ETH zürich

Global Sensitivity Analysis with Dependent Inputs

Philippe Wiederkehr


Risk, Safety &
Uncertainty Quantification

Report Data Sheet

Client:	–
Contract Ref.:	Master Thesis Philippe Wiederkehr
Contact:	Prof Dr. Bruno Sudret, Dr. Stefano Marelli
Address:	Stefano-Granscini-Platz 5
	CH-8093 Zürich
	Switzerland
ETH Project:	Master Thesis Philippe Wiederkehr

Report Ref.	R-MTWiederkehr-001
Title:	Global Sensitivity Analysis with Dependent Inputs
Authors:	Philippe Wiederkehr
Date:	September 20, 2018
Signature:	

Abstract

Sobol' indices are a well established measure for global sensitivity analysis on models with independent input variables. However, several alternative sensitivity measures have been proposed to deal with dependent (correlated) inputs. Among others, Kucherenko et al. (2012) and Caniou (2012) have proposed two different generalisations of the Sobol' indices. The first uses a direct decomposition of variance, evaluatable with a double loop Monte Carlo estimation, while the latter uses a High Dimensional Model Representation as structural substitute of the actual model. The two approaches lead to different indices whose interpretation differs and is sometimes not trivial. In the context of this thesis, the two methods are implemented and applied onto models with increasing complexity to assess how the indices change for different dependence structures. The resulting indices are interpreted and discussed in order to understand the evolution of the values for varying correlation. Dependence is modeled by a Gaussian copula. For expensive-to-evaluate models the double loop Monte Carlo estimation of Kucherenko indices might in general not be feasible. Instead of the actual model, a cheap-to-evaluate surrogate, built using polynomial chaos expansion, is applied.

Keywords: Sensitivity analysis, Dependence, Surrogate modelling

Contents

1 Introduction

1.1 Need for sensitivity analysis

Sensitivity analysis is a useful tool often used nowadays in industry and academics. However, for a non-practitioner it is normally unclear what sensitivity analysis is about. The following simple example should serve as an introduction to this research field. Imagine a construction material supplier starting his company. After the first two years he checks his bookkeeping: he has a gravel and a clay quarry, each equipped with the needed excavation machines, trucks for transport and workers. The books tell him how much materials he extracts from each quarry every month and the money he makes selling it to the manufacturers. Naturally, he does not get the same amount of material every month, depending on many factors like obstacles in the ground, condition of the machines, health of his workers, weather conditions and demand from the manufacturers. In other words, each of his income contributions has uncertainties: for every quarry there exists an expected value of output μ, a variance σ^2 and a likelihood for certain values around the mean, described by a probability density function (pdf). As a result, his income is also varying every month and from time to time the entrepreneur's monthly income sinks below a critical level and he cannot pay his bills without taking from his reserves. To avoid these scenarios in the future, he wants to stabilise his income by decreasing its monthly variance. To determine in which of his income sources he should put the most effort to reduce the variance he poses the following question: the uncertainty of which contribution influences the income variance the most?

Needless to say, the suppliers example is simplified by an additive model to calculate the output (the farmer's income). Nevertheless, the same question arises frequently in engineering practice, where mathematical models have become very important to describe processes or phenomena. For example, in civil engineering models are used to calculate deflections in decisive parts of structures and the maximal stresses the materials have to resist. For complex structures analytical solutions are not feasible. Instead numerical methods such as the finite element method (FEM) are often used. Available FEM software usually provide the user with an interface, where he enters information (the input variables) about the structure (such as geometry, materials, cross-sections, degrees of freedom) and the loads (such as position, direction, absolute value) and then calculates the deflections and stresses (the outputs) based on this information. Naturally, those results already answers many of the engineer's questions. However, in order to understand the behaviour of the structural system, one needs to know how much the single inputs (or subsets of them) influence the output and in which way (structural, interaction or correlation with other inputs).

Not only structural engineering but many other engineering and research areas use mathematical models. The input values are often known only to some degree of uncertainty

and are therefore described as random variables (Sudret, 2007). The goal is then to understand which of those input variables influences the output variance the most and which do not. These questions are addressed by the sensitivity analysis.

1.2 Global sensitivity analysis

An easy and very intuitive way to investigate the influence of input parameters is to look at partial derivatives of the model at certain points of interest. This is called local sensitivity analysis, since the results show the sensitivity of the model around a certain point. In contrast to local sensitivity analysis, global sensitivity analysis (GSA) provides information about the influence due to variation over the whole input spectrum. The goals of GSA are the following: to determine which inputs contribute the most to the output variance, which ones are non-influential and, in some cases, understand the interaction and dependency structure of the underlying computational model. The results in turn are primarily used to achieve two objectives. One is variance reduction of the output by reducing the variance of the most influential inputs. The other is model simplification by fixing non-influential inputs to their expected value, thus decreasing the amount of variables in the model. Lastly, through GSA one can also achieve a deeper understanding of the model behaviour by interpreting the dependency and interaction structure.

According to Iooss and Lemaître (2014), there are three different types of GSA: screening techniques, importance measures and deep exploration of sensitivity. The goal of screening methods is to identify non-influential inputs using only a small number of model runs. This allows a model simplification before going on to apply more elaborate and costly analyses. However, they might not always be successful.

The importance measures rely instead on fitting a model to the output and studying this fitted model. For linear models, the Pearson's Correlation Coefficient ρ, often referred to as linear correlation coefficient, the Standard Regression Coefficient (SRC) and Partial Correlation Coefficient (PCC) provide importance measures for an input variable. If the model is not linear but still monotonic, a rank transformation can be applied to the sample sets. Afterwards, the Spearman's Correlation Coefficient ρ_S, the Standard Rank Regression Coefficient (SRRC) and the Partial Rank Correlation Coefficient (PRCC) can be defined analogously to the linear case. However, in practice the models are frequently non-linear and non-monotonic. In this case, a decomposition of the output variance can lead to helpful importance indices for single inputs or even subsets of inputs.

Deep exploration methods aim to give a better understanding of sensitivity. Graphical and smoothing techniques, like the scatterplots or parallel coordinate plots provide more information on interactions between parameters. However, analysing the behaviour of an experimental system or a long running computational model may be very costly. In order to cope with this, metamodel-based methods were developed. The goal of metamodelling

is to approximate a complex model by a surrogate which has good prediction capabilities at negligible computational cost.

Ferretti et al. (2016) reviewed databases of high impact factor journals like Science and Nature to find out how often and what types of sensitivity analyses are used in scientific publications. They found that the majority of sensitivity analyses are either local or one factor-at-a-time (OAT) analyses. Even though the share of GSA is growing over the years, traditional techniques are still prevailing. The authors suppose the reason for this is the complexity of GSA methods. However, simpler methods are often based on assumptions that frequently do not suit the problem. Thus, using simple methods like OAT can lead to incorrect results. Methods to derive correct indices do, however, demand more expertise. Furthermore, the interpretation and handling of those indices is not always obvious. Scientific work on the application and interpretation of those indices may further increase the share of appropriate GSA in scientific papers and practice.

1.3 Importance indices for independent input variables

Due to their applicability for complex models, the variance-based indices are widely used today to analyse models. For independent input variables, the so-called ANOVA (ANalysis Of VAriance) leads to the Sobol' indices (Sobol', 1993; Homma and Saltelli, 1996; Sobol', 2001). Those are unambiguous and provide helpful information on the importance of input variables by allocating a share of the total variance to each input variable or its interactions with other variables (see Sections 2.2 and 2.3). Using those indices, it is easy to spot important and non-influental variables and detect interactions between variables. Sobol' indices of all orders added together equal 1, clearly showing they are effective shares of the total output variance.

1.4 Dependence and correlation

In reality, however, there often exists some kind of dependence relation between (two or more) input variables. For example, components may be produced by the same machine, materials are tested by the same company or estimations on different quantities are made by one expert. If there is any kind of relationship between variables, one talks about dependence. Correlation describes a specified relationship between two variables. Since the result of a mathematical function will naturally be correlated with the inputs, different correlation measures were already mentioned in Section 3. To describe the relationship between input variables, the linear correlation coefficient is often used and therefore will be examined in the following.

The linear correlation coefficient ρ is arguably the most used measure of dependence. It

is assessed by the normalized covariance between two variables X_1 and X_2:

$$\rho_{12} = \frac{\mathrm{Cov}[X_1, X_2]}{\sigma_1 \sigma_2}, \tag{1}$$

where $\sigma_i = \sqrt{\mathrm{Var}[X_i]}$ is the standard deviation of X_i, $i = 1, \ldots, 2$. The coefficient lies between -1 and 1. If ρ_{12} is larger than 0, one can expect a large value of X_2 if a large value of X_1 is observed, where large indicates above the mean. A negative value of ρ_{12}, on the other hand, means that if X_1 lies above its mean, X_2 tends to lie below its expected value. In the case of no covariance between the samples of X_1 and X_2, the linear correlation coefficient is 0. Figure **??** shows two sampled data sets of $(X_1, X_2) \sim \mathcal{U}^2[0, 1]$ for different values of linear correlation.

There are two interesting facts to conclude from Figure **??**. One is that, in case of correlation, for a certain value of X_1, X_2 is not uniformly distributed between 0 and 1 anymore. The distribution of the correlated variable X_2 changes. Since the resulting distribution depends on a certain value of $X_1 = x_1^*$, it is called conditional distribution. The other conclusion is that through correlation, be it positive or negative, the input space shrinks. For no correlation the sample points were distributed evenly in the whole square $[0, 1]^2$ whereas they are constrained in the case of correlation.

In case the samples of X_1 and X_2 are both monotonically increasing, one could also talk about a high correlation. However, the linear correlation coefficient may not show this, since it detects, as the name implies, only linear correlation. Therefore, Spearman introduced the rank correlation coefficient, also called Spearman's ρ, denoted by ρ_S. In order to compute this measure, first points in each sample set are assigned a rank in the sample. The smallest value gets rank 1 and the largest rank N (for a sample of size N). This corresponds to replacing the random variable X_i by $F_{X_i}(X_i) \in [0, 1]$, where F_{X_i} is the cumulative distribution function of X_i. The linear correlation coefficient of the ranks of the sample points results in ρ_S:

$$\rho_S = \frac{\mathrm{Cov}[F_{X_1}(X_1), F_{X_2}(X_2)]}{\sqrt{\mathrm{Var}[F_{X_1}(X_1)]\,\mathrm{Var}[F_{X_2}(X_2)]}}. \tag{2}$$

The correlations between input variables can be collected and summarised in a matrix, the diagonal of which is equal to 1. The entry (i, j) with $i \neq j$ represents the correlation between X_i and X_j. Since $\rho_{ij} = \rho_{ji}$ the matrix is symmetric. Eq. (3) shows a generic linear correlation matrix for N variables:

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & \cdots & \rho_{1N} \\ \rho_{21} & 1 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \rho_{N-1,N} \\ \rho_{N1} & \cdots & \rho_{N,N-1} & 1 \end{pmatrix}. \tag{3}$$

Of course, it is also possible to collect the covariances in the same manner. The covariance

matrix is symmetric as well and denoted as $\mathbf{\Sigma}$.

Dependence between input variables can have a strong influence on the outcomes of statistical analyses. With an easy example Caniou (2012) managed to show that it is crucial to include dependence, if existent. Neglecting dependence and simply calculating Sobol' indices for models with dependent input variables can result in a over- or underestimation of risk. An underestimation can result in an wrong allocation of resources for risk reduction, whereas an overestimation may lead to oversizing of a structure and unneeded expenses.

1.5 Copula theory

The probabilistic distribution of random variables $\mathbf{X}$ can be displayed by their multivariate joint probability density function (PDF) or its integral, the cumulative distribution function (CDF). However, in engineering practice such a representation is not common. Instead, the data set of each variable is processed separately and described by a marginal distribution. The question of dependence or correlation is only addressed later and often even ignored. In case of independence between variables this may not be problematic, but as mentioned, it will lead to wrong results if correlation is present. The goal of copula theory is to produce a joint CDF from the marginal distributions of the random variables (Nelsen, 2006):

$$F_{\mathbf{X}}(\mathbf{X}) = \mathcal{C}[F_{X_1}(X_1), \ldots, F_{X_M}(X_M)]. \tag{4}$$

There virtually exists an infinite amount of copulas. For a given data set, the copula can be inferred using standard statistical tools, e.g. maximum likelihood estimation. Otherwise the copula structure may be postulated within certain parametric families. The Gaussian copula belongs to the class of elliptical copulas and allows for an easy isoprobabilistic transform, the Nataf-transform (Lebrun and Dutfoy, 2009). In case the marginal distributions are all Gaussian, their linear correlation matrix is equal to the matrix of the copula parameters. Moreover, this copula is already included in UQLab, the used software in the context of this thesis. For these reasons, the Gaussian copula is used to represent dependence.

1.6 Importance indices for dependent input variables

As mentioned above and shown in Figure **??**, because of correlation between X_1 and X_2, for a certain value of X_1 the distribution of X_2 will not be the same as its original anymore. If one analyses a mathematical model with the variables $\mathbf{X} = (X_1, X_2, X_3)$, where X_1 and X_2 are correlated, but X_3 is not, it seems obvious that the importance of X_1 and X_2 should change (compared to a fully uncorrelated case) because of correlation. A change of X_1 will not only change the output through its structural and interactive con-

tributions, but also through its correlation with X_2. However, if this results in a decrease or increase in importance depends on the model and the nature of the correlation.

Up to date there is no single commonly shared vision on how the influence of correlation should be included in sensitivity analysis. Whereas in the independent case the definition of importance is clear and the variance-based sensitivity indices allow a meaningful ranking of the variables, the introduction of dependency leads to many problems. For one, the methods to define importance indices for independent variables, like ANOVA, are not formulated generally and are therefore not applicable if correlation is present. Further, generating samples for their estimation may be complicated by correlation. Additionally, as mentioned, it is not clear how correlation should influence the importance of variables or what a meaningful sensitivity index would be. The definition of importance indices is a primary issue.

One of the first to propose a solution to this problem was Borgonovo (2007). The idea is to analyse the change in the output distribution caused by the variation of an input variable. The resulting index is moment-independent and definitively quantifies the influence of a variable. However, such an index does not distinguish between structural, interactive and correlative influences of the variable. Additionally, the interpretation of the index is not straightforward compared to the variance-based ones, since variance is a well-defined and often handled parameter in uncertainty quantification.

1.7 Goal of the thesis

For the reasons mentioned in the last section, this thesis will analyse methods leading to variance-based indices. Caniou (2012) and Kucherenko et al. (2012) both aim to generalise the Sobol' indices for dependent variables. Each approach starts from a different definition of the classic Sobol' indices and their resulting indices differ, highlighting the fact that importance is not a clearly defined quantity in the presence of correlation. Furthermore, the interpretation and usefulness of those indices is not apparent.

This thesis aims to produce a source code for the two methods, provide possible interpretations of the indices and clarify the difficulty of global sensitivity anaylsis for dependent input variables. By implementing the approaches by Caniou (2012) and Kucherenko et al. (2012) into UQLab (Marelli and Sudret, 2014) and comparing them to each other as well as to the classical Sobol' indices, the advantages and disadvantages of each approach will be discussed.

2 Investigated approaches

2.1 Overview

The Sobol' indices (Sobol', 1993; Homma and Saltelli, 1996) allow a partial allocation of the output variance to each input variable. The variance being a well interpretable and understandable quantity has lead to the Sobol' indices being used frequently in GSA. The functional decomposition of variance via an High Dimensional Model Representation (HDMR) is often referred to as ANOVA (ANalysis Of VAriance). However, it is only formulated for independent input variables. Tackling this issue, Li et al. (2010) and Kucherenko et al. (2012) proposed two different methods of variance decomposition for correlated input variables, both generalising the Sobol' indices in different ways. To clearly show the relationship between each new approach and the classical Sobol' indices, first two ways of defining the Sobol' indices are presented in Sections 2.2 and 2.3. Afterwards, the methodology of the two approaches for correlated inputs is presented in Sections 2.4 resp. 2.5. As mentioned, this section will focus on the methodology of the approaches, while details on implementation will be given in Section 3.

2.2 Sobol' indices using an HDMR

The original idea behind the Sobol' indices is to represent the model as a sum of component functions with increasing dimensionality. Such a representation is called High Dimensional Model Representation (HDMR). For independent variables $\mathbf{X}$ and a square-integrable model $\mathcal{M}$ with finite variance there exists a unique decomposition (Sobol', 1993; Le Gratiet et al., 2017):

$$Y = \mathcal{M}(\mathbf{x}) = \mathcal{M}_0 + \sum_{i=1}^{M} \mathcal{M}_i(x_i) + \sum_{1 \le i < j \le M} \mathcal{M}_{ij}(x_i, x_j) + \cdots + \mathcal{M}_{12\ldots M}(\mathbf{x}) \,, \tag{5}$$

where $\mathcal{M}_0$ is a constant (indeed, it is the mean of Y) and does not depend on any variable $\mathbf{X}$, X_i depicts a single variable of $\mathbf{X} = (X_1, \ldots, X_M)$ and the component functions are orthogonal to each other. Using this HDMR, which is called the Sobol'-Hoeffding decomposition, the definition of importance indices is straightforward. The first order index of X_i is the covariance of the function solely depending on X_i and the output Y normalized by the total output variance:

$$S_i = \frac{\mathrm{Cov}[\mathcal{M}_i(x_i), Y]}{\mathrm{Var}[Y]} \,, \tag{6}$$

where $\mathrm{Cov}[\bullet, \bullet]$ depicts the covariance between the two arguments and $\mathrm{Var}[\bullet]$ is the variance of the argument. Since Y can be expressed as in Eq. (5), it also contains $\mathcal{M}_i(x_i)$.

Therefore, the covariance term can be split up:

$$S_i = \frac{\text{Var}[\mathcal{M}_i(x_i)]}{\text{Var}[Y]} + \frac{\text{Cov}[\mathcal{M}_i(x_i), (Y - \mathcal{M}_i(x_i))]}{\text{Var}[Y]} \,. \tag{7}$$

The aforementioned orthogonality of the component functions implies that there is no covariance between them. Hence, the second term in Eq. (7) goes to zero and the first order index of variable X_i is simply the variance of the component function solely depending on X_i divided by the total variance:

$$S_i = \frac{\text{Var}_i[\mathcal{M}_i(x_i)]}{\text{Var}[Y]} \,. \tag{8}$$

The term "first order" highlights that this index depicts the influence of one single variable X_i. The same can be done with the bivariate component functions, leading to the so-called second order indices:

$$S_{ij} = \frac{\text{Var}_{ij}[\mathcal{M}_{ij}(x_i, x_j)]}{\text{Var}[Y]} \,. \tag{9}$$

The second order index represents the effect of interaction between X_i and X_j. Extending this idea onto more than two variables, it is possible to define a $|\mathbf{u}|$-th order index of subset $\mathbf{u}$:

$$S_{\mathbf{u}} = \frac{\text{Var}_{\mathbf{u}}[\mathcal{M}_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})]}{\text{Var}[Y]} \,. \tag{10}$$

Using these observations, the total variance can be formulated as:

$$\text{Var}[Y] = \sum_{i=1}^{M} \text{Var}[\mathcal{M}_i(x_i)] + \sum_{1 \le i < j \le M} \text{Var}[\mathcal{M}_{ij}(x_i, x_j)] + \cdots + \text{Var}[\mathcal{M}_{12\ldots M}(\mathbf{x})] \,. \tag{11}$$

The variance of the sum of the component functions is, in fact, the sum of the variances of the decomposition. Thus, the Sobol' indices add up to 1 and are effectively shares of the total variance caused by the component functions.

Various combinations of those indices can be computed. Often used is the so-called total index of X_i, combining all Sobol' indices including subscript i:

$$\begin{aligned} S_i^T = S_i + \sum_{j \neq i} S_{ij} + \sum_{1 \le j < k \le M,\, \{j,k\} \neq i} S_{ijk} + \cdots = \sum_{i \in \mathbf{w}} S_{\mathbf{w}} = \\ = \frac{1}{\text{Var}[Y]} \sum_{i \in \mathbf{w}} \text{Var}[\mathcal{M}_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}})] \,. \end{aligned} \tag{12}$$

The total index includes the structural and all interaction effects of an input variable. The total indices of two interacting variables both fully include the interaction effects. As a result, summing all total indices counts interactive effects multiple times and therefore will be greater than (or equal to, if no interactions exist) 1.

The closed index of a subset $\mathbf{u}$ is in fact not considered a Sobol' index, but also used in

this context. It is defined as follows:

$$S_{\mathbf{u}}^{clo} = \frac{\mathrm{Var}_{\mathbf{u}}[\mathcal{M}_{\mathbf{u}}(\mathbf{x}_{\mathbf{u}})]}{\mathrm{Var}[Y]} + \sum_{\mathbf{w} \subseteq \mathbf{u}} \frac{\mathrm{Var}_{\mathbf{w}}[\mathcal{M}_{\mathbf{w}}(\mathbf{x}_{\mathbf{w}})]}{\mathrm{Var}[Y]} \,. \tag{13}$$

It is similar to the $|\mathbf{u}|$-th order index of subset $\mathbf{u}$ but also includes all lower order indices of variables $\mathbf{X}_{\mathbf{u}}$. It is called "closed" because it includes all first order indices of the variables in $\mathbf{X}_{\mathbf{u}}$ and the interactions between them, but no interactive effects with variables not included in the subset.

2.3 Sobol' indices as expectation values

A more intuitive way to define the importance of input X_i is to analyse how the model output Y changes for different values of variable X_i:

$$S_i = \frac{\mathrm{Var}_i[Y|X_i]}{\mathrm{Var}[Y]} \,. \tag{14}$$

The value of Y for a certain value of X_i can be calculated as the average of the model evaluations from a sample of $\mathbf{X}_{\sim i}$ and a given $X_i = x_i^*$, where $\mathbf{X}_{\sim i}$ are all variables except X_i. In mathematical terms, this is $\mathrm{E}_{\sim i}[\mathcal{M}(X_i = x_i^*, \mathcal{X}_{\sim i})] = \mathrm{E}_{\sim i}[Y|X_i]$, where $\mathrm{E}[\bullet]$ describes the expected value of the argument and $\mathcal{X}_{\sim i}$ is a sample of $\mathbf{X}_{\sim i}$. Using this, Eq. (14) can be rewritten:

$$S_i = \frac{\mathrm{Var}_i[\mathrm{E}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]} \,. \tag{15}$$

This is the first order Sobol' index. It measures how the expected value of Y varies for different values of X_i. It can be calculated directly in a double loop Monte Carlo (MC) estimation: for each realisation x_i^* of random variable X_i, the expectation $\mathrm{E}_{\sim i}[Y|x_i^*]$ is calculated from a sample of the remaining variables $\mathcal{X}_{\sim i}$ and x_i^*, forming the inner loop. The same is repeated for varying values of X_i, which constitutes the outer loop.

If not only one, but two variables are fixed at certain values, second order indices can be computed analogously. To make them complementary to the first order indices and only represent the interaction effects between the two variables, the first order indices have to be subtracted:

$$S_{ij} = \frac{\mathrm{Var}_{i,j}[\mathrm{E}_{\sim\{i,j\}}[Y|X_i, X_j]]}{\mathrm{Var}[Y]} - S_i - S_j \,. \tag{16}$$

Consequently, for a subset $\mathbf{u}$ of indices it is possible to define the $|\mathbf{u}|$-th order index:

$$S_{\mathbf{u}} = \frac{\mathrm{Var}_{\mathbf{u}}[\mathrm{E}_{\sim \mathbf{u}}[Y|\mathbf{X}_{\mathbf{u}}]]}{\mathrm{Var}[Y]} - \sum_{\mathbf{w} \subset \mathbf{u}} S_{\mathbf{w}} \,. \tag{17}$$

The total index of X_i is obtained as the sum of all Sobol' indices including i:

$$S_i^T = \sum_{i \in \mathbf{u}} S_{\mathbf{u}} \,. \tag{18}$$

According to the law of total variance (Cramer and Kamps, 2017), the variance of the output can be split up as follows:

$$\mathrm{Var}[Y] = \mathrm{Var}_i[\mathrm{E}_{\sim i}[Y|X_i]] + \mathrm{E}_i[\mathrm{Var}_{\sim i}[Y|X_i]] \,. \tag{19}$$

Normalizing it by the output variance, one finds the sum of two shares:

$$1 = \frac{\mathrm{Var}_i[\mathrm{E}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]} + \frac{\mathrm{E}_i[\mathrm{Var}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]} \,. \tag{20}$$

Since the first term is the first order index of X_i (see Eq. (15)), the second term has to include all combined effects of the remaining variables $\mathbf{X}_{\sim i}$. It is thus the total index of subset $\sim i$:

$$1 = S_i + S_{\sim i}^T \,, \tag{21}$$

where:

$$S_i = \frac{\mathrm{Var}_i[\mathrm{E}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]} \,. \tag{22}$$

$$S_{\sim i}^T = \frac{\mathrm{E}_i[\mathrm{Var}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]} \,. \tag{23}$$

Interchanging the subscripts i and $\sim i$ in Eq. (23), the total index of X_i can also be expressed as an expectation:

$$S_i^T = \frac{\mathrm{E}_{\sim i}[\mathrm{Var}_i[Y|\mathbf{X}_{\sim i}]]}{\mathrm{Var}[Y]} \,. \tag{24}$$

This index represents the expected variance of Y, when only X_i is varied. This index can also be computed in a double-loop. However, this time in the inner loop the values of $\mathbf{X}_{\sim i}$ are fixed and the variance of Y for changing values of X_i is computed. This variance is averaged over different values of $\mathbf{X}_{\sim i}$.

As in the previous section, the closed index of subset $\mathbf{u}$ can be formulated:

$$S_{\mathbf{u}}^{clo} = \frac{\mathrm{Var}_{\mathbf{u}}[\mathrm{E}_{\sim \mathbf{u}}[Y|\mathbf{X}_{\mathbf{u}}]]}{\mathrm{Var}[Y]} \,. \tag{25}$$

In Kucherenko et al. (2012) this formulation is called first order effect of subset $\mathbf{u}$, since it depicts the first order index of $\mathbf{X}_{\mathbf{u}}$ as if they were one single variable.

Table 1 summarizes the different definitions of the Sobol' indices and the closed indices obtained by the two introduced approaches, namely as expectations and by using an HDMR. Interestingly, it turns out that the two approaches are formally equivalent for independent variables and do yield the same values. The direct comparison shows that the

component function $\mathcal{M}_i(x_i)$ of the HDMR can be interpreted as the conditional expectation $\mathrm{E}_{\sim i}[Y|X_i]$.

Table 1: Sobol' indices as expectations and using an HDMR

	As expectations	Using an HDMR
$S_i \cdot \mathrm{Var}[Y]$	$\mathrm{Var}_i[\mathrm{E}_{\sim i}[Y\|X_i]]$	$\mathrm{Var}[\mathcal{M}_i(x_i)]$
$S_i^T \cdot \mathrm{Var}[Y]$	$\mathrm{E}_{\sim i}[\mathrm{Var}_i[Y\|\mathbf{X}_{\sim i}]]$	$\sum_{i\in\mathbf{w}} \mathrm{Var}[\mathcal{M}_\mathbf{w}(\mathbf{x}_\mathbf{w})]$
$S_\mathbf{u} \cdot \mathrm{Var}[Y]$	$\mathrm{Var}_\mathbf{u}[\mathrm{E}_{\sim\mathbf{u}}[Y\|\mathbf{X}_\mathbf{u}]] - \mathrm{Var}[Y] \cdot \sum_{\mathbf{w}\subset\mathbf{u}} S_\mathbf{w}$	$\mathrm{Var}[\mathcal{M}_\mathbf{u}(\mathbf{x}_\mathbf{u})]$
$S_\mathbf{u}^{clo} \cdot \mathrm{Var}[Y]$	$\mathrm{Var}_\mathbf{u}[\mathrm{E}_{\sim\mathbf{u}}[Y\|\mathbf{X}_\mathbf{u}]]$	$\mathrm{Var}[\mathcal{M}_\mathbf{u}(\mathbf{x}_\mathbf{u})] + \mathrm{Var}[Y] \cdot \sum_{\mathbf{w}\subset\mathbf{u}} S_\mathbf{w}$

2.4 ANCOVA

The ANCOVA (ANalysis of COVAriance), also called SCSA (Structural and Correlated Sensitivity Analysis), was first introduced by Li et al. (2010). As the name suggests, this method aims to produce helpful sensitivity indices for correlated input variables based on the decomposition provided by ANOVA (see Section 2.2). The basic assumption is that there exists an HDMR of the investigated model $Y = \mathcal{M}(\mathbf{X})$. This assumption is critical, because due to the dependency between variables, there does not exist a unique HDMR for a model. The reason for this is that because of correlation between variables the component functions cannot be orthogonal anymore. As a result, there does not exist a unique decomposition (Chastaings et al., 2012). There are, however, ways to cope with this issue. Mara et al. (2015) proposed to do a Rosenblatt transform for every variable X_i, each respecting X_i as the "independent" variable on which the others depend. Thus, M (the amount of variables) sets of independent variables $\mathbf{U}^i$, $i = 1, \ldots, M$ are created. Finally, from each set of variables $\mathbf{U}^i$ and a fitted function g_i, such that $g_i(\mathbf{U}^i) = \mathcal{M}(\mathbf{X}) = Y$, the Sobol' indices of the underlying variable X_i are calculated.

Another approach, proposed by Caniou (2012) and the one pursued in this thesis, is to set up an HDMR assuming *independence* between the variables. In this case the HDMR in Eq. (5) is still a unique functional decomposition of the model. After the HDMR is set up, the importance of X_i can again be defined as the covariance between $\mathcal{M}_i(x_i)$ and Y and the term can be split up as in Eq. (7). However, since the HDMR will be applied on samples of correlated variables, the statement of zero covariance between the different terms does not hold anymore. In this case, the second term in Eq. (7) is not zero and there are two contributions to S_i. These two summands were proposed by Li et al. (2010) as the *uncorrelative* (or structural) and *correlative* importance indices of X_i.

Analysing this method, Caniou (2012) realised that the second term $(Y - \mathcal{M}_i(x_i))$ includes terms $\mathcal{M}_{\mathbf{u}}(\mathbf{x_u})$, where $i \in \mathbf{u}$. Covariance between $\mathcal{M}_i(x_i)$ and $\mathcal{M}_{\mathbf{u}}$ is not only due to correlation but can also be due to the fact that $\mathbf{u}$ contains i. For this reason, he proposes to split the covariance term up into an interactive and correlative term. As a result of this extension, the first order index now reads:

$$S_i = S_i^U + S_i^I + S_i^C \,, \tag{26}$$

where S_i^U, S_i^I and S_i^C represent the uncorrelative, interactive and correlative index of X_i respectively. They are defined as:

$$S_i^U = \frac{\mathrm{Var}[\mathcal{M}_i(x_i)]}{\mathrm{Var}[Y]} \,, \tag{27}$$

$$S_i^I = \frac{\mathrm{Cov}[\mathcal{M}_i(x_i), \sum_{i\in\mathbf{u}} \mathcal{M}_{\mathbf{u}}(\mathbf{x})]}{\mathrm{Var}[Y]} \tag{28}$$

and:

$$S_i^C = \frac{\mathrm{Cov}[\mathcal{M}_i(x_i), \sum_{i\notin\mathbf{v}} \mathcal{M}_{\mathbf{v}}(\mathbf{x})]}{\mathrm{Var}[Y]} \,. \tag{29}$$

where $\{\mathcal{M}_{\mathbf{u}}, \mathcal{M}_{\mathbf{v}}\} \in (Y - \mathcal{M}_i)$ and $\mathbf{u} \cap \mathbf{v} = \emptyset$. This split aims to separate the effects of X_i as detailed as possible. The interactive index S_i^I includes the structural influence of X_i in $\mathcal{M}_i(x_i)$ and $\sum_{i\in\mathbf{u}} \mathcal{M}_{\mathbf{u}}(\mathbf{x})$ as well as correlation effects between the terms. Thus, some confusion between interactive and correlative influence of X_i remains. Since for total indices, such lower order indices are summed up, this confusion only grows stronger. For this reason, the definition of proper total ANCOVA indices remains open.

2.5 Direct decomposition of variance and MCS estimation

The Sobol' indices introduced in Sections 2.2 and 2.3 are formulated for independent inputs only. The introduction of correlation brings the need for a new definition of importance indices. Since there exists no unique HDMR for dependent variables (see Section 2.4), it seems reasonable to generalise the Sobol' indices based on the formulation as expectations given in Section 2.3. Kucherenko et al. (2012) proposed to define sensitivity indices using the direct decomposition of variance with the law of total variance. If the input variables $\mathbf{X}$ are divided into two complementary subsets $\mathbf{X_u}$ and $\mathbf{X}_{\sim\mathbf{u}}$, the total variance is assigned as the following sum (Cramer and Kamps, 2017):

$$\mathrm{Var}[Y] = \mathrm{Var}_{\mathbf{u}}[\mathrm{E}_{\sim\mathbf{u}}[Y|\mathbf{X_u}]] + \mathrm{E}_{\mathbf{u}}[\mathrm{Var}_{\sim\mathbf{u}}[Y|\mathbf{X_u}]] \,. \tag{30}$$

As seen in Section 2.3, the first summand is the closed index of subset $\mathbf{u}$. Thus, the second term has to represent the total index of subset $\sim\mathbf{u}$. For $\mathbf{u} = i$ and normalized by $\mathrm{Var}[Y]$ one ends up with Eq. (21). The first order and total indices are again formulated as in

Eq.s (22) and (24):

$$S_i = \frac{\text{Var}_i[\text{E}_{\sim i}[Y|X_i]]}{\text{Var}[Y]}\,. \tag{31}$$

$$S_i^T = \frac{\text{E}_{\sim i}[\text{Var}_i[Y|\mathbf{X}_{\sim i}]]}{\text{Var}[Y]}\,. \tag{32}$$

They are formally the same definitions as im the independent case, however their estimation through Monte Carlo estimation becomes more complex. Because of dependence, fixing X_i to a certain value will change the distribution of the variables correlated to X_i. For a given $X_i = x_i^*$, the other variables $\mathbf{X}_{\sim i}$ must be sampled conditionally, producing the sample $\hat{\mathcal{X}}_{\sim i}$. The averaged value of the model runs $\mathcal{M}(x_i^*, \hat{\mathcal{X}}_{\sim i})$ is $\text{E}_{\sim i}[Y|X_i = x_i^*]$. This is the new, adjusted inner loop for Kucherenko's indices and the difference to the classical Sobol' indices. This procedure is repeated for different values of X_i, forming the outer loop, and the variance of the expectations is calculated. Normalizing by the total variance results in the first order index S_i.

Analogously, when $\mathbf{X}_{\sim i}$ are fixed to certain values $\mathbf{x}_{\sim i}^*$, the remaining variable X_i has to be sampled conditionally, producing $\hat{\mathcal{X}}_i$. The variation of $\mathcal{M}(\hat{\mathcal{X}}_i, \mathbf{x}_{\sim i}^*)$ is averaged over different values of $\mathbf{x}_{\sim i}^*$ and normalized by $\text{Var}[Y]$, resulting in the total index S_i^T. The steps of calculation are the same as for the Sobol' indices, but the samples have to be set up conditionally. For arbitrary marginal distributions and copulas this is not a trivial matter.

2.6 Optimisation of the double loop MCS estimation

In practice, it is expensive, in terms of computation time to run the computational models hundreds of millions of times in an MC double loop. The precision of the estimations improves for increasing loop sizes and can produce highly varying results for smaller loop sizes. In order to shorten computation time, Kucherenko et al. (2012) replaces the inner loop by an estimator and therefore only runs "one loop". This method leads to accurate results for a sample size of $N = 2^{13} = 8192$. This is a considerable improvement over the double loop approach with $(N_{il} \cdot N_{ol} + N_{var})$ samples, where N_{il} is the size of the inner loop, N_{ol} is the size of the outer loop and N_{var} is the amount of samples used for the estimation of the total variance.

The optimisation used in the context of this thesis, is to substitute an HDMR for the actual model. Those are usually evaluable way more efficiently. To do this, the HDMR can be set up, as for the ANCOVA indices, for independent variables and then used to calculate the model outputs of the conditional samples. The accuracy of the HDMR evaluations depends on the amount of samples used for the experimental design. If still the double loop approach is taken, the MCS estimation of the sensitivity indices continues to depend on the loop sizes, but the calculations are done much faster. Notably, the HDMR is an

approximation of the actual model. The goal is to get a good predictor with only a relatively small amount of model runs of the actual model. The accuracy of the HDMR can be measured in various ways. The value used in this thesis is the relative generalisation error:

$$\epsilon_{gen} = \frac{\mathrm{E}[(\mathcal{M}(\mathbf{X}) - \mathcal{M}^{HDMR}(\mathbf{X}))^2]}{\mathrm{Var}[\mathcal{M}(\mathbf{X})]}\,. \tag{33}$$

The difference between the actual model output and the one from the surrogate is squared to get positive numbers and then averaged over many samples. This expectation is normalized by the variance of the actual model, which is also a quadratic value. If the relative generalisation error goes towards zero, the model provides good predictions. To get a good model, the experimental design has to be sufficiently large, which depends on the model complexity. Values are presented in Section 3.6.

2.7 Interpretation issues

As mentioned in Section 2.2, the Sobol' indices represent a share of the total variance caused by the respective variable or subset of variables. The first order indices only include structural influence of X_i in the model $\mathcal{M}(\mathbf{x})$ whereas the total indices additionally include all interaction effects of said variable. In case variable X_i interacts with other variables, the total index S_i^T is larger than the first order index S_i.

The application of the Sobol' indices will be shown for the following example taken from Sudret and Marelli (2016). The object of interest is a basic structural element, the simply supported beam (see Figure 1). For a homogeneous beam with Young's modulus E, length L and a rectangular cross-section of width b and height h, the midspan deflection V under a uniform load p can be calculated as follows:

$$V = \frac{5}{32}\frac{pL^4}{E\,bh^3}\,. \tag{34}$$

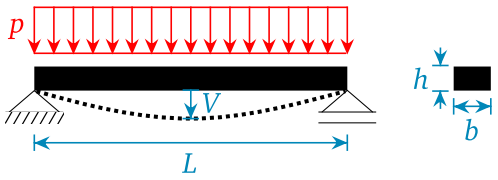


Figure 1: Simply supported beam under uniform load

The input variables are considered independent and are lognormally distributed $X_i \sim \mathcal{LN}(\lambda_i, \zeta_i)$, $i \in \{b, h, L, E, p\}$ with means and standard deviations shown in Table 2. UQLab allows for a quick and easy calculation of the Sobol' indices if the model and the input parameters are provided. The Sobol' indices are shown in the last two columns of Table 2. The expected value of the deflection V is 0.150 mm and the total variance is $2.44 \cdot 10^{-3}$ mm^2.

Table 2: Mean and standard deviations of the input variables as well as the Sobol' indices

Variable	Mean μ	Std. deviation σ	First order index S_i	Total index S_I^T
b	0.15 m	7.5 mm	0.0263	0.0295
h	0.3 m	15 mm	0.2456	0.2645
L	5 m	50 mm	0.0137	0.0190
E	30'000 MPa	4'500 MPa	0.2438	0.2608
p	0.01 MN/m	0.002 MN/m	0.4325	0.4564

Using these indices, the input variables can be sorted by their importance. In this example, the first order and total indices of a variable have the same rank. The load p is the most important parameter with nearly half of the output variance based on its variation. The Young's modulus E and the height h of the beam are about equally important, each contributing about a quarter of the output variance. The width b and length L of the beam are almost non-influential to the system's variance. Suppose the goal is to reduce the variance of the deflection in order to decrease the probability of reaching the service limit state. According to the Sobol' indices, reducing the variance of p should influence the output variance the most. If the standard deviation of p can be decreased by 15% to 0.0017 MN/m, the total variance drops to $2.14 \cdot 10^{-3}$ mm^2, a decrease of about 8.5%. The same decrease of standard deviation for selected other variables results in a total variance of $2.24 \cdot 10^{-3}$ mm^2 for E and $2.42 \cdot 10^{-3}$ mm^2 for L. This shows that the variables with higher Sobol' indices have a stronger influence on the model variance. The variables with small indices have a negligible effect and can be set to a constant value to simplify the model without great impact on the result.

In contrast to the Sobol' indices, the interpretation of Kucherenko's and Caniou's indices (see Table 3) is not obvious. What do certain values mean? Can Kucherenko's total indices be smaller than first order indices? Can ANCOVA indices be negative? If so, why Such questions are unanswered yet, although there is much potential for further insight into how dependence influences total variance. And specifically for the two investigated approaches, there is more to explore, including: do the two approaches yield the same values? If not, why not, since they did so for the uncorrelated case? And will variables be ranked in the same order by both indices? A better understanding of the indices can lead to an overall better understanding of the computational model and its variables. Such knowledge allows for more precise allocation of resources when aiming for variance reduction. Through application of the two approaches on different model functions and the discussion of the results in Section 4 answers to some of those questions or new impulses for further work should be found.

Table 3: Methods leading to variance-based importance indices for correlated variables

Method	Indices of X_i	Requirements
ANCOVA	$S_i^U = \frac{\text{Var}[\mathcal{M}_i(x_i)]}{\text{Var}[Y]}$ $S_i^I = \frac{\text{Cov}[\mathcal{M}_i(x_i), \sum_{i \in \mathbf{u}} \mathcal{M}_{\mathbf{u}}(\mathbf{x})]}{\text{Var}[Y]}$ $S_i^C = \frac{\text{Cov}[\mathcal{M}_i(x_i), \sum_{i \notin \mathbf{v}} \mathcal{M}_{\mathbf{v}}(\mathbf{x})]}{\text{Var}[Y]}$	HDMR of the investigated model
Direct Decomposition	$S_i = \frac{\text{Var}_i[\text{E}_{\sim i}[Y\|X_i]]}{\text{Var}[Y]}$ $S_i^T = \frac{\text{E}_{\sim i}[\text{Var}_i[Y\|\mathbf{X}_{\sim i}]]}{\text{Var}[Y]}$	Knowledge about conditional distributions and possibility to sample from them

3 Implementation and Validation

3.1 Overview

The methods for correlated variables (see Sections 2.4 and 2.5) were implemented into UQLab (Marelli and Sudret, 2014), a software developed at ETH Zürich. UQLab is a suitable tool to easily perform different tasks of uncertainty quantification such as probabilistic input modelling, polynomial chaos expansion (PCE), reliability analysis and sensitivity analysis. So far, sensitivity analyses can only be performed for independent input variables. Nevertheless, this software provides many useful features and the methods implemented within the scope of this master's thesis should eventually be included into UQLab. For these reasons MATLAB with the UQLab software was used in this thesis.

3.2 Direct decomposition of variance

The general formulation of the Kucherenko indices is the same as for the Sobol' indices. However, as seen in Section 2.5, the difference lies in the samples, from which the expectations and variances are calculated. Due to correlation, fixing one variable to a certain value will influence the distribution of the remaining ones. Therefore, they have to be sampled conditionally. It is not obvious, how an arbitrary joint distribution of variables with any marginal distribution types would be influenced by fixing one variable to a value. As mentioned in the introduction, in this work dependence will be modelled using the marginal distributions of the random variables and a Gaussian copula. Using the Nataf transform (Lebrun and Dutfoy, 2009), the distributions can be transformed into standard normal space. There, a joint Gaussian distribution with covariance matrix $\mathbf{\Sigma}_G$ can be conditioned to sample the remaining variables, if one (or more) is fixed to a certain value x_i^* (resp. $\mathbf{x}_\mathbf{u}^*$). Because a Gaussian copula is applied, its parameter matrix $\mathbf{\Sigma}_C = \mathbf{\Sigma}_G = \mathbf{\Sigma}$. It is assumed that all the inputs with distribution type and moments, as well as the correlation matrix between the inputs $\mathbf{\Sigma}$ are provided. Based on this, the following two conditional sampling algorithms were developed and implemented.

3.2.1 Conditioning on one variable

For the first order index of X_i the model $\mathcal{M}(\mathbf{X}) = Y$ has to be evaluated on samples where $\mathbf{X}_{\sim i}$ are conditioned on $X_i = x_i^*$:

$$S_i = \frac{\mathrm{Var}_i[\mathrm{E}_{\sim i}[Y|X_i]]}{\mathrm{Var}[Y]}. \tag{35}$$

The variable X_i is sampled N_{ol} times, where N_{ol} is the amount of outer loop samples. Then, each of the realisations x_i^* is transformed into standard normal space, resulting

in z_i^*, where its influence on the joint distribution of $\mathbf{Z}_{\sim i} = \Phi^{-1}(F_{\mathbf{X}_{\sim i}}(\mathbf{X}_{\sim i}))$ (i.e. $\mathbf{X}_{\sim i}$ in normal space) can be determined. Remember that $\mathbf{Z} = (Z_i, \mathbf{Z}_{\sim i})$ has a multivariate zero-mean joint Gaussian distribution, meaning $\boldsymbol{\mu} = \mathbf{0}$ and a given correlation matrix $\boldsymbol{\Sigma}$. For $Z_i = z_i^*$ the distribution parameters are modified and the variables $\hat{\mathbf{Z}}_{\sim i}$ are sampled conditionally according to those, the "ˆ" circumflex accent (hat) symbolizing the conditional nature of the distribution. Such a sample combined with z_i^* forms a conditional sample $\mathcal{Z}_{|_{x_i^*}}$. At this point, one inner loop is completed. The amount of sample points is depicted by N_{il}. This procedure is repeated for different values of x_i^*, forming the outer loop. Notably, the sample $\mathcal{Z}_{|_{x_i^*}}$ is still in standard space. The transformation into the actual distributions happens separately: each variable Z_j is transformed from $\mathcal{N}(0,1)$ into $F_{X_j}(x_j)$. The algorithm is summarised as follows:

1. Generate N_{ol} samples of $X_i \sim F_{X_i}(x_i)$ from the marginal distribution. This is the outer loop sample for the double loop Monte Carlo estimation.

   For each variable X_i, $i = 1, \ldots, M$**:**

2. Transform each realisation $X_i = x_i^*$ into normal space:

$$z_i^* = \Phi^{-1}(F_{X_i}(x_i^*))\,. \tag{36}$$

3. Since a Gaussian Copula is applied, the other inputs can now be sampled with the conditional distribution:

$$\hat{\mathbf{Z}}_{\sim i} \sim \mathcal{N}(\mu_{\mathbf{Z}_{\sim i}} + \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{ii}^{-1}(z_i^* - \mu_{Z_i})\,, \quad \boldsymbol{\Sigma}_{\sim i \sim i} - \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_{ii}^{-1} \boldsymbol{\Sigma}_i^T)\,. \tag{37}$$

   where $\boldsymbol{\Sigma}_i$ is the i-th column vector of $\boldsymbol{\Sigma}$ without σ_i^2, $\boldsymbol{\Sigma}_{\sim i \sim i}$ is the $((M-1) \times (M-1))$ covariance matrix of all variables except X_i and $\boldsymbol{\Sigma}_{ii}$ is the variance of X_i (in standard normal space equal to 1). Since $\mathbf{Z}_{\sim i}$ and Z_i are in the normal space, their expected values are 0. Eq. (37) can then be simplified into:

$$\hat{\mathbf{Z}}_{\sim i} \sim \mathcal{N}(\boldsymbol{\Sigma}_i z_i^*\,, \quad \boldsymbol{\Sigma}_{\sim i \sim i} - \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_i^T)\,. \tag{38}$$

   Interestingly, setting $X_i = x_i^*$ does condition both the mean and the variance of $\mathbf{X}_{\sim i}$. However, the conditioned variance does not depend on the value of the conditioning variable, only the mean does. Changing the mean of the normal distribution will simply shift the distribution curve along the input axis. Using this fact in combination with Eq. (38) one can write the conditional variable as:

$$\hat{\mathbf{z}}_{\sim i} = \boldsymbol{\Sigma}_i z_i^* + \hat{\mathbf{z}}_{\sim i}^0\,, \tag{39}$$

   where $\hat{\mathbf{Z}}_{\sim i}^0 \sim \mathcal{N}(0\,, \quad \boldsymbol{\Sigma}_{\sim i \sim i} - \boldsymbol{\Sigma}_i \boldsymbol{\Sigma}_i^T)\,.$

4. Combining z_i^* and a sample of $\hat{\mathbf{Z}}_{\sim i}$ delivers a conditional sample for $X_i = x_i^*$ in

normal space:

$$\mathcal{Z}_{|x_i^*} = \begin{bmatrix} \hat{z}_{11} & \cdots & z_i^* & \cdots & \hat{z}_{1M} \\ \vdots & \ddots & \vdots & \iddots & \vdots \\ \hat{z}_{k1} & \cdots & z_i^* & \cdots & \hat{z}_{kM} \\ \vdots & \iddots & \vdots & \ddots & \vdots \\ \hat{z}_{N_{il}1} & \cdots & z_i^* & \cdots & \hat{z}_{N_{il}M} \end{bmatrix} .$$

where N_{il} is the amount of conditional samples in the inner loop.

5. Each random variable is separately transformed back into its marginal distribution space:

$$x_{j|_{x_i^*}} = F_{X_j}^{-1}(\Phi(z_j)), \qquad j = 1, \ldots, M . \tag{40}$$

6. The conditional sample for $X_i = x_i^*$ stands. The procedure is repeated for all realisations of X_i and consequently also for each random variable.

Using Eq. (39) allows one to reducethe sampling time. Instead of producing a new conditional sample for each realisation of X_i (i.e. N_{ol} conditional samples), only one sample $\hat{\mathbf{z}}_{\sim i}^0$ is produced with a mean equal to 0 and then shifted for each realisation by $\mathbf{\Sigma}_i z_i^*$, where $z_i^* = \Phi^{-1}(F_{X_i}(x_i^*))$. The implemented MATLAB function to sample conditioned on one variable is called $getKUCHsamp1.m$. It produces a $(N_{ol} \times M)$ cell-array. At each position (i, j) a sample is stored with $\mathbf{X}_{\sim j}$ sampled conditioned on $X_j = x_{ij}$, being the i-th point of a sample of $X_j \sim F_{X_j}(x_j)$.

3.2.2 Conditioning on all variables except one

To compute the total index of variable X_i, the model has to be evaluated on samples with X_i conditional on $\mathbf{X}_{\sim i} = \mathbf{x}_{\sim i}^*$:

$$S_i^T = \frac{\mathrm{E}_{\sim i}[\mathrm{Var}_i[Y|\mathbf{X}_{\sim i}]]}{\mathrm{Var}[Y]} . \tag{41}$$

For unconditioned variables, it does not matter if they are sampled from their joint distribution $F_{\mathbf{X}}$ and then transformed into normal space or sampled from $\mathcal{N}^M(\mathbf{0}, \mathbf{\Sigma})$ right away. For this reason, in this algorithm $\mathbf{Z}_{\sim i}$ are sampled in Gaussian space. For each realisation $\mathbf{Z}_{\sim i} = \mathbf{z}_{\sim i}^*$, the remaining variable $\hat{Z}_i$ can then be sampled conditionally (Kucherenko et al., 2012). The realisation $\mathbf{z}_{\sim i}^*$ combined with a sample of $\hat{Z}_i$ forms a conditional sample in the standard normal space. That is the inner loop. This is repeated for each realisation of $\mathbf{Z}_{\sim i}$, forming the outer loop. As before, the transformations from normal into the actual distribution is done for each variable separately. The algorithm is summarised as follows:

1. Generate N_{ol} samples of $\mathbf{Z}$ in the standard normal space with zero-mean and covariance matrix $\mathbf{\Sigma}$:

$$\mathbf{Z} \sim \mathcal{N}^M(\mathbf{0}, \mathbf{\Sigma})\,. \tag{42}$$

For each variable Z_i, $i = 1, \ldots, M$**:**

2. For each realisation of the other variables $\mathbf{Z}_{\sim i} = \mathbf{z}^*_{\sim i}$, the chosen variable can be conditionally sampled with:

$$\hat{Z}_i \sim \mathcal{N}(\mu_{Z_i} + \mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}(\mathbf{z}^*_{\sim i} - \boldsymbol{\mu}_{\mathbf{Z}_{\sim i}})\,, \quad \mathbf{\Sigma}_{ii} - \mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}\mathbf{\Sigma}^T_{\sim i})\,. \tag{43}$$

where $\mathbf{\Sigma}_{\sim i}$ is the i-th row vector of $\mathbf{\Sigma}$ without σ_i^2, $\mathbf{\Sigma}_{\sim i\sim i}$ is the $((M-1)\times(M-1))$ covariance matrix of all variables except Z_i and $\mathbf{\Sigma}_{ii}$ is the variance of Z_i (in standard space equal to 1). Since $\mathbf{Z}_{\sim i}$ and the unconditioned Z_i are in the standard normal space with zero-mean, their expected values are 0. Eq. (43) can be rewritten as:

$$\hat{Z}_i \sim \mathcal{N}(\mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}\mathbf{z}^*_{\sim i}\,, \quad 1 - \mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}\mathbf{\Sigma}^T_{\sim i})\,. \tag{44}$$

Again using the fact that only the mean depends on the conditioning values, the variable can be expressed as a sum:

$$\hat{z}_i = \mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}\mathbf{z}^*_{\sim i} + \hat{z}^0_i\,, \tag{45}$$

where $\hat{Z}^0_i \sim \mathcal{N}(0\,, \quad 1 - \mathbf{\Sigma}_{\sim i}\mathbf{\Sigma}^{-1}_{\sim i\sim i}\mathbf{\Sigma}^T_{\sim i})\,.$

3. Combining $\mathbf{z}^*_{\sim i}$ and a sample of $\hat{Z}_i$ delivers the conditional sample for $\mathbf{X}_{\sim i} = \mathbf{x}^*_{\sim i} = F^{-1}_{\mathbf{X}_{\sim i}}(\Phi^{\mathrm{M-1}}(\mathbf{z}^*_{\sim i}))$ in the normal space:

$$\mathcal{Z}|_{\mathbf{x}^*_{\sim i}} = \begin{bmatrix} z^*_1 & \cdots & z^*_{i-1} & \hat{z}_{i1} & z^*_{i+1} & \cdots & z^*_M \\ \vdots & \iddots & \vdots & \vdots & \vdots & \ddots & \vdots \\ z^*_1 & \cdots & z^*_{i-1} & \hat{z}_{ik} & z^*_{i+1} & \cdots & z^*_M \\ \vdots & \ddots & \vdots & \vdots & \vdots & \iddots & \vdots \\ z^*_1 & \cdots & z^*_{i-1} & \hat{z}_{iN_{il}} & z^*_{i+1} & \cdots & z^*_M \end{bmatrix}\,.$$

4. Each column j of this sample is now transformed into the respective marginal distribution of X_j:

$$x_j|_{\mathbf{x}^*_{\sim i}} = F^{-1}_{X_j}(\Phi(z_j)), \qquad j = 1, \ldots, M\,. \tag{46}$$

In this algorithm, the same trick as for the first algorithm is applied: instead of sampling the conditioned variable $\hat{Z}_i$ for every realisation of the outer loop (N_{ol} times), $\hat{Z}^0_i$ is sampled once with zero mean and the conditioned correlation matrix. The conditional

mean for each realisation of $\mathbf{Z}_{\sim i}$ is then added to this sample, producing all outer loop samples in standard space. Similar to the function used to sample conditional on one variable, $getKUCHsamp2.m$ produces a $(N_{ol} \times M)$ cell-array. This time, at each position (i, j) a sample is stored with X_j conditioned on $\mathbf{X}_{\sim j} = \mathbf{x}^*_{\sim j}$, being the i-th realisation of $\mathbf{X}_{\sim j}$ in the outer loop.

3.3 ANCOVA

As seen in Section 2.4, the ANCOVA indices are calculated based on the components of an HDMR of the actual model. However, for correlated input variables, there does not exist a unique HDMR. Caniou (2012) circumvents this issue by setting up the HDMR for independent variables and then just using it as a structural representation of the model. This approach is followed. A PCE is set up for independent variables and used as HDMR of the actual model. A PCE is a projection of the model response onto polynomials, that are orthonormal to each other with respect to the marginal distributions. Therefore, a PCE is unique for independent variables (Xiu, 2010) and defined by the marginal distributions of the variables and the coefficients and the indices (degrees) of the polynomials. To compute the ANCOVA indices, the HDMR's component functions fitting the descriptions in Section 2.4 have to be isolated. This can be done by analysing the indices $\boldsymbol{\alpha}$. The following notation will be used. The subset $\mathbf{w}_i$ contains the indices of polynomials solely containing X_i ($\alpha_i > 0$ and $\alpha_{j\neq i} = 0$), that together make up $\mathcal{M}_i(x_i)$. Subset $\mathbf{u}_i$ contains the indices of polynomials, that contain X_i ($\alpha_i > 0$) and some other variable ($\sum_{j\neq i} \alpha_j > 0$). Finally, subset $\mathbf{v}_i$ includes all other polynomial indices, i.e. the indices of all polynomials that do not include X_i ($\alpha_i = 0$). Table 4 shows a scheme of the three subsets for $i = 1$. Each index subset ($\mathbf{w}$, $\mathbf{u}$ and $\mathbf{v}$) and the corresponding coefficients form a new custom PCE. Those are used to calculate the variances and covariances needed for the indices. The following list summarises the procedure to calculate the indices:

1. Create one sample $\mathcal{X}^I$ from the marginal distributions: $F^I_{\mathbf{X}} = \prod_{i=1}^{M} F_{X_i}(x_i)$
   and one sample $\mathcal{X}^C$ from the correlated distr.: $F_{\mathbf{X}} = \mathcal{C}[F_{X_1}(x_1), \ldots, F_{X_M}(x_M)]$
2. Build a PCE from the independent sample $\mathbf{x}^I$ and $\mathcal{M}(\mathbf{x}^I) = y^I \rightarrow \mathcal{M}^{PCE}$ and calculate its output variance with the correlated sample $\text{Var}[y^{PCE}] = \text{Var}[\mathcal{M}^{PCE}(\mathcal{X}^C)]$

   For each variable X_i, $i = 1, \ldots, M$**:**
3. Identify the $\boldsymbol{\alpha}$-subsets $\mathbf{w}_i$, $\mathbf{u}_i$ and $\mathbf{v}_i$ (for an example see Table 4)
4. Build new custom PCE's for each subset: $\mathcal{M}^{PCE}_{\mathbf{w}_i}$, $\mathcal{M}^{PCE}_{\mathbf{u}_i}$ and $\mathcal{M}^{PCE}_{\mathbf{v}_i}$ and calculate $\text{Var}[\mathcal{M}^{PCE}_{\mathbf{w}_i}(\mathcal{X}^C)]$, $\text{Cov}[\mathcal{M}^{PCE}_{\mathbf{w}_i}(\mathcal{X}^C), \mathcal{M}^{PCE}_{\mathbf{u}_i}(\mathcal{X}^C)]$ and $\text{Cov}[\mathcal{M}^{PCE}_{\mathbf{w}_i}(\mathcal{X}^C), \mathcal{M}^{PCE}_{\mathbf{v}_i}(\mathcal{X}^C)]$ from the correlated sample.

Table 4: Subsets $\mathbf{w}_1$, $\mathbf{u}_1$ and $\mathbf{v}_1$ of $\boldsymbol{\alpha}$.

$\boldsymbol{\alpha}$	X_1	X_2	...	...	X_M
	1	0	0	...	0
$\mathbf{w}_1$	⋮	0	...	...	0
	5	0	...	...	0
	$\alpha_1 > 0$		$\alpha_{j>1} = 0$		
	1	0	2	...	2
$\mathbf{u}_1$	⋮	1	5	...	...
	3	4	...	...	1
	$\alpha_1 > 0$		per row: $\sum_{j=2}^{M} \alpha_j > 0$		
	0	1	5	...	...
$\mathbf{v}_1$	0	...	...	...	...
	0	...	...	...	...
	$\alpha_1 = 0$				

5. Obtain the sensitivity indices S_i^U, S_i^I and S_i^C of X_i by dividing the respective terms by Var$[y^{PCE}]$

3.4 Validation: Direct decomposition

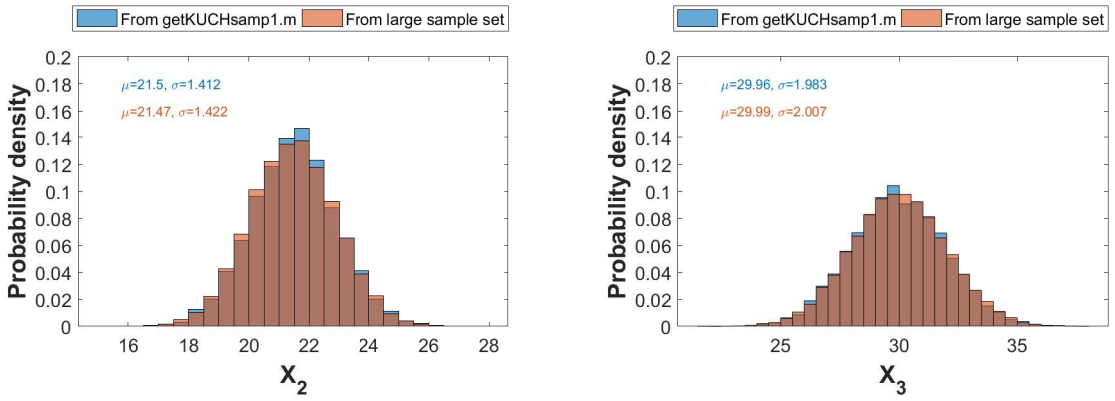
The essential parts of the present implementation of Kucherenko's approach are the sampling methods explained in Section 3.2. Once the conditioned samples are set up, the calculation of the indices consists of ascertaining their expected values or variances, which is a rather straightforward matter compared to the conditioning. For this reason, the implemented sampling algorithms are validated. This is done by comparing the distribution of a sample from the code to the distribution of selected points from a large sample.

3.4.1 Conditioning on one variable

Running the MATLAB code $getKUCHsamp1.m$ produces a cell-array containing the conditioned samples (see also Section 3.2.1). One of those samples (j, i), conditioned on $X_i = x_i^*$, is chosen and the distributions of the conditioned variables are plotted as histograms. The reference sample is set up as follows. A large sample with 10^6 points of $\mathbf{X}$ is obtained from the joint probability distribution $F_{\mathbf{X}}(\mathbf{X})$. From this large sample only the points with x_i close to the reference value x_i^*, i.e. with $0.99x_i^* < x_i < 1.01x_i^*$, are selected, thereby "conditioning" the sample on $X_i \approx x_i^*$. From this sample, the distributions of the variables $\mathbf{X}_{\sim i}$ are plotted for comparison. Below are comparisons for three

variables $X_1 \sim \mathcal{N}(10, 2)$, $X_2 \sim \mathcal{N}(20, 2)$ and $X_3 \sim \mathcal{N}(30, 2)$ and different cases of linear correlations between them.

Case 1: Positive correlation between X_1 and X_2 and no correlation between X_1 and X_3: In this first case, the variables X_1 and X_3 are not correlated while X_1 and X_2 have strong positive correlation $\rho_{12} = 0.7$. The distributions of X_2 and X_3 are conditioned on $X_1 = 12.11$. Figure 2 shows the histograms of the conditioned samples and their statistic moments.



(a) Conditioned distribution of X_2

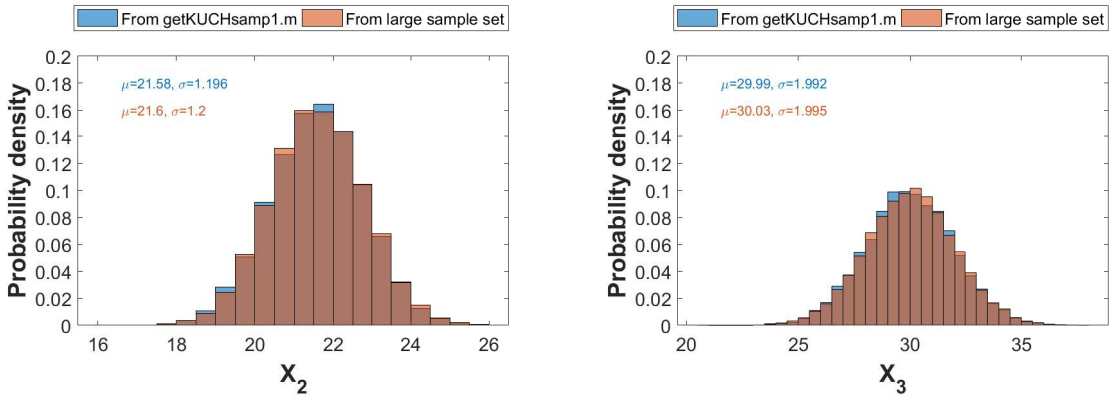
(b) Conditioned distribution of X_3

Figure 2: Distributions of X_2 and X_3 conditioned on $X_1 = 12.11$ from implemented code $getKUCHsamp1.m$ (blue) and reference (red) with distribution parameters.

The value of X_1, on which was conditioned, lies above its mean $\mu_1 = 10$. As a result, it is to be expected, that the positively correlated X_2 will also tend to lie above its mean. This condition is met as shown in Figure 2a: the mean of the conditioned sample is at 21.46 and thereby above the mean $\mu_2 = 20$. The change in standard deviation is harder to estimate. Nevertheless, it can be predicted that the standard deviation should decrease, since X_2 is now restricted to a certain area by X_1. This is also the case: the standard deviation decreased from 2 to 1.43. Moreover, no matter the conditioning value of X_1, the standard deviations of X_2 and X_3 are constant at the values presented in the plots, which is a correct feature as mentioned in Section 3.2.1. The histograms of the samples from $getKUCHsamp1.m$ (blue) fit the references (red) well, they lay right on top of each other. The distribution of X_3 in Figure 2b still matches its marginal $\mu_3 = 30$, $\sigma_3 = 2$. Since there is no correlation, conditioning on any value of X_1 should not influence the distribution of X_3 and, in fact, it does not.

Case 2: Negative correlation between X_1 and X_2 and positive correlation between X_2 and X_3: In this case, X_2 is correlated negatively to X_1, $\rho_{12} = -0.8$, and positively

to X_3, $\rho_{23} = 0.4$. The distributions of X_2 and X_3 are conditioned on $X_1 = 8.02$. Figure 3 shows the conditioned samples as histograms with their moments.



(a) Conditioned distribution of X_2

(b) Conditioned distribution of X_3

Figure 3: Distributions of X_2 and X_3 conditioned on $X_1 = 8.02$ from implemented code $getKUCHsamp1.m$ (blue) and reference (red) with distribution parameters.

The conditioning value 8.02 of X_1 lies below its mean $\mu_1 = 10$. The negatively correlated X_2 should therefore tend to lie above its mean $\mu_2 = 20$ and its standard deviation should be decreased by the conditioning. As Figure 3a shows, the produced data meets these criteria: the mean of the conditioned X_2 is 21.6 and its standard deviation 1.20. The distribution of X_3 (see Figure 3b) is unaltered by the conditioning on X_1: the mean and standard deviation still are at their original values. This behaviour is to be expected, since X_1 and X_3 are not correlated, even tough both are correlated to X_2. Table 5 shows more conditioning values of X_1 with the corresponding moments from the code-generated sample and the reference samples of X_2 and X_3. No matter the conditioning value of X_1, the standard deviation is always the same. The values from the implemented code match the reference values well.

3.4.2 Conditioning on all variables except one

The second developed MATLAB sampling code $getKUCHsamp2.m$ produces samples where one of the variables is conditioned on all the others (see also Section 3.2.2). The conditioning on more than one variable is not as trivial as the conditioning on one. Nevertheless, some predictions can be made and a reference sample shows if the produced samples are reasonable. The same variables as in Section 3.4.1 are used but this time X_1 is conditioned on X_2 and X_3 for different values and cases of correlation. The reference sample is enlarged to 10^7 points since the reduction through conditioning on two variables will decrease the number of selected samples more drastically.

Table 5: Moments of X_2 and X_3 for different conditioning values of X_1 for $\rho_{12} = -0.8$ and ρ_{23}=0.4

	Code generated		Reference	
Conditioning value of X_1	mean	std.	mean	std.
9.69	$\mu_2 = 20.2$	$\sigma_2 = 1.21$	$\mu_2 = 20.2$	$\sigma_2 = 1.20$
	$\mu_3 = 30.0$	$\sigma_3 = 2.00$	$\mu_3 = 30.0$	$\sigma_3 = 2.00$
11.5	$\mu_2 = 18.8$	$\sigma_2 = 1.21$	$\mu_2 = 18.8$	$\sigma_2 = 1.20$
	$\mu_3 = 30.0$	$\sigma_3 = 2.00$	$\mu_3 = 30.0$	$\sigma_3 = 2.00$
10.1	$\mu_2 = 19.9$	$\sigma_2 = 1.21$	$\mu_2 = 19.9$	$\sigma_2 = 1.20$
	$\mu_3 = 30.0$	$\sigma_3 = 2.00$	$\mu_3 = 30.0$	$\sigma_3 = 2.00$
10.7	$\mu_2 = 19.4$	$\sigma_2 = 1.21$	$\mu_2 = 19.4$	$\sigma_2 = 1.20$
	$\mu_3 = 30.0$	$\sigma_3 = 2.00$	$\mu_3 = 30.0$	$\sigma_3 = 2.00$

Case 1: Positive correlation between X_1 and X_2 and no correlation between X_1 and X_3 In this case, X_1 is only correlated to X_2 with $\rho_{12} = 0.7$. The distribution of X_1 is conditioned on $X_2 = 23.27$ and $X_3 = 29.12$. Figure 4 shows the results.

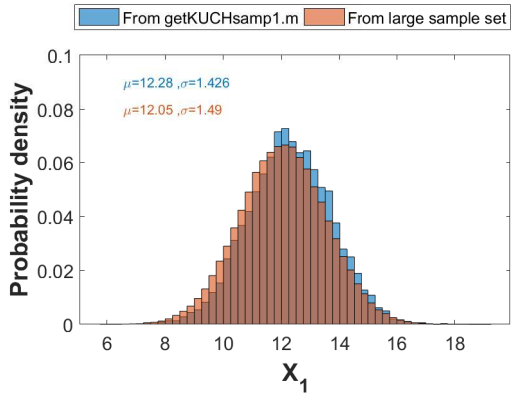


Figure 4: Distribution of X_1 conditioned on $X_2 = 23.11$ and $X_3 = 28.80$ from implemented code $getKUCHsamp2.m$ (blue) and reference (red) with distribution parameters.

The conditioning value of X_2 lies above its mean whereas the one of X_3 lies below its own one. However, since only X_2 is correlated (positively) to X_1, only the former should have influence. Figure 4 shows that X_1 does in fact tend to be below its mean $\mu_1 = 10$. Moreover, the standard deviation is decreased to about $1.4-1.5$. The samples from the code (blue) coincide with the reference from the large sample (red).

Case 2: Positive correlation between X_1 and X_2 and negative correlation between X_1 and X_3 In this second case, X_1 is conditioned on X_2 and X_3 while being correlated to both. The conditioning values are $X_2 = 17.82$ and $X_3 = 30.70$ and the correlation coefficients are $\rho_{12} = 0.7$ and $\rho_{13} = -0.3$. The results are shown in Figure 5.

This time, the conditioning value of X_2 lies below its mean and the one of X_3 lies above its own one. Since X_2 is correlated to X_1 more strongly, it has a stronger influence and X_1 tends to lie below its mean, too, with a conditioned mean around 8.7. Naturally, the standard distribution is reduced as well to about 1.3. The values from the code (blue) match the reference (red) derived from the large sample well. Table 6 shows more conditioning values of X_2 and X_3 with the corresponding moments from the code-generated sample and the reference samples of X_1. Notably, the standard deviation is constant, no matter the conditioning values and the moments of conditional samples match the reference values well.

Overall, the samples from the implemented code show the expected tendencies and fit the reference well. The mean is shifted in the supposed direction and the standard deviation takes on the same value when conditioned on X_1, no matter the conditioning value.

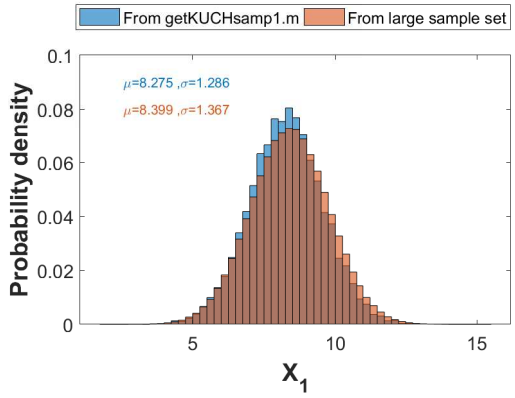


Figure 5: Distribution of X_1 conditioned on $X_2 = 17.82$ and $X_3 = 30.70$ from implemented code $getKUCHsamp2.m$ (blue) and reference (red) with distribution parameters.

The distribution parameters are apparently altered correctly by the codes. Therefore, the codes are considered validated.

3.5 Validation: ANCOVA

The validation of the ANCOVA method is done based on the resulting indices. The algorithm from Section 3.3 is applied on an example given in Caniou (2012) and then compared with his results. The used mathematical model is depicted in Eq. (47), where all variables are distributed $X_i \sim \mathcal{N}(0.5, 1)$, $i = 1, \ldots, 5$. A Gaussian Copula with the linear correlation matrix in Eq. (48) is used.

$$Y = \mathcal{M}(\mathbf{X}) = X_1 + X_2 + X_3 + X_4 + X_5 \tag{47}$$

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.6 & 0.2 & 0 & 0 \\ 0.6 & 1 & 0 & 0 & 0 \\ 0.2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0.2 & 1 \end{pmatrix} \tag{48}$$

The model is simple. It is additive, only introducing structural contributions from all variables but no interaction. The linear correlation matrix is set up so as to create two independent groups of variables. X_1 is strongly correlated to X_2 and weakly to X_3. Those two variables themselves are only correlated to X_1. X_4 and X_5 are correlated weakly. The

Table 6: Moments of X_1 for different conditioning values of X_2 and X_3 for $\rho_{12} = 0.7$ and $\rho_{13} = -0.3$

		Code generated		Reference	
Conditioning values		mean	std.	mean	std.
X_2	X_3				
17.5	33.1	$\mu_1 = 7.23$	$\sigma_1 = 1.30$	$\mu_1 = 7.57$	$\sigma_1 = 1.36$
21.1	28.8	$\mu_1 = 11.1$	$\sigma_1 = 1.30$	$\mu_1 = 11.0$	$\sigma_1 = 1.38$
19.6	26.5	$\mu_1 = 10.8$	$\sigma_1 = 1.30$	$\mu_1 = 10.6$	$\sigma_1 = 1.36$
22.4	27.9	$\mu_1 = 12.3$	$\sigma_1 = 1.30$	$\mu_1 = 12.0$	$\sigma_1 = 1.38$

results from the reference Caniou (2012) are shown in Table 7a and the ones produced by the developed MATLAB-code in Table 7b. Since there does not exist any interaction in the model, the interactive indices of all variables S_i^I are all equal to 0 and not listed in the tables.

Variables	S_i	S_i^U	S_i^C
X_1	0.24	0.13	0.11
X_2	0.24	0.13	0.11
X_3	0.19	0.13	0.06
X_4	0.16	0.13	0.03
X_5	0.16	0.13	0.03
Sum	1.00	0.65	0.03

(a) Indices from the reference thesis (Caniou, 2012)

Variables	S_i	S_i^U	S_i^C
X_1	0.26	0.14	0.12
X_2	0.23	0.14	0.09
X_3	0.17	0.14	0.03
X_4	0.17	0.14	0.03
X_5	0.17	0.14	0.03
Sum	1.00	0.70	0.30

(b) Indices calculated by the code

Table 7: ANCOVA indices for additive model with 5 variables.

Clearly, the tables show different indices. Nevertheless, they agree in two points: firstly, the uncorrelated contribution, measured by S_i^U, is equal for all variables and secondly, X_4 and X_5 are overall equally important to the variance of the model. But the reference indices attribute the same importance to X_1 and X_2 while the computed indices show X_1 as the most important variable. Additionally, the reference presents X_3 as more important than X_4 and X_5. However, the computed indices are the same for X_3, X_4 and X_5. Those differences seem to prove the code wrong.

Form the intuition, all variables should have the same uncorrelated contribution. Thus, the ranking of the variables must be based on S_i^C, which is based on covariances. Since all variables have the same marginal variance, this directly translates to the amount of correlation a variable has. X_1 correlates the most with X_2 and X_3 and should therefore be the most important variable. Second most important should be X_2 having a linear correlation coefficient $\rho_{12} = 0.6$. Those two are followed by all other variables with equal indices because they all correlate with one other variable with $\rho = 0.2$. The implemented algorithm produces values that meet this prior guess. Since there is no clear argumentation on the reasonableness of the values in Caniou (2012) and the results produced by the developed algorithm meet the expectations, the code is considered validated. This example is picked up again in Section 4, with more argumentation for the author's values.

Since the ANCOVA method uses a PCE as surrogate model, one has to ensure it is accurate enough. Therefore, the experimental design has to be sufficiently large. For the simple models, 200 model runs were used, since the computational cost was negligible. In the following Section 3.6 the needed experimental design for a complex model will be determined, thereby proving the used size for the simple models as adequate or even excessive.

3.6 Double loop MCS estimation using PCE

The estimation of the Kucherenko indices via a double loop MCS can be improved by replacing the actual model by an HDMR (see Section 2.6), which decreases the computation time considerably. Since the ANCOVA approach uses a PCE instead of the actual model to separate specific terms, the same model can be used for this estimation. In order to determine the needed amount of samples to get a good predictor, a PCE is set up for increasing sizes of experimental designs. For every PCE the relative generalisation error in Eq. (33) is calculated for a validation set of correlated samples. If the error is smaller than $0.5\% = 5 \times 10^{-3}$, the sample size is considered sufficiently large.

The computation time is not a major issue for the simple models in Sections 4.2 - 4.5 nor for the Ishigami function in Section 4.7. However, the truss model in Section 4.8 is computationally much more expensive. The calculation of the Kucherenko indices took approximately 4.5 hours for loop sizes of $1{,}000$, resulting in $1{,}000{,}000$ model runs per index. However, tests on simpler models showed, that this loop size is not sufficient and the resulting indices can strongly vary. Therefore, this model is to be replaced by a PCE. The model code calculates the midspan deflection of the truss in Figure 17 and has ten different input variables. Six of them, the point loads, are considered dependent (see Eq. (76)). Figure 6 shows the relative generalisation error for experimental designs of size 10 to 400 and a validation set of $10{,}000$ points.

It can be seen that already for a small amount of model runs for the experimental design,

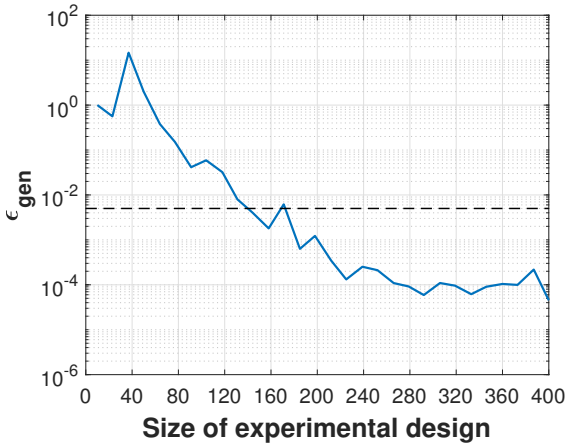


Figure 6: Relative generalisation error of the PCE of the truss model for increasing experimental design size.

the error gets very small. With about 200 samples, the error is around 2×10^{-3} and therefore small enough. Consequently, for the PCE of the truss model an experimental design with 200 samples is used.

4 Results and Discussion

4.1 Procedure

In this section the implemented methods are applied on different mathematical models of increasing complexity. Low complexity models in Sections 4.2 - 4.5 should allow to understand the effect of changes only in structural, correlation or interaction parts. This understanding should eventually be used to analyse more complex models in Sections 4.7 and 4.8. The simple models start with a projector, that gives out one of the input variables. The additive models introduce more structural components and are used to reproduce the results from the references the approaches are taken from. The interactive model and the Ishigami function eventually bring in interactions between the variables, thereby including all effects separable by the available sensitivity indices. Finally, the truss model simulates an application of the theory on an engineering problem and in order to reduce computation time, a surrogate model is used. The results of the models are analysed in the respective sections and discussed twice in Sections 4.6 and 4.9.

To clearly distinguish the indices resulting from different methods, only the classical Sobol' indices are denoted as S_i and S_i^T. The indices from the direct decomposition (Kucherenko et al., 2012) are denoted as K_i and K_i^T and the ones resulting from ANCOVA (Caniou, 2012) as $A_i = A_i^U + A_i^I + A_i^C$. In the cases of no correlation, the Sobol' indices are identified through post-processing of the PCE of the model, set up on a experimental design of 200 samples. For the calculation of the Kucherenko indices, both the inner and the outer loop have a length of 10^4. For the ANCOVA indices also 200 samples are used (see Section 3.6) for the experimental design of the PCE and 10^6 samples for the estimation.

4.2 Projector

The first function is the projector in Eq. (49). It takes $\mathbf{X} = (X_1, X_2, X_3)$ and simply returns X_1. The marginal distributions of the input variables are $X_i \sim \mathcal{N}(0, 1)$, $i = 1, 2, 3$. The sensitivity indices are calculated for different correlations between the variables. For no correlation also the Sobol' indices are calculated as reference. To model dependence, a Gaussian copula is used. Since the marginal distributions are all Gaussian as well, the linear correlation coefficients are identical to the parameters of the copula.

$$Y = \mathcal{M}(\mathbf{X}) = X_1 \tag{49}$$

Case 1 No correlation: In a first step, no correlation is assumed. In this case the Sobol' indices are calculated to serve as reference values for the Kucherenko and ANCOVA

Table 8: Projector Case 1: Uncorrelated sensitivity indices

	S_i	S_i^T	K_i	K_i^T	$A_i = A_i^U$
X_1	1.00	1.00	1.00	1.00	1.00
X_2	0.00	0.00	0.00	0.00	0.00
X_3	0.00	0.00	0.00	0.00	0.00

indices. The results are shown in Table 8.

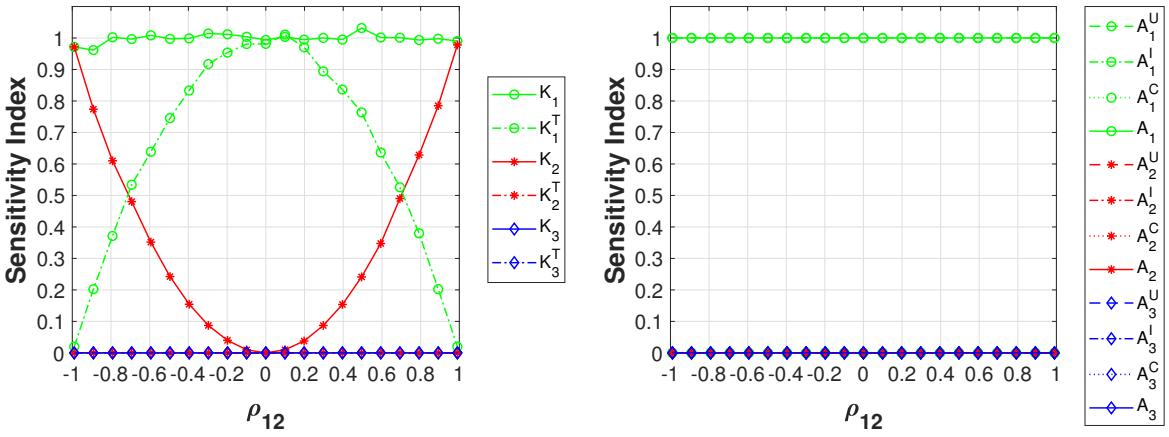
$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{50}$$

The Kucherenko and ANCOVA indices do indeed yield the same importance allocation as the Sobol' indices. Variable X_1 is the only contributing variable. Since there is neither interaction nor correlation, its total index is equal to the first order index.

Case 2 Varying correlation between X_1 and X_2: Now, to isolate the influence of correlation between X_1 and X_2, the linear correlation coefficient ρ_{12} is varied from -1 to 1, while X_3 stays independent (see Eq. (51)). The resulting indices are shown in Figure 7.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & 0 \\ \rho_{12} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad -1 < \rho_{12} < 1 \tag{51}$$

Figure 7a shows the Kucherenko indices of the three variables for a change in ρ_{12}. They are always positive and the total index of a variable is everywhere smaller or equal to its first order index. At zero covariance, the first order indices meet the corresponding total indices, since there is no interaction between the variables and at this point no correlation either. Unsurprisingly, the first order and total effect of X_3 are always at 0, reflecting the fact that X_3 does not have any influence on the variance of the model, which is correct. Additionally, the first order index of X_1 always stays at 1, since a change in X_1 results in the same amount of change in Y. The other curves allow for more interesting observations. Through correlation, X_2 gains influence on Y. It does not matter if the correlation is positive or negative because only the variance of the actually influencing variable X_1 is important and the total variance stays constant at $\text{Var}[Y] = \text{Var}[X_1] = 1$. At the extremes $\rho_{12} = \pm 1$, X_2 is as important as X_1, since they behave exactly alike. K_1^T



(a) Kucherenko indices

(b) ANCOVA indices

Figure 7: Projector Case 2: Sensitivity indices for varying linear correlation between X_1 and X_2

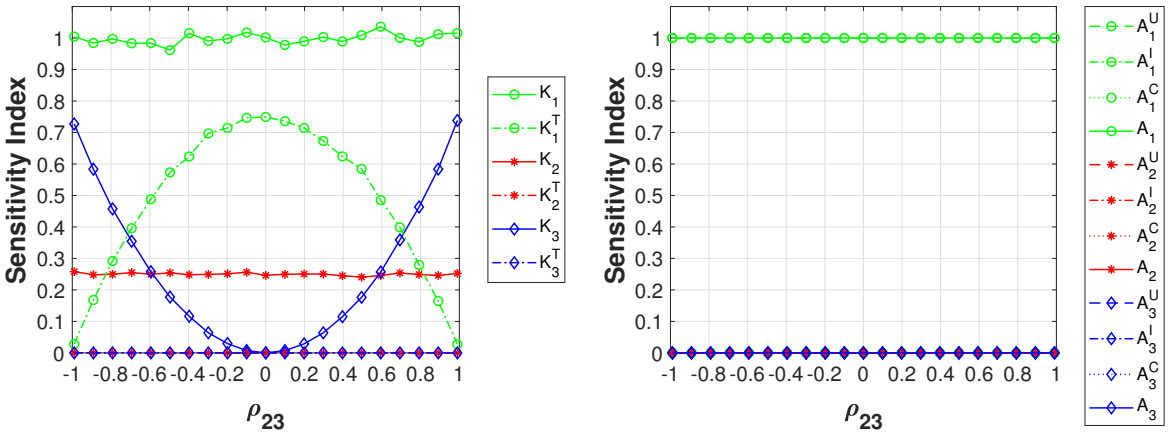
decreases with increasing absolute value of correlation at the same rate as K_2 increases and reaches 0 at the extremes. For perfect correlation this is to be expected based on the analytical formulas in Section 2.3. K_2^T always stays at zero because if the other variables X_1 and X_3 are fixed to a value, a change in X_2 will not influence Y.

The ANCOVA indices in Figure 7b show different aspects. X_1 is the only variable influencing the variance of the outcome according to these indices. This is due to the ANCOVA indices being based on a structural representation of the model for zero correlation. In the case of the projector, only X_1 has a structural, independent influence. Its uncorrelative index A_1^U is therefore always equal to 1. Since there are no component functions of the other variables (i.e. they are zero), there is no covariance between them and other component functions and the indices are all 0.

Case 3 Positive correlation between X_1 and X_2 and varying between X_1 and X_3: In this case, the third variable is added to the system by correlating it to X_1 as shown in Eq. (52). The correlation between X_1 and X_2 is fixed at 0.5. In order to keep X_2 and X_3 uncorrelated and the correlation matrix positive definite, ρ_{13} is only varied between -0.85 and 0.85. The resulting indices are shown in Figure 8.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.5 & \rho_{13} \\ 0.5 & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix} \quad -0.85 < \rho_{13} < 0.85 \tag{52}$$

As can be seen in Figure 8, all indices are positive and the total indices are smaller or equal to the corresponding first order indices. The first order index of X_1 is again equal



(a) Kucherenko indices (b) ANCOVA indices

Figure 8: Projector Case 3: Sensitivity indices for varying linear correlation between X_1 and X_3 and constant correlation $\rho_{12} = 0.5$

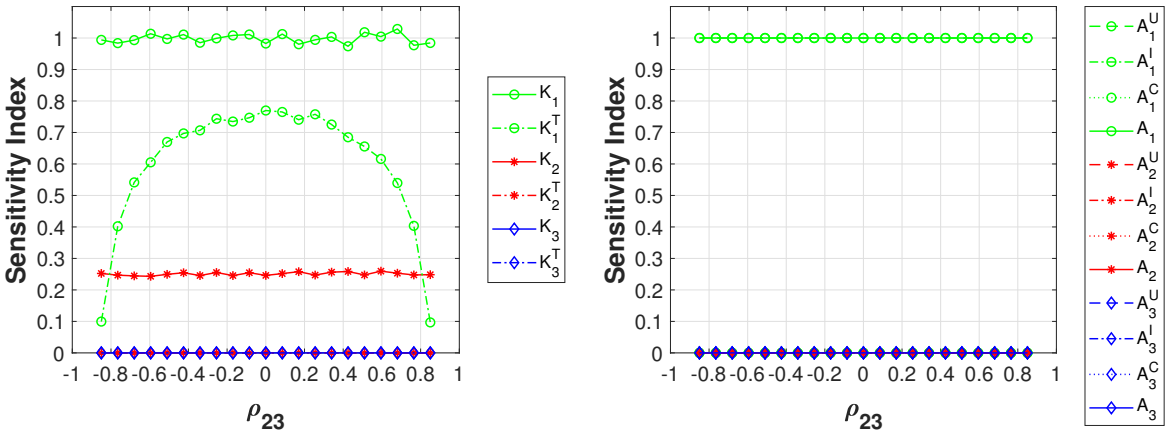
to 1 for all correlation values due to the same reason as before (see **Case 2**). However, at zero correlation ρ_{13}, the first order indices of X_1 and X_2 are not equal to their total indices because those variables are still correlated. This constant correlation between X_1 and X_2 leads to K_2 being constant at 0.25. The increase in correlation $|\rho_{13}|$ leads, similarly to before, to an increase of K_3 and a decrease of K_1^T of the same amount. For $|\rho_{13}| > 0.5$ X_3 is more important for the total variance than X_2 which seems reasonable because its correlation with X_1 is stronger than the one of X_2.

The ANCOVA indices in Figure 8b show the same as for **Case 1**. For variables without independent, structural contribution, the component functions will be zero and the variables will have no influence. A_1^U (and with it the sum of the indices A_1) is constant at 1 and the other indices stay at zero.

Case 4 Positive correlation between X_1 and X_2 and varying between X_2 and X_3: In this last case of the projector model, ρ_{12} is again fixed at 0.5, but X_3 is correlated to X_2 (see Eq. (53)). The results are shown in Figure 9.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.5 & 0 \\ 0.5 & 1 & \rho_{23} \\ 0 & \rho_{23} & 1 \end{pmatrix} \quad -0.85 < \rho_{23} < 0.85 \tag{53}$$

Figure 9a shows, as expected, that X_3 does not have any influence on the total variance since it is uncorrelated to X_1. The curve of K_1^T behaves like in **Case 1** excepts it starts lower due to the correlation $\rho_{12} = 0.5$. K_2 stays constant at 0.25, as it did in **Case 3**.



(a) Kucherenko indices

(b) ANCOVA indices

Figure 9: Projector Case 4: Sensitivity indices for varying linear correlation between X_1 and X_3 and constant correlation $\rho_{12} = 0.5$

The ANCOVA indices in Figure 9b do not produce new insight compared to the other cases.

4.3 Additive Model 1

In order to introduce structural influence from more than one variable, the additive models from Kucherenko et al. (2012) (*Test case 1. Linear Model*) and Caniou (2012) (*Equal structural contribution of correlated parameters*) are used. In this section, the model from Caniou (2012) is analysed, while the one from Kucherenko et al. (2012) is analysed in the next Section 4.4.

The model in Eq. (54) has 5 input variables X_i, $i = 1, \ldots, 5$, which all have marginal distributions $\mathcal{N}(0.5, 1)$. In the following, cases of linear correlation given in the reference thesis will be applied. Since a Gaussian copula is applied and the marginal distributions are all Gaussian as well, the linear correlation coefficients are identical to the parameters of the copula.

$$Y = \mathcal{M}(\mathbf{X}) = X_1 + X_2 + X_3 + X_4 + X_5 \tag{54}$$

Case 1 No correlation: First, no correlation between any of the variables is assumed. Table 9 shows the sensitivity indices resulting from each method.

The methods all yield the same indices. Notably, all variables contribute $1/5$ to the models variance, since they all have the same marginal variance and have equal structural roles. There does not exist interaction nor correlation.

Table 9: Additive Model 1 Case 1: uncorrelated sensitivity indices

	S_i	S_i^T	K_i	K_i^T	A_i	A_i^U	A_i^I	A_i^C
X_1	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0
X_2	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0
X_3	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0
X_4	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0
X_5	0.2	0.2	0.2	0.2	0.2	0.2	0.0	0.0

Table 10: Additive Model 1 Case 2: Sensitivity indices for correlation

	K_i	K_i^T	A_i	A_i^U	A_i^I	A_i^C
X_1	0.47	0.09	0.26	0.14	0.0	0.11
X_2	0.37	0.09	0.23	0.14	0.0	0.09
X_3	0.21	0.14	0.17	0.14	0.0	0.03
X_4	0.21	0.14	0.17	0.14	0.0	0.03
X_5	0.21	0.14	0.17	0.14	0.0	0.03

Case 2 Positive correlation: Second, the correlation is set equal to the example in the reference (see Eq. (55)). This case was already used in the validation of the code for the ANCOVA indices in Section 3.5. Table 10 shows the Kucherenko and ANCOVA indices.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.6 & 0.2 & 0 & 0 \\ 0.6 & 1 & 0 & 0 & 0 \\ 0.2 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0.2 & 1 \end{pmatrix} \tag{55}$$

The first order Kucherenko indices K_i lead to the following ranking of the variables: X_1 is the most important variable with a sensitivity index of 0.47. It is followed by variable X_2 with $K_2 = 0.37$ and finally by variables X_3, X_4 and X_5, which all get a first order index of 0.21. Notably, those numbers do add up to more than 1. Additionally to the structural contribution, Kucherenko's first order index also increases with correlation effects. However, the total index decreases with growing correlation. This effect can be observed in Table 10: the total indices of X_1 and X_2 are smaller than the ones of X_3, X_4 and X_5, which are identical. It is known from the case of independence, that the total Sobol' index of a variable is larger, if it has strong interactions. However, in this model, no interactions are apparent. In this case, the ranking of the variables according to the first order index is reasonable.

The ANCOVA indices A_i lead to the same ranking of variables. The meaning of the values was already discussed in Section 3.5: the structural contribution is the same for all variables, but the correlative contributions differ depending on how strongly correlated a variable is. The PCE of the model is set up assuming independence. As a result it will look similar to:

$$\mathcal{M}^{HDMR}(\mathbf{X}) = X_1 + X_2 + X_3 + X_4 + X_5 = \sum_{i=1}^{5} \mathcal{M}_i^{HDMR}(X_i). \tag{56}$$

The total variance is calculated using this model with a sample of correlated variables. The variance can therefore be decomposed as:

$$\text{Var}[\mathcal{M}^{HDMR}(\mathbf{X})] = \text{Var}[X_1 + X_2 + X_3 + X_4 + X_5]. \tag{57}$$

Eq. (63) in the next section shows how to decompose this variance for the case of 3 variables. The variance of the sum of 5 variables will be decomposed analogously:

$$\text{Var}[\mathcal{M}^{HDMR}(\mathbf{X})] = \sum_{i=1}^{5} \text{Var}[X_i] + 2 \sum_{1=i<j}^{5} \text{Cov}[X_i, X_j]. \tag{58}$$

For the case at hand this yields $\text{Var}[Y] = 5 \cdot 1 + 2 \cdot (0.6 + 0.2 + 0.2) = 7$. The uncorrelated indices are calculated by dividing $\text{Var}[\mathcal{M}_i^{HDMR}(x_i)]$ by the total variance. For the model in Eq. (56) this results in the marginal variances divided by the total variance. One obtains $1/7 \approx 0.14$ for each variable. The value of A_1^C is derived in the following:

$$A_1^C = \frac{\text{Cov}[X_1, X_2 + X_3 + X_4 + X_5]}{\text{Var}[Y]} = \frac{0.8}{6} \approx 0.13\,, \tag{59}$$

where:

$$\begin{aligned} &\text{Cov}[X_1, X_2 + X_3 + X_4 + X_5] = \text{E}[(X_1 - \text{E}[X_1])(\sum_{i=2}^{5} X_i - \text{E}[\sum_{i=2}^{5} X_i])] = \\ &= \text{E}[\sum_{i=2}^{5} (X_1 - \text{E}[X_1])(X_i - \text{E}[X_i])] = \sum_{i=2}^{5} \text{E}[(X_1 - \text{E}[X_1])(X_i - \text{E}[X_i])] = \\ &= \sum_{i=2}^{5} \text{Cov}[X_1, X_i] = \text{Cov}[X_1, X_2] + \text{Cov}[X_1, X_3] = \rho_{12}\sigma_1\sigma_2 + \rho_{13}\sigma_1\sigma_3. \end{aligned} \tag{60}$$

The MC estimation matches the analytical value. Overall, the ANCOVA indices allow for a more detailed interpretation of the first order indices. Additionally, since there does not exist any interaction, the sum of the first order indices equals 1, making them interpretable as shares of the total variance, in contrast to the Kucherenko indices.

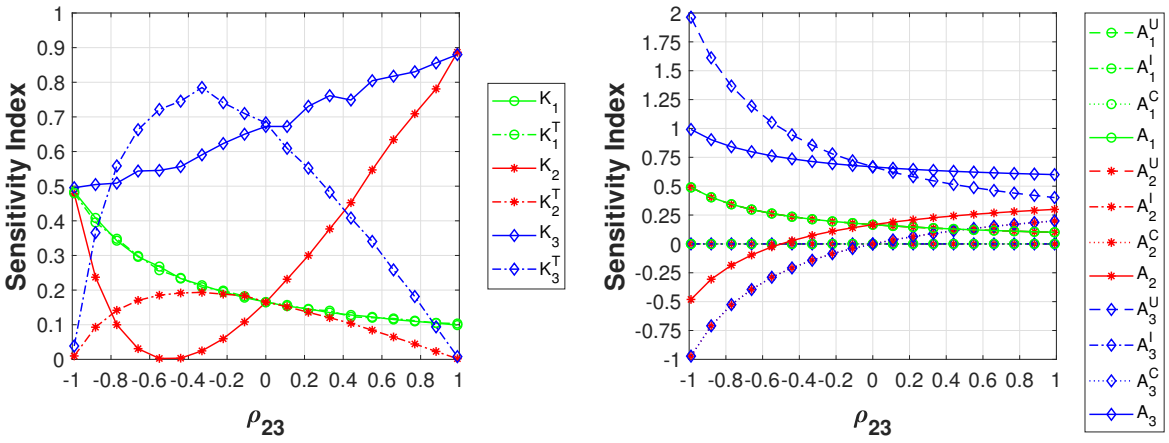
4.4 Additive Model 2

In order to introduce structural influence from more than one variable, the additive models from Kucherenko et al. (2012) (*Test case 1. Linear model*) and Caniou (2012) (*Equal structural contribution of independent / correlated parameters*) are used. In this section, the model from Kucherenko et al. (2012) in Eq. (61) is analysed.

This model has 3 input variables, of which X_1 and X_2 are marginally distributed $\mathcal{N}(0,1)$ and X_3 is marginally distributed $\mathcal{N}(0,2)$. As in the reference paper, X_1 is independent and the linear correlation between X_2 and X_3 will be varied from -1 to 1 (see Eq. (62)). Again, since a Gaussian copula is applied on Gaussian marginals, the linear correlation coefficients are equal to the copula parameters.

$$Y = \mathcal{M}(\mathbf{X}) = X_1 + X_2 + X_3 \tag{61}$$

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho_{23} \\ 0 & \rho_{23} & 1 \end{pmatrix} \quad -1 < \rho_{23} < 1 \tag{62}$$



(a) Kucherenko indices

(b) ANCOVA indices

Figure 10: Additive Model 2: Sensitivity indices for different values of linear correlation between X_2 and X_3

Figure 10 shows the resulting sensitivity indices. The Kucherenko indices in Figure 10a are always positive and match the ones in the reference paper Kucherenko et al. (2012). At zero correlation the total indices meet the corresponding first order indices. However, the indices $K_1 = K_1^T$ and $K_2 = K_2^T$ take values around 0.167 whereas $K_3 = K_3^T$ lies around 0.667. This is due to the larger marginal variance of X_3. The total variance of the

model can be decomposed as follows:

$$\begin{aligned}\text{Var}[X_1 + X_2 + X_3] = \text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] + 2\,\text{Cov}[X_1, X_2] + \\ + 2\,\text{Cov}[X_1, X_3] + 2\,\text{Cov}[X_2, X_3]\end{aligned} \tag{63}$$

At zero correlation, the covariance between the variables vanishes. The total variance becomes $\text{Var}[Y] = \text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] = 1^2 + 1^2 + 2^2 = 1 + 1 + 4 = 6$. Hence, for zero correlation the variance share of X_3 is $\text{Var}[X_3]/\text{Var}[Y] = 4/6 \approx 0.667$ and 4 times larger than the ones of X_1 and X_2 with each $1/6 \approx 0.167$. At the extremes $\rho_{23} = \pm 1$ the total indices of the correlated variables K_2^T & K_3^T equal zero, as expected. The first order indices, though, take on different values at the extremes. The values can be explained using Eq.s (63) and (1). For $\rho_{23} = -1$ the total variance reads $\text{Var}[Y] = \text{Var}[X_1] + \text{Var}[X_2] + \text{Var}[X_3] + \text{Cov}[X_2, X_3] = 1^2 + 1^2 + 2^2 + 2 \cdot (-1) \cdot 1 \cdot 2 = 1 + 1 + 4 - 4 = 2$. One thing to note here is that the total variance decreases due to negative correlation. Another thing is, that $1/2$ of the variance is due to X_1 and the other half due to X_2 and X_3, explaining them all being at 0.5 for $\rho_{23} = -1$. At the other extreme $\rho_{23} = 1$, the same decomposition gives $\text{Var}[Y] = 1^2 + 1^2 + 2^2 + 2 \cdot 1 \cdot 1 \cdot 2 = 1 + 1 + 4 + 4 = 10$. Notably, the positive correlation leads to an increase in total variance. Now $1/10$ of it is due to X_1 and $9/10$ are due to X_2 and X_3, as shown in Figure 10a. At both extremes one could ask, why X_2 and X_3 do not each have half of their combined effect, meaning 0.25 each at $\rho_{23} = -1$ and 0.45 at $\rho_{23} = 1$. This is due to the Kucherenko indices allocating the full correlative effects to each of the correlated variables. Additionally, at such extreme correlation values, the variables basically act like only one random variable X_{23} which contributes half of the total variance. The interval $-0.8 \leq \rho_{23} \leq 0$ stands out because here $K_i^T \geq K_i$, $i = 2, 3$ unlike anywhere else. Furthermore, K_2^T and K_3^T both peak between -0.5 and -0.4 while K_2 reaches its minimum. Overall, K_2^T behaves similarly to K_3^T, except its values are 4 times smaller, the same difference in scale the variables X_2 and X_3 have in variance ($\sigma_2^2 = 1$, $\sigma_3^2 = 4$). In this area $-0.5 < \rho_{23} < -0.4$, presumably, the variance of X_2 gets counteracted by the variance of X_3, since $\sigma_3^2 = 2\sigma_2$. As a consequence, the first order index of X_2 drops to zero. At the same time, K_3^T exceeds K_3 due to this "interaction" with X_2. For this model with these specific marginals and correlations, the extreme point lies here. From here, K_i^T, $i = 1, 2$ drops again. Finally, $K_1^T = K_1$ for all values of ρ_{23}, since X_1 does not have interactions nor correlation with the other two variables. Its decrease from $\rho_{23} = -1$ to $\rho_{23} = 1$ is due to growth of total variance, resulting in a smaller share for X_1's constant variance.

The ANCOVA indices yield a seemingly dissimilar plot (see Figure 10b). Please note the different scaling of the y-Axis in the two plots for increased readability. Nevertheless, there are two similarities. The uncorrelative index A_1^U (and with it the first order index A_1) show a similar curve as $K_1 = K_1^T$. Moreover, at zero correlation the same first order indices are achieved with both approaches. At $\rho_{23} = 0$ all correlative indices (A_1^C, A_2^C and A_3^C) are zero, as they should be. In fact, A_1^C is zero everywhere since it is not correlated to

another variable. X_2 and X_3 only being correlated with each other share the same curve of their correlative index. Also, all interactive indices are zero, depicting the expected, since there is no interaction in the current model. The uncorrelative indices seem to behave curiously at first sight. To understand them, a polynomial HDMR of the current model is used. It is assumed that $\mathcal{M}^{HDMR} = X_1 + X_2 + X_3$. Using the definition of the uncorrelative indices (Eq. (27)), one can formulate the uncorrelative index of variable X_3:

$$A_3^U = \frac{\text{Var}[\mathcal{M}_3^{HDMR}(X_3)]}{\text{Var}[Y]} = \frac{\text{Var}[X_3]}{\text{Var}[Y]} \tag{64}$$

The variance of X_3 is known from its marginal distribution and the total variance selected values of ρ_{23} was derived above. The result is $A_3^U = 4/2 = 2$ for $\rho_{23} = -1$, $A_3^U = 4/6 \approx 0.667$ for $\rho_{23} = 0$ and $A_3^U = 4/10 = 0.4$ for $\rho_{23} = 1$. These are the same values as in Figure 10b. The remaining question is for an explanation of the behaviour of the correlative indices A_2^C and A_3^C. This can be explained using their definition in Eq. (29). For variable X_3 the numerator is the covariance between $\mathcal{M}_3^{HDMR}(X_3) = X_3$ and $\Sigma_{i \notin \mathbf{v}} \mathcal{M}_{\mathbf{v}}^{HDMR}(\mathbf{X_v}) = X_1 + X_2$. Such a covariance can be separated as follows:

$$\text{Cov}[X_3, X_1 + X_2] = \text{Cov}[X_3, X_1] + \text{Cov}[X_3, X_2] \tag{65}$$

In the case at hand, there only exists correlation (and therefore covariance) between X_2 and X_3. To get the correlative index, the covariance has to be normalized by the output variance. Doing this, one obtains $A_3^C = (-1) \cdot 2 \cdot 1/2 = -1$ for $\rho_{23} = -1$, $A_3^C = 0 \cdot 2 \cdot 1/6 = 0$ for $\rho_{23} = 0$ and $A_3^C = 1 \cdot 2 \cdot 1/10 = 0.2$ for $\rho_{23} = 1$. These are also the values in Figure 10b. Finally, the first order indices A_1, A_2 and A_3 represent the sum of the different influences of the respective variable. The overall curve trend of spreading towards $\rho_{23} = -1$ and compacting towards $\rho_{23} = 1$ can be explained by the larger values of the total variance at positive correlation, to which the covariance terms are normalized. It is also interesting to see that the correlative indices of X_2 and X_3 are identical. The fact that X_2 has no first order effect at $\rho_{23} = -0.5$ is also displayed.

4.5 Interactive Model

The model in Eq. (66) introduces interaction on a level of low complexity. X_1 and X_2 are multiplied with each other and X_3 is added to the product. The marginal distributions of the variables are $X_i \sim \mathcal{N}(0.5, 1)$, $i = 1, 2, 3$ in order to not have zero mean. This is important because the ANCOVA indices are based on structural contribution. A distribution of X_1 with zero mean would cause the structural contribution of X_2 to be zero, since it is weighted by the mean of its multiplier X_1. This model shows the behaviour of combined correlation and interaction. It is again tested for different cases of linear correlation with a Gaussian copula.

$$Y = \mathcal{M}(\mathbf{X}) = X_1 X_2 + X_3 \tag{66}$$

Table 11: Interactive Model Case 1: uncorrelated sensitivity indices

	S_i	S_i^T	K_i	K_i^T	A_i	A_i^U	A_i^I	A_i^C
X_1	0.1	0.5	0.1	0.5	0.1	0.1	0.0	0.0
X_2	0.1	0.5	0.1	0.5	0.1	0.1	0.0	0.0
X_3	0.4	0.4	0.4	0.4	0.4	0.4	0.0	0.0

To make estimations and explore the development of the indices, the total variance will be derived analytically. For this model a variance decomposition can be done as follows:

$$\text{Var}[Y] = \text{Var}[X_1X_2 + X_3] = \text{Var}[X_1X_2] + \text{Var}[X_3] + 2\text{Cov}[X_1X_2, X_3]. \tag{67}$$

Case 1 No correlation: In order to get an understanding of this model, first the uncorrelated case will be investigated. The sensitivity indices are calculated the same way as for the last model and displayed in Table 11.

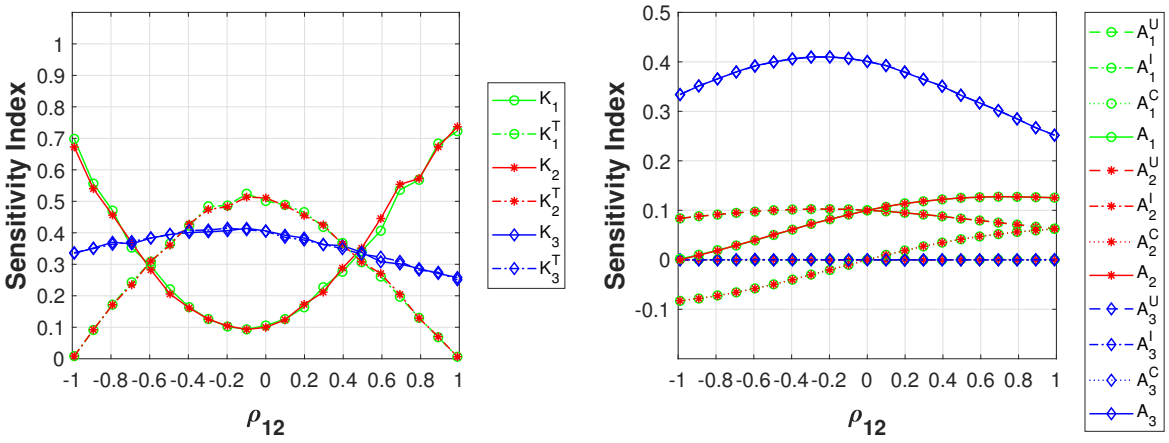
The first order indices derived from the methods for correlated variables K_i and A_i agree with the Sobol' indices S_i. Variable X_3 structurally contributes 40% to the total variance and thereby the most. X_1 and X_2 each only contribute 10% structurally, but their interaction, measured by $S_i^T - S_i = K_i^T - K_i = 0.4$, $i = 1, 2$, contributes 40%, as much as X_3. Including their interaction, X_1 and X_2 are more important than X_3. The ANCOVA indices $A_i = A_i^U$ do not yield more insight on the importance on the variables. However, $A_i^I = 0$, $i = 1, 2$ reminds of the fact, that the interactive index only comes into existence through correlation. The values in Table 11 can also be explained using Eq. (67). In case of independence, the first order index of X_3 (equalling the total index) reads:

$$A_3 = \frac{\text{Var}[X_3]}{\text{Var}[Y]} = \frac{1}{(1 \cdot 1 + 1 \cdot 0.5^2 + 1 \cdot 0.5^2) + 1} = \frac{1}{2.5} = 0.4 \tag{68}$$

Case 2 Varying correlation between X_1 and X_2: In this case the effect of correlating interactive variables is investigated. Therefore, X_1 and X_2 are correlated, whereas X_3 stays independent (see Eq. (69)). The resulting Kucherenko and ANCOVA indices are shown in Figure 11.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & 0 \\ \rho_{12} & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \quad -1 < \rho_{12} < 1 \tag{69}$$

The Kucherenko indices in Figure 11a show that for all values of correlation between X_1 and X_2, the curves of K_1 and K_2 and the ones of K_1^T and K_2^T match perfectly. Since



(a) Kucherenko indices (b) ANCOVA indices

Figure 11: Interactive Model Case 2: Sensitivity indices for different values of linear correlation between X_1 and X_2

in this case, X_3 has neither interactions nor correlation, its total index always equals its first order index. The first order index depicts the structural contribution of a variable. Since the means of X_1 and X_2 as well as their structural contribution are the same, it is unsurprising that their first order indices are equal. With this in mind and adding the fact, that X_1 and X_2 only interact and correlate with each other, the equality of total indices becomes obvious. The total indices of X_1 and X_2 are zero at $|\rho_{12}| = 1$, since fixing X_2 and X_3 to certain values determines the value of X_1. As a consequence, the conditional samples contain constant values and, therefore, the model evaluations are constant as well, the variance is zero. The same concept applies for X_2. The total indices K_1^T and K_2^T raise to their Sobol' counterparts at $\rho_{12} = 0$, where they are larger than their first order indices, as explained in **Case 1**. For increasing $|\rho_{12}|$, the total indices decrease while the first order indices increase. Notably, the point were the values become identical is not the same for $\rho_{12} \to 1$ and $\rho_{12} \to -1$. This is due to the total variance not changing equally for negative and positive correlation because covariance plays a role as shown in Eq. (**??**). For the case at hand with an independent X_3 and varying ρ_{12}, one obtains a total variance of $\text{Var}[Y] = \text{Var}[X_1 X_2] + \text{Var}[X_3]$, where $\text{Var}[X_1 X_2] = 2 \cdot 0.5^2 \cdot \rho_{12} + 0.5^2 + 0.5^2 + 1^2 + \rho_{12}^2$. As a result, one gets the total variance as $\text{Var}[Y] = (\rho_{12}^2 + 0.5\rho_{12} + 1.5) + 1$, which will take different values for positive or negative values of ρ_{12} because of the linear term $0.5\rho_{12}$. At the negative extreme $\rho_{12} = -1$ the total variance is $\text{Var}[Y] = 2+1 = 3$: the contribution of X_1 and X_2 $(2/3)$ is double the one of X_3 $(1/3)$. At the other extreme $\rho_{12} = 1$, the variance is $\text{Var}[Y] = 3 + 1 = 4$: X_1 and X_2 contribute $3/4$ of the variance and the rest comes from X_3. Interestingly, for certain negative values of ρ_{12}, X_3 is allegedly more important than X_1 and X_2. Interactive and correlative effects of X_1 and X_2 seemingly counteract each other. Additionally, the fact that $K_1 = K_2 = 2$ and at $\rho_{12} = -1$, $K_1^T = K_2^T = 0$ and $K_1 + K_3 = 1$ leads to the assumption, that at the extremes $K_1 = K_2$ include all effects, structural, correlation and interaction. It seems that from $\rho_{12} = 0$ to $|\rho_{12}| = 1$,

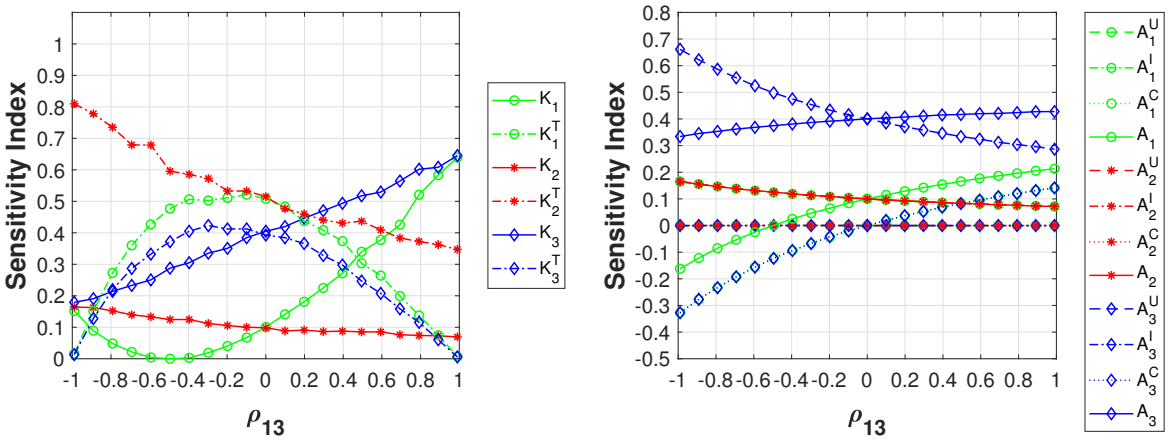
Kucherenko's first order index also starts including the interaction effects.

A first glance at the ANCOVA indices in Figure 11b reveals that X_3 is more important than the other variables. This is presumably due to the fact, that in the HDMR of the model X_1 and X_2 are weighted by each others means ($\mu_1 = \mu_2 = 0.5$) whereas X_3 gets double the structural influence with a coefficient of 1. Overall, A_3 behaves similar to K_3 and the first order indices of all three variables match the Sobol' indices at zero correlation. Looking at the indices in detail, one realises that some are constantly zero, namely A_3^I, A_3^C and, more interestingly, A_1^I and A_2^I. A look at the actual values in the results yields, that A_1^I and A_2^I are not 0 but extremely small, even negligible. The covariance between the terms including X_1 and the ones including X_1 and X_2 is just very small. In contrast, the covariance between $\mathcal{M}_1^{HDMR}(X_1)$ and $\mathcal{M}_2^{HDMR}(X_2)$ is of considerable size and therefore displayed as A_1^C resp. A_2^C of equal value in Figure 7b. The values of A_1 and A_2 at $\rho_{12} = -1$ are noteworthy. The two indices A_i^U and A_i^C cancel out for $i = 1, 2$, leading to an assumption, that only X_3 influences the model variance on "first order" level, neglecting interaction. This seems to contradict the Kucherenko indices completely, which tell that the first order effects of X_1 and X_2 are dominant. There are two reasons. One was mentioned before: Kucherenko's first order indices include interaction effects if correlation is present. The other lies in the HDMR: as mentioned, for the present model, X_1 and X_2 are weighted by each others means in their first order component functions. For $\rho_{12} = -1$, those two terms cancel out and only the interaction term stays, which not respected in the ANCOVA analysis.

Case 3 Varying correlation between X_1 and X_3: In this case X_1 correlates with X_3 and interacts with X_2 at the same time. The indices for varying ρ_{13} are shown in Figure 12.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix} \quad -1 < \rho_{13} < 1 \tag{70}$$

The Kucherenko indices of X_1 and X_3 in Figure 12a are very similar to the ones of X_2 and X_3 from the Additive Model 2. The reason for this similarity is obvious: in both cases there is a sum of two correlated variables. The smaller values of K_3 and K_3^T arise from the smaller marginal variance for this case. The total indices of X_1 and X_2 are larger than the first order indices due to the interaction between them. $K_3 = K_3^T$ for $\rho_{13} = 0$. X_3 contributes 40% of the total variance at this point, X_1 and X_2 each 10% and their interaction 40%. The total variance is obtained as $\text{Var}[Y] = \text{Var}[X_1 X_2] + \text{Var}[X_3] = (0.5^2 \cdot 1^2 + 0.5^2 \cdot 1^2 + 1^2 \cdot 1^2) + 1^2 = 1.5 + 1 = 2.5$, also proving that X_3 contributes 40%.



(a) Kucherenko indices (b) ANCOVA indices

Figure 12: Interactive Model Case 3: Sensitivity indices for different values of linear correlation between X_1 and X_3

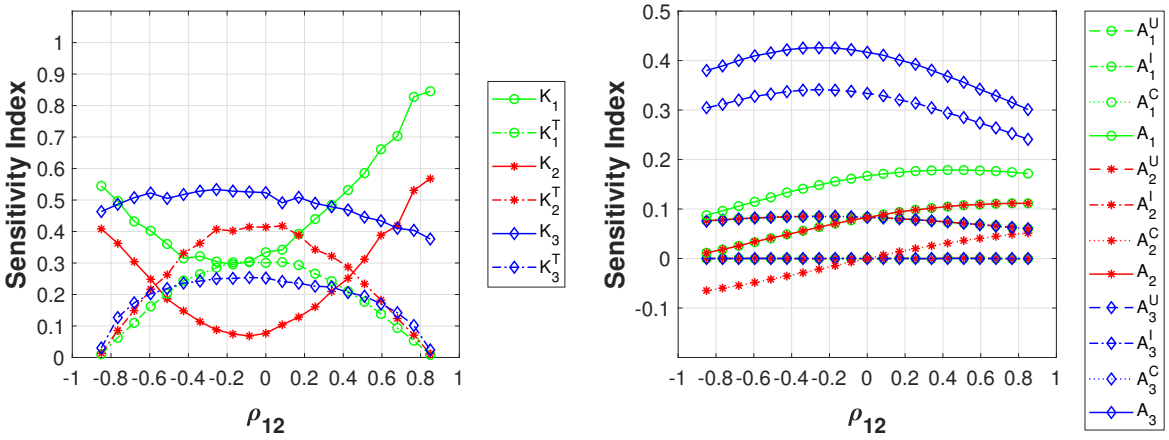
For an arbitrary ρ_{13} the total variance is obtained as $\mathrm{Var}[Y] = \mathrm{Var}[X_1X_2] + \mathrm{Var}[X_3] + 2\mathrm{E}[X_2]\mathrm{Cov}[X_1, X_3] = 1.5 + 1 + 2 \cdot 0.5\rho_{13} = 1.5 + 1 + \rho_{13}$. This shows that $\mathrm{Var}[Y]$ decreases for negative correlation. At $\rho_{13} = -1$, the total indices of the (extremely) correlated variables decrease to zero, whereas the first order indices all meet slightly below 0.2: about 20% of the total variance is due to X_1 and X_3 and the remaining 80% are due to X_2 and its interaction with X_1. The extrema at correlation values of -0.5 and -0.3 presumably arise from the same circumstances as in Section 4.4. The similarities of the situations is obvious: in both cases these peaks happen with a sum of two correlated variables.

The ANCOVA indices in Figure 12b show a similar curve for the first order index of the independent X_2, only consisting of A_2^U. Once again it is noted, that A_2^I being zero does not imply, that X_2 has no interaction effects. A_1^U behaves exactly like A_2^U. But the contribution of correlation, measured by A_1^C, alters the curve of A_1. The largely negative terms towards $\rho_{13} = -1$ once again arise from growing negative covariance and shrinking total variance.

Case 4 Varying correlation between X_1 and X_2 and positive correlation between X_1 and X_3: In the third case of the interactive model X_1 is linked to X_3 by constant correlation $\rho_{13} = 0.5$ and to X_1 by varying correlation $-0.85 < \rho_{12} < 0.85$. The variables

X_2 and X_3 remain uncorrelated.

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & \rho_{12} & 0.5 \\ \rho_{12} & 1 & 0 \\ 0.5 & 0 & 1 \end{pmatrix} \quad -0.85 < \rho_{12} < 0.85 \tag{71}$$



(a) Kucherenko indices (b) ANCOVA indices

Figure 13: Interactive Model Case 4: Sensitivity indices for different values of linear correlation between X_1 and X_2

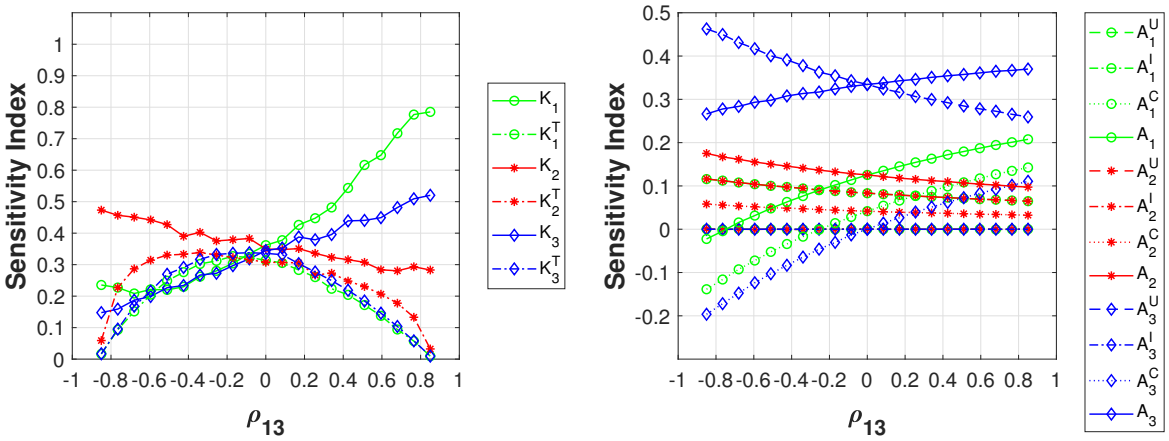
The Kucherenko indices K_2 and K_2^T in Figure 13a behave similarly to **Case 2**, where ρ_{12} also varied between -1 and 1. The indices of X_1 are similar as well, but the constant correlation $\rho_{13} = 0.5$ alters both its indices. In essence, K_1 lies higher up and K_1^T is lowered to a degree, that the effect of interaction with X_2 does not suffice to make up the decrease due to correlation. The constant correlation between X_1 and X_3 also lowers K_3^T to run below K_3. The trends in the extremes $|\rho_{12}| = \pm 1$ can easily be explained. The extreme correlation exists between the multiplicands X_1 and X_2, increasing the variance of the product $\mathrm{Var}[X_1 X_2]$ and therefore the total variance $\mathrm{Var}[Y]$. Since X_1 and X_2 contribute more to the total variance at the extremes their first order indices increase. However, for negative correlation, the product $X_1 X_2$ is likely to be counteracted by an X_3 of the opposite sign, therefore decreasing the variance of the overall outcome.

Also the ANCOVA indices in Figure 13 resemble the ones from **Case 2**. In the case at hand, X_3 gains a non-zero correlative index A_3^C due to constant correlation to X_1. The slight alteration is due to the change in the total variance for different values of ρ_{12}. The interactive index is still constantly at zero, resulting in a first order index $A_3 = A_3^U + A_3^C$, that is larger than before. Like in the related case, the interactive indices are zero, or just negligibly small. The correlative index A_1^C lies comparably higher since it now includes correlation to X_3, A_2^C shows the same values as in the related **Case 2**. The behaviour of

the uncorrelated indices is the same. The resulting first order indices now lie at different values with respect to their Kucherenko counterparts.

Case 5 Varying correlation between X_1 and X_3 and positive correlation between X_1 and X_2:

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.5 & \rho_{13} \\ 0.5 & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix} \quad -0.85 < \rho_{13} < 0.85 \tag{72}$$



(a) Kucherenko indices

(b) ANCOVA indices

Figure 14: Interactive Model Case 5: Sensitivity indices for different values of linear correlation between X_1 and X_3

As in the previous case, the Kucherenko indices in Figure 14a resemble the ones from an already investigated case, this time **Case 3**. Index K_1 lies higher due to correlation and K_1^T is lowered, the same explanation applies as in **Case 4**. Nevertheless, for correlation values between -0.6 and -0.1 the total index lies above the first order index, indicating a stronger effect of the interaction than the correlation. K_2 lies notably higher due to the constant correlation to X_1 whereas its total index falls below it indicating stronger correlation effects for X_2. The values at the extremes can again be explained by assuming the behaviour of the total variance. For $|\rho_{13}| \to 1$, a positive product X_1X_2 is more likely to be added to a positive X_3. In contrast, for negative correlation, a negative product X_1X_2 is likely to be added to a positive X_3 and vice versa. Thereby the total variance is decreased and the constantly correlated X_2 gains the most influence.

The ANCOVA indices of X_3 in Figure 14b are almost equal to the ones in **Case 3**, differences are most likely due to a shift in the total variance $\mathrm{Var}[Y]$. The uncorrelated indices of X_1 and X_2 are also similar to before. Also, all interactive indices are zero. However, due to the constant correlation to X_1, A_2^C is now non-zero and the resulting A_2 is larger,

displaying the fact, that X_2 gained influence through correlation with X_1. The same applies for X_1: due to the constant correlation $\rho_{12} = 0.5$ its correlative index A_1^C now lies above A_3^C.

4.6 Discussion of the low complexity models

The models in Sections 4.2 - 4.5 show several properties of the two sensitivity measures under scrutiny. The projector model reveals fundamental differences between the ANCOVA and Kucherenko indices. On the one hand, the ANCOVA indices favour the independent, structural contributions of the variables. A variable that does not contribute in this way is regarded as unimportant for the total variance. If the model itself does not change in terms of structure, interactions and the variables keep their dependence structure, the ANCOVA indices stay the same. On the other hand, the Kucherenko indices represent the apparent influence of variables. As shown by the projector model, depending on correlation with X_1, the variance of X_2 was coupled with the total variance, giving the appearance of influence. For the case of $\rho_{12} = 1$, K_2 is as large as K_1 but since $\mathrm{Var}[Y] = \mathrm{Var}[X_1]$, a reduction of $\mathrm{Var}[X_2]$ does not reduce the total variance. Nevertheless, independent variables receive identical first order indices from both methods, even if correlation between other variables is present in the model (see X_1 in Figure 10, X_3 in Figure 11 and X_2 in Figure 12).

In principle, interactions will increase the total index K_i^T whereas correlation increases the first order index K_i and decreases the total index at the same time. These observations enable one to explain several behaviours in the Kucherenko plots above. However, in additive models the total index can exceed the first order index in presence of correlation (see Figures 10a and 12a). Such unusually high values of Kucherenko indices happen often for ρ slightly smaller than 0. Kucherenko et al. (2012) offers analytical expressions of the indices for the model in Eq. (61). A proper explanation in terms of correlation and interaction was attempted in Section 4.4. It is assumed by the author of this thesis, that correlation causes an "additive interaction" effect between variables.

The ANCOVA indices offer a detailed split of the first order indices. Notably, in case of correlation, they do not behave like their Kucherenko counterparts. One reason is that negative linear correlation leads to negative covariance and consequently negative values of A_i^C. Those counteract (by definition) positive values of A_i^U. Another reason is that due to the setup of the PCE and the definition of the ANCOVA indices, for models with hardly any interaction effects, the ANCOVA first order indices always add up to one (see Section 4.3 and 4.4). Also important to note is that the interactive indices A_i^I are not to be confused with higher order Sobol' indices. For zero correlation, those terms do not exist (see Table 11). In presence of correlation, those terms depict the covariance between $\mathcal{M}_i(x_i)$ and $\sum_{i\in\mathbf{u}} \mathcal{M}_\mathbf{u}(\mathbf{X_u})$, resulting in a "first order interaction and correlation

coefficient" that can be very small (see Section 4.5 **Case 2**). Finally, the sum of the effects $A_i = A_i^U + A_i^I + A_i^C$ must not be looked at isolated, the benefit of these indices lies in the split of information they provide.

4.7 Ishigami Function

As a further step in complexity, the Ishigami function in Eq. (73) is tested. This function is a benchmark application example in sensitivity analysis, since it includes strongly non-linear, non-monotonic terms, multiple variables, two of which interacting, and different coefficients for the addends. The coefficients are chosen according to Kucherenko et al. (2012) to be $a = 7$ and $b = 0.1$. The variables are all uniformly distributed between $-\pi$ and π: $X_i \sim \mathcal{U}(-\pi, \pi)$, $i = 1, 2, 3$. The same dependence as in the reference paper is applied, namely the linear correlation ρ_{13} is varied from -1 to 1 through a Gaussian copula (see Eq. (74)). In this case, where the marginal distributions are not Gaussian, the linear correlation matrix does not equal the copula parameters anymore. To go from one to the other, a transform has to be applied. For different values of copula parameters, X_1 and X_3 were sampled and their linear correlation coefficient was computed. Using linear interpolation, the transformation function is built. As can be seen in Figure 15, the two correlation values are very similar.

$$Y = \mathcal{M}(\mathbf{X}) = \sin(X_1) + a \sin^2(X_2) + b X_3^4 \sin(X_1) \tag{73}$$

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0 & \rho_{13} \\ 0 & 1 & 0 \\ \rho_{13} & 0 & 1 \end{pmatrix} \quad -1 < \rho_{13} < 1 \tag{74}$$

The Kucherenko indices in Figure 16a match the values in Kucherenko et al. (2012). It can be seen that K_2 equals K_2^T for all values of ρ_{13}. This is to be expected because X_2 does not have interaction nor correlation, only structural influence. At zero correlation the indices equal their Sobol' equivalents, with K_1 at 0.3, K_3 at 0 and K_1^T and K_3^T larger than the according first order indices due to interaction. Interestingly, the total indices of X_1 and X_3 dive below the respective first order indices at the same value of $|\rho_{13}| = 0.6$. At this point the correlation is stronger than the interaction effect. The more or less constant value of K_1 indicates that the first order index of X_1 is not affected strongly by correlation. In contrast, the first order effect of X_3 is negligible for $\rho_{13} = 0$ and grows stronger for increasing $|\rho_{13}|$. At the extremes $|\rho_{13}| = 1$, but practically everywhere, X_2 influences the variance of Y more than X_1 or X_3. In fact, here X_1 and X_2 contribute $1/3$ of the model variance, while X_2 causes $2/3$. This is due to the high coefficient the term

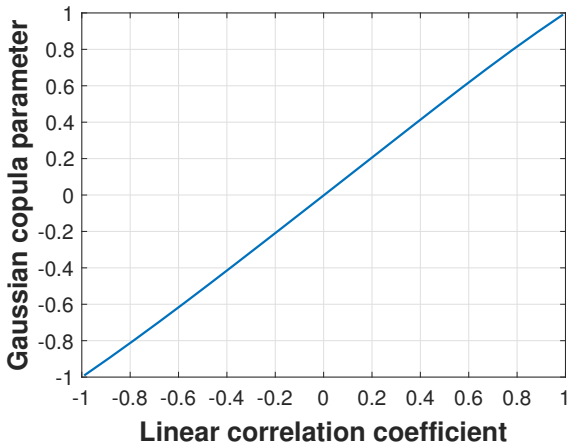
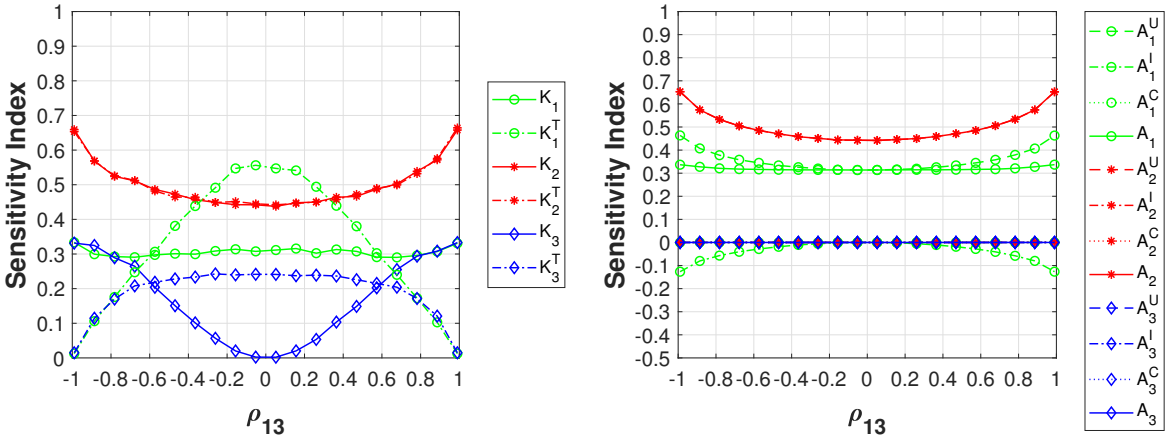


Figure 15: Relationship between the linear correlation coefficient and the Gaussian copula parameter for uniformly distributed marginals



(a) Kucherenko indices

(b) ANCOVA indices

Figure 16: Ishigami function: Sensitivity indices for different values of linear correlation between X_1 and X_3

$\sin(X_2)$ is multiplied by. Nevertheless, for the interval $-0.33 < \rho_{13} < 0.33$ X_1 has the largest influence on the output variance.

The ANCOVA indices in Figure 16b show the same behaviour of the first order index of X_2, which is represented by A_2^U (and therefore A_2). Also, the first order index of X_1 is very similar to the one from Kucherenko but split up into different components. One can see that the uncorrelated index of X_1 grows for an increase in $|\rho_{13}|$. The increase of A_1^U and A_2^U for growing correlation $|\rho_{13}| \to 1$ leads to the assumption, that the total variance decreases, consequently enlarging the ratio $\frac{\text{Var}[\mathcal{M}_i^{HDMR}(X_i)]}{\text{Var}[Y]}$, $i = 1, 2$. This is indeed the case: for $|\rho_{13}| = 1$ one obtains $\text{Var}[Y] = 9.5$, while for $\rho_{13} = 0$ one obtains $\text{Var}[Y] = 13.7$. The interactive effect index A_1^I, which also includes the correlation effect between X_1 and X_3, is strongest for extreme correlation values but mostly close to zero. Overall, the variation in A_1^U and A_1^I more or less cancel out and the first order index of

X_1 stays constant at $1/3$. The indices of X_3 are constantly at zero, displaying the fact, that it has no independent, structural influence, since the term including X_3 is weighted by $\mu_1 = 0$.

4.8 Structural truss model

As an application of the methods, an FEM model of a truss structure was used. The description of the geometry and the input variables is shown in Figure 17, where the properties of the bars in the lower and upper chord (E_1, A_1) are assumed different from the diagonal bars (E_2, A_2). The code is a MATLAB-based finite-element solver calculates the midspan deflection u. The model and the loads with their distributions and parameters are already implemented into UQLab and are used unaltered for the set up of the PCE for ANCOVA and the double loop for Kucherenko's indices. There exist 10 input variables (Lee and Kwak, 2006; Blatman and Sudret, 2011):

- the Young's moduli of the two types of bars in Pa:
  $\{E_1,\, E_2\} \sim \mathcal{LN}(2.1 \cdot 10^{11}, 2.1 \cdot 10^{10})$,
- the cross sectional areas in m^2: $\{A_1,\, A_2\} \sim \mathcal{LN}(2.0 \cdot 10^{-3}, 2.0 \cdot 10^{-4})$
- and the six loads P_i, $i = 1, \ldots, 6$ in N: $\{P_i\} \sim \mathcal{G}(5.0 \cdot 10^4, 7.5 \cdot 10^3)$, $i = 1, \ldots, 6$.

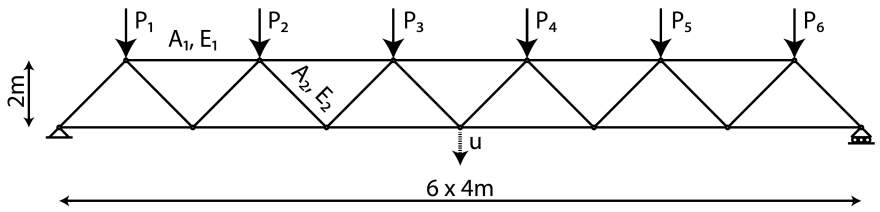


Figure 17: Situation of the truss loaded by $P_1 - P_6$

Case 1 No correlation First the sensitivity analysis is done for no correlation between the variables, to get a idea of the variance contributions. The Sobol' indices are displayed in Table 12.

It is clear that the properties of the chords (E_1, A_1) are the most important with indices $S_i = 37\%$, whereas the web elements hardly influence the variance of the midspan deflection. Moreover, the loads that are closer to the midspan contribute more to its variance. Still, the central loads P_3 and P_4 only contribute 7.7%. Since $S_i = S_i^T$, $i = 1, \ldots, 10$, interactions between variables seem not to have great influence on the variance of the midspan deflection.

Case 2 Correlated loads In reality, point loads do not exist. A load is normally applied on a certain area or, in the two-dimensional case, on a width. However, the truss model

only considers point loads on the given nodes. In order to combine the two (distributed loads in reality with point loads in the FEM model) the loads are correlated. As a result, a high load on one node will imply tendentially higher loads on close nodes. The important parameter in this coupling is the distance between the nodes. The given correlation values are assumed to be the Gaussian copula's parameters and not linear correlations between the Gumbel distributed load variables. In fact, a numerical analysis on the two correlation parameters (see Figure 18) has shown, that for positive correlation values the both measurements are virtually identical. Since the chosen correlation coefficients are positive, they are set to be the Gaussian copula's parameters without transformation. The correlation matrix is set up using an exponential correlation kernel, to ensure a positive definite correlation matrix. The correlation of the two loads at the nodes i and j is given by:

$$\rho_{ij} = e^{-\frac{d_{ij}}{\lambda}}, \tag{75}$$

where d_{ij} is the distance between the nodes and λ is a correlation length. With $d_{ij} = 1 \cdot n$, where n is the amount of nodes from i to j, and $\lambda = 3$ the correlation matrix of the loads reads:

$$\boldsymbol{\rho} = \begin{pmatrix} 1 & 0.72 & 0.51 & 0.37 & 0.26 & 0.19 \\ 0.72 & 1 & 0.72 & 0.51 & 0.37 & 0.26 \\ 0.51 & 0.72 & 1 & 0.72 & 0.51 & 0.37 \\ 0.37 & 0.51 & 0.72 & 1 & 0.72 & 0.51 \\ 0.26 & 0.37 & 0.51 & 0.72 & 1 & 0.72 \\ 0.19 & 0.26 & 0.37 & 0.51 & 0.72 & 1 \end{pmatrix}, \tag{76}$$

giving a correlation coefficient of 72% for neighbouring nodes, which is assumed to be reasonable.

As mentioned in Section 3.6, the FEM's code evaluation time for the double loop of lengths 10^3 (totally 10^6 model evaluations) is 4.5 hours. In order to improve the estimation and keep the evaluation time within realistic bounds, a PCE is set up on 200 samples and used as cheap-to-evaluate structural surrogate. To give a reference, the time needed for the set up of such a PCE and $1{,}000{,}000$ model runs is about 170 seconds. This is a considerable reduction of computation time.

The estimation of the ANCOVA indices is, as for the other models, based on evaluations on 10^6 samples. For the Kucherenko indices also the same loop sizes as for the other models, twice 10^4, are applied. The calculation took 1.5 hours and the resulting indices are presented in Table 12.

The Kucherenko indices distribute importance differently in this case. While the material properties still get about the same indices, the loads gained a lot of importance through correlation. The central loads now have first order indices over 50% and even the out-

Table 12: Sensitivity indices of the truss model. Sobol' indices for **Case 1** without correlation and Kucherenko and ANCOVA indices for **Case 2** with correlated loads.

	Independent case		Dependent case					
Variables	S_i	S_i^T	K_i	K_i^T	A_i	A_i^U	A_i^I	A_i^C
E_1	0.368	0.373	0.351	0.365	0.243	0.243	0	0
E_2	0.012	0.012	0.013	0.013	0.008	0.008	0	0
A_1	0.366	0.372	0.371	0.380	0.243	0.243	0	0
A_2	0.012	0.012	0.013	0.013	0.008	0.008	0	0
P_1	0.005	0.005	0.247	0.003	0.023	0.003	0	0.012
P_2	0.037	0.038	0.431	0.012	0.084	0.025	0	0.060
P_3	0.077	0.077	0.567	0.026	0.139	0.051	0	0.088
P_4	0.076	0.077	0.568	0.027	0.138	0.051	0	0.088
P_5	0.036	0.036	0.430	0.012	0.083	0.024	0	0.059
P_6	0.005	0.005	0.259	0.002	0.023	0.003	0	0.020

ermost loads get about 25% each. Those are much higher numbers than before. The smaller total indices of the loads also indicate a strong presence of correlation. Since the Sobol' indices suggest little to no interaction effects in the model, the first order indices can be regarded as a good measurement of importance. However, since the total indices get significantly enlarged by correlation, their sum is obtained as $\sum_i K_i = 3.25$. The indices do definitely not represent shares of the total variance anymore.

The ANCOVA indices show a similar picture in terms of ranking of the variables. The loads closer to the midspan contribute more, but overall the loads have higher indices than in the uncorrelated case. However, the properties of the chords still seem dominant for the variance of the deflection. Notably, the uncorrelated indices A_i^U show values similar to the uncorrelated ones, highlighting that additional importance arises from the added correlation measured in A_i^C. The interactive effects are negligible. It was already observed that interactions hardly influence the variance and even if there does exist some small interaction, the covariance needed for A_i^C is often negligibly small (compare Section 4.5). Lastly, there is a slight but consistent change observable in the uncorrelated indices: they are all smaller than their Sobol' counterparts. Since there are no considerable interaction effects, the sum of the ANCOVA indices should result in 1, as already discussed in Section 4.6. Since the first order indices of the loads get larger, the others have to compensate this by getting smaller. The sum $\sum_i A_i$ is in fact very close to one.

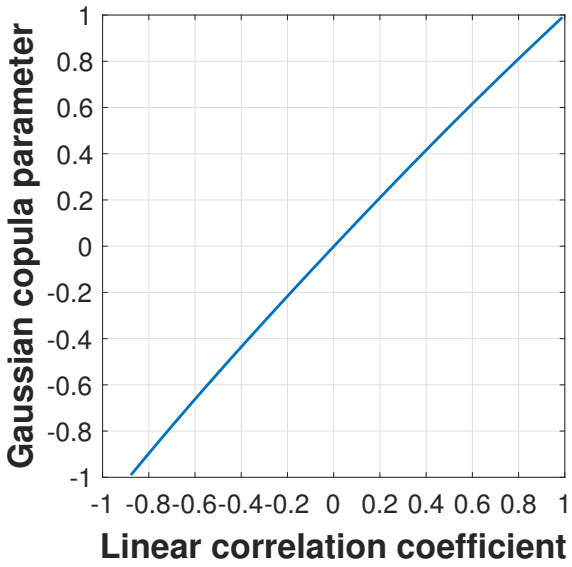


Figure 18: Relationship between the linear correlation coefficient and the Gaussian copula parameter for Gumbel distributed marginals

4.9 Discussion

This section is an extension of the discussion started in Section 4.6. The Ishigami function in Section 4.7 is an additional step in complexity compared to the Sections 4.2 - 4.5. Strong non-linearity and interactions coupled with dependence do not allow for analytical derivations of the partial variances and covariances. But the trends witnessed earlier are present as well. If interactions exist and the correlation is weak $|\rho| \to 0$, Kucherenko's total index is larger than the corresponding first order index. At some point, the correlation effect exceeds the interaction effect and K_i^T falls below the K_i. This phenomenon is observed for positive and negative values of correlation. The Ishigami function even displays some interactive index A_1^I, which is rare to see. Interestingly the index is negative for both positive and negative values of correlation. Comparing the two plots, the similar behaviour of K_2 and A_2 can be seen.

The truss model in Section 4.8 with ten input variables of various marginal distributions serves as an application of the methods on a real computational code. Two important things are to note here. First, complex computational codes take a long time to evaluate. The time needed can quickly exceed realistic time spans available for an analysis. The exploitation of a surrogate model via PCE provides a way to calculate the double loop Monte Carlo estimation of Kucherenko's indices with sufficient accuracy but in comparably no time. Second, respecting dependence may lead to different allocation of sensitivity importance. The significance can be easily demonstrated by putting the example into context.

Imagine the truss is the structural system of an old bridge over a small river in a rural area. As part of the renovation of the road, the safety of the bridge is assessed. The local

construction office produces an FEM code of the structural system with measurements of the input variables such as in the previous section. It turns out that the expected midspan deflection is not too large. However, the variance caused by the uncertainties in the inputs may lead to an excess of the set limit and suspended load of the river might get stuck. In order to decrease the variance, a global sensitivity analysis is performed. They use a software to get the Sobol' indices, notably for independent loads. It turns out, that the main sources of variance are the uncertainties in the material properties. Manipulating the loads would have been feasible, but since it has apparently no strong effect, this is not an option. Instead, two possibilities remain: either the material properties are standardised by replacing or reinforcing elements or tearing the bridge down and build a new one. Either way, the measure is expensive, cumbersome and the bridge might be a construction site for some time.

If the GSA would have been done for dependent loads, the results would be different. Remember how and why correlation was set up. In fact, it is a more realistic representation of the loads. For either, the Kucherenko or ANCOVA indices, the significance of the loads would become apparent. A reduction of the load variance would, in fact, lead to a decrease of the total variance. Moreover, this is easy to accomplish by allowing only one direction to drive at the time. Because dependence is considered and included in the analysis an effective but much cheaper measure to reach the goal has been found.

Another question that needs to be answered at this point, is which priorisation is now correct. Kucherenko's indices elect the central loads as most important, whereas the ANCOVA indices still select the material properties. In a quick study, once the variance of P_3 and once the variance of A_1 were decreased by 50%. The decrease of the sectional area variance lead in fact to a greater decrease of the total variance. This indicates that the balancing of the ANCOVA indices leads to a correct ranking of the indices. Kucherenko indices of correlated variables might exceed the ones of independent, but factually more important variables.

5 Conclusion and Outlook

The development and application of computational models has become crucial in academia and industry. They allow the explanation of phenomena in mathematical terms and are used to make predictions. However, in practice every variable in a mathematical formula represents an event with an expectation value and a variance. As a result, the output of the mathematical model is also tainted with uncertainties. The goal of global sensitivity analysis is, on one hand, to simplify those models by fixing unimportant variables that do not influence the variance of the model output and, on the other hand, to find the important variables for the total variance, in order to reduce it. For models with independent input variables, well established sensitivity measures are the Sobol' indices S_i and S_i^T. Those so-called sensitivity indices are commonly interpreted as follows: a large first order index indicates a strong structural influence of the corresponding variable on the total variance. If a variable interacts with other variables within the model, its total index increases and exceeds the first order index. The total index is never smaller than the first order index since it is the sum of the first order index and higher order indices. All of those are ratios of variances and therefore must be positive. In order to reduce the total variance, it is usually rewarding to reduce the variance of the one variable with the largest total index.

In the context of this thesis another effect comes into play, namely dependence between input variables. There exist various approaches to handle this situation, all leading to different indices, that are sometimes hard to interpret. Kucherenko et al. (2012) and Caniou (2012) both formulated generalisations of the Sobol' indices for dependent inputs. The first formulates them as expectation values. Those can be computed in a double loop or with an estimator for the inner loop and, consequently, just one loop. The second utilises an HDMR set up for the case of independence to modify the ANOVA to work for dependent inputs, resulting in the so-called ANCOVA. In case of independence, both methods yield the Sobol' indices. Those methods have been implemented into MATLAB using the UQLab software, which was developed at the ETH Zurich, and validated.

In Section 4 the two methods are applied on mathematical models of increasing complexity and different cases of dependence, which is modelled by correlation between variables and a Gaussian copula. Using low complexity models, the occurring effects are isolated and the changes in the indices can be explained and understood. Following, the methods are applied on a more complex function and an FEM code of a truss, simulating a real life application of the methods. In the discussions, the behaviour of the indices is analysed and explained as far as possible. The ANCOVA indices A_i are based on independent, structural contributions of the variables in the HDMR and are formulated for first order effects, which they split up into uncorrelative, interactive and correlative contributions. The uncorrelated index A_i^U represents the structural role of the variable in the model.

The interactive index A_i^I is a "first order interaction" measure, that for the investigated models mostly stayed small. It must not be mistaken for an effective interaction measure such as higher order Sobol' indices. The correlative index A_i^C measures the influence of a variable on terms it is not included in, thus only by correlation. The indices A_i^I and A_i^C are calculated through covariance of specific terms, which results in negative terms for negative correlation. For models where interaction effects have no influence on the total variance, the ANCOVA indices add up to 1.

The Kucherenko indices K_i and K_i^T represent apparent influences of variables on the total variance. Even variables without structural role or effective influence on the model output can be attributed importance through correlation with an influencing variable. The first order index increases with correlation, whereas the total index decreases. As long as the total index is larger than the corresponding first order index, the effect of interaction is stronger than the one of correlation. The case of correlated interactive variables is interesting. Kucherenko's indices interpret $|\rho| \to 1$ as an increase of the first order and decrease of the interactive effects of the correlated interacting variables, whereas the ANCOVA indices split the effects up, which can sometimes sum up to 0. It seems as if Kucherenko's first order index does not allow for a clear split between interaction and correlation effects, especially for stronger correlation.

The discussion of the truss shows that neglecting dependence can lead to wrong or just inefficient and costly decisions. It is essential to include dependence, if present, in a sensitivity analysis. It has to be noted, that the example of the truss was simplified by the fact that there were no strong interaction effects on the total variance. Interaction effects are hard to measure for both investigated methods. The total Kucherenko indices K_i^T only inform about the ratio of interaction to correlation effect, whereas the interactive ANCOVA index A_i^C is an interaction effect on first order basis and fails to detect even strong interaction effects in many cases. At this point, the comparison of the Sobol indices $S_i^T - S_i$ still informs the best about present interaction effects. The optimal strategy proposed by the author is to calculate the Sobol' and the ANCOVA indices. The first inform about interaction effects, while the latter can quantify and split up the present correlation effects. The exploitation of higher order ANCOVA indices to get information on interaction effects suggests itself at this point.

Global sensitivity analysis, especially with dependent inputs, will undoubtedly only find broader application, if the calculation and interpretation of the indices is approachable and clearly defined. This thesis aims to make a step into this direction. Another crucial obstacle is computation time and therefore cost. The ANCOVA method is set up for an HDMR and provides fast results for little amounts of model runs. The estimation of Kucherenko's indices through the Monte Carlo double loop is computationally extremely expensive and in general not feasible. In order to bring down this cost the same HDMR as for the ANCOVA can be used as structural surrogate of the actual model and combined

with the double loop. The investigation and understanding for multiple dependence cases for the models was prioritized to the implementation of the Kucherenko estimator. Nevertheless, the calculation of accurate index estimates is still considerably faster. In the context of this thesis, polynomial chaos expansions of the models were used since they are available in UQLab.

Looking ahead, there are many possibilities. For one, Kucherenko's estimator of the inner loop can be implemented and compared to the HDMR double loop method. Whichever method is chosen, both ways to calculate sensitivity indices for dependent input variables should be implemented into UQLab to make them accessible to more UQ analysts and researchers. Additionally, the search for a better handling of dependence in sensitivity analysis continues. The Shapley values (Owen, 2013; Iooss and Prieur, 2017) provide one possibility, the mentioned approach by Mara et al. (2015) using the Rosenblatt transform another. The exploitation of higher order ANCOVA indices is also interesting since they might add information on interaction effects. Overall, this field at the border between mathematics and engineering still leaves much to be improved and discovered.

References

Blatman, G. and B. Sudret (2011). Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *J. Comput. Phys 230*, 2345–2367.

Borgonovo, E. (2007). A new uncertainty importance measure. *Reliab. Eng. Sys. Safety 92*, 771–784.

Caniou, Y. (2012). *Global sensitivity analysis for nested and multiscale models*. Ph. D. thesis, Université Blaise Pascal, Clermont-Ferrand.

Chastaings, G., F. Gamboa, and C. Prieur (2012). Generalized Hoeffding-Sobol decomposition for dependent variables – application to sensitivity analysis. *Electronic Journal of Statistics 6*, 2420–2448.

Cramer, E. and U. Kamps (2017). *Grundlagen der Wahrscheinlichkeitsrechnung und Statistik*. Springer Spektrum, 4^{th} edition.

Ferretti, F., A. Saltelli, and S. Tarantola (2016). Trends in sensitivity analysis practice in the last decade. *Science of the Total Environment 568*, 666–670.

Homma, T. and A. Saltelli (1996). Importance measures in global sensitivity analysis of non linear models. *Reliab. Eng. Sys. Safety 52*, 1–17.

Iooss, B. and P. Lemaître (2014). A review on global sensitivity analysis methods. *arXiv preprint arXiv:1404.2405*.

Iooss, B. and C. Prieur (2017). Shapley effects for sensitivityanalysis with dependent inputs: comparisons with sobol' indices, numerical estimation and applications. Technical report.

Kucherenko, S., A. Tarantola, and P. Annoni (2012). Estimation of global sensitivity indices for models with dependent variables. *Comput. Phys. Comm. 183*, 937–946.

Le Gratiet, L., S. Marelli, and B. Sudret (2017). *Metamodel-based sensitivity analysis: polynomial chaos expansions and Gaussian processes*, Chapter 8, Handbook on Uncertainty Quantification (Ghanem, R. and Higdon, D. and Owhadi, H. (Eds.). Springer.

Lebrun, R. and A. Dutfoy (2009). A generalization of the Nataf transformation to distributions with elliptical copula. *Prob. Eng. Mech. 24*(2), 172–178.

Lee, S. and B. Kwak (2006). Response surface augmented moment method for efficient reliability analysis. *Structural Safety 28*, 261–272.

Li, G., H. Rabitz, P. Yelvington, O. Oluwole, F. Bacon, C. Kolb, and J. Schoendorf (2010). Global sensitivity analysis for systems with independent and/or correlated inputs. *J. Phys.Chem. 114*, 6022–6032.

Mara, T., S. Tarantola, and P. Annoni (2015). Non-parametric methods for global sensitivity analysis of model output with dependent inputs. *Enviroenmental Modeling & Software 72*, 173–183.

Marelli, S. and B. Sudret (2014). UQLab: A framework for uncertainty quantification in Matlab. In *Vulnerability, Uncertainty, and Risk (Proc. 2nd Int. Conf. on Vulnerability, Risk Analysis and Management (ICVRAM2014), Liverpool, United Kingdom)*, pp. 2554–2563.

Nelsen, R. (2006). *An introduction to copulas* (2nd ed.), Volume 139 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Owen, A. (2013). Sobol' indices and shapley value. Technical report, Standford University.

Sobol', I. (1993). Sensitivity estimates for nonlinear mathematical models. *Math. Modeling & Comp. Exp. 1*, 407–414.

Sobol', I. (2001). Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. *Math. Comput. Simul. 55*(1-3), 271–280.

Sudret, B. (2007). *Uncertainty propagation and sensitivity analysis in mechanical models – Contributions to structural reliability and stochastic spectral methods*. Université Blaise Pascal, Clermont-Ferrand, France. Habilitation à diriger des recherches, 173 pages.

Sudret, B. and S. Marelli (2016). Lecture notes on uncertainty quantification in engineering. Technical report, ETH Zurich.

Xiu, D. (2010). *Numerical methods for stochastic computations – A spectral method approach*. Princeton University press.