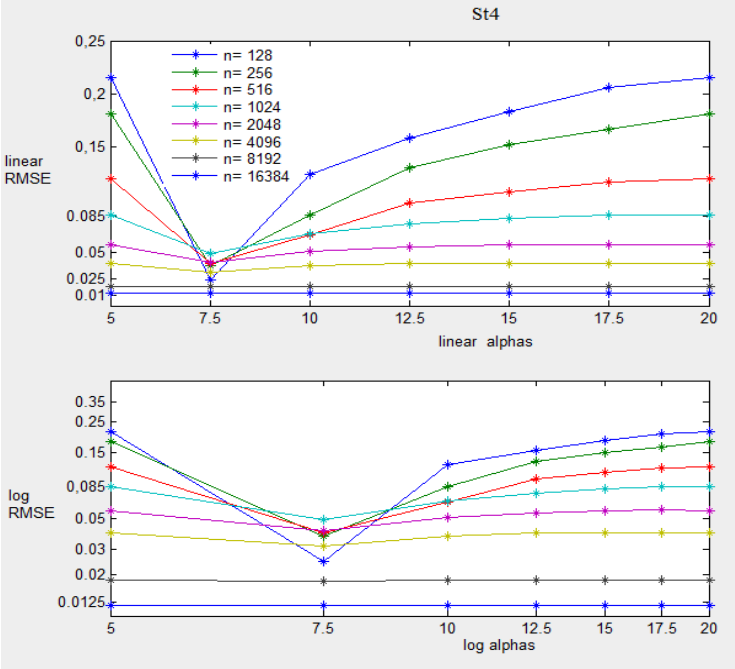
Fig. 4.14



Graphic of the trend of S_{T3} RMSE

Fig. 4.15 shows the perfect trend of S_{T4}. At small sample sizes when, the aperture is alpha=5, the cone captures the best points inside, but we may have empty cones for some points and for some directions. That's why the estimate improves as the cone aperture grows. But from alpha=7.5 the best point have already been captured and the estimate does not change. When the sample size is high (8192 and 16384) the aperture does not influence the estimate because the best points have already been inside the cone.

Fig. 4.15



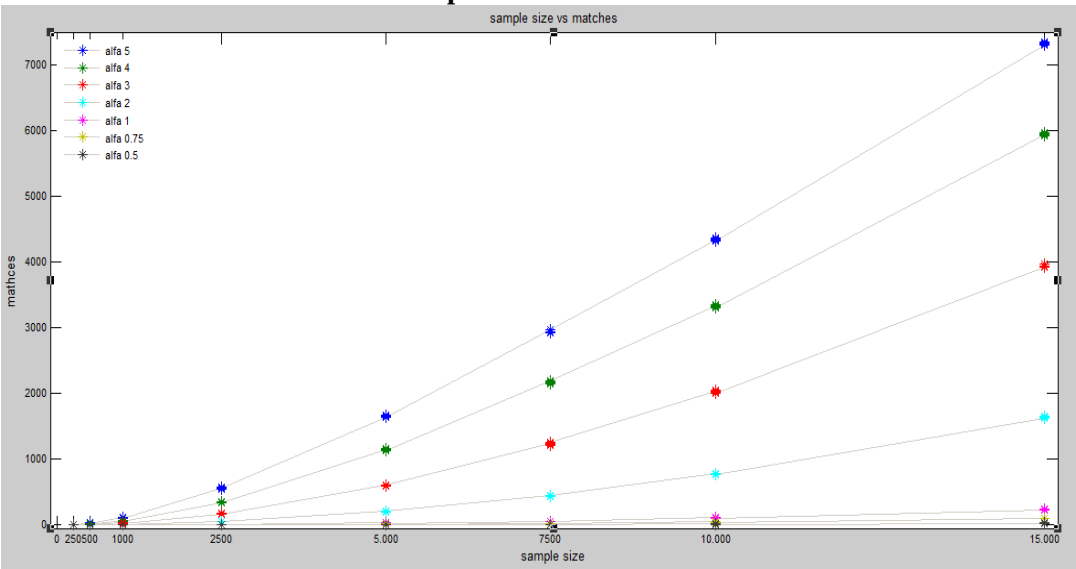
Graphic of the trend of S_{T4} RMSE

To give an idea of how many points we can find inside the cone at different apertures, as the sample size grows, we calculated those points and plotted them in Fig. 4.16.

We called "matches" the average of the number of good points, for each sensitivity index, captured by the cone and used to calculate the four indices altogether.

As the sample size increases Fig. 4.16 shows that the number of good points captured by the cone grows, and it grows to a greater extent with the increasing of the aperture of the cone. At sample size 15000 we found on average 7320 points inside the cone, in correspondence of alpha=5.

Fig. 4.16
sample size vs matches



Graphic of the trend of the number of good points captured by the cone at different aperture as the sample size grows

3.4 Conclusions

In SA the estimation of higher order effects and, in particular, total effects from given data still remains an open problem.

In this chapter we presented a new technique for estimating variance–based total sensitivity indices from given data because we wanted to try to give a contribution for the solution of this problem.

We created a unique algorithm, that we called "***cone approach***", and investigated its use in estimation of total indices. The algorithm has been implemented in a Matlab code for further tests and outcome analysis.

During our work, we specifically focused on total effects estimation, being aware of the concrete possibility of not being able to find a final solution. Thus, the core feature of the study was to investigate and define a possible approach to the problem.

In SA, the Jansen formula is efficiently used to calculate total effects indices when data are designed ad hoc. But when data are given, this formula cannot be applied because of the impossibility to build other independent points along the same direction of the one of interest. The cone approach tries to overcome this basic limit, searching the best points to be used in the Jansen formula on the base of their radial "distance" from the needed direction, i.e. those points X_i, whose angle between them and the projection of the others on the direction i is the smallest.

The Jansen formula was then applied to obtain numerical results to be analyzed in detail. In our knowledge no previous studies followed this method and our tests represent the first tentative for this kind of analysis.

Matlab software was used to create new scripts for numerically implementing our approach and made several tests.

We first tested the cone approach with a simple additive function, the Corner Peak function and performed it at a given aperture of the cone, alpha=5. The method gave good estimates especially when the number of input factors were small. Fig. 4.3 showed that when $k = 4$ the estimates show higher variability and overestimation above the analytical value. The performance of the cone approach is worsening with increasing of the number of input factors.

We noticed that the cone approach performance does not completely meet our expectations when the complexity of the function used is higher and the number of the factors increase. The numerical results, obtained performing the Ishigami function, unfortunately, confirmed that.

So we made some reflections.

Does a larger cone aperture increase the risk of augmenting the errors? Are the best points identified by the code good for increasing the accuracy of the estimates? We found that this depends on the test case and the input factor analyzed. In the case of complex functions, why the indices do not converge to analytical value as the sample size grows? Perhaps, there could be the risk that a considerable number of new points are located close to the boundary of the cone and can therefore introduce further error.

With a large sample, and with a small aperture of the cone, the small amount of points inside the cone (even two or three) are very good and the results are good. Increasing the aperture, we run the risk of including other points which deteriorate the estimate. So, apparently there is no optimal number of points that yields a good estimate.

Despite these results, in our opinion the defined "***cone approach***" still remains a valid approach to be investigated and studied. It is in our intentions to make further research on it.

We first intend to test the code implemented in MATLAB for the present thesis, on other test functions at increasing complexity. We want to understand the way in which the performance of the total indices changes when we are dealing with different situations: different sample sizes and different apertures.

Furthermore we intend to review the code. Maybe we can improve it or make it more powerful.

We are working on the cone approach and we will work on it until we find answers to our questions.

CHAPTER 4

QUASI RANDOM BALANCE DESIGNS

In this chapter we present a new approach for the estimation of the first order effects in GSA given a specific sample design. This method adopts the RBD approach published by Tarantola *et al*., (2007) for the computation of first order sensitivity indices in association to Quasi–Random numbers. For this reason we first introduce the Random Balance Design and the Quasi Random numbers.

4.1 An introduction to Random Balance Designs

4.1.1 History

Designed experiments are used in a very wide range of applications (industrial, biological and agricultural experiments) since they aim at evaluating the performance of

a system, or optimizing its performance in terms of one or more output responses. Typically these experiments involve many potentially important factors of which only a few are expected to have active effects and those effective factors are not known a priori. In such experiments, the experimenter's endeavor is to minimize the number of runs to identify the active factors, those having a strong effect on the output, for efficient utilization of resources and minimization of cost and time. The basic problem here is how to identify these few active factors in an efficient way. It is impossible to investigate thoroughly all factors under consideration and knowing every main effect can be wasteful because non–significant factors are not usually of interest. We require some means of making the available number of computer runs and the number of factors compatible.

We can succinctly summarize this difficulty of experimental design in simulation as too many factors and too few runs. This situation where many effects are unimportant is called *effect–sparsity* (Box & Meyer, 1986) and assumed that only a few dominant effects actually affect the response.

To address this problem, one approach is to use a so–called *supersaturated design*, an increasingly popular tool for screening factors in the presence of *effect–sparsity*. A supersaturated design is namely an experimental design whose run size is not large enough for estimating all the main effects represented by the columns of the design matrix, that is a design with k factors and n observations where $n < k$ (the number of runs is smaller than the number of factors). If a first–order model is assumed (i.e. a model without interactions between factors) and if the number of significant factors is expected to be small, a supersaturated design can save considerable cost.

The advantage of these designs is that they reduce the experimental cost drastically. Because of their run size economy, these designs can be broadly exploited to screen active factors when experimentation is expensive and the number of potentially active factors is large.

The construction of supersaturated designs dates back to Satterthwaite (1959) and Booth&Cox (1962). The former suggested constructing such designs by randomization procedure (random balance designs).

4.1.2 An overview of the approach

In an experimental design, given two variables x_1 and x_2, we say that "x_1 has *exact balance* with respect to x_2 if the exact distribution of x_2 is the same for each value of x_1" (Satterthwaite, 1959).

A *random design* is one for which "a random sampling process is used to choose all or some of the elements of a design matrix" $X=[x_1,x_2]'$. If we carry out a series of experiments, good practice would be to conduct them in random order. The order of those experiments is in fact a further variable, i.e. x_3 in addition to the previous x_1 and x_2. Our design matrix X will have then three variables and its third column, the order numbers of the experiment, are selected by an appropriate random sampling process. Here the design is a random design with respect to x_3, the order variable but it is a fixed design with respect to the other variables x_1 and x_2.

The latter variable x_3 is said to have *random balance* with respect to x_1 and x_2 "if the random sampling process used to select the x_3 values associated with any one combination of (x_1, x_2) is identical to the random sampling process used to select the values of x_3 associated with every other combination of (x_1, x_2), no matter of the type of random sampling process used to select the specific input variable values" (Satterthwaite, 1959).

This sampling technique is used in the random balance design method, illustrated in the next sections. This method combines Satterthwaite random balance design for the sampling with the estimator used for the Fourier Amplitude Sensitivity Test (FAST), another technique of sensitivity analysis. Therefore, before describing the random balance design method, we provide an introduction to FAST.

4.2 The FOURIER AMPLITUDE SENSITIVITY TEST

The Fourier Amplitude Sensitivity Test (FAST) is one of the earliest method developed to estimate variance–based sensitivity indices and it offers an alternative approach to compute first–order effects.

FAST was originally proposed by Cukier *et al*. in the 1970s to perform sensitivity analysis of a chemical computer model (Cuckier *et al*, 1973). FAST is computationally efficient and it is independent of any assumption about the model structure; furthermore it works for monotonic and non–monotonic models. The core feature of FAST is that it explores the multidimensional space of the input parameters by a search curve which scans the whole parameters' space. The multidimensional integrations over the input space is thus replaced by a one-dimensional integration. The sensitivity coefficients of FAST are calculated from the terms in the Fourier decomposition of the model output. Unlike other global SA methods such as the Standardized Regression Coefficient (SRC), Correlation Ratio (Pearson) or Partial Correlation Coefficient (PCC) mentioned previously, FAST computes the "main effect" contribution of each input factor to the variance of the output estimating the same statistical quantity given by

$$\frac{\mathrm{Var}_{Xi}\left[E(Y \mid X_i)\right]}{\mathrm{Var}(Y)} \tag{4.1}$$

where Y denotes the output variable, X_i denotes an input factor, $E(Y \mid X_i)$ denotes the expectation of Y conditional on a fixed value of X_i and the variance is taken over all possible values of X_i.

4.2.1 The method

The main idea behind FAST is to convert the $k-$dimensional integral over X into one–dimensional integral in s by using appropriate transformation functions.

Let us consider the following model $y = f(\mathbf{x})$ where y is the output variable and $\mathbf{x} = x_1, x_2, ..., x_k$, $\forall i = 1, ..., k$, is a random vector with a given joint probability density function $p(\mathbf{x}) = p(x_1, x_2, ..., x_k)$ assumed to be known, the output y therefore is also a random variable.

The r^{th} moment of y is defined as a multi–dimensional integral:

$$\left\langle y^{(r)} \right\rangle = \int_{\Omega} f^{r}(x_1, x_2, ..., x_n)\, p(x_1, x_2, ..., x_n)\, d\,\mathbf{x} \tag{4.2}$$

where $\Omega = (\mathbf{x} \mid 0 \le x_i \le 1; \quad i = 1, ..., k$ is the domain of the input factors (for simplicity chosen as the unit hypercube).

Cukier *et al*. (1978) started from the integral in (4.2) to compute sensitivity indices using a multidimensional Fourier transformation of f to decompose the variance of y. Because of the computational complexity of the multi-dimensional integration, they perform the Fourier analysis along a search curve that explores the input domain and that can be defined by the following set of parametric equations

$$X_i(s) = G_i\left(\sin(\omega_i s)\right) \qquad \forall i = 1, 2, ..., k, \tag{4.3}$$

where G_i are transformation functions that will provide a uniformly distribution in the unit hypercube, s is a scalar variable varying over the range $-\infty < s < \infty$ and

$\{\omega_i\}, \forall i = 1, 2, ..., k$ is a set of different frequencies each associated with an input factor. The x_i are now expressed in terms of the parametric equations G_i.

As s varies, all the parameters change simultaneously along the curve and each x_i oscillates periodically at the corresponding frequency ω_i whatever G_i is. The output y will present different periodicities combined with the different frequencies ω_i and it is highly influenced by the i^{th} parameter if the amplitude of the periodic oscillation of y at frequency ω_i is high. Through a Fourier decomposition of $f(s)$[4] we can see the contribution of each individual x_i to the total output variance.

The search curve drives arbitrarily close to any point x of the input domain if and only if we use a set of incommensurate frequencies; in this case the search curve is space–filling according to the ergodic Weyl's theorem (1938) and the r^{th} moment of y in (4.1) can be computed by an integral over one-dimensional domain

$$\bar{y}^{(r)} = \lim_{T\to\infty} \frac{1}{2T} \int_{-T}^{+T} f^{\,r}(x_1(s),\ x_2(s),...,\ x_n(s))\ ds \tag{4.4}$$

Thanks to Weyl's theorem (Whey, 1938), which implies the equivalence between the function expressed in terms of the x_i and the parameter equation expressed in terms of s, we can write

$$\left\langle y^{(r)} \right\rangle \equiv \bar{y}^{(r)} \tag{4.5}$$

[4] From now on $f(s) = f(x_1(s), x_2(s), ..., x_n(s))$

The output variance $V(Y)$ of the model can be computed by evaluating one–dimensional integrals as follows

$$V(Y) = \left\langle \mathrm{y}^{(2)} \right\rangle - \left\langle \mathrm{y}^{(1)} \right\rangle^2 \equiv \bar{y}^{(2)} - (\bar{y}^{(1)})^2 \tag{4.6}$$

The space–filling property is only an idealization as the frequencies ω_i cannot be in practice truly incommensurate. The finite precision of computers allows the use of rational numbers for the frequencies, i.e. a commensurate set of frequencies. When such a set is used, there exists a finite positive rational number T, such that:

$$f(s) = f(s+T) \tag{4.7}$$

i.e. the curve describes a closed path. This introduces an approximation to Weyl's theorem, which means that equation (4.5) no longer holds.

Cukier *et al*. (1973) showed that if the ω_i's are positive integers, $T = 2\pi$. Since $f(s)$ is a periodic function of s with period 2π within the finite interval $(-\pi, \pi)$, equation (4.4) becomes

$$\bar{y}^{(r)} = \frac{1}{2\pi} \int_{-\pi}^{+\pi} f^{r}(s)\, ds \tag{4.8}$$

and the variance of the output is estimated as follows

$$\hat{V}(Y) = \bar{y}^{(2)} - (\bar{y}^{(1)})^2 = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f^{\,2}(s)\; ds - \left[\frac{1}{2\pi}\int_{-\pi}^{+\pi} f\,(s)\; ds\right]^2 \tag{4.9}$$

We now expand $f(s)$ in a Fourier series

$$y = f(s) = \sum_{j=-\infty}^{+\infty} \left\{ A_j \cos js + B_j \sin js \right\} \tag{4.10}$$

where the Fourier coefficients A_j and B_j are defined by

$$A_j = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f\,(s)\cos js\, ds \tag{4.11}$$

$$B_j = \frac{1}{2\pi}\int_{-\pi}^{+\pi} f\,(s)\sin js\, ds \tag{4.12}$$

over the domain of integer frequencies $j \in Z = \{-\infty,...,-1,0,1,...,+\infty\}$. The spectrum of the Fourier series expansion of $f(s)$ is defined as $\Lambda_j = A_j^2 + B_j^2$ with $j \in Z$. Since $f(s)$ is a real–valued function A_j, B_j and Λ_j have the following properties

$$A_{-j} = A_j, \quad B_{-j} = -B_j, \quad \Lambda_{-j} = \Lambda_j \tag{4.13}$$

By evaluating the spectrum for the fundamental frequencies ω_i and its higher harmonics $p\omega_i$, $p=1,2,3,...$ we can estimate the portion of output variance arising from the uncertainty of input factor X_i

$$\hat{V}_i = \sum_{p\in Z^0}^{+\infty} \Lambda_{p\omega_i} = 2\sum_{p=1}^{+\infty} \Lambda_{p\omega_i} \tag{4.14}$$

Thanks to Parseval's theorem, a Fourier–based estimate of the total variance can be obtained:

$$\hat{V}(Y) = \sum_{j\in Z^0} \Lambda_j = 2\sum_{j=1}^{+\infty} \Lambda_j \tag{4.15}$$

The ratio $V_i/V(Y)$ is the estimate of the main effect in FAST. Its magnitude reflects the influence that the i^{th} input factor has on the output and it does not in principle depend on the choice of the set of frequencies used in the computations.

4.3 The RANDOM BALANCE DESIGN method

The RBD method was proposed to overcome the computational cost of the classical FAST, which increases with the number of factors. In fact, RBD remains computationally cheap for the estimation of first–order sensitivity indices even for models with many factors.

The RBD procedure combines Satterthwaite's sampling designs (Satterthwaite, 1959) with the Fourier estimator used in FAST (Tarantola *et al*., 2006).

In classic FAST a quite complex algorithm is needed to set the frequencies such that they are free of interferences, to avoid bias of the sensitivity estimates. The sampling design of RBD overcomes this problem. In the RBD method, we explore the input space using a unique frequency ω and input variables are distinguished by taking random permutations of the coordinates of the sample points to explore as much as possible the entire input space (otherwise the curve would explore only the diagonal of the input space). Usually the frequency ω is an arbitrary integer, set to 1 for simplicity. Let us select N points on variable s over $(-\pi, \pi)$. Let $s_{i_1}, s_{i_2}..., s_{i_N}$ denote the random permutations on the set $\{1, ..., N\}$, the experimental design $X_i(s_{ij})$ is

$$X_i = G_i(\sin(\omega s_{ij})), \quad \forall\, i = 1,..., k \quad \text{and} \quad \forall\, j = 1,..., N \tag{4.16}$$

that provides a different permutation for each factor X_i. The model is then run N times over the sample size:

$$Y(s_j) = f(X_1(s_{1j}), X_2(s_{2j}),..., X_k(s_{kj})) \quad \forall\, j = 1,..., N \tag{4.17}$$

The values of the model output $Y(s_j)$ are reordered, $Y^R(s_j)$, such that the corresponding values $X_i(s_{ij})$ are ranked in increasing order. By doing so the harmonic content of X_i propagates through

$$Y \quad \text{to} \quad Y^R(s_j) \tag{4.18}$$

The sensitivity of Y to X_i is quantified by the Fourier spectrum of the reordered model output:

$$F(\omega) = \left| \frac{1}{\pi} \sum_{j=1}^{N} Y^R(s_j) \exp(-i\,\omega\, s_j) \right|^2 \tag{4.19}$$

evaluated at $\omega = 1$ and its higher harmonics $\omega = 2$, $\omega = 3$ up to a maximum M compatible with the sample size N:

$$\hat{V}_i = \mathrm{Var}[E(Y \mid X_i)] = \sum_{l=1}^{M} F(\omega)|_{\omega=l} = \sum_{l=1}^{M} F(l) \tag{4.20}$$

where $\hat{V}_i$ is an estimate of the main effect V_i, i.e. the nominator of the main effect for factor X_i. This procedure is repeated for all the other factors, whereby the same set of model outputs is simply reordered according to $X_i(s_{ij})$ and (4.19) and (4.20) are used to estimate V_i, $i = 2,...,k$.

With the use of permutations, the total cost is kept down to N, instead of $\sim k * N$ (like in Sobol' and FAST)

With respect to the FAST method, the RBD one has several advantages: The main advantage is that it is relatively easy to implement, and the sample size N, being independent of the number of factors *k*, can lead to a considerable saving in computer time for expensive models.

A disadvantage of the RBD method is that it allows the computation of first–order terms only; we can use the sum of these to check if the model is additive. If the sum is noticeably smaller than 1, we must use another method to compute interactions or total–effect terms.

4.4 QUASI RANDOM NUMBERS

Many problems in numerical analysis are concerned with high dimensional integrals and sensitivity analysis is one of them. While the regular grid methods are very efficient for low dimensional integrands, they become computationally impractical when the number of dimensions increases and thus the number of required integrand evaluations grows exponentially. This effect is known as "*the curse of dimensionality*". (Thompson *et al*., 1998).

The Monte Carlo (MC) method provides a direct approach for performing simulation and integration. It is simple, direct and easy to use. MC integration converges at a rate $O\left(1/N^{1/2}\right)$, where N is the number of sampled points, that is independent of the dimension of the integral. For this reason is the only viable method for a wide range of high–dimensional problems. The price for its robustness is that the rate of convergence attained by MC is rather slow. (Metropolis, 1987)

The result of this combination of ease of use, wide range of applicability and slow convergence, is that an enormous amount of computer time is spent on MC computation. (Caflish, 1998)

A higher rate of convergence can be obtained by using deterministic uniformly distributed sequences, so-called Quasi–Random (QR) sequences. Methods based on the usage of such sequences are known as Quasi Monte Carlo (QMC). Asymptotically, QMC can provide the rate of convergence $O\left(1/N\right)$. (Caflish, 1998)

QR sequences are a deterministic alternative to random sequences (Kuipers and Niederreiter, 1974; Hua and Wang, 1981; Niederreiter, 1992; Zaremba, 1968). QR sequences are designed to provide better uniformity than random sequences, and hence higher rate of convergence. Uniformity of a sequence is measured in terms of its *discrepancy* (see below) and for this reason QR sequeces are also called *Low Discrepancy Sequencies* (LDS). Some have objected to the name “Quasi–Random” since these sequences are intentionally not random.

For sufficiently large N, QMC should always outperform MC. However, in practice such sample sizes quite often are infeasible, especially when high dimensional problems are concerned. Many numerical experiments demonstrated that the advantages of QMC can disappear for high–dimensional problems. There were claims that the degradation in performance of QMC occurs at $n \leq 12$ (Bratley, 1992). In contrast, other papers reported the superiority of QMC over MC for some integrands with $n = 360$ (Paskov, 1995). Some explanations for such inconsistent results were given using the notion of the effective dimension (Caflisch, 1997). This notion is based on the ANalysis Of Variance (ANOVA). It was shown how the ANOVA components are linked to the effectiveness of QMC integration methods (Lemieux, 2000).

The efficiency of MC methods is determined by the properties of the random numbers (Kucherenko, 2012) but the limiting factor in accuracy is that samples generated randomly tend to have clusters and gaps (Saltelli *et al*., 2008). The reason is that subsequent points are generated independently. Since subsequent points know nothing about each other there is some small chance that they will lie very close together, as new points are added randomly. They do not necessarily fill the gaps between previously generated sampled points. Where a cluster of points occurs, function values in that neighborhood are overemphasized in statistical analysis (Caflish, 1998).

Where a gap arises, function values within that gap are not sampled for statistical analysis. The net effect is that mean values estimated with random samples have an uncertainty that diminishes slowly as $1/\sqrt{N}$. To reduce an estimated uncertainty by a factor of 10, the analyst must increase N by a factor of 100. (Saltelli *et al*., 2008)

4.4.1 Regular grid, MC and QMC sampling methods

The regular grid of points seems to be an efficient way for the integral evaluation. For up to 4 dimensions it works better or not worse than random sampling. For dimensions higher than 4, regular grid is not practical. The points in the regular grid are centered into each cell of the grid.

There are three problems for the use of regular grid for evaluation of integrals:

- The problem of dimensionality ("the curse of dimensionality"). It's been discussed in the previous section.
- It is not possible to incrementally enlarge the size of the grid and at the same time keep the grid uniform. This means that with a uniform grid approach it is not possible to have a termination criterion that can be invoked incrementally.
- The concavity bias. The regular grid generates small errors that add up, whereas random sampling generates big errors that cancel on average. Details can be found in Dupire & Savine (1998).

Figure 4.1

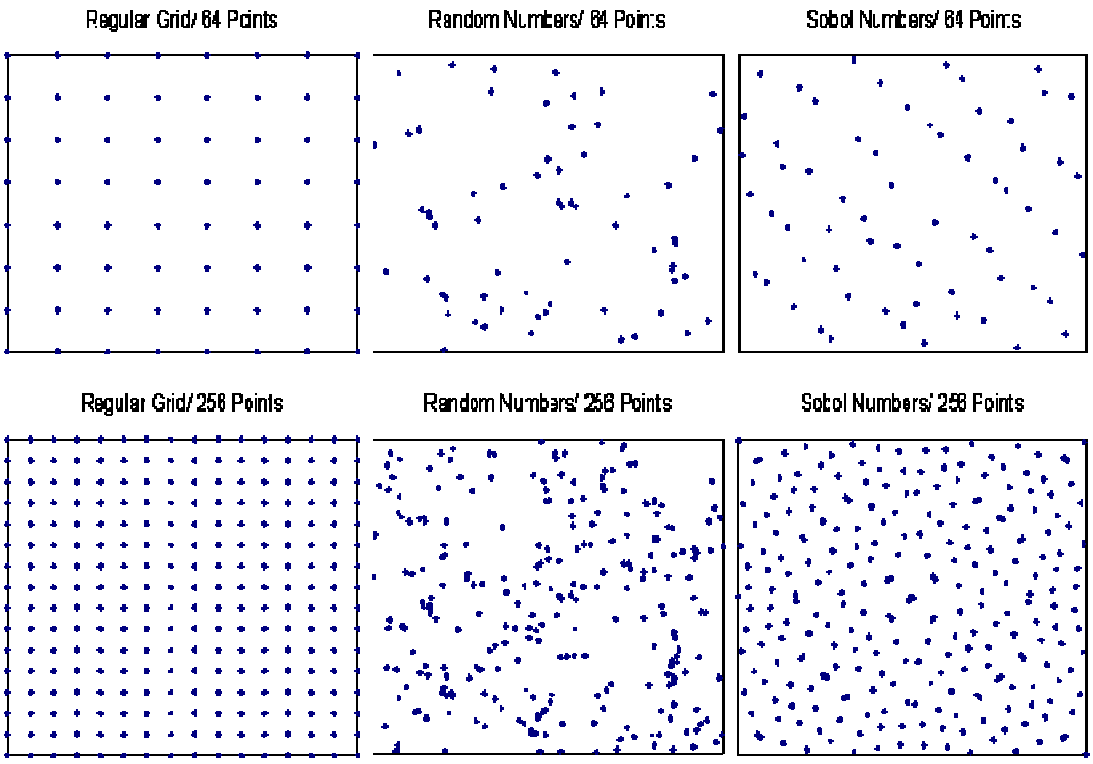
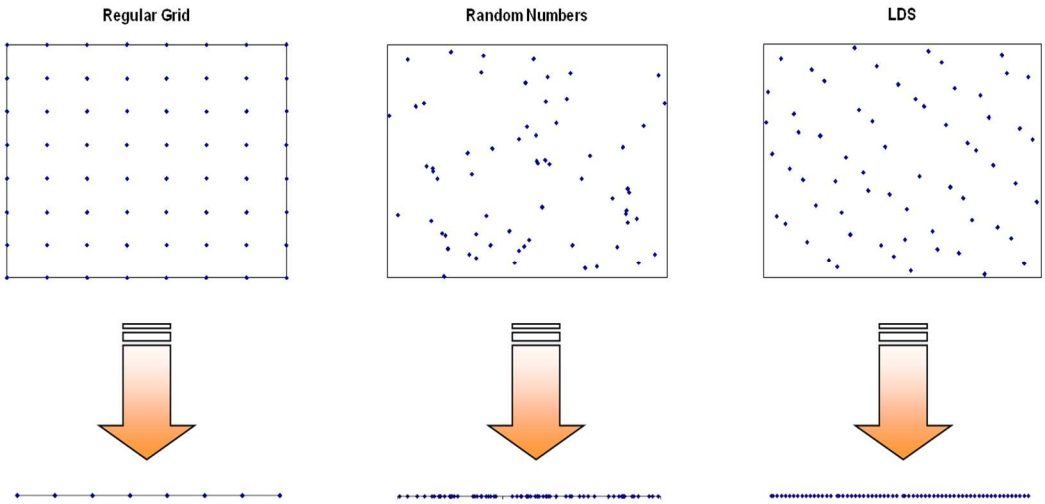


Fig 4.1 illustrates the difference between the regular grid, random sampling and Quasi Monte Carlo sampling. Distribution of N=64 (upper row) and N=256 (lower row) point in two dimensions.

The random and QR sampling methods do not have this kind of problems. QR sequences are specifically designed to place sample points as uniformly as possible. Unlike random numbers, successive low discrepancy points "know" about the position of previously sampled points and fill the gaps between them. (Kucherenko, 2012) (see Fig. 4.1)

Figure 4.2 illustrates the difference between the regular grid, the random sampling and QM sampling. It also illustrate an important property of LDS, that the projection of the *k*–dimensional LDS on the *s*–dimensional subspace has good uniform distributions. In particular, it explains the efficiency of QMC methods in high dimensions for many practical problems; LDS sampling gives much better way of arranging *N* points in *k*–dimensions.

Figure 4.2



Three different ways of arranging N=64 points in two dimensions.

4.4.2 A particular QMC sequence: the Sobol' LP_τ

There are a few commonly used QR sequences. Different principles are used for their construction by Halton, Faure, Sobol, Niederreiter and others. (Kucherenko, 2012). Many practical studies have proven that the Sobol' QR sequence is in many aspects superior to others (Paskov, 1995; Sobol,1998).

In sensitivity analysis, and in the present thesis, the Sobol' LP_τ quasi–random sequence is used (Sobol', 1967). All LP_τ sequences have an asymptotic uniform distribution as $N \to \infty$; to a certain extent, the uniformity of the sampled points is also observed for fairly small N. The sequences also satisfy additional uniformity properties (Sobol', 1976), called property A and property A', defined as:

Property A: for any k–dimensional sequence of length 2^d there is exactly one point in each 2^d hypercubes that result from subdividing the unit hypercube along each of its length extensions into half.

Property A': for any k–dimensional sequence of length 4^d there is exactly one point in each 4^d hypercubes that result from subdividing the unit hypercube along each dimension into four equal parts.

4.5 Quasi Random Balance Design

We present now a new approach for the estimation of the main effects in GSA which is based on RBD and quasi random numbers, enhancing the precision of the first order effect estimation.

This classical RBD method suffers from two problems:

- the estimated sensitivity indices are often quite loose
- small sensitivity indices are biased,

Given the set of N equally spaced values of variable s in $(-\pi, \pi)$, we construct a permutation of this set in a quasi–random fashion. Let $s_{i_1}, s_{i_2} \ldots, s_{i_N}$ denote the quasi random permutation obtained. Then, the experimental design $X_i(s_{ij})$ is given by

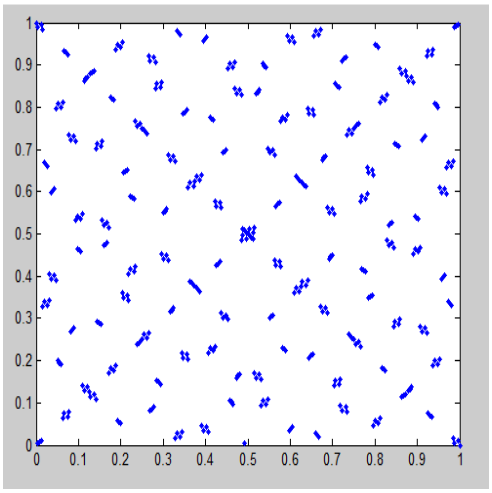
$$X_i = G_i(\sin(\omega s_{ij})), \quad \forall\, i = 1,..., k \quad \text{and} \quad \forall\, j = 1,..., N \tag{4.21}$$

and the procedure follows the standard RBD approach.

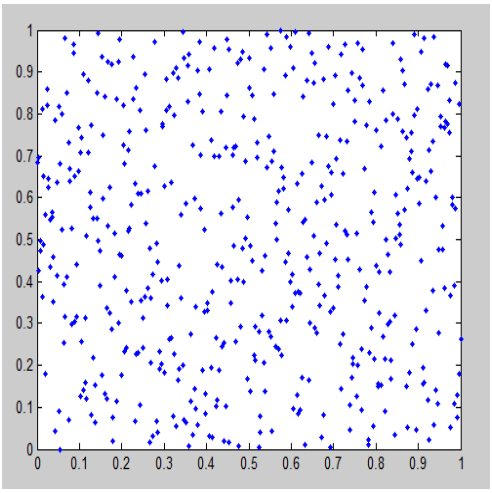
Thanks to the more homogeneous covering of the space offered by quasi–random numbers, the quasi–random permuted sample has better space–filling properties than the simple permuted sample (Fig. 4.3 and Fig. 4.4). This property enhances the quality of the estimates of sensitivity indices, as we show on the numerical test cases described in the next section.

Fig. 4.3

Quasi Random permutions



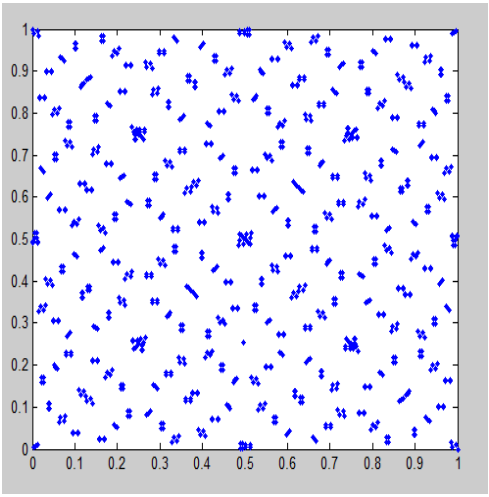
Random Permutations



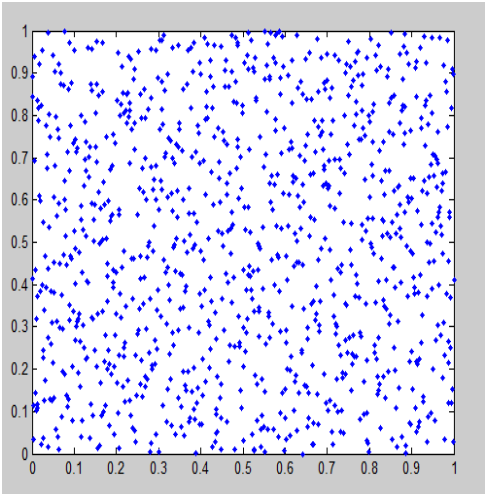
Coverage of the input space when using random and quasi-random permutations. N=512.

Fig. 4.3

Quasi Random permutions



Random Permutations



Coverage of the input space when using random and quasi-random permutations. N=1024.

In appendix A.3 we have included the new MATLAB code specifically created to perform QRBD method. We tested it using the five test functions which are described in details in the next section.

4.5.1 Analytical Test cases

This section illustrates the application of the proposed RBD method and the new QRBD (RBD using QRP) in the calculation of the first order indices.

We applied these methods to the following test functions for which the analytic values for S_i are available:

- *Linear function*
- *Corner Peak function*
- *Ishigami function*
- *Sudret function*
- *G–function*

We used the existing RBD algorithm (see appendix A.2) implemented in a MATLAB code for the estimation of first order indices and we tested the QRBD variant used on a sample generated by a QRP against the RBD method. We applied QRBD and RBD for each of the above functions at different sample sizes starting from 128 to 1024 at steps of two. Each of the test functions can be defined at a given set of input factors. We decided to set their number to $k = 3$ for the linear function, the Corner Peak function and the Ishigami function and to $k=8$ for G–function in order to test the methodology under different complexity. The results are reported below and plotted in figures for each function.

- ***Linear function***

$$Y = x_1 + x_2 + x_3 \tag{4.22}$$

We started our application of the two aforementioned methods with the simplest linear function with uniform distribution between (0,1). The analytical values for first order indices are given by $S_1 = S_2 = S_3 = 1/3$

The sensitivity results for this function are summarized in Fig. 4.4.

Both methods converge to the analytical value although in very different ways.

RBD estimates show a very high variability for all the three sensitivity indices especially for small sample sizes. This variability reduces when the sample size increases. Furthermore, for smaller sample sizes, the true value presents a slight over–estimation.

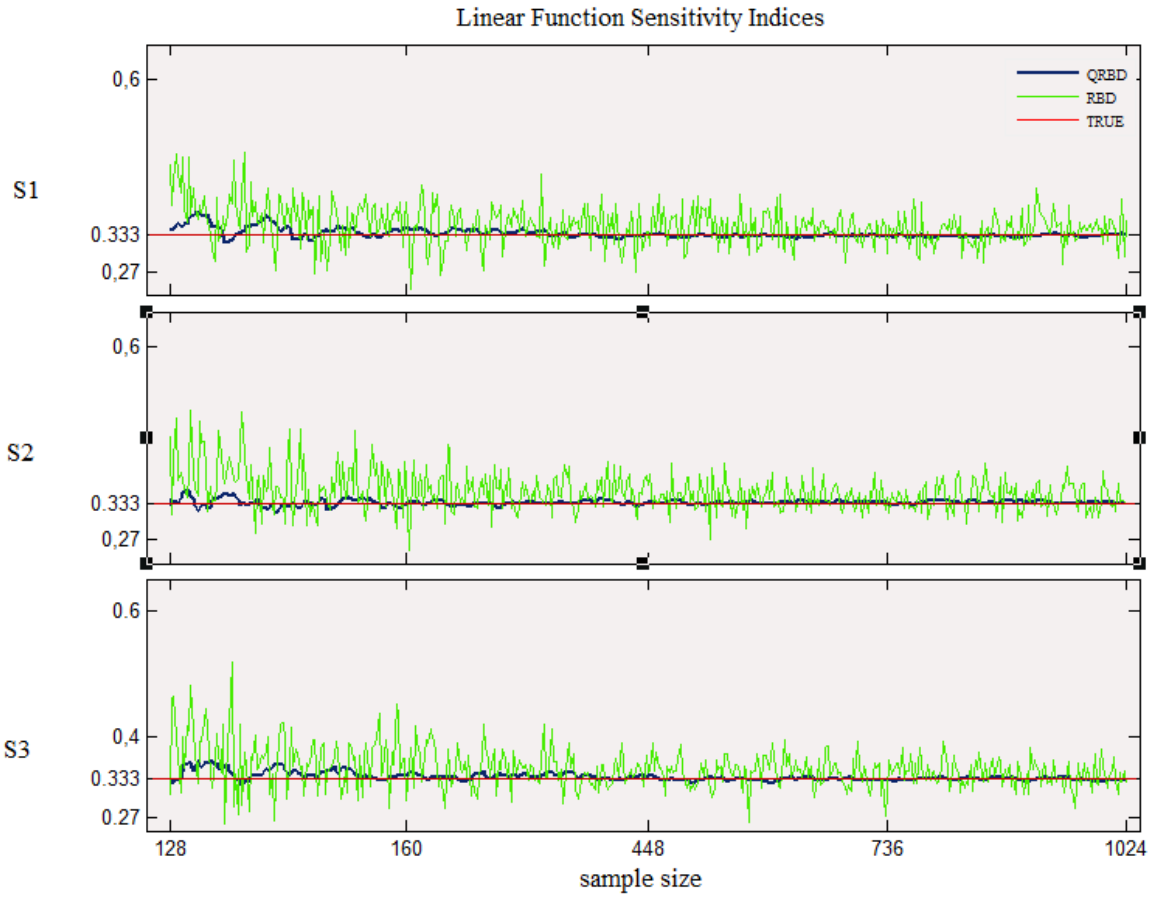
QRBD tends exactly to the analytical value, and it has a much smaller variability than RBD. We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.1 for RBD and QRBD.

Table 4.1

RMSE		
	QRBD	**RBD**
S1	0.0035	0.0282
S2	0.0025	0.0261
S3	0.0027	0.0231

The numerical results obtained for QRBD from running the MATLAB code, show that at a sample size of 1024 the value of the three sensitivity indices S_1, S_2, S_3 was respectively 0.332972, 0.333268 and 0.332919. In conclusion, QRBD is found to be superior than RBD.

Figure 4.4



Graphics of the three first order sensitivity indices against the sample size for the linear function test case with 3 factors. Analytical value are shown by a red line

- ***Corner Peak function***

$$Y = \left(1 + \sum_{i=1}^{k} a_i X_i\right)^{-(k+1)} \tag{4.23}$$

This is an additive function, introduced by Genz (1987) to test the method for multi–dimensional integration. The independent input factors are uniformly distributed over (0,1) and the constants a_i determine the sharpness and the location of the difficulty and they represent the weight of X_i. These functions are characterized by a peculiarity represented by several kinds of peaks.

We first consider the case in which the input factors are associated with the same parameter $a_1 = a_2 = a_3 = 0.01$. They contribute the same of output variance. The determined sensitivity analytical values are $S_1 = S_2 = S_3 = 1/3$.

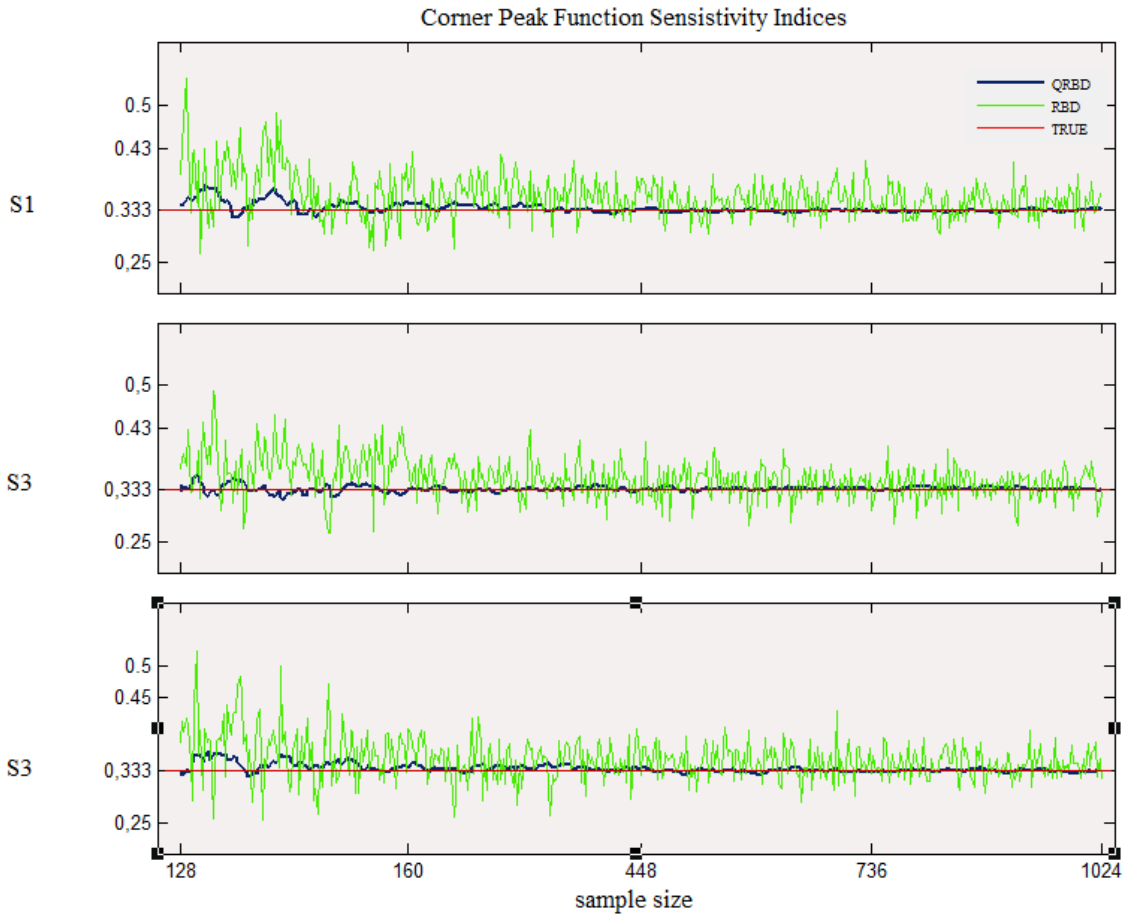
As we can see in Fig. 4.5 in the RBD method the estimated sensitivity indices show an evident variability slightly over the analytical value and it reduces as the sample size increases. This bias is due to the small output uncertainty driven by the small values of the coefficients a_i. When QRBD is performed the indices show much better results, as in the previous function. The indices converge quite rapidly to the analytical values, more rapidly than RBD. We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.2 for RBD and QRBD.

For some ranges of sample size (e.g. between 128 and 200) the estimates are slightly above the analytic values, for some other ranges they are slightly below them). Also in this case the numerical results obtained by QRBD from running the MATLAB code proved that at a sample of 1024 the value of the three sensitivity indices S_1, S_2, S_3 was respectively 0.332905, 0.333192, 0.332867. In conclusion, also in this case QRBD is found to be superior than RBD.

Table 4.2

RMSE		
	QRBD	**RBD**
S1	0.0035	0.0266
S2	0.0025	0.0269
S3	0.0027	0.0264

Figure 4.5



Graphics of the three first order sensitivity indices against the sample size for the linear function test case with 3 factors. Analytical value are shown by a red line

- ***Ishigami function***

$$Y = \sin X_1 + a \sin^2 X_2 + b X_3^4 \sin X_1 \tag{4.24}$$

This is a non–monotonic function uniformly distributed in $(-\pi, \pi)$. It has three input factors X_1, X_2, X_3 and the value of the parameters a and b are assumed to be $a = 7$ and $b = 0.1$. The analytical values of the sensitivity indices are $S_1 = 0.3138$, $S_2 = 0.4424$, $S_3 = 0$. The main peculiarity of this model is the dependence on X_3: there is no additive effect on Y but there is an interaction between X_3 and X_1. (Ishigami & Homma, 1990). Indeed, S_1 and S_2 account jointly for 76% of the variance, hence 24% of the variance must be due to higher order effects.

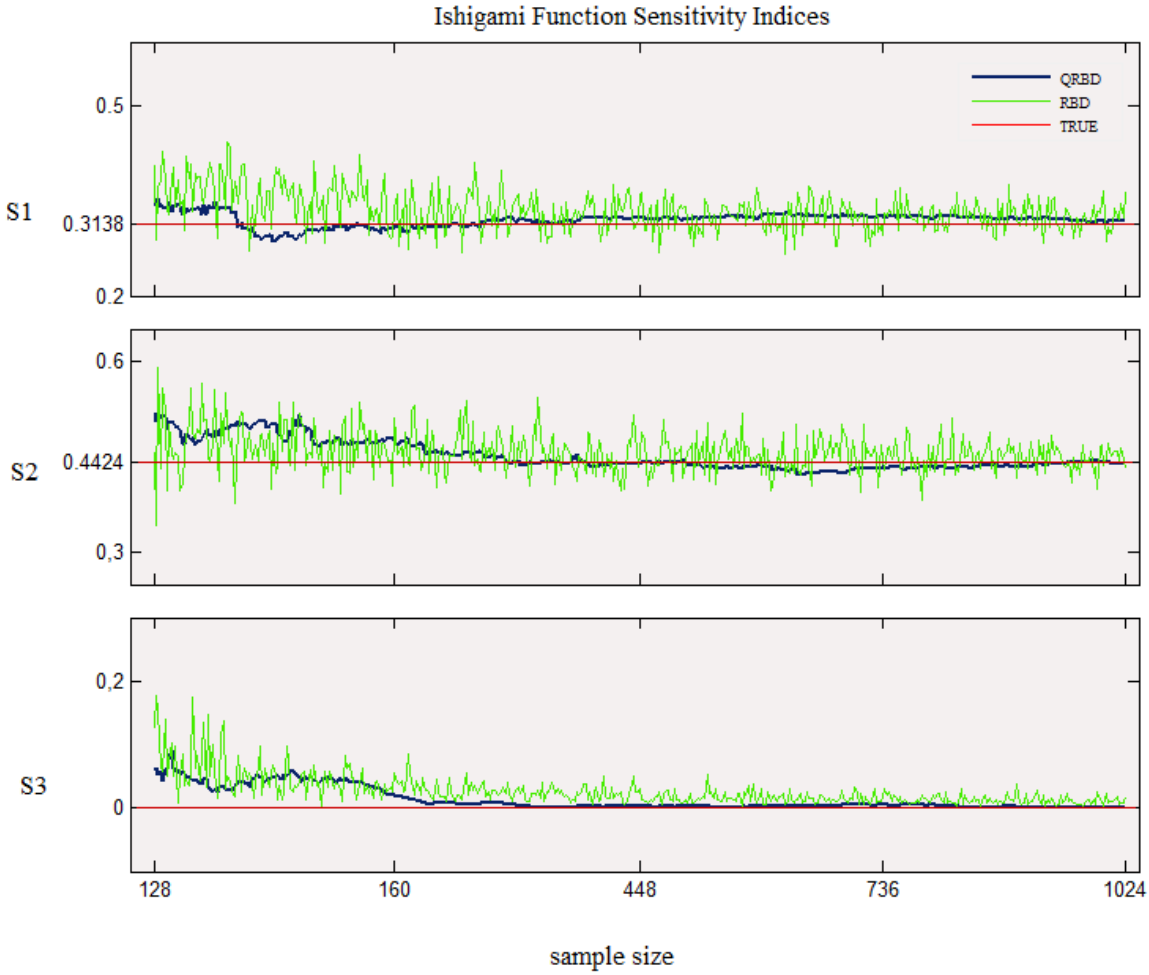
As in the previous test cases, when RBD is performed with Ishigami function we find the same variability slightly above the analytical value, especially for small samples. We notice that for the third index, S_3, the variability is considerably reduced from a sample size of about 160. We found from the numerical results that this index is almost zero near the sample size 1024.

QRBD is sometimes over and sometimes under estimated for different sample size ranges: S_2 and S_3 are slightly overestimated up to a sample size of about 200, but they converge to the analytical value with increasing sample size. At increasing sample size S_1 shows slower oscillations around the analytical value than the other sensitivities.

QRBD converges more rapidly than RBD. We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.2 for RBD and QRBD.

With respect to the previous two test cases, when the estimates are calculated with Ishigami functions show a higher variability. This is because this function is more complex and present interaction between X_3 and X_1.

Figure 4.6



Graphics of the three first order sensitivity indices against the sample size for the Ishigami function test case with 3 factors. Analytical value are shown by a red line

Table 4.3

RMSE		
	QRBD	**RBD**
S1	0.0098	0.0255
S2	0.0088	0.0258
S3	0.0028	0.0174

- ***Sudret function***

$$Y = \frac{1}{2^k}\prod_{i=1}^{k}(3X_i^2 + 1) \tag{4.25}$$

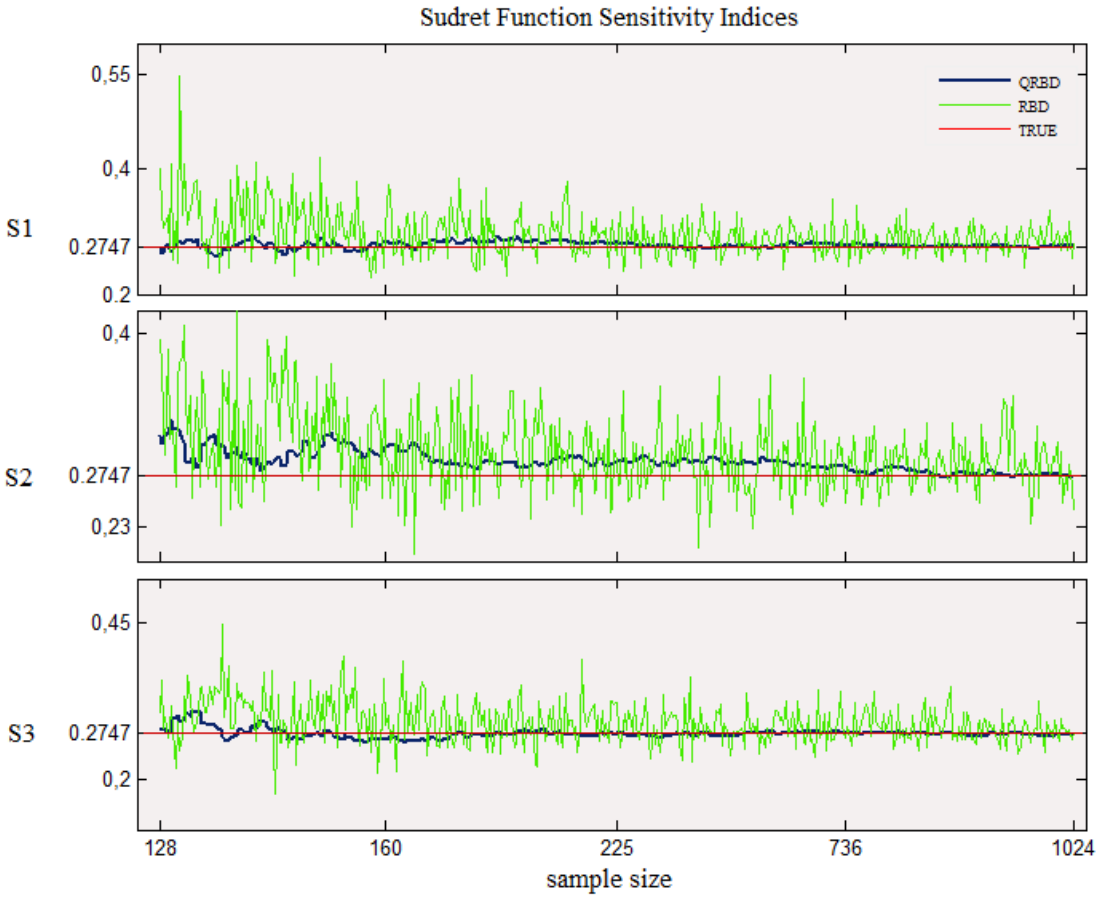
This is a polynomial function studied by Sudret (2008) where X_i are independent and identical distributed uniform random variables over (0,1). The exact global sensitivity indices can be determined as

$$S_i = \frac{5^{-s}}{(6/5)^k - 1} = 0.2747 \qquad s = 1,\cdots,k \tag{4.26}$$

For k=3 this function represents a sixth–order polynomial function which is a product form of three quadratic polynomials. When performed with RBD, the sensitivity indices present higher variability with respect to the other test functions, especially for S_2. The indices converge very slowly to the analytical values: at sample size 1024 (the maximum we tested) we still notice the presence of variability. As a curiosity, at the sample size of 1024 $S_3 = 0.274696$.

Despite in QRBD S_2 is over estimated until sample size about 900, the method performs better than RBD in terms of variability of the estimates at all sample sizes. We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.2 for RBD and QRBD.

Figure 4.7



Graphics of the three first order sensitivity indices against the sample size for the Sudret function test case with 3 factors. Analytical value are shown by a red line

Table 4.4

RMSE		
	QRBD	**RBD**
S1	0.0038	0.0267
S2	0.0095	0.0265
S3	0.0032	0.0267

- ***G–Function***

Introduced by Sobol' (2001) this function has been widely used to validate the methods for the sensitivity analysis.

This is a non–monotonic function whose analytical expression takes the following form

$$Y = \prod_{i=1}^{k} g_i(X_i) \text{ where } g_i(X_i) = \frac{|4X_i - 2 + a_i|}{a_i + 1} \tag{4.27}$$

The input factors X_i are uniformly distributed in the range (0,1). $a_i \geq 0$ are real–valued non–negative deterministic constants and are chosen to specify the role of the corresponding input parameter X_i since the range of uncertainty of $g_i(X_i)$ depends exclusively on the value of a_i.

$$1 - \frac{1}{1+a_i} \leq g_i(X_i) \leq 1 + \frac{1}{1+a_i} \tag{4.27}$$

The lower a_i the greater the importance of X_i. The table below gives some examples of the range of $g_i(X_i)$ and the relative importance of the input parameter X_i for the values of a_i

a_i	**Parameter** X_i	**Range of** $g_i(X_i)$
0	Very important	$0 \leq g_i(X_i) \leq 2$
1	Important	$0.5 \leq g_i(X_i) \leq 1.5$
9	Non–important	$0.9 \leq g_i(X_i) \leq 1.1$
99	Non–significant	$0.99 \leq g_i(X_i) \leq 1.01$

The partial variances of the first order are given by

$$Var_i = Var(E[Y \mid X_i]) = \frac{1}{3(1+a_i)^2} \tag{4.28}$$

$$Var(Y) = -1 + \prod_{i=1}^{k}(1+Var_i) \tag{4.29}$$

This allows us to calculate analytically the first order sensitivity indices for the G–function. We tested G–function at different values of a_i as follows:

a) $a_1 = a_2 = a_3 = 0$ all the input factors very important

b) $a_1 = 0 \quad a_2 = 1 \quad a_3 = 9$ the input factors are in decreasing order of importance

c) $a_1 = 99 \quad a_2 = 0 \quad a_3 = 9$ the input factors are in random order of importance

d) $a_1 = 99 \quad a_2 = 99 \quad a_3 = 99$ all the input factors are equally non–important

The values of S_i calculated at different sets of a_i combinations are plotted in the following Figures.

All the Figures below show that both QRBD and RBD methods provide estimates that converge to analytical value as the sample size increases.

Unlike the previous functions, when performing G–function with QRBD and when the factors are equally non–significant (Fig.4.10) or are all very important (Fig. 4.8), the estimates have a higher variability than before. This variability is, however, smaller than in RBD.

When the input factors are in decreasing order of importance, case b), we notice a higher variability if the input factors are important and the index is quite overestimated until sample size 736. We notice a much lower variability if the input factors are very important and non–important. In this case the curves of the two estimates lie on the

straight line of the true value. We notice only a slight overestimation for S_3 when performed with RBD method.

When the input factors are in random order of importance, case c), we can immediately se from Fig. 4.11 a perfect alignment of S_2 (the input factor is very important) with the straight line of the analytical value. S_1 and S_3 are also good estimates.

We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.5 for RBD and QRBD.

Table 4.5

RMSE								
	$a_1 = a_2 = a_3 = 0$		$a_1 = 0,\ a_2 = 1,\ a_3 = 9$		$a_1 = 99, a_2 = 0, a_3 = 9$		$a_1 = 99,\ a_2 = 99,\ a_3 = 99$	
	QRBD	**RBD**	**QRBD**	**RBD**	**QRBD**	**RBD**	**QRBD**	**RBD**
S1	0.0095	0.0229	0.0059	0.0097	0.0052	0.0182	0.0181	0.0241
S2	0.0097	0.0221	0.0231	0.0255	0.0024	0.0022	0.0249	0.0261
S3	0.0120	0.0230	0.0040	0.0192	0.0028	0.0191	0.0226	0.0250

Figure 4.8

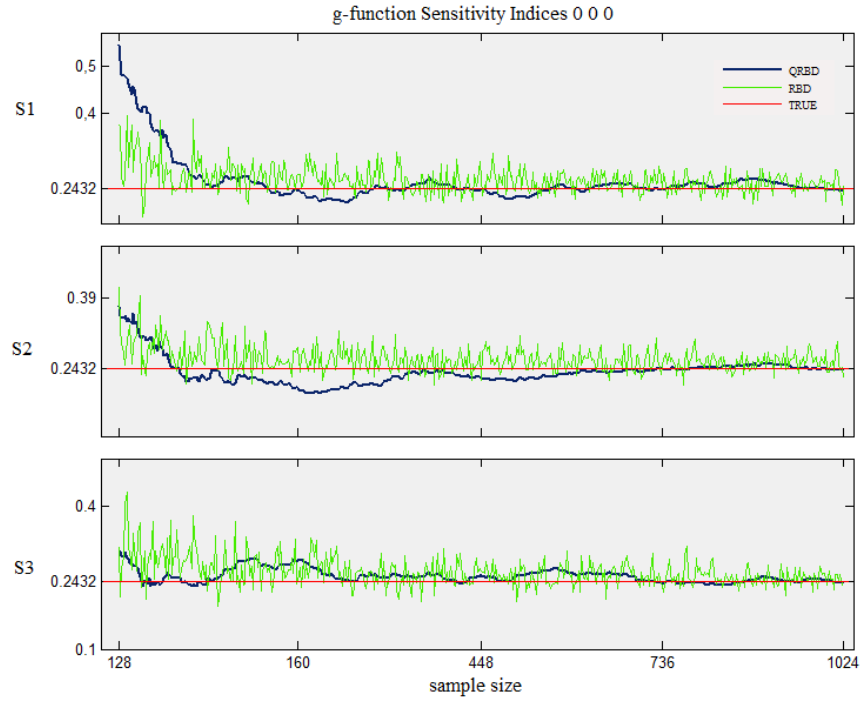


Figure 4.9 Graphics of the three first order sensitivity indices against the sample size for the G-Function function test case with 3 factors, $a_1 = a_2 = a_3 = 0$. Analytical value are shown by a red line

Figure 4.9

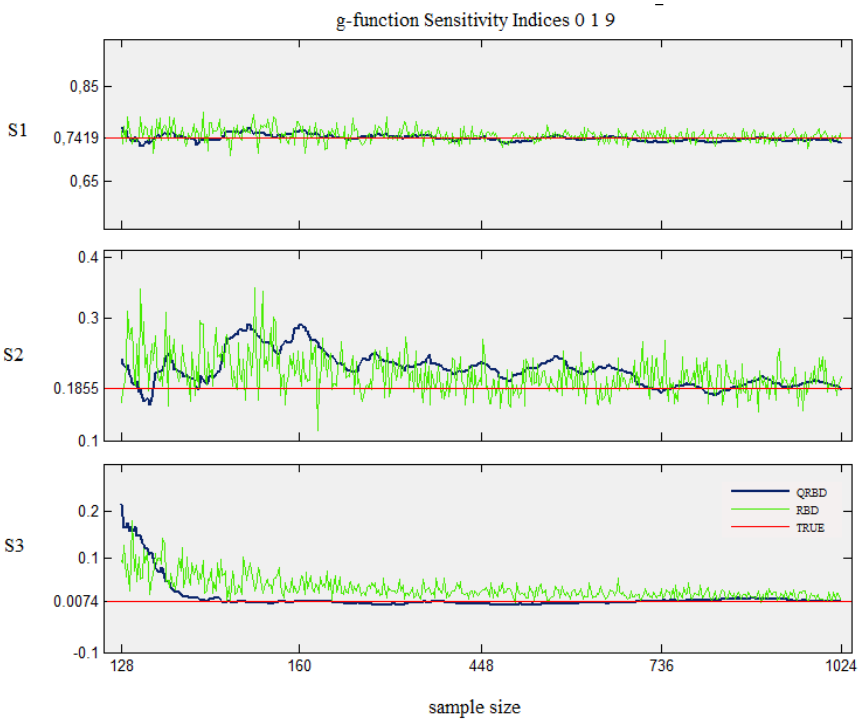


Figure 4.9 Graphics of the three first order sensitivity indices against the sample size for the G-Function function test case with 3 factors, $a_1 = 0 \quad a_2 = 1 \quad a_3 = 9$ Analytical value are shown by a red line

Figure 4.10

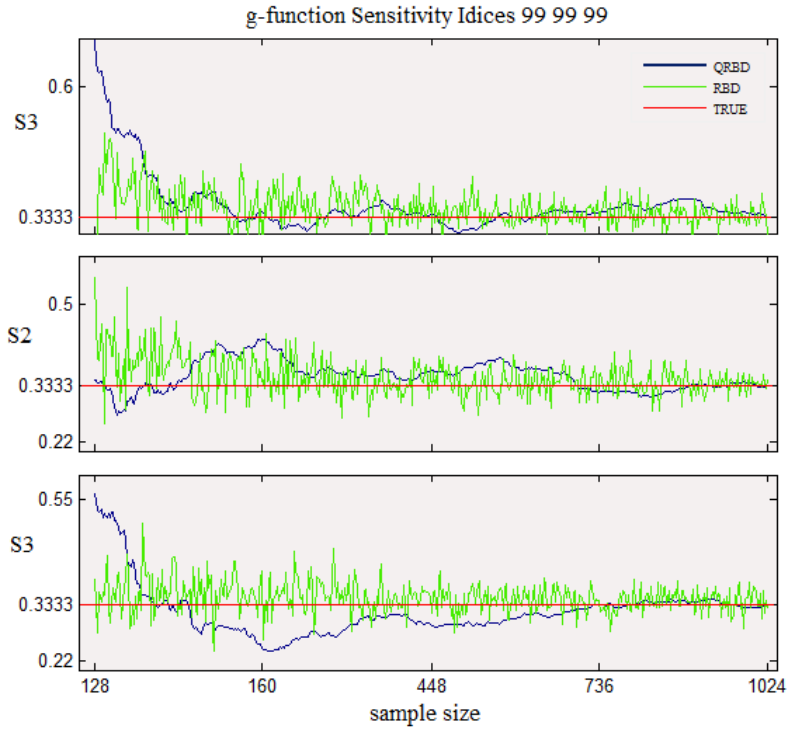


Figure 4.10 Graphics of the three first order sensitivity indices against the sample size for the G-Function function test case with 3 factors, $a_1 = 99 \quad a_2 = 99 \quad a_3 = 99$. Analytical value are shown by a red line

Figure 4.11

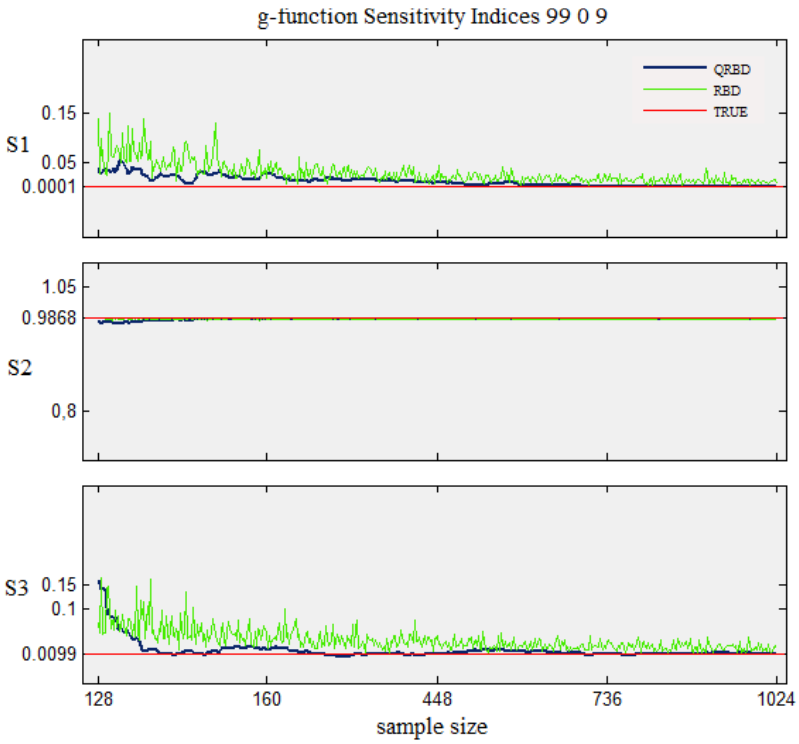


Figure 4.11 Graphics of the three first order sensitivity indices against the sample size for the G-Function function test case with 3 factors, $a_1 = 99 \quad a_2 = 0 \quad a_3 = 9$. Analytical value are shown by a red line

We decided to G–function with a number of input factors *k=8,* in order to test the methodology under different complexity. The results are reported below and plotted in figures for each function.

- ***G–Function*** *with* $k = 8$

In the case in which the dimensionality is $k = 8$, the parameters a_i are assumed to have following values: $a_1 = 0,\ a_2 = 1,\ a_3 = 4.5,\ a_4 = 9,\ a_5 = 99,\ a_6 = 99,\ a_7 = 99,\ a_8 = 99$, the input factor are in decreasing order of importance.

The main variance contributions of each input factor in the sensitivity analysis are summarized as follows

$$S_1 = 0.7162,\ S_2 = 0.1791,\ \ S_3 = 0.237,\ \ S_4 = 0.0072,\ \ S_5 = S_6 = S_7 = S_8 = 7.2\times10^{-5}.$$

The values of S_i, , are plotted in the following Figures.

All the Figures below show that both QRBD and RBD methods provide estimates that converge to analytical value as the sample size increases.

The difference from one index to another depends on the importance of the input factor. If the input factor is very important, non–important and non–significant, the correspondent estimates, S_1 (Fig. 4.12), S_4 (Fig. 4.13), S_5 (Fig. 4.14), S_6 (Fig. 4.14), S_7 (Fig. 4.15) and S_8 (Fig. 4.15) show a reduced variability for both RBD and QRBD method if we compare them with the previous functions. The curves of those estimates are aligned to the straight line of the analytical value, especially from sample 160. We notice a slight overestimation for S_6 (Fig. 4.14) performed with RBD.

When the input factor is important, the correspondent estimates S_2 (Fig. 4.12) and S_3 (Fig. 4.13) show a higher variability then the previous estimates. S_2 present a slight overestimation above the analytical value for both RBD and QRBD methods while S_3 present a slight underestimation for QRBD method until sample 736 and a slight overestimation for RBD method

Figure 4.12

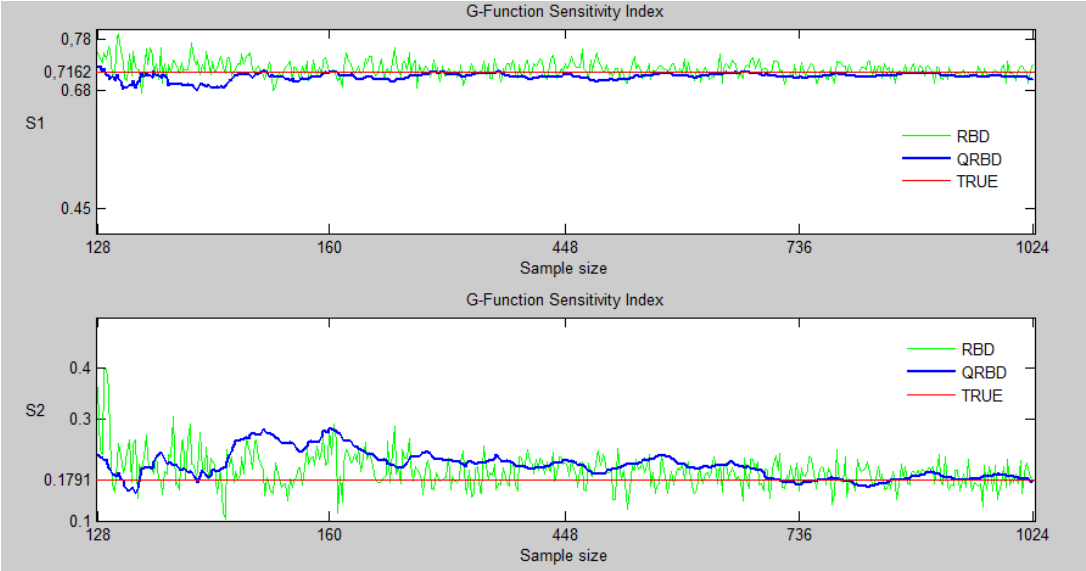


Figure 4.12 Graphics of S_1, and S_2 against the sample size for the G-Function with $k = 8$ and $a_1 = 0,\ a_2 = 1$. Analytical value are shown by a red line

Figure 4.13

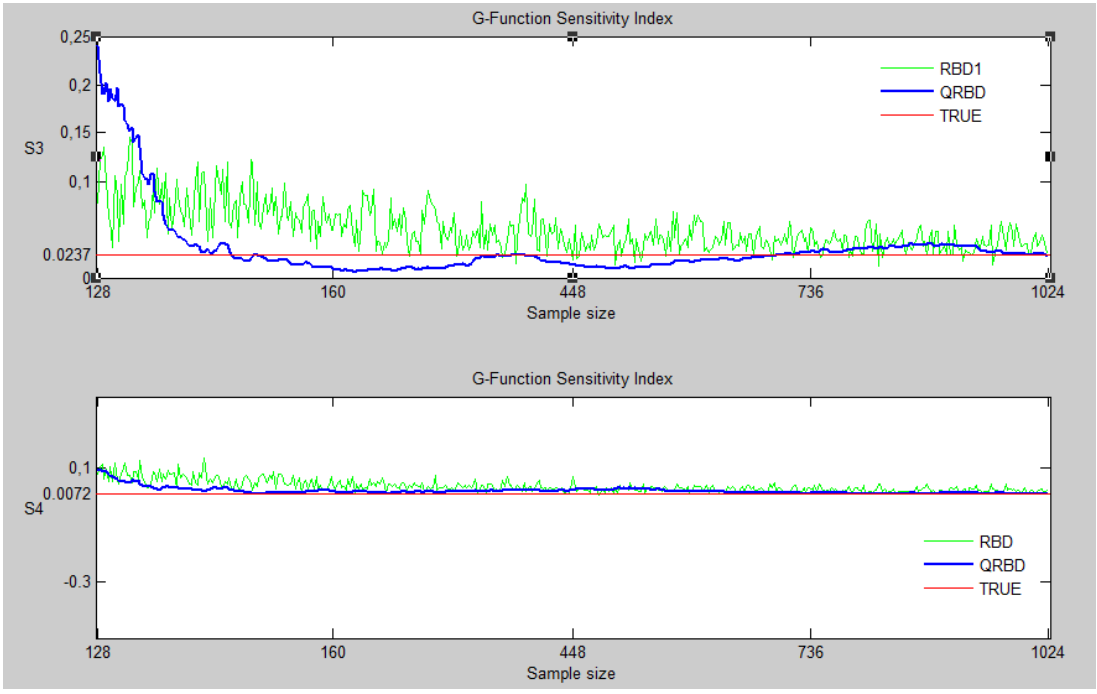


Figure 4.13 Graphics of S_3, and S_4 against the sample size for the G-Function with $k = 8$ and $a_3 = 4.5,\ a_4 = 9$. Analytical value are shown by a red line

Figure 4.14

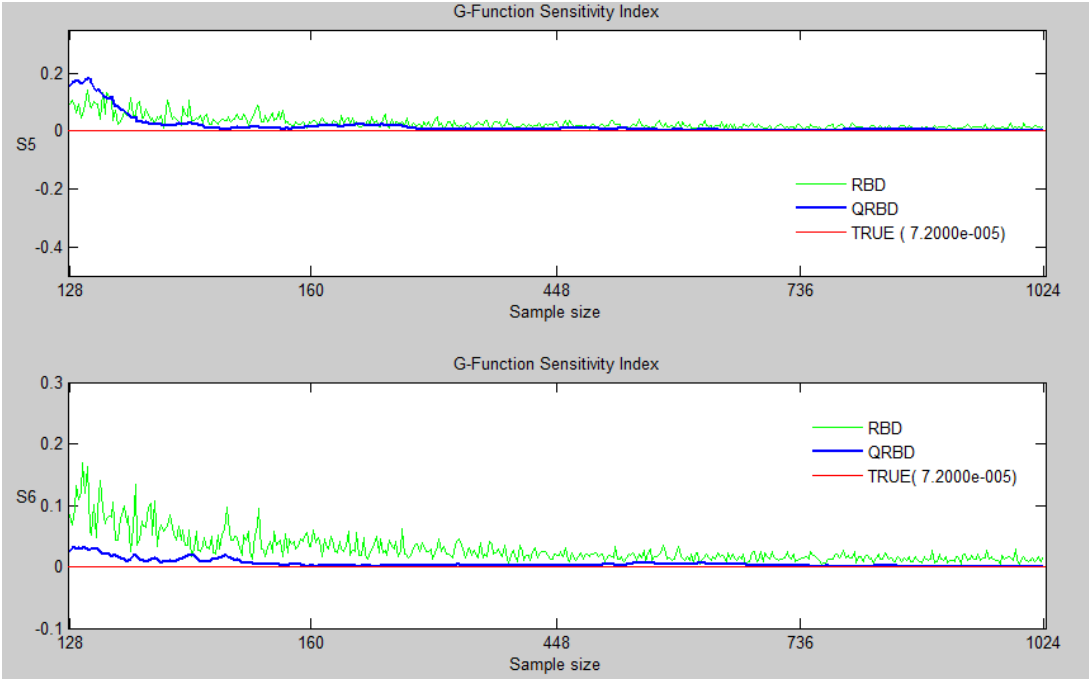


Figure 4.14 Graphics of S_5, and S_6 against the sample size for the G-Function with $k=8$ and $a_5 = 99,\ a_6 = 99$. Analytical value are shown by a red line

Figure 4.15

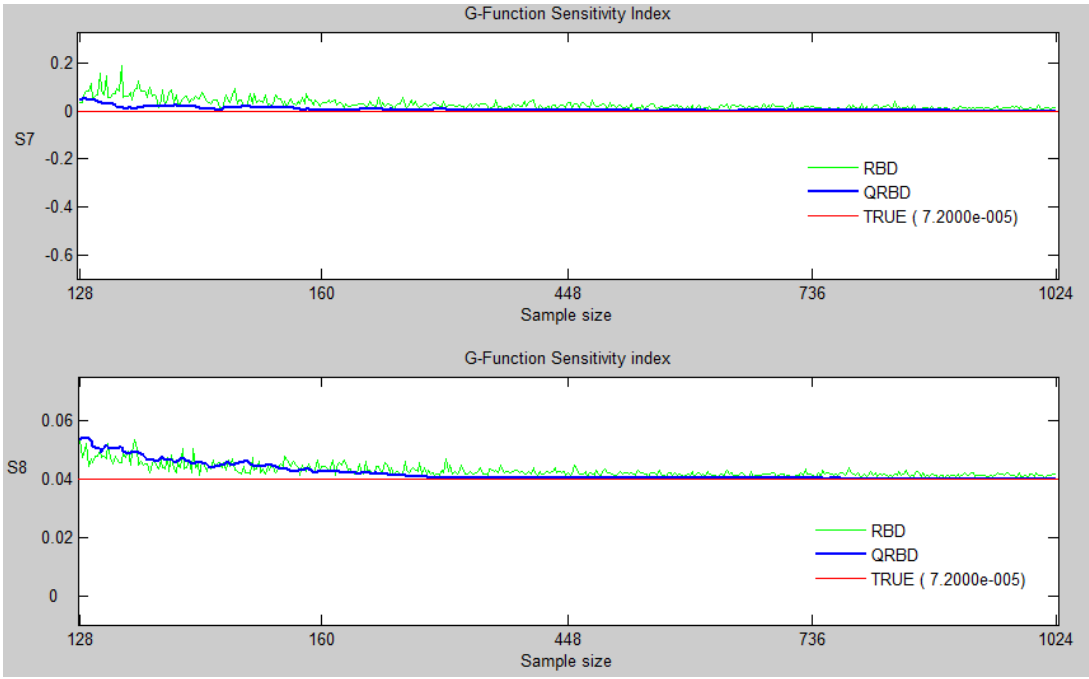


Figure 4.15 Graphics of S_7, and S_8 against the sample size for the G-Function with $k=8$ and $a_7 = 99,\ a_8 = 99$. Analytical value are shown by a red line

We computed the standard deviation of the estimates in the sample size range (200–1024) obtaining the following values summarized in Table 4.6 for RBD and QRBD.

Table 4.6

RMSE		
$a_1 = 0,\ a_2 = 1,\ a_3 = 4.5,\ a_4 = 9,\ a_5 = 99,\ a_6 = 99,\ a_7 = 99,\ a_8 = 99$		
	QRBD	**RBD**
S1	0.0091	0.0129
S2	0.0223	0.0263
S3	0.0076	0.0200
S4	0.0093	0.0187
S5	0.0050	0.0175
S6	0.0026	0.0167
S7	0.0020	0.0180
S8	0.0017	0.0172

In conclusion, we can say that both the methods give good estimates when performed with G–function with 8 parameters. QRBD is found to be slightly superior than RBD in terms of variability around the analytical value.

4.6 Conclusions

QR sequences are a deterministic alternative to random sequences and are specifically designed to place sample points as uniformly as possible. They are designed to provide better uniformity and hence higher rate of convergence.

In this chapter we presented a new approach, for the estimation of the first order effects in GSA. This method is based on the RBD approach, where random permutations are used, in association to Quasi–Random permutation.

We test the new QRBD method against the existing RBD and applied both techniques using five test functions.

Although the two methods only differ in the way permutations are taken, we discovered them to perform very differently.

We proved that as soon as QMC samples are used to generate the permutations the quality of the estimates obtained with QRBD increases significantly thanks to the more homogeneous covering of the space.

This property enhances the quality of the estimates of sensitivity indices, as we showed by numerical test cases. We conducted the investigation on five test functions of increasing complexity and from all of them we proved that QRBD method provide estimates which much lower variability around the analytical value and better convergence.

On the basis of our firstly results, the QRBD method would appear as a better estimation technique. Further analysis and tests will be conducted to confirm these early evidences, but what we have been able to stress is clearly a primary indications.

The improved outcomes obtained in recent years in SA thanks to more detailed and planned input sample spaces showed as this field of analysis is still a very promising and unexplored one.

Actually the interest of some major SA researchers points on this issue.

Hence, investing time and theory in the sample design pays off.

CHAPTER 5

CONCLUSION

Variance–based sensitivity analysis often requires a considerable number of model executions, one for each sample point considered. It is custom practice to try to reduce the number of executions at most, in order to obtain estimates of sensitivity indices of a given accuracy.

Our intentions in preparing this thesis were to attempt to overcome one of the drawbacks of variance–based measures in sensitivity analysis, i.e. to minimize the computational cost for achieving the required accuracy of sensitivity measures.

When model runs are already available, i.e. the input points and the corresponding model outputs are “given”, the analyst wishes to use them for the sensitivity analysis in order to save computational time. Therefore, techniques should be available to estimate sensitivity indices for given data.

The initial objective of the present thesis was to devise a “given data” approach for the estimation of total sensitivity indices, which takes into consideration the overall effect of interactions among model inputs.

No “given data” methodology is available today for estimating total sensitivity indices.

We created a unique algorithm, that we called “***cone approach***”, and investigated its use in estimation of total indices. The algorithm has been implemented in a Matlab code for further tests and outcome analysis.

Unfortunately the performance of the cone approach was worsening with increasing of the complexity of the test functions used to perform it, and with increasing the number of input factors.

Although the results did not completely meet our expectations, this new technique is worthy of further study and analysis. In our opinion the defined “***cone approach***” still remains a valid approach to be investigated and studied.

We are still working on this innovative approach and we will work on it until we find answers to our questions. We should remember that in SA the estimation of higher order effects and, in particular, total effects from given data still remains an open issue.

Needing additional time in order to continue our research on the cone approach we wanted to focus on a method of more immediate treatment.

We developed a new approach for the estimation of the first order effects given a specific sample design, named Quasi Random Balance Design. This method is based on the RBD approach, where random permutations are used, in association to Quasi–Random permutation.

QR sequences are a deterministic alternative to random sequences and are specifically designed to place sample points as uniformly as possible. They are designed to provide better uniformity and hence higher rate of convergence.

We tested the new QRBD method against the existing RBD and applied both techniques using five test functions. Although the two methods only differ in the way permutations are taken, we discovered them to perform very differently.

We proved that as soon as QMC samples are used to generate the permutations the quality of the estimates obtained with QRBD increases significantly thanks to the more homogeneous covering of the space. The investigation conducted on the five test

functions proved that QRBD method provide estimates which much lower variability around the analytical value and better convergence.

On the basis of our first results, the QRBD method would appear as a better estimation technique. Further analysis and tests will be conducted to confirm these early evidences, but what we have been able to stress is clearly a primary indications.

There are a number of future research ideas to improve, on one hand, and extend, on the other, the methods developed in this thesis.

APPENDIX

This appendix provide basic MATLAB codes used to compute sensitivity indices in according to the methods used in the present thesis.

A.1

The "***cone approach***"

```
function St=conesti(x,y)

[n,k]=size(x);

Alfa=10;

yp=zeros(n,k);
St=zeros(1,k);

for i=1:n
  z=abs(x-ones(n,1)*x(i,:));
  znorm=z./(sqrt(sum(z.^2,2))*ones(1,k));
  cone = znorm >= cos(Alfa*pi/180);

  for j=1:k
    if ~isempty(find(cone(:,j), 1)),      cc=0;
      while cc==0,
        [~,elem]=max(znorm(:,j));
        if cone(elem,j)==1,
           cc=1;
        end
      end
      yp(i,j)=y(elem);
    else
    end
  end

end
VY=var(y);

for j=1:k
    len=n-sum(isnan(yp(:,j)));
    St(j)=nansum((y-yp(:,j)).^2)/4/len/VY;
end
```

A.2

RBD method

```
function S = RBD(k,repl,N)


M=6;

S=[];
rand('state',sum(100*clock))

for r=1:repl

    sens=[];

    s0=[-pi:2*pi/N:pi]';

    for z=1:k;
        p(z,:)=randperm(N);
    end

    s=s0(p)';

    x=.5+asin(sin(s))/pi;

    % y=f(x)

    for i=1:k

        [ss,ind]=sort(s(:,i));
        yr=y(ind);


        spectrum=(abs(fft(yr))).^2/N;
        V=sum(spectrum(2:N));
        si(i)=2*sum(spectrum(2:M+1))/V;

    end
    S=[S; si];
end

return
```

A.3

QRBD method

```
function S = QRBD(k,repl,N)



M=6;

T=1;
fre=1;
S=[];
rand('state',sum(100*clock))

for r=1:repl

    sens=[];

    s0=[-pi:2*pi/N:pi]';

    for i=1:N
    p = floor(1+N*(1/(2*N)+sobolseq(i+(r-1)*N,k)));
    end


    s=s0(p);

    x=.5+asin(sin(s))/pi;

    %---test model
    % the user has to insert his/her own test model here

    for t=1:T
       for i=1:k


          [ss,ind]=sort(s(:,i));
          yr=y(ind);


          spectrum=(abs(fft(yr))).^2/N;
          V=sum(spectrum(2:N));
          si(i)=2*sum(spectrum(2:M+1))/V;

        end

     S=[S; si];
     end
end

return
```

Bibliography

Box G. E. and Meyer R. D. (1986) *An analysis for unreplicated fractional factorials.* Technometrics **28**, 11–18.

Box G. E. P. and Draper N. R. (1987). *Empirical model-building and response surfaces*. New York: Wiley. (DOE ref)

Bratley P., Fox B. L., Niederreiter H., (1992) *Implementation and tests of low discrepancy sequences, ACM Transactions on Modeling and Computer Simulation* 2 (3) 195 – 213.

Caflisch R. (1998) *Monte Carlo and Quasi–Monte Carlo methods*. Cambridge University Press, 1998, Acta Numerica pp. 1–49

Cariboni, J., Gatelli, D., Liska, R., Saltelli, A., (2007). *The role of sensitivity analysis in ecological modelling*. Ecol. Model. 203, 167–182.

Chan K. , Saltelli A., Tarantola S. *Sensitivity Analysis Of Model Output: Variance-Based Methods Make The Difference*, Proceedings of the 1997 Winter Simulation Conference ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

Cooper J. N., (2005) *Quasirandom arithmetic permutations*. Journal of Number Theory, 114(1):153–169

Cukier R. I., Levine H. B. and Schuler K. E. (1978) *Nonlinear sensitivity analysis of multiparameter model systems*. J. Copmput. Phus. **26**, 1 – 42

Cukier RI, Fortuin CM, Shuler KE, Petschek AG, Schaibly, JH. (1973) *Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients: theory*. J Chem Phys; 59:3873-8

Daniel C. (1958) *On varying one factor at a time*. Biometrics **14**, 430–431

Daniel C. (1973) *One–at–a–time–plans*. J. Am. Statist. Assoc. **68**, 353–360

Draper D., Pereira A., Prado P., Saltelli A., Cheal R., Eguilior S., Mendes B. and Tarantola S. (1999) *Scenario and parametric uncertainty in GESAMAC: A methodological study in nuclear waste disposal risk assessment*. Comput. Phys. Commun. **117**, 142 – 155

Dupire B., Savine A., (1998) *Dimension Rduction and other Ways of Speeding Monte carlo simulation in Dupire (Eds), Monte Carlo - Methodologies and Applications for Pricing and risk Management*, risk Books, pp. 51–63

Fisher R. A. (1935) *The design of Experiments*. Oliver & Boyd, Edimburgh

Helton J. C. and Burmaster D. E. (1996) *Treatment of aleatory and epistemic uncertainty in performance assessments for complex systems*. Guest editorial, Special Issue of Reliab. Engng. Sust. Safety **54**, 91–94

Helton J. C., Garner J. W., McCurley R. D. and Rudeen D. K. (1991) *Sensitivity analysis techniques and results for the performance assessment at the waste isolation pilot plant*. Sandia National Laboratories Report SAND90–7103

Helton J.C. (1993) *Uncertainty and sensitivity analysis techniques for use in performance assessment for radioactive waste disposal*. Reliab. Engng. Syst. Safety 42, 327–367

Helton J.C., Davis F.J., *Latin hypercube sampling and the propatation of uncertainty in analyses of complex systems*. Reliability Engineering and System Safety 81 (2003) 23-69

Homma T. and Saltelli A. (1994) *Global sensitivity analysis of nonlinear models. Importance measures and Sobol' sensitivity indices*. JRC Rep. EUR 16052 EN, Luxembourg

Homma T. and Saltelli A. (1996) *Importance measures in global sensitivity analysis of nonlinear models*. Reliability Engineering and System Safety **52**, 1–17.

Hora S. and Iman R. (1986) *A comparison of maximum/bounding and Bayesian/Monte Carlo for fault tree uncertainty analysis*. Report SAND85-2839, Sandia Laboratories.

Huber P. (1981) *Robust Statistics*. Wiley, New York

Iman R. and Hora S. (1990) *A robust measure of uncertainty importance for use in fault tree system analysis*. Risk Analysis **10**(3), 401–403.

Iman R. I. and Helton J. C. (1988) *A comparison of uncertainty and sensitivity analysis techniques for computer models*. Risk Anal. **8**, 71–90

Ishigami T. and Homma T., *An importance qualification technique in uncertainty analysis for compute models*, in: Proceedings of the Isuma '90. First International Symposium on Uncertainty Modelling and Analysis, University of Maryland, 3–5 December 1996.

Ishigami T. and Homma T., *An importance qualification technique in uncertainty analysis for compute models*. Proceedings of the Isuma '90. First International Symposium on Uncertainty Modelling and Analysis, University of Maryland, 3–5 December 1996

Isserlis, L.. *On the partial correlation ratio*. Part I: Theoretical. Biometrika,Vol. 10 (2–3): 391–411, 1914.

Jansen M.J.W. , Analysis of variance designs for model output, Computer Physics Communications 117 (1999) 35–43.

Jansen M.J.W., Rossing W.A.H., Daamen R.A., Monte Carlo estimation of uncertainty contributions from several independent multivariate sources, in: J. Gasmanand, G. van Straten (Eds.), Predictability and Nonlinear Modelling in Natural Sciences and Economics, Kluwer Academic Publishers, Dordrecht, 1994, pp. 334-343

K. Chan, A. Saltelli, S. Tarantola, Sensitivity Analysis Of Model Output: *Variance-Based Methods Make The Difference*, Proceedings of the 1997 Winter Simulation Conference ed. S. Andradóttir, K. J. Healy, D. H. Withers, and B. L. Nelson

Kucherenko S., Personal communication (2012)

Kucherenko S., Feil B., Shah N., Mauntz W. *Application of Global Sensitivity Indices for Measuring the Effectiveness of Quasi-Monte Carlo Methods*

M. Frechet. *Sur le coefficient, dit de correlation et sur la correlation en general*. Revue de l'Institut International de Statistique, 1(4):16{23, 1934.

McKay M. D., Beckman, R. J. and Conover, W. J. (1979) *A comparison of three methods of selecting values of input variables in the analysis of output from a computer code*. Technometrics. **21**(2), 239–245.

Metropolis, N. (1987). "*The beginning of the Monte Carlo method*". *Los Alamos Science* (1987 Special Issue dedicated to Stanislaw Ulam): 125–130.

Morris M. D., Moore L. M. and McKay M. D. (2006) *Sampling plans based on balanced incomplete block designs for evaluating the importance of computer model inputs*. Journal of Statistical Planning and Inference, 136(9):3203–3220

Morris M. D., Moore L. M. and McKay M. D. (2008) *Using orthogonal arrays in the sensitivity analysis of computer models*. Technometrics, 50(2):205–215

National Research Council's Science and Decisions: Advancing Risk Assessment (2009)(controllare pag. 96 e97)

Oakley, J. and A. O'Hagan (2004). *Probabilistic sensitivity analysis of complex models: a Bayesian approach*. Journal of the Royal Statistical Society B **66**, 751–769

Oreskes N., Shrader–Frechette K. and Belitz K. (1994) *Verification, validation and confirmation of numerical models in the earth sciences*. Science **263**, 64 –646

Paruolo P., Saltelli A., and Saisana M.. *Ratings and rankings: Voodoo or science?* Arxiv preprint arXiv:1104.3009, 2011.

Paskov, S. H. and Traub, J. F. (1995), *Faster evaluation of financial derivatives*, J. Portfolio Management, 22(1), 113-120.

Pearson, K. . *On the partial correlation ratio*. Proc. Royal Soc. London A, 91:492{498, 1915.

Plischke E., Borgonovo E., L. Smith C., (2013) *Global Snsitivity Measures from Given Data*. European Journal of Operational Research 226 536–550

Rabitz H. (1989) *System analysis at molecular scale*. Science **246**, 221–226

Rabitz H., Aliş Ö . F., Shorter J. and Shim K. (1999) *Efficient input–output model representations*. Comput. Phys. Commun. **117**, 11–20.

Risk Newsletter (1987) *Defining Risk editorial note*, Risk Newsletter, **7**(3), 5.Saltelli *et al.*, 2009

Rosen R. (1991) *A comprehensive inquiry into the nature, origin and fabrication of life*. Columbia University Press, New York

Saltelli A. and Marivoet J. (1990) *Non–parametric statistics in sensitivity analysis for model output: a comparison of selected techniques*. Reliab. Engng. Syst. Safety **28**, 229–253

Saltelli A. and Sobol' I. M. (1995 a) *About the use of rank transformation in sensitivity analysis of model output. Rwliabi. Engng*. Syst. Safety **50**, 225–239

Saltelli A. *Sensitivity analysis for importance assessment*. Risk Analysis, Vol. 22, N. 3, 2002

Saltelli A. *Making best use of model evaluations to comupute sensitivity indices*. Computer Physics Communications 2002, 145, 280-297

Saltelli A., Andres T. H. and Homma T. (1993) *Sensitivity analysis of model output: an investigation of new techniques*. Comput. Stat. Data Anal. **15**, 211–238

Saltelli A., Annoni P., Azzini I., Campolongo F., Ratto M., and TarantolaS.. *Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index*. Comput. Phys. Commun., 181(2):259{270, 2010.

Saltelli A., Planas C., Tarantola S. and Campolongo F. *The concept of sensitivity with applications to official statistics*. Forthcoming in ROS (2013)

Saltelli A., Ratto M., Tarantola S., Campolongo F. , *Sensitivity Analysis for Chemical Models*

Saltelli A., Tarantola S. and Chan K. (1999). *Quantitative model–independent method for global sensitivity analysis of model output*. Technometrics **41**(1), 39–56.

Saltelli A., Tarantola S. and Chan K.. *Presntenting results from model based studies to decision-makers:can sensitivity analysis be a defogging agent?* Risk analysis, vol. 18 N. 6, 1998

Saltelli, A., Chan, K. and Scott, M. (Eds) (2000a). *Sensitivity Analysis*. John Wiley & Sons, Ltd, New York.

Saltelli, A., S. Tarantola, F. Campolongo and M. Ratto (2004). *Sensitivity Analysis in Practice: A Guide to Assessing Scientific Models*. John Wiley & Sons, Ltd.

Satterwaite F.E. *Random balance experimentation*. Technometrics 1959;1(2):11-37

Sobol' I.M., Quasi-Monte Carlo methods, Progress in Nucl. Energy 24(1)(2)(3) 1990 55-61

Sobol' I. (2001) *Global sensitivity indices for nonlinear mathematicalmodels and their Monte Carlo estimates*, Mathematics and Computers in Simulation 55(1) 271–280

Sobol' I. M. (1966) *On the method of Haar series in the theory of multidimensional quadratures*, Internat. Congress. Mthem., Moscow, 4, 48.

Sobol' I. M. (1967) *On the distribution of points in a cube and the approximate elvaluation of integrals.* USSR Comput. Maths. Math. Phys., 7, N4, 86–112

Sobol' I. M. (1976) *Uniformly distributed sequences with additional uniformly properties*. USSR Comput. Maths. Math. Phys., 16, N5, 236–242

Sobol' I. M. (1990b) *Sensitivity estimates for non–linear mathematical models*. Matematicheskoe Modelirovanie **2**, 112 – 118 (in Russian) [Transl. *Sensitivity analysis for non – linear mathematical models*. Math. Modelling Comput. Exp. **1**, 407–414 (1993)]

Sobol' I. M. (1998) *On Quasi–Monte Carlo integration. Matematics and Computers in Simulation* 47 103–112

Sobol' I. M. *Numerical Monte Carlo Methods*, Nauka, Moscow, 1973 (Russian)

Sobol' I. M., Turchaninov V. I., L. Levitan Yu., Shukhman B. V.. (1992) *Quasirandom sequence generators*. Keldysh Institute of applied mathematics, Russian Accademy of Sciences – Moscow

Sobol' I.M., *Sensitivity analysis for non-linear mathematical models*, Mathematical Modelling and Computational Experiment 1 (1993) 407–414; Translated from Russian: I.M. Sobol', Sensitivity estimates for nonlinear mathematical models, Matematicheskoe Modelirovanie 2 (1990) 112–118

Sudret B. (2008) *Global sensitivity analysis using polynomial chaos expansions*, Reliability Engineering & System Safety 96(7) 825–837

Tarantola S., Gatelli D., Mara T.A. (2006) *Random balance designs for the estimation of first–order global sensitivity indices*. Reliability Engineering and System Safety ;91:717-27

Thompson J.F., Soni B. K . & Weatherill N.P. (1998). *Handbook of Grid Generation*. CRC-Press

Tomovic R. and Vukobratovic M. (1972) *General Sensitivity Theory*. Elsevier, New York

Tomovic R. and Vukobratovic M. (1972) *General Sensitivity Theory*. Elsevier, New York

Turanyi T. (1990) *Sensitivity analysis of complex kinetic systems. Tools and applications*. J. Math. Chem. **5**, 203–248

Weyl H., 1938 Am. J. Math. 60,889