Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs

Marouane Il Idrissi, Vincent Chabridon, Bertrand Iooss

▶ **To cite this version:**

Marouane Il Idrissi, Vincent Chabridon, Bertrand Iooss. Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs. Environmental Modelling & Software, 2021. hal-03106452v3

HAL Id: hal-03106452
https://hal.science/hal-03106452v3

Submitted on 18 May 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Developments and applications of Shapley effects to reliability-oriented sensitivity analysis with correlated inputs

Marouane Il Idrissi[a,b,c], Vincent Chabridon[a,b], Bertrand Iooss[a,b,c,d]

[a] *EDF Lab Chatou, 6 Quai Watier, 78401 Chatou, France*
[b] *SINCLAIR AI Lab., Saclay, France*
[c] *Institut de Mathématiques de Toulouse, 31062 Toulouse, France*
[d] *Corresponding Author - Email: bertrand.iooss@edf.fr - Phone: +33130877969*

Abstract

Reliability-oriented sensitivity analysis methods have been developed for understanding the influence of model inputs relative to events which characterize the failure of a system (e.g., a threshold exceedance of the model output). In this field, the target sensitivity analysis focuses primarily on capturing the influence of the inputs on the occurrence of such a critical event. This paper proposes new target sensitivity indices, based on the Shapley values and called "target Shapley effects", allowing for interpretable sensitivity measures under dependent inputs. Two algorithms (one based on Monte Carlo sampling, and a given-data algorithm based on a nearest-neighbors procedure) are proposed for the estimation of these target Shapley effects based on the ℓ^2 norm. Additionally, the behavior of these target Shapley effects are theoretically and empirically studied through various toy-cases. Finally, the application of these new indices in two real-world use-cases (a river flood model and a COVID-19 epidemiological model) is discussed.

Keywords: sensitivity analysis, reliability analysis, Sobol' indices, Shapley effects, input correlation

1. Introduction

Nowadays, numerical models are extensively used in all industrial and scientific disciplines to describe physical phenomena (e.g., systems of ordinary differential equations in ecosystem modeling, finite element models in structural mechanics, finite volume schemes in computational fluid dynamics) in order to design, analyze or optimize various processes and systems. These numerical models are often useful from either a scientific standpoint (e.g., by improving the understanding of modeled physical phenomena) or from an engineering standpoint (e.g., to better assist a decision-taking process). In addition to this tremendous growth in computational modeling and simulation, the identification and treatment of the multiple sources of uncertainties has become an essential task from the early design stage to the whole system life cycle. As an example, such a task is crucial in the management of complex systems such as those encountered in energy exploration and production (De Rocquigny et al., 2008) and in sustainable resource development (Beven, 2008).

In addition, the emergence of global sensitivity analysis (GSA) of model outputs played a fundamental role in the development and enhancement of these numerical models (see, e.g., Pianosi et al. (2016); Razavi et al. (2021) for recent reviews). Mathematically, if the model inputs (resp. output)

are denoted by X (resp. Y) and the model is written $G(\cdot)$, such as

$$Y = G(X), \tag{1}$$

GSA aims at understanding the behavior of Y with respect to (w.r.t.) $X = (X_1, \ldots, X_d)^\top$ the vector of d inputs. GSA has been extensively used as a versatile tool to achieve various goals: for instance, quantifying the relative importance of inputs regarding their influence on the output (a.k.a. "ranking"), identifying the most influential inputs among a large number of inputs (a.k.a. screening) or analyzing the input-output code (i.e., the numerically modeled phenomenon) behavior (Saltelli et al., 2008; Iooss and Lemaître, 2015).

When complex systems are critical or need to be highly safe, numerical models can also be of great help for risk and reliability assessment (Lemaire et al., 2009). Indeed, to track potential failures of a system (which could lead to dramatic environmental, human or financial consequences), numerical models allow a simulation of its behavior far from its nominal one (see, e.g., Richet and Bacchi (2019) in flood hazard assessment). In such a context, analytical or experimental approaches can be inappropriate, too expensive, or too difficult to perform. Based on numerical simulations, the tail behavior of the output distribution can be studied and typical *risk measures* can be estimated (Rockafellar and Royset, 2015). Among others, the probability that the output Y exceeds a given threshold value $t \in \mathbb{R}$, given by $\mathbb{P}(Y > t)$ and often called a *failure probability*, is widely used in many applications. When $\{Y > t\}$ is a rare event (i.e., associated to a very low failure probability), advanced sampling-based or approximation-based techniques (Morio and Balesdent, 2015) are required to accurately estimate the failure probability. In this very specific context, dedicated sensitivity analysis methods have been developed, especially in the structural reliability community (see, e.g., Wu (1994); Song et al. (2009); Wei et al. (2012)). In such a framework, called *reliability-oriented sensitivity analysis* (ROSA) (Chabridon, 2018; Perrin and Defaux, 2019), the idea is to provide importance measures dedicated to the problem of rare event estimation.

Formally, standard GSA methods mostly focus on quantities of interest (QoI) characterizing the central part of the output distribution (e.g., the variance for Sobol' indices (Sobol, 1993), the entire distribution for moment-independent indices (Borgonovo, 2007)), while ROSA methods focus on risk measures and their associated practical difficulties (e.g., costly to estimate, inducing a conditioning on the distributions, non-trivial interpretation of the indices). Following Raguet and Marrel (2018), ROSA methods can be categorized regarding the type of study they consider, i.e., according to the following two categories:

- *target sensitivity analysis* (TSA) aims at measuring the influence of the inputs (considering their entire input domain) on the *occurrence* of the failure event. This means considering the following random variable, defined by the indicator function of the failure domain: $\mathbb{1}_{\{G(X)>t\}}$;
- *conditional sensitivity analysis* aims at studying the influence of the inputs on the *conditional* distribution of the output $Y|\{G(X) > t\}$, i.e., exclusively within the critical domain. By Eq. (1), a conditioning also appears on the inputs' domain.

Various indices have been proposed to tackle these two types of studies (see, e.g., Li et al. (2012); Wei et al. (2012); Perrin and Defaux (2019); Marrel and Chabridon (2021)). The present paper is

dedicated to ROSA (under the assumption that the QoI is a failure probability) and focuses on a TSA study. However, a new consideration for TSA is addressed in the present work: the possible statistical dependence between the inputs.

Indeed, most of the common GSA methods (and it is similar for the ROSA ones) have been developed under the assumption of independent inputs. As an example, the well-known Sobol' indices (Sobol, 1993) which rely on the so-called functional analysis of variance (ANOVA) and Hoeffding decomposition (Hoeffding, 1948), can be directly interpreted as shares of the output variance that are due to each input and combination of inputs (called "interactions") as long as the inputs are independent.

When the inputs are dependent, the inputs' correlations dramatically alter the interpretation of the Sobol' indices. To handle this issue, several approaches have been investigated in the literature. For instance, Jacques et al. (2006) proposed to estimate indices for groups of correlated inputs. However, this approach does not allow for a quantification of the influence of individual inputs. Amongst other similar works, Li et al. (2010); Chastaing et al. (2012) proposed to extend the functional ANOVA decomposition to a more general one (e.g., taking the covariance into account). However, the indices obtained for these approaches can be negative, which limits their practical use due to interpretability challenges (i.e., as a share of the output's variance). In addition to this, other works (see, e.g., Xu and Gertner (2008); Mara and Tarantola (2012)) considered a Gram–Schmidt procedure to decorrelate the inputs and proposed to estimate two kinds of contributions for each variable (an uncorrelated one and a correlated one). These developments finally resulted in the proposition of a set of four Sobol' indices (instead of the two standard ones which are the first-order index and total index in the independent case) which enable the correlation effects to be fully captured in a GSA (Mara et al., 2015). Despite this achievement, this approach remains difficult to implement in practice (see Benoumechiara and Elie-Dit-Cosaque (2019) for extensive studies). Finally, the VARS approach (Do and Razavi, 2020) (allowing a thorough analysis of the inputs-output relationships) can handle input correlation but is out of scope of the present work which only focuses on variance-based sensitivity indices, directly computed from the numerical model.

Recently, another method has been developed by considering another type of indices: the *Shapley effects*. The initial formulation originates from the "Shapley values" developed in the field of Game Theory (Shapley, 1953; Osborne and Rubinstein, 1994). The underlying idea is to fairly distribute both gains and costs to multiple players working cooperatively. By analogy with the GSA framework, the inputs can be seen as the players while the overall process can be seen as attributing shares of the output variability to the inputs. Considering the variance of the output in a GSA formulation leads to the so-called "Shapley effects" proposed by Owen (2014). In the same vein, Owen and Prieur (2017); Iooss and Prieur (2019); Benoumechiara and Elie-Dit-Cosaque (2019) bridge the gap between Sobol' indices and Shapley effects while illustrating the usefulness of these new indices to handle correlated inputs in the GSA framework.

Thus, the present work attempts to to extend the use of Shapley effects to the ROSA context. Overall, the main objective is to develop a ROSA index which enables TSA to be performed (i.e., capturing the influence of the inputs on a risk measure, typically a failure probability here) under the constraint of dependent inputs. This work relies on the use of recent promising results and numerical

tools (both in field of TSA Spagnol (2020) and Shapley effects' estimation Broto et al. (2020)).

The outline of this paper is the following. Section 2 is devoted to a pedagogical introduction of the statistical dependence challenges for variance-based sensitivity indices, that can be solved by Shapley effects. Section 3 presents a new formulation of TSA, based on Shapley effects leading to the novel target Shapley effects, while Section 4 develops two algorithms for their estimation. Section 5 provides illustrations on simple toy-cases which give analytical expressions of the target Shapley effects, allowing deeper appreciation of their behavior. Section 6 applies these new sensitivity indices to two use-cases: a simplified model of a river flood and an epidemiological model applied to the COVID-19 pandemic. Finally, Section 7 gives conclusions and research perspectives.

Throughout this paper, the mathematical notation $\mathbb{E}(\cdot)$ (resp. $\mathbb{V}(\cdot)$) will represent the expectation (resp. variance) operator.

2. Variance-based sensitivity analysis with dependent inputs: the Shapley solution

While devoted to computer experiments, GSA has close connections with multivariate data analysis and statistical learning (Christensen, 1990; Hastie et al., 2002). Indeed, in all these topics, one important issue is often to provide a weight to some variables (the inputs) w.r.t. its impact on another variables (the outputs). Depending on the domain, such a weight can either be called a "sensitivity index" or an "importance measure". A very convenient way is to base these weights on the ANOVA (analysis of variance) decomposition (Christensen, 1990; Sobol, 1993) of the output variance. Indeed, such a decomposition provides a natural division of the output's variance in shares attributed to each input. The principle of the "variance-based sensitivity indices" (Saltelli et al., 2008) consists then in understanding how to separate the contribution of each X_i from the variance of Y. However, due to potential statistical dependencies between inputs, this decomposition cannot be directly performed. Starting from a simple example of a linear model, chosen for pedagogical purposes, this section provides a reminder on this topic while illustrating the important potential of Shapley effects in practice.

2.1. Understanding the correlation issues via the linear model case

In this section, the aim is to quantify the relative importance of d scalar inputs X_j $(j = 1, \ldots, d)$ by fitting on a data sample (coming from the model Eq. (1)) a linear regression model so as to predict a scalar output Y:

$$Y(X) = \sum_{j=0}^{d} \beta_j X_j + \epsilon \,, \tag{2}$$

where $X_0 = 1$, $\beta = (\beta_0, \ldots, \beta_d)^\top \in \mathbb{R}^{d+1}$ is the effects vector and $\epsilon \in \mathbb{R}$ the model's error of variance σ^2. If a sample of inputs and outputs $(\mathbf{X}^n, \mathbf{Y}^n) = \left(X_1^{(i)}, \ldots, X_d^{(i)}, Y^{(i)}\right)_{i=1,\ldots,n}$ is available (with $n > d$), the Ordinary Least Squares method (see, e.g., Christensen (1990)) can easily be used to estimate the parameters β and σ^2 in the linear regression model in Eq. (2). Moreover, one obtains the predictor $\widehat{Y}(x^*)$ of Y at any prediction point x^*. An important validation metric of this model is the classical *coefficient of determination* given by:

$$R^2_{Y(X)} = \sum_{i=1}^{n} \left[\widehat{Y}(X^{(i)}) - \bar{Y}\right]^2 \Big/ \left[Y^{(i)} - \bar{Y}\right]^2 \tag{3}$$

where $\bar{Y}$ is the output empirical mean. $R^2_{Y(X)}$ represents the percentage of output variability explained by the linear regression model of Y on X. Finally, from Eq. (2), the variance decomposition expresses as:

$$\mathbb{V}(Y) = \sum_{j=1}^{d} \beta_j^2 \mathbb{V}(X_j) + 2 \sum_{k>j} \beta_j \beta_k \mathrm{Cov}(X_j, X_k) + \sigma^2 \ . \tag{4}$$

In the specific case of independent inputs, the covariance terms cancel and the standard ANOVA (i.e., $\mathbb{V}(Y) = \sum \beta_j^2 \mathbb{V}(X_j) + \sigma^2$) is obtained. Then, global sensitivity indices, called Standardized Regression Coefficients (SRC), can be directly computed:

$$\mathrm{SRC}_j = \beta_j \sqrt{\mathbb{V}(X_j)/\mathbb{V}(Y)} \ . \tag{5}$$

The estimation of the SRC is made by replacing the terms in Eq. (5) by their estimates. Interestingly, this metric for relative importance is signed (thanks to the regression coefficient sign), giving the sense of variation of the output w.r.t. each input. Moreover, SRC_j^2 represents a share of variance and the sum of all the SRC_j^2 approaches R^2 (i.e., the amount of explained variance by the linear model). Note that, in a perfect linear regression model (i.e., without any random error term ϵ), SRC_j is equal to the linear Pearson's correlation coefficient between X_j and Y (denoted by $\rho(X_j, Y)$). Note also that the ANOVA and SRC^2 extend to the functional ANOVA and Sobol' indices in the general (non-linear model) case (see Appendix A).

When the inputs are dependent, the main concern is to allocate the covariance terms in Eq. (4) to the various inputs. In this case, the Partial Correlation Coefficient (PCC) has been promoted in GSA (Helton et al., 2006; Saltelli et al., 2008) as a substitute to the SRC, in order to cancel the effects of other inputs when allocating the weight of one input X_j in the variance of Y:

$$\mathrm{PCC}_j = \rho(X_j - \widehat{X_{-j}}, Y - \widehat{Y_{-j}}) \tag{6}$$

where X_{-j} is the vector of all the d inputs except X_j, $\widehat{X_{-j}}$ is the prediction of the linear model expressing X_j w.r.t. X_{-j} and $\widehat{Y_{-j}}$ is the prediction of the linear model Y w.r.t. X_{-j}. However, PCC is not a right sensitivity index of the input. Indeed, it consists in measuring the linear correlation between Y and X_j by fixing X_{-j}, and is then a measure of the linearity (and not the importance) between the output and one input.

Instead of controlling other inputs X_{-j} such as done in the PCC, the Semi-Partial Correlation Coefficient (SPCC) quantifies the proportion of the output variance explained by X_j after removing the information brought by X_{-j} (on X_j) (Johnson and LeBreton, 2004):

$$\mathrm{SPCC}_j = \rho(X_j - \widehat{X_{-j}}, Y) \ . \tag{7}$$

SPCC can also be expressed by using the relation $\mathrm{SPCC}_j^2 = R^2_{Y(X)} - R^2_{Y(X_{-j})}$, which clearly shows that SPCC gives the additional explanatory power of the input X_j in the linear regression model of Y on X. However, the SPCC of highly correlated inputs will be small, despite their "real" explanatory power on the output. This aspect seems to be the main drawback of SPCC and probably explains its lack of popularity for GSA purposes.

To give an intuitive view of the limitations induced by the *multicollinearity* of the inputs (i.e., when inputs are linearly correlated to each other), Venn diagrams can be used (see, Figure 1) in the case of two inputs, X_1 and X_2, and one output Y. From Figure 1, the coefficient of determination can be written as:

$$R^2_{Y(X_1,X_2)} = \frac{a+b+c}{a+b+c+\sigma^2} \,, \tag{8}$$

where $a+b+c+\sigma^2$ is equal to the variance of Y and $a+b+c$ represents the part of explained variance by the regression model (with $b=0$ in the uncorrelated case). In this elementary example, the previously introduced sensitivity indices are given by Clouvel (2019):

$$\begin{array}{lcllcl} \mathrm{SRC}_1^2 & = & (a+b)/(a+b+c+\sigma^2) \,, & \mathrm{SRC}_2^2 & = & (c+b)/(a+b+c+\sigma^2) \,, \\ \mathrm{PCC}_1^2 & = & a/(a+\sigma^2) \,, & \mathrm{PCC}_2^2 & = & c/(c+\sigma^2) \,, \\ \mathrm{SPCC}_1^2 & = & a/(a+b+c+\sigma^2) \,, & \mathrm{SPCC}_2^2 & = & c/(a+b+c+\sigma^2) \,. \end{array} \tag{9}$$

Thus, one can understand the limitations of SRC, PCC and SPCC when correlation is present: the variance share which comes from the correlation between inputs (i.e., the b value in Figure 1 - right) is allocated two times with the SRC but not allocated at all with SPCC, while PCC does not represent any variance sharing.

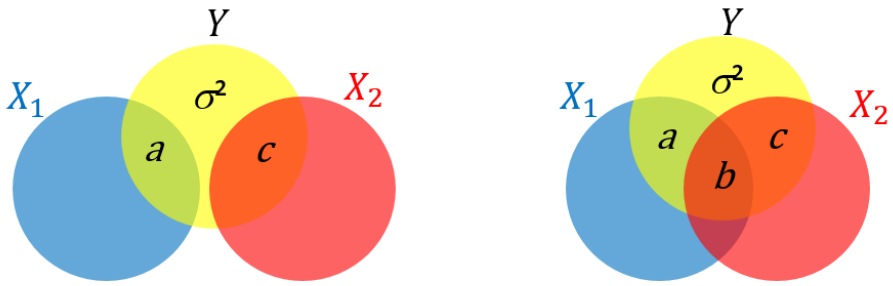

Figure 1: Inspired from Clouvel (2019). Illustration scheme of the effect of two inputs X_1 and X_2 on an output variable Y when they are: uncorrelated (left) or correlated (right).

The three problems above can be solved by using another sensitivity index which finds a way to partition the R^2 among the d inputs: the LMG (Lindeman et al., 1980; Grömping, 2006) (acronym based on the authors' names, i.e., "Lindeman - Merenda - Gold") uses sequential sums of squares from the linear model and obtains an overall measure by averaging over all orderings of inputs. Mathematically, let A be a subset of indices in the set of all subsets of $\{1,\ldots,d\}$ and $X_A = (X_j : j \in A)$ a group of inputs. LMG is based on the measure of the elementary contribution of any given variable X_j to a given subset model $Y(X_A)$ by the increase in R^2 that results from adding that predictive variable to the regression model:

$$\mathrm{LMG}_j = \frac{1}{d!} \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\ldots,d\}}} \left[R^2_{Y(X_{v\cup\{j\}})} - R^2_{Y(X_v)} \right] \tag{10}$$

with v the indices entered before j in the order π. In Eq. (10), the sum is performed over all the

permutations of $\{1, \ldots, d\}$. For the case of two inputs (see Figure 1), we can easily show that:

$$\text{LMG}_1 = (a + b/2)/(a + b + c + \sigma^2) \ , \ \text{LMG}_2 = (c + b/2)/(a + b + c + \sigma^2) \ . \tag{11}$$

Then, in the LMG framework, the $R^2_{Y(X_1,X_2)}$ has been perfectly shared into two parts with an equitable distribution of the b term between X_1 and X_2.

This allocation principle exactly corresponds to the application of the *Shapley values* (Shapley, 1953) on the linear model. This attribution method has been primarily used in cooperative game theory, allowing for a cooperative allocation of resources between players based on their collective production (see Appendix B for a more formal definition). The Shapley values solution consists in fairly distributing both gains and costs to several actors working in coalition. In situations when the contributions of each actor are unequal, it ensures that each actor gains as much or more as they would have from acting independently. Now, if the actors are identified with a set of inputs and the value assigned to each coalition is identified to the explanatory power of the subset of model inputs composing the coalition, one obtains the LMG in Eq. (10).

2.2. Shapley effects

In the general case, when no assumption is made on the model $G(\cdot)$ (see Eq. (1)), variance-based sensitivity indices have been developed (see, Sobol (1993); Saltelli et al. (2008)) and applied to perform a GSA of complex models (see, e.g., Nossent et al. (2011)).

When the inputs are assumed to be independent, they allow the variance of the model output to be decomposed according to each possible subsets of inputs (called "Sobol' indices"). They are a means to measure the individual effects of inputs, as well as the effect of their interaction (see, Appendix A for the theoretical details).

When the inputs are effectively dependent, the Sobol' indices lose their inherent interpretation (i.e., decomposition in individual and interaction effects). To remedy this drawback, Owen (2014) recently proposed game theoretic GSA indices, in the same fashion as the LMG indices (see, Eq. (10)), inspired by the Shapley values of cooperative games. They are defined by:

$$Sh_j = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(\text{val}(A \cup \{j\}) - \text{val}(A)\right), \tag{12}$$

where $\text{val}(A)$ is called *cost* function (or *value* function) assigned to a subset $A \in \mathcal{P}_d$ of inputs, $\mathcal{P}_d$ denotes the set of all possible subsets of $\{1, \ldots, d\}$, $\{-j\}$ denotes the set of indices $\{1, \ldots, d\} \setminus j$ and $|A|$ denotes the cardinal number of A.

For GSA purposes, Owen (2014) proposes to use the "closed Sobol' indices" as the value function in Eq. (12):

$$\text{val}(A) = S^{\text{clos}}_A = \frac{\mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_A\right]\right)}{\mathbb{V}\left(G(X)\right)}, \tag{13}$$

The attribution properties of the Shapley values applied to this particular cost function, leads to the

definition of the *Shapley effects*:

$$Sh_j = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(S^{\text{clos}}_{A \cup \{j\}} - S^{\text{clos}}_{A} \right). \tag{14}$$

These indices allow for a quantification of the importance of each input, which intrinsically takes into account both interaction and dependence effects on the output's variance. Moreover, two important properties of the Shapley effects allow for their interpretation: they sum up to one and are non-negative. Thus, they can be considered as a decomposition of the output's variance. They allow for input ranking by attributing to each input a percentage of the variable of interest's variance. These indices have been extensively studied by Owen and Prieur (2017); Iooss and Prieur (2019). An alternative way of defining the Shapley effects has also been proposed, by taking the following cost function:

$$\text{val}(A) = \frac{\mathbb{E}\left[\mathbb{V}\left(G(X)|X_{\overline{A}}\right)\right]}{\mathbb{V}\left(G(X)\right)} \tag{15}$$

where $\overline{A} = \{1, \ldots, d\} \setminus A$. This alternative definition leads to an equivalent definition of the Shapley effects (Eq. (14)), as outlined by Song et al. (2016), and allows for additional estimation methods.

In order to illustrate Shapley effects' attributions, one can first consider a model with three inputs $X = (X_1, X_2, X_3)^\top$. From Eq. (14), one has:

$$\begin{aligned} Sh_1 = & \frac{1}{3} S^{\text{clos}}_1 \\ & + \frac{1}{6}\Big[\left(S^{\text{clos}}_{\{1,2\}} - S^{\text{clos}}_{2} \right) + \left(S^{\text{clos}}_{\{1,3\}} - S^{\text{clos}}_{3} \right) \Big] \\ & + \frac{1}{3} \left(S^{\text{clos}}_{\{1,2,3\}} - S^{\text{clos}}_{\{2,3\}} \right). \end{aligned}$$

If the three inputs are assumed to be independent, this result leads to:

$$Sh_1 = S_1 + \frac{1}{2} S_{\{1,2\}} + \frac{1}{2} S_{\{1,3\}} + \frac{1}{3} S_{\{1,2,3\}}$$

where one can notice that the Shapley effects' decomposition consists in allocating the initial Sobol' index, plus an equal share of the interaction effects between all the inputs. However, if dependence between inputs is assumed, this behavior cannot be clearly illustrated, except when a linear model $G(\cdot)$ is assumed (see Subsection 2.1).

The quantity $\left(S^{\text{clos}}_{A \cup \{j\}} - S^{\text{clos}}_{A} \right)$ can be interpreted as being a quantification of the *marginal effects of the input j in relation to the subset of variables A*. It is heavily linked to the notion of *marginal contributions* of cooperative games, aiming at quantifying the bargaining power of a player in an allocation process (Brandenburger, 2007). If S^{clos}_{A} is believed to contain the initial effects of the inputs in A, plus their interaction effects, and any effect due to their dependence structure, then the increment $\left(S^{\text{clos}}_{A \cup \{j\}} - S^{\text{clos}}_{A} \right)$ quantifies the initial effect of the input j, its interaction effects with the inputs in A, and the effects due to their dependence. Then, the Shapley attribution weighs all the marginal effects, in order to assess the effective influence of the involved inputs through their marginal contributions, in the same fashion as the LMG indices for a linear model, as depicted in Section 2.1.

It is important to note that the above mentioned alleged decomposition of S_A^{clos} cannot be verified due to the lack of a univocal functional variance decomposition when inputs are dependent. However, the interpretation of the Shapley effects does not rely on the chosen cost function to be meaningful, but rather on its ability to quantify marginal contributions. Empirical studies and analytical studies show that the choice of S_A^{clos} as a cost function remains pertinent, even when inputs are dependent (see, Iooss and Prieur (2019)).

3. Reliability-oriented Shapley effects for target sensitivity analysis

3.1. A brief overview of reliability-oriented sensitivity analysis

When focusing on complex systems, one often needs to prepare for possible critical events, which potentially have a low occurence probability but lead to a system failure. Such failures may have dramatic human, environmental and economic consequences, depending on the context. The fields of reliability assessment and risk analysis (Lemaire et al., 2009; Richet and Bacchi, 2019), aim to prevent these failures. Mathematically, a reliability problem focuses on a *risk measure* computed from the tail of the variable of interest's distribution (Rockafellar and Royset, 2015). Performing sensitivity analysis in such a context requires the use of dedicated tools, which have been developed by various authors under the denomination of "reliability-oriented sensitivity analysis" (ROSA) (see, e.g., Perrin and Defaux (2019); Derennes et al. (2021); Marrel and Chabridon (2021)). A large panel of ROSA methods have been proposed in the structural reliability community such as, for example, several variance-based approaches (see, e.g., Morio (2012); Wei et al. (2012); Perrin and Defaux (2019); Chabridon et al. (2020)) and moment-independent approaches (see, e.g., Cui et al. (2010); Li et al. (2012); Derennes et al. (2021)). From the GSA community, several extensions have also been proposed in order to study risks, or reliability measures. The contrast-based indices proposed by Fort et al. (2016) are, amongst others, an example of a versatile tool which can handle several types of QoI. They were applied in the works of Browne et al. (2017); Maume-Deschamps and Niang (2018) in quantile-oriented formulations. Other formulations such as the quantile-based global sensitivity measures (Kucherenko et al., 2019) or other indices related to dependence measures (Raguet and Marrel, 2018; Marrel and Chabridon, 2021) have been proposed.

In the context of reliability assessment, a typical risk measure is the *failure probability* given by:

$$p_t^Y \overset{\text{def}}{=} \mathbb{P}\left(Y > t\right) = \mathbb{P}\left(G(X) > t\right) = \mathbb{E}\left[\mathbb{1}_{\{G(X)>t\}}(X)\right] = \mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X)\right] \tag{16}$$

where $t \in \mathbb{R}$ represents a threshold characterizing the state of the system. Typically, the event $\{Y > t\}$ denotes a failure event (i.e., the system described by the model $G(X)$ enters a failure state). As for $\mathcal{F}_t$, it represents the input failure domain, i.e., $\mathcal{F}_t \overset{\text{def}}{=} \{X \mid G(X) > t\}$.

Performing a ROSA study poses a few challenges: firstly, the variable of interest here is not directly Y anymore, but rather a binary random variable whose occurrence is characterized by the indicator function $\mathbb{1}_{\mathcal{F}_t}(X)$; secondly, in practice, these failure events are typically "rare events", associated to a low failure probability which might be difficult to estimate in practice through typical sampling methods (Morio and Balesdent, 2015); thirdly, the type of study one desires to perform has to be reinterpreted regarding the new QoI. Regarding this last point, Raguet and Marrel (2018) focus on

two types of studies when dealing with critical events: the first one, called *target sensitivity analysis* (TSA), aims at catching the influence of the inputs on the occurrence of the failure event, while the second one, called "conditional sensitivity analysis" aims at studying the influence of the inputs once the threshold value has been reached (i.e., within the failure domain). The present paper is dedicated to ROSA (under the assumption that the QoI is a failure probability given by Eq. (16)) and aims at developing tools for TSA.

To illustrate this paradigm in plain text, one can refer to the study of the water level in a river protected by a dyke. From the traditional GSA point of view, the central question would be "Which inputs influence the water level?", while in the TSA paradigm, one focuses more on the question "Which inputs influence the occurrence of a flood?". Note that this particular example is studied in depth in Subsection 6.1.

When the inputs are assumed to be independent, a first category of sensitivity indices dedicated to TSA are the "target Sobol' indices" whose first formulation has been proposed by Li et al. (2012). In this document, the original Sobol' indices, in the TSA context, are denoted:

$$\text{T-S}_A = \sum_{B \subseteq A} (-1)^{|A|-|B|} \frac{\mathbb{V}\left(\mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X) \mid X_B\right]\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)} \tag{17}$$

where $\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right) = p_t^Y(1-p_t^Y)$. Similarly, the closed Sobol' indices (see Appendix A) are defined as follows:

$$\text{T-S}_A^{\text{clos}} = \frac{\mathbb{V}\left(\mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X) \mid X_A\right]\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)}. \tag{18}$$

Several estimation schemes for these indices have been proposed when dealing with rare failure events (Wei et al., 2012; Perrin and Defaux, 2019). To illustrate the behavior of these indices, one can consider a linear model given by $Y = X_1 + X_2 + X_3$, with $X = (X_1, X_2, X_3)^\top$, three standard Gaussian random variables assumed to be independent. The left plot of Figure 2 represents the probability density function (pdf) of Y, along with four different threshold values, corresponding to four different failure probability levels. The right plot of Figure 2 presents the different values of T-S_A, w.r.t. the threshold t. Note that, when the inputs are assumed to be independent, the second-order Sobol' indices (i.e., when $|A| = 2$) verify T-$S_{\{1,2\}}$ = T-$S_{\{1,3\}}$ = T-$S_{\{2,3\}}$. One can additionally remark that as soon as t induces a low or high failure probability (i.e., "close" to 0 or 1), the third-order (i.e., $|A| = 3$) closed Sobol' index for TSA increases, indicating high interaction between the three inputs. Note that studying this behavior falls under the conditional sensitivity analysis paradigm, which is out of the scope of this paper. However, the acknowledgment of this phenomenon remains important for better understanding the proposed indices' behavior.

Another category of sensitivity indices dedicated to TSA fall under the category of "moment-independent" ROSA indices. Among others, one can mention the two indices proposed by Cui et al. (2010) which are given by:

$$\eta_A = \frac{1}{2}\mathbb{E}\left[\left|p_t^Y - p_t^{Y|X_A}\right|\right] \tag{19a}$$

$$\delta_A = \frac{1}{2}\mathbb{E}\left[\left(p_t^Y - p_t^{Y|X_A}\right)^2\right] \tag{19b}$$

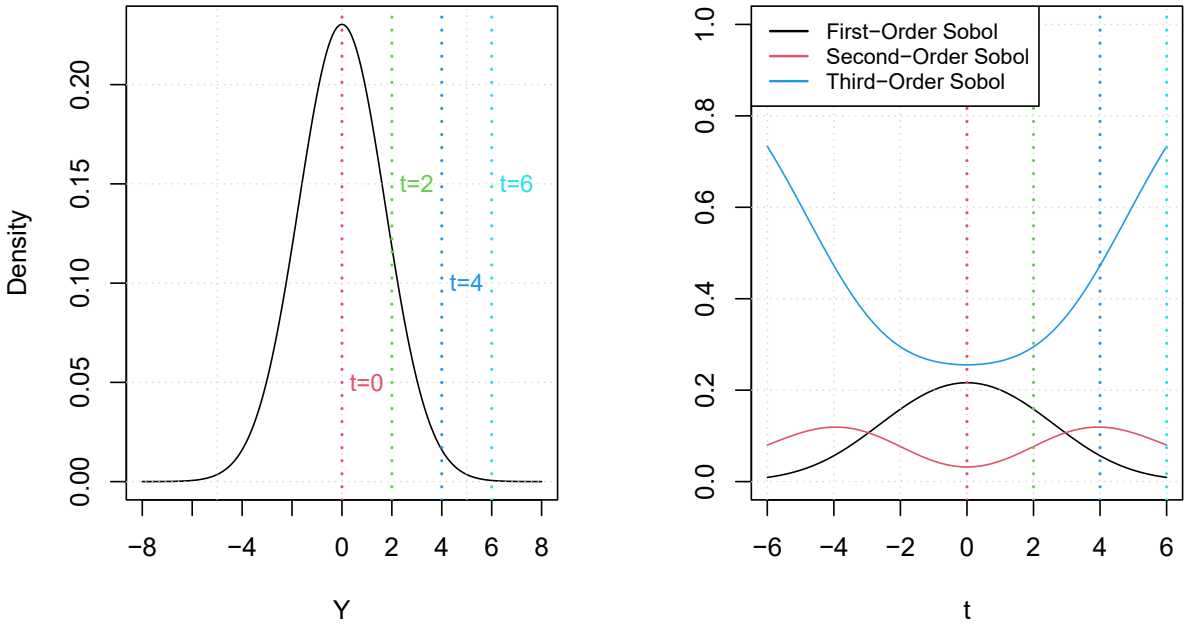

Figure 2: Probability density function of the output with four different threshold values (left) and the related target Sobol' indices (right) for Y being the sum of three independent Gaussian random variables.

where $p_t^{Y|X_A}$ denotes the conditional failure probability when X_A is fixed. Note that, if η_A does not require any independence assumption for a meaningful quantification of input influence, it is known to be difficult to estimate in practice (Derennes et al., 2021). As for δ_A, it is simply proportional to the target closed Sobol' index given in Eq. (18). Note that an extension of δ_A has been proposed in Li et al. (2016) for correlated inputs. It relies on a similar orthogonalization procedure strategy as proposed by Mara and Tarantola (2012) for usual Sobol' indices. However, as mentioned previously, this tends to increase the number of estimated indices to properly interpret the inputs' influence.

The following section aims at introducing the *distance-based TSA indices*, while highlighting their links with existing TSA indices. New TSA indices inspired from the Shapley values (see, Section 2.2) are then proposed.

3.2. Distance-based TSA indices

As outlined by several authors (Fort et al., 2016; Raguet and Marrel, 2018), it can be noted that $\mathbb{V}\left(\mathbb{E}\left[Y|X_A\right]\right) = \mathbb{E}\left[\left(\mathbb{E}\left[Y|X_A\right] - \mathbb{E}\left[Y\right]\right)^2\right]$. This equality can be interpreted as the expected squared distance between two expectations, and thus allows to apprehend closed Sobol' indices (see Eq. (13)) as a particular case of distance-based indices. This broader point of view has been adopted by Fort et al. (2016) to provide a generalization of the Sobol' indices using *contrast functions*.

By applying a similar idea for TSA, one can extend the standard T-S_A and η_A definition to more general cases based on distances. One can then define more general distance-based TSA indices, relative to a subset of inputs $A \in \mathcal{P}_d$, as follows:

$$\text{T-S}_A^{\mathcal{D}} = \frac{\mathbb{E}\left[\mathcal{D}\left(p_t^Y, p_t^{Y|X_A}\right)\right]}{\mathbb{E}\left[\mathcal{D}\left(p_t^Y, p_t^{Y|X}\right)\right]} \tag{20}$$

where $\mathcal{D}(\cdot,\cdot)$ can be any distance function. Links can be made between this definition and the indices presented previously, through the use of specific distance functions. For example, by choosing the distance derived from the ℓ^1 norm (i.e., the absolute difference), one can remark that the corresponding distance-based TSA index is proportional to the η_A index:

$$\text{T-S}_A^{\ell^1} = \frac{\mathbb{E}\Big[\big|\mathbb{E}[\mathbb{1}_{F_t}(X)] - \mathbb{E}[\mathbb{1}_{F_t}(X)|X_A]\big|\Big]}{\mathbb{E}\Big[\big|\mathbb{1}_{\mathcal{F}_t}(X) - \mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X)\right]\big|\Big]} = \frac{2}{\mathbb{E}\Big[\big|\mathbb{1}_{\mathcal{F}_t}(X) - \mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X)\right]\big|\Big]}\eta_A \tag{21}$$

Moreover, by using the distance derived from the ℓ^2 norm (i.e., the squared difference), one can remark the resulting distance-based TSA indices are equal to the closed Sobol' index for TSA, as defined in Eq. (18):

$$\text{T-S}_A^{\ell^2} = \text{T-S}_A^{\text{clos}} = \frac{\mathbb{V}\Big(\mathbb{E}\big[\mathbb{1}_{\mathcal{F}_t}(X) \mid X_A\big]\Big)}{\mathbb{V}\Big(\mathbb{1}_{\mathcal{F}_t}(X)\Big)}. \tag{22}$$

As outlined, the distance-based TSA indices are intimately related to existing ones (i.e., η_A and the closed Sobol' indices for TSA), and can thus be seen as a broader class of indices. Moreover, they are relevant candidates as cost functions for defining Shapley values inspired TSA indices, following a similar line of thinking as in Section 2.2.

3.3. ($\mathcal{D}$)-target Shapley effects

In this subsection, a novel family of TSA indices is proposed, namely the ($\mathcal{D}$)-target Shapley effects. As briefly mentioned previously, these indices are constructed by taking distance-based TSA indices, defined in Eq. (20), as cost functions in a Shapley attribution procedure (see Eq. (12)). For a specific input $j \in \{1,\dots,d\}$, its ($\mathcal{D}$)-target Shapley effects can be defined as being:

$$\text{T-Sh}_j^{\mathcal{D}} = \frac{1}{d}\sum_{A\subset\{-j\}} \binom{d-1}{|A|}^{-1} \Big(\text{T-S}_{A\cup\{j\}}^{\mathcal{D}} - \text{T-S}_A^{\mathcal{D}}\Big) \tag{23}$$

where $\{-j\} = \{1,\dots,d\} \setminus j$. The main property allowing for a clear interpretation of the ($\mathcal{D}$)-target Shapley effects is the following:

Property 1 (($\mathcal{D}$)-target Shapley effects decomposition)**.** *Let $A \in \mathcal{P}_d$, and $val(A) = \text{T-}S_A^{\mathcal{D}}$. For any distance function $\mathcal{D}(.,.)$, the following property holds:*

$$\sum_{j=1}^{d} \text{T-}Sh_j^{\mathcal{D}} = 1. \tag{24}$$

It is important to note that this decomposition property does not rely on any independence assumption about the probabilistic model of the inputs. However, in order to ensure a meaningful interpretation of these indices, (i.e., as a percentage of a statistical dispersion), one needs to ensure that the T-Sh_j are non-negative, for all $j = 1,\dots,d$.

By choosing $\text{T-S}_A^{\ell^1}$ as a cost function (i.e., $\mathcal{D}(x,y) = |x-y|$), one can then define the (ℓ^1)-target

Shapley effect associated to a variable $j \in \{1, \dots, d\}$ as being:

$$\text{T-Sh}_j^{\ell^1} = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(\text{T-S}_{A \cup \{j\}}^{\ell^1} - \text{T-S}_A^{\ell^1} \right). \tag{25}$$

These indices are non-negative (see the proof in Appendix C.1) which allows the (ℓ^1)-target Shapley effects to be interpreted as the percentage of the mean absolute deviation of the indicator function (i.e., $\mathbb{E}\Big[\big|\mathbb{1}_{\mathcal{F}_t}(X) - \mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X)\right]\big|\Big]$), allocated to each input X_j, $j \in \{1, \dots, d\}$.

By choosing $\text{T-S}_A^{\ell^2}$ as a cost function (i.e., $\mathcal{D}(x, y) = (x - y)^2$), the (ℓ^2)-target Shapley effect associated to the variable $j \in \{1, \dots, d\}$ can be defined as:

$$\text{T-Sh}_j^{\ell^2} = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(\text{T-S}_{A \cup \{j\}}^{\ell^2} - \text{T-S}_A^{\ell^2} \right). \tag{26}$$

Being also non-negative (see the proof in Appendix C.2), they can be interpreted as a percentage of the variance of the indicator function allocated to the input X_j, $j \in \{1, \dots, d\}$. Moreover, using Eq. (22), by analogy with Eqs. (13) and (15) (in a similar fashion as the alternate cost function proposed by Song et al. (2016)), if one chooses to define the cost function $\text{val}(A)$ as being:

$$\text{T-E}_A \overset{\text{def}}{=} \frac{\mathbb{E}\Big[\mathbb{V}\big(\mathbb{1}_{\mathcal{F}_t}(X) | X_{\overline{A}}\big)\Big]}{\mathbb{V}\Big(\mathbb{1}_{\mathcal{F}_t}(X)\Big)} \tag{27}$$

with $\overline{A} = \{1, \dots, d\} \setminus A$, then one has an equivalent way of defining the (ℓ^2)-target Shapley effect.

In the following, the (ℓ^2)-target Shapley effect $\text{T-Sh}_j^{\ell^2}$ will be referred to as "the" target Shapley effect and denoted T-Sh_j:

$$\text{T-Sh}_j \overset{\text{def}}{=} \text{T-Sh}_j^{\ell^2}. \tag{28}$$

4. Estimation methods and practical implementation of target Shapley effects

The estimation of the target Shapley effects Eq. (26) can be done into two distinct steps:

- **Step #1**: estimation of the *conditional elements*, i.e., the estimation of $\text{T-S}_A^{\ell^2}$ or T-E_A for all $A \in \mathcal{P}_d$;
- **Step #2**: an *aggregation procedure*, i.e., a step to compute the T-Sh_j by plugging in the previous estimations of Step #1 in Eq. (26).

In the following, two estimation methods are proposed: the first one based on a Monte Carlo sampling procedure, and the second one based on a nearest-neighbor approximation technique.

4.1. Monte Carlo sampling-based estimation

This procedure, introduced in Song et al. (2016) for the estimation of Shapley effects, relies on a Monte Carlo estimation of the conditional elements. It requires the ability to sample from the marginal distributions of the inputs (i.e., P_{X_A} for all $A \subseteq \{1, \dots, d\} \setminus \emptyset$), as well as from all the conditional

distributions (i.e., $P_{X_{\overline{A}}|X_A}$, for all possible subsets of inputs A). Additionally, one also needs to be able to evaluate the model $G(\cdot)$ which is usually the case in the context of uncertainty quantification of numerical computer models (ignoring the potential difficulties related to the cost of a single evaluation of $G(\cdot)$) (De Rocquigny et al., 2008).

In order to estimate a conditional element $\text{T-S}_A^{\ell^2}$, one needs to randomly draw several i.i.d. samples:

- an i.i.d. sample of size N drawn from P_X and denoted by $(X^{(1)}, \ldots, X^{(N)})$;
- another i.i.d. sample of size N_v drawn from P_{X_A} and denoted by $(X_A^{(1)}, \ldots, X_A^{(N_v)})$;
- for each element $X_A^{(i)}, i = 1, \ldots, N_v$, a corresponding sample of size N_p drawn from $P_{X_{\overline{A}}|X_A}$ given that $X_A = X_A^{(i)}$ and denoted by $(\widetilde{X}_i^{(1)}, \ldots, \widetilde{X}_i^{(N_p)})$.

Then, the Monte Carlo estimator of $\text{T-S}_A^{\ell^2}$ can be defined as:

$$\widehat{\text{T-S}}_{A,\text{MC}} = \frac{\sum_{i=1}^{N_v} \left(\frac{1}{N_p} \sum_{j=1}^{N_p} \mathbb{1}_{\mathcal{F}_t}(\widetilde{X}_i^{(j)}, X_A^{(i)}) - \widehat{p}_t^Y \right)^2}{(N_v - 1)\widehat{p}_t^Y (1 - \widehat{p}_t^Y)} \tag{29}$$

with

$$\widehat{p}_t^Y = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\mathcal{F}_t}(X^{(i)}). \tag{30}$$

Finally, the aggregation procedure gives:

$$\widehat{\text{T-Sh}}_{j,\text{MC}} = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(\widehat{\text{T-S}}_{A \cup \{j\},\text{MC}} - \widehat{\text{T-S}}_{A,\text{MC}} \right). \tag{31}$$

Thus, one gets that $\widehat{\text{T-Sh}}_{j,\text{MC}}$ is an unbiased consistent estimator of T-Sh_j.

Algorithm 1 provides a detailed description on how to implement this estimator in practice. This estimation method requires $(N + d! \times (d-1) \times N_v \times N_p)$ calls to the numerical model $G(\cdot)$. Its empirical convergence w.r.t. N_v is illustrated in Appendix E.1. As expected, this first estimation method can become quite expensive in practice. Moreover, numerical models usually encountered in industrial studies can be costly-to-evaluate, which can strongly limit the use of such a method in practice.

Another algorithm has been proposed in Song et al. (2016), by leveraging an equivalent definition of the Shapley allocations, as an arithmetic mean over all the $d!$ permutations of $\{1, \ldots, d\}$. In the same fashion as in Eq. (10), it writes:

$$\widehat{\text{T-Sh}}_{j,\text{MC}} = \frac{1}{m} \sum_{\substack{\pi \in \text{permutations} \\ \text{of } \{1,\ldots,d\}}} \left(\widehat{\text{T-S}}_{v \cup \{j\},\text{MC}} - \widehat{\text{T-S}}_{v,\text{MC}} \right) \tag{32}$$

with v being the indices before j in the order π. In Eq. (32), the sum is not performed over all the permutations of $\{1, \ldots, d\}$ but only on m randomly chosen permutations. By sampling $m < d!$ permutations, one can drive the computational cost of this algorithm to $(N + m \times (d-1) \times N_v \times N_p)$ calls to $G(\cdot)$, for a less precise, but still convergent estimator.

Algorithm 1: Target Shapley effects estimation by a Monte Carlo procedure.

Input: G, t, d, N, N_v, N_p, `simJoint`, `simMarginal`, `simConditional`

Output: $(\widehat{\text{T-Sh}}_{j,\text{MC}})_{j=1,\dots,d}$

/* Sample from the joint distribution */

$(X^{(1)}, \dots, X^{(N)}) \leftarrow$ `sim_joint`(N)

/* Estimate the failure probability */

$\widehat{p}_t^Y \leftarrow \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{\{G(X^{(i)})>t\}}(X^{(i)})$

/* For every subsets of inputs */

for $A \in \mathcal{P}_d$ **do**

  /* Sample from the marginal distribution */

  $(X_A^{(1)}, \dots, X_A^{(N_v)}) \leftarrow$ `simMarginal`(A, N_v)

  /* For every element of the marginal distribution sample */

  for $i = 1, \dots, N_v$ **do**

    /* Sample from the conditional distribution given the element of the marginal distribution */

    $(\widetilde{X}_i^{(1)}, \dots, \widetilde{X}_i^{(N_p)}) \leftarrow$ `simConditional`$(\overline{A}, N_p, X_A^{(i)})$

  /* Compute the conditional element */

  $\widehat{\text{T-S}}_A \leftarrow \frac{1}{N_v - 1} \sum_{i=1}^{N_v} \left(\frac{1}{N_p} \sum_{j=1}^{N_p} \mathbb{1}_{\mathcal{F}_t}(\widetilde{X}_i^{(j)}, X_A^{(i)}) - \widehat{p}_t^Y \right)^2 \times \frac{1}{\widehat{p}_t^Y (1 - \widehat{p}_t^Y)}$

/* Aggregation step */

for $j = 1, \dots, d$ **do**

  $\widehat{\text{T-Sh}}_{j,\text{MC}} \leftarrow 0$

  for $A \subset \{-j\}$ **do**

    /* Apply the Shapley weights to every computed increments */

    $\widehat{\text{T-Sh}}_{j,\text{MC}} += \frac{1}{d} \binom{d-1}{|A|}^{-1} \left(\widehat{\text{T-S}}_{A \cup \{j\}} - \widehat{\text{T-S}}_A \right)$

4.2. Given-data estimation using a nearest-neighbor procedure

A "given-data" estimation method has been introduced by Broto et al. (2020) to estimate the Shapley effects. This method can be seen as an extension of the Monte Carlo estimator when only a single i.i.d. input-output sample is available. This method is appropriate when the input distributions are not known or when the numerical model $G(\cdot)$ is not available anymore. The main idea behind this method is to replace the exact samples from the conditional distributions $P_{X_{\overline{A}}|X_A}$ by approximated ones based on a non-parametric nearest-neighbor procedure.

Let $\left(X^{(1)}, \dots, X^{(N)}\right)$ be an i.i.d. sample of the inputs X and $A \in \mathcal{P}_d \setminus \{\emptyset, \{1:d\}\}$. Let $k_N^A(l, n)$ be the index such that $X_A^{\left(k_N^A(l,n)\right)}$ is the n-th closest element to $X_A^{(l)}$ in $\left(X_A^{(1)}, \dots, X_A^{(N)}\right)$. Note that, if two observations are at an equal distance from $X_A^{(l)}$, then one of the two is uniformly randomly selected. Finally, one can define an estimator of the equivalent cost function defined in Eq. (27):

$$\widehat{\text{T-E}}_{A,\text{KNN}} = \frac{\sum_{l=1}^{N} \left(\frac{1}{N_s - 1} \sum_{i=1}^{N_s} \left[\mathbb{1}_{\mathcal{F}_t}\left(X^{\left(k_N^{\overline{A}}(l,i)\right)} \right) - \frac{1}{N_s} \sum_{h=1}^{N_s} \mathbb{1}_{\mathcal{F}_t}\left(X^{\left(k_N^{\overline{A}}(l,h)\right)} \right) \right]^2 \right)}{N \widehat{p}_t^Y (1 - \widehat{p}_t^Y)}. \tag{33}$$

Under some mild assumptions, Broto et al. (2020) showed that this estimator does asymptotically converge towards T-E$_A$. With estimates for the conditional elements, one can then define the following

plug-in estimator:

$$\widehat{\text{T-Sh}}_{j,\text{KNN}} = \frac{1}{d} \sum_{A \subset \{-j\}} \binom{d-1}{|A|}^{-1} \left(\widehat{\text{T-E}}_{A\cup\{j\},\text{KNN}} - \widehat{\text{T-E}}_{A,\text{KNN}} \right) \tag{34}$$

where $\widehat{p}_t^Y$ is the empirical mean of $\mathbb{1}_{\mathcal{F}_t}(X)$ on the i.i.d. sample. Algorithm 2 represents the procedure for this given-data estimator. Its empirical convergence w.r.t. the sample size is illustrated in Appendix E.2. This method is less computationally expensive (in terms of model evaluations) compared to the Monte Carlo sampling-based method, since no additional model evaluation, other than the ones in the i.i.d. sample, is required in order to produce estimates of the target Shapley effects. Since the samples of the conditional and marginal distributions are approximated by a non-parametric procedure, this method also reduces the possible input modeling error (e.g., in the context of ill-defined input distributions), at the cost of less accurate estimates. Another constraint is due to the fact that the input-output sample has to be i.i.d. which prevents it from being used, for instance, in advanced orthogonal designs of computer experiments.

Algorithm 2: Target Shapley effects estimation by a nearest-neighbor procedure.

Input: X, Y, t

Output: $(\widehat{\text{T-Sh}}_{j,\text{KNN}})_{j=1,\dots,d}$

/* Estimate the failure probability */

$\widehat{p}_t^Y \leftarrow \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{\{G(X^{(i)})>t\}}(X^{(i)})$

/* For every subsets of inputs */

for $A \in \mathcal{P}_d$ **do**

  /* Sample of X_A */

  $X_A \leftarrow (X_i^{(j)})_{j=1,\dots,n}^{i\in A}$

  for $i = 1, \dots, N$ **do**

    /* For each row i of X_A, find the N_s nearest rows in X */

    $(\widetilde{X}_i^{A,(j)})_{j=1,\dots,N_s} \leftarrow \texttt{KNN}(X_A^{(i)}, X, N_s)$

  /* Compute the conditional element */

  $\widehat{\text{T-E}}_A \leftarrow \sum_{l=1}^N \left(\frac{1}{N_s-1} \sum_{i=1}^{N_s} \left[\mathbb{1}_{\mathcal{F}_t}\left(\widetilde{X}_l^{A,(i)}\right) - \frac{1}{N_s}\sum_{h=1}^{N_s} \mathbb{1}_{\mathcal{F}_t}\left(\widetilde{X}_l^{A,(h)}\right) \right]^2 \right) \times \left(N \widehat{p}_t^Y (1-\widehat{p}_t^Y)\right)^{-1}$

/* Aggregation step */

for $j = 1, \dots, d$ **do**

  $\widehat{\text{T-Sh}}_{j,\text{MC}} \leftarrow 0$

  for $A \subset \{-j\}$ **do**

    /* Apply the Shapley weights to every computed increments */

    $\widehat{\text{T-Sh}}_{j,\text{MC}} += \frac{1}{d}\binom{d-1}{|A|}^{-1} \left(\widehat{\text{T-E}}_{A\cup\{j\}} - \widehat{\text{T-E}}_A \right)$

In Broto et al. (2020), a random permutation algorithm, homologous to Eq. (32), has been developed, which allows for reducing the overall complexity of the method, which, for the sake of conciseness, is not developed in this paper.

4.3. Software and reproducibility of results

The algorithms described in the preceding subsections have been implemented in the `sensitivity` R package (Iooss et al., 2021). More precisely, the `shapleyPermEx()` (sampling-based algorithm) and

`sobolshap_knn()` (given-data algorithm) functions can be directly used for the estimation of the target Shapley effects. In the applications of Section 6, only the `sobolshap_knn()` function is used for numerical tractability. Appendix D provides some minimal code examples for the implementation of the Monte Carlo (see, Section 4.1) and nearest neighbors estimation procedure (see, Section 4.2), along with their random permutation variants.

All further results can be accessed on a GitLab[1] repository, along with the data used in the following sections. R code files are available, with explicit code, along with all custom-made functions, in order to reproduce the results presented in this paper. The procedures for the theoretical approximations of Section 5 are made available, along with the data-simulation functions for the flood case in Subsection 6.1. The two datasets used for Subsection 6.2 are also available. Finally, all the figures can be reproduced by simply re-running the different `RMarkdown` files in the aforementioned GitLab repository.

5. Analytical results using a linear model with Gaussian inputs

To illustrate the behavior of the proposed indices, a first toy-case involving a linear model and multivariate Gaussian inputs is presented. In this setting, analytical results can be derived for the marginal distributions of all subsets of inputs, their conditional distribution, and the distribution of the output given a subset of inputs. Subsequently, analytical formulas can be obtained for both the target Sobol' indices and the target Shapley effects.

Let $(\beta_0, \beta) = (\beta_0, \beta_1, \ldots, \beta_d)^\top \in \mathbb{R}^{d+1}$, $\mu = (\mu_1, \ldots, \mu_d)^\top \in \mathbb{R}^d$ and $\Sigma \in \mathcal{M}_d(\mathbb{R})$ a full-rank symmetric $(d \times d)$ matrix. Assume that $X \sim \mathcal{N}_d(\mu, \Sigma)$, and that the model output writes

$$Y = \beta_0 + \beta^\top X. \tag{35}$$

Then, one has $Y \sim \mathcal{N}\left(\beta_0 + \beta^\top \mu, \beta^\top \Sigma \beta\right)$ and, for any $A \in \mathcal{P}_d$, $(Y|X_A = x_A) \sim \mathcal{N}\left(\widetilde{\mu}_A, \widetilde{\Sigma}_A\right)$ with

$$\widetilde{\mu}_A = \beta_0 + \beta_A^\top x_a + \beta_{\overline{A}}^\top(\mu_{\overline{A}} + \Sigma_{A,12}\Sigma_{A,22}^{-1}(x_a - \mu_A)), \quad \widetilde{\Sigma}_A = \beta_{\overline{A}}^\top(\Sigma_{A,11} - \Sigma_{A,12}\Sigma_{A,22}^{-1}\Sigma_{A,21})\beta_{\overline{A}}.$$

Moreover, one also can recall that

$$(X_{\overline{A}}, X_A)^\top \sim \mathcal{N}_d\left(\begin{pmatrix}\mu_{\overline{A}} \\ \mu_A\end{pmatrix}, \Sigma_A = \begin{pmatrix}\Sigma_{A,11} & \Sigma_{A,12} \\ \Sigma_{A,21} & \Sigma_{A,22}\end{pmatrix}\right)$$

with the partitions of Σ_A having sizes $\begin{pmatrix}(d-|A|)\times(d-|A|) & (d-|A|)\times|A| \\ |A|\times(d-|A|) & |A|\times|A|\end{pmatrix}$. Evaluating these results requires some numerical approximations of the theoretical values of T-Sh$_j$ for all $j = 1, \ldots, d$. This has been achieved by using standard multidimensional integration tools, and more specifically, the function `adaptIntegrate()` from the `cubature` package of the R software has been used, with a fixed error tolerance set to 10^{-8}. This allowed the study of simple toy-cases in order to validate the behavior of the target Shapley effects.

[1] https://gitlab.com/milidris/review_l2tse

In the following, the inputs are first assumed to be independent, and are studied w.r.t. the threshold t. Then, a toy-case involving linear correlation between inputs (driven by a parameter ρ) is studied. Finally, a last toy-case aims at studying the proposed indices' behavior in the presence of an exogenous input.

5.1. Independent standard Gaussian inputs

The first toy-case can be specified by:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \right), \quad Y = \sum_{i=1}^{3} X_i. \tag{36}$$

In this case, the three inputs are equally important in terms of defining Y, but they should also be equally important for the variable of interest $\mathbb{1}_{\mathcal{F}_t}(X)$, as assessed by the target Sobol' indices defined as in Eq. (18).

From Li et al. (2012) and Lemaitre (2014), one can easily deduce that the first-order (FO) target closed Sobol' indices are all equal to each other. Thus, one has:

$$\text{T-S}_{\text{FO}} \overset{\text{def}}{=} \text{T-S}_1^{\ell^2} = \text{T-S}_2^{\ell^2} = \text{T-S}_3^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(\frac{t-X}{\sqrt{2}}\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)}, \tag{37}$$

while the second-order (SO) target closed Sobol' indices are given by:

$$\text{T-S}_{\text{SO}} \overset{\text{def}}{=} \text{T-S}_{\{1,2\}}^{\ell^2} = \text{T-S}_{\{1,3\}}^{\ell^2} = \text{T-S}_{\{2,3\}}^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(t - X'\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)} \tag{38}$$

where $\Phi(.)$ is the standard Gaussian cumulative distribution function (cdf), $X \sim \mathcal{N}(0,1)$ and $X' \sim \mathcal{N}(0,2)$. Finally, one can also show that the third-order (TO) target closed Sobol' indices are equal to:

$$\text{T-S}_{\text{TO}} \overset{\text{def}}{=} \text{T-S}_{\{1,2,3\}}^{\ell^2} = 1. \tag{39}$$

From Eqs. (37), (38), and (39), and from Property 1, one can deduce that:

$$\text{T-Sh}_1 = \text{T-Sh}_2 = \text{T-Sh}_3 = \frac{1}{3}. \tag{40}$$

Additionally, as the inputs are independent, interpreting the original target Sobol' indices (i.e., Eq. (17)) is meaningful, and they are equal to:

$$\text{T-S}_i = \text{T-S}_{\text{FO}}, \forall i \in \{1,2,3\} \tag{41}$$

$$\text{T-S}_{\{i,j\}} = \text{T-S}_{\text{SO}} - 2\text{T-S}_{\text{FO}}, \forall i,j \in \{1,2,3\}, i \neq j \tag{42}$$

$$\text{T-S}_{\{1,2,3\}} = \text{T-S}_{\text{SO}} - 3\left(\text{T-S}_{\text{FO}} + \text{T-S}_{\text{SO}}\right). \tag{43}$$

The target Sobol' indices are illustrated in Figure 3 (right). One can remark that, focusing on the indicator variable of interest $\mathbb{1}_{\mathcal{F}_t}(X)$ instead of the model output Y leads to interaction effects between

the inputs, as outlined in Section 3.1. The target Shapley effects, however, remain constant for all threshold values t. Such a behavior is expected: it highlights the fact the target Shapley effects do not report the interaction effects as the target Sobol' indices would. The proposed indices rather summarize (in the sense of the Shapley values allocation) the target Sobol' indices into a single index. Their goal is not to report on the "types of effects" (i.e., correlation or interaction), but rather provide a global index which sums up each input's importance.

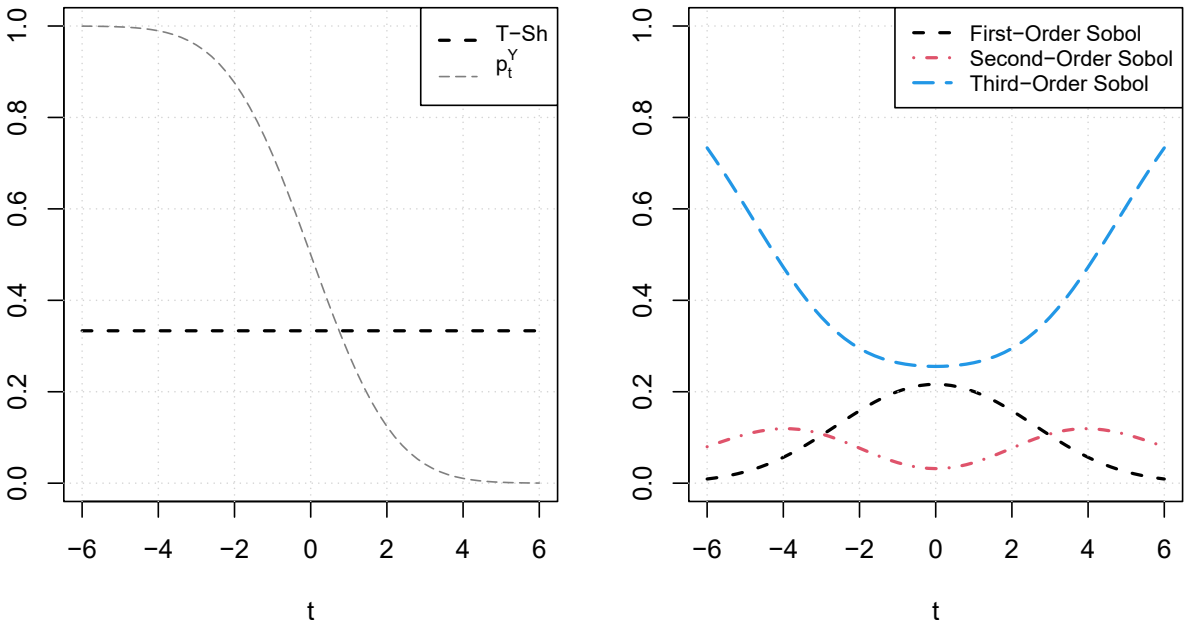

Figure 3: Target Shapley effects (left) and target Sobol' indices (right) for the linear model with standard independent multivariate Gaussian inputs, w.r.t. t.

5.2. Correlated Gaussian inputs with unit variance

The behavior of the target Shapley effects are now studied when a linear dependence is added to the inputs. Since Property 1 still holds without any condition on the dependence structure on the input variables, these indices remain interpretable as a percentage of the output's variance. The following model is studied:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho \\ 0 & \rho & 1 \end{pmatrix} \right), \quad Y = \sum_{i=1}^{3} X_i. \tag{44}$$

where $-1 < \rho < 1$. In this scenario one has:

$$\text{T-S}_1^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(\frac{t-X}{\sqrt{2(1+\rho)}}\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)}, \tag{45}$$

$$\text{T-S}_2^{\ell^2} = \text{T-S}_3^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(\frac{t-X(1+\rho)}{\sqrt{2-\rho^2}}\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)}, \tag{46}$$

$$\text{T-S}_{\{1,2\}}^{\ell^2} = \text{T-S}_{\{1,3\}}^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(\frac{t-X'}{\sqrt{1-\rho^2}}\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)}, \tag{47}$$

$$\text{T-S}_{\{2,3\}}^{\ell^2} = \frac{\mathbb{V}\left(\Phi\left(t-X''\right)\right)}{\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right)} \tag{48}$$

where $X \sim \mathcal{N}(0,1)$, $X' \sim \mathcal{N}(0, 1+(1+\rho)^2)$ and $X'' \sim \mathcal{N}\left(0, 2(1+\rho)\right)$.

From these results, one can directly remark that $\text{T-Sh}_2 = \text{T-Sh}_3$. Note that the values of the target Shapley effects can also be obtained by combinations of target Sobol' indices (see Eq. (26)). These results are illustrated in Figure 4. For fixed threshold values t, the target Shapley effects of the correlated inputs X_2 and X_3 increases when ρ increases. This is an expected behavior since, in this case:

$$\mathbb{V}\left(\mathbb{1}_{\mathcal{F}_t}(X)\right) = \Phi\left(\frac{t}{\sqrt{3+2\rho}}\right)\left(1-\Phi\left(\frac{t}{\sqrt{3+2\rho}}\right)\right), \tag{49}$$

and subsequently, for a fixed t, the variance of the variable of interest will grow with ρ, as illustrated in Figure 5. This increase in variance due to the correlation between X_2 and X_3 is then attributed through T-Sh_2 and T-Sh_3, which increase with ρ. On the other hand, T-Sh_1 decreases accordingly, to accommodate Property 1.

In Figure 4, the behavior of the indices w.r.t. ρ is illustrated. T-Sh_1 is predominantly above T-Sh_2 and T-Sh_3 when ρ is negative, and below when it is positive. This can be explained by the fact that, X_2 and X_3 cancel each other out when their correlation is negative, thus lowering the value of $\text{T-S}_{\{2,3\}}$ below $\text{T-S}_{\{1,2\}}$ and $\text{T-S}_{\{1,3\}}$, automatically increasing T-Sh_1 in accordance to Property 1. On the other hand, for positive values of ρ, $\text{T-S}_{\{2,3\}}$ is higher than $\text{T-S}_{\{1,2\}}$ and $\text{T-S}_{\{1,3\}}$, which in turn corresponds to T-Sh_1 being lower than $\text{T-Sh}_2 = \text{T-Sh}_3$.

5.3. Quantifying the importance of an exogenous input in the Gaussian setting

In this toy-case, inspired by Lemaitre (2014), the following model is considered:

$$\begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \end{pmatrix} \sim \mathcal{N}_4\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & \rho \\ 0 & 0 & 1 & 0 \\ 0 & \rho & 0 & 1 \end{pmatrix}\right), \quad Y = X_1 + 6X_2 + 4X_3 \tag{50}$$

where X_4 is an exogenous input, but correlated to X_2, which is the most important variable in terms of variance contribution, due to its higher linear coefficient. The threshold is fixed at $t = 16$. This scenario allows for the verification of how the target Shapley effects attribute the importance of X_4

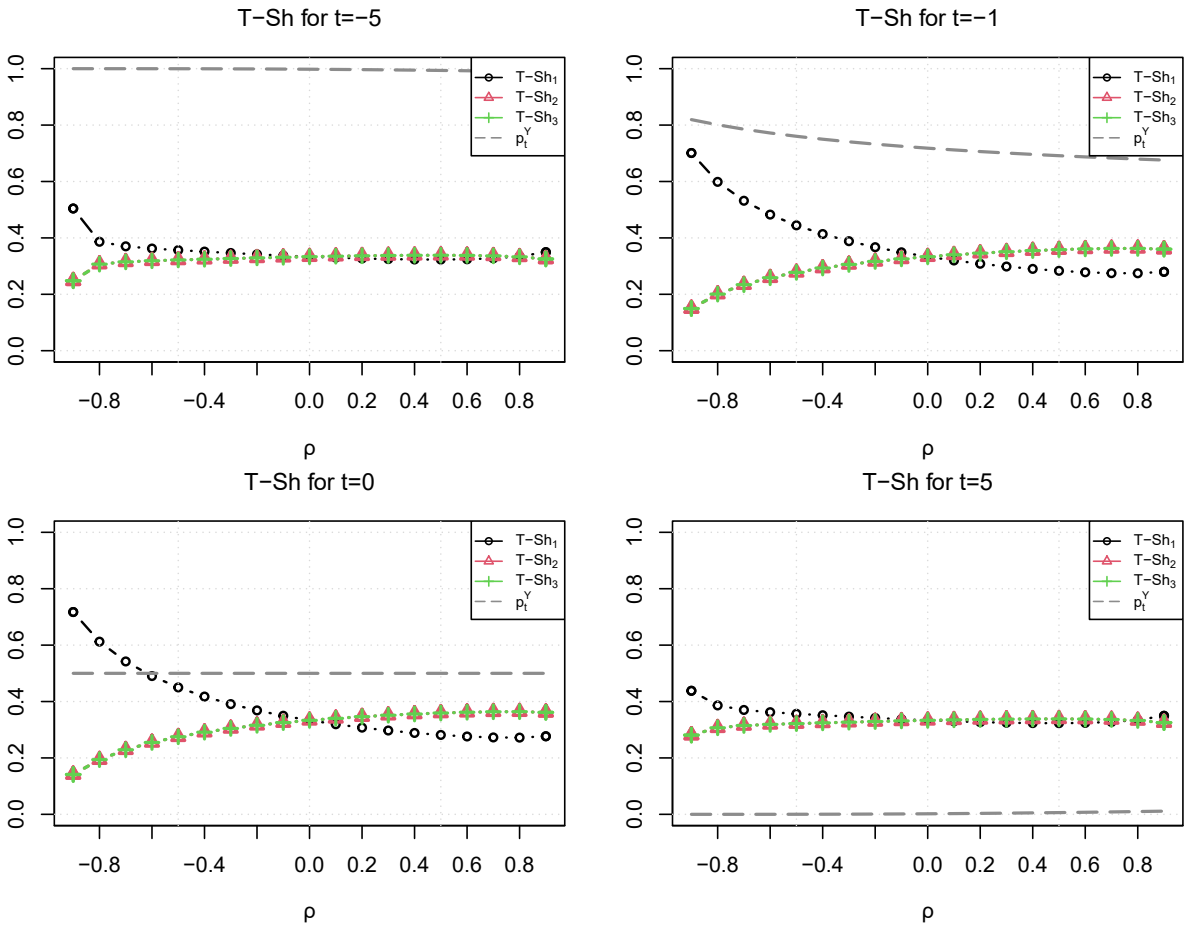

Figure 4: Evolution, w.r.t. ρ and for various threshold values, of the target Shapley effects of correlated Gaussian standard inputs in a linear model.

which is correlated with an endogenous input, even though it does not appear in the model. In the results given in Figure 6, one can remark that T-Sh$_4$ increases when ρ approaches either 1 or -1, despite the fact that it has no direct causal effect on the model $G(\cdot)$.

5.4. Discussion on causal relationship assessment

As outlined in Subsection 5.3, one can remark that, in the context of highly correlated inputs, the proposed indices fail to provide insights on the causal relationships of an input on the model's output: an exogenous input may receive a share of the output's variance. This behavior is intrinsically due to the Shapley values allocation method, and has been highlighted in Iooss and Prieur (2019) in the case of the Shapley effects. This particularity makes the interpretation of the Shapley effects, and subsequently the target Shapley effects, quite delicate. A prior investigation of the correlation structure of the data, through, for instance, estimated correlation matrices or any tool dedicated to multicollinearity diagnostics such as the variance inflation factor (Fox and Monette, 1992) is strongly advised, along with an input validation process, ensuring the absence of exogenous inputs in the TSA study. Other allocations systems, such as the *proportional values* (Ortmann, 2000), are purposefully

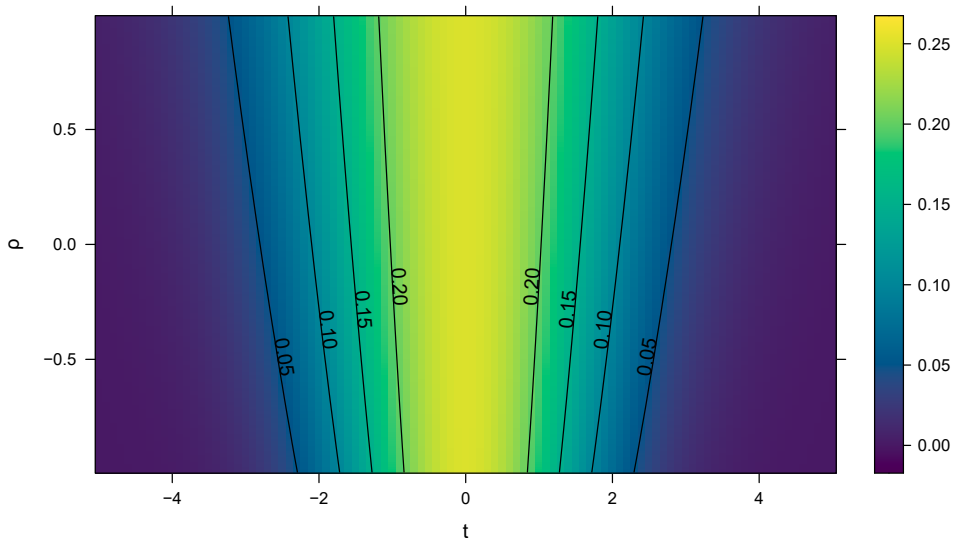

Figure 5: Variance of $\mathbb{1}_{\mathcal{F}_t}(X)$ w.r.t. ρ and t for correlated Gaussian standard inputs with a linear model.

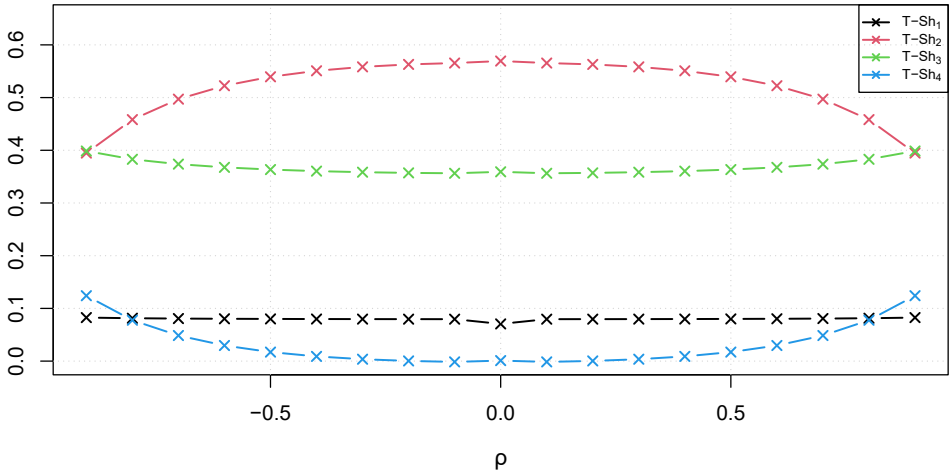

Figure 6: Target Shapley effects for the Gaussian Linear model with an exogenous input, w.r.t. ρ.

designed in order to highlight causal effects (Feldman, 2005), but fall out of the scope of this document and are not developed further.

6. Applications

In this section, two models related to real-world phenomena which include dependent random inputs are studied in the context of TSA.

6.1. A simplified flood model

The target Shapley effects are firstly computed on a simplified model of a river flood (Lemaitre, 2014; Iooss and Lemaître, 2015). This model's goal is to simulate the behavior of a river's water level,

and to compare it to a fixed dyke height. After a strong simplification of the one-dimensional Saint-Venant equation (with uniform and constant flow rate), the maximal annual water level h is modeled as:

$$h = \left(\frac{Q}{BK_s \sqrt{\frac{Z_m - Z_v}{L}}} \right)^{\frac{3}{5}}, \tag{51}$$

while the model output writes:

$$Y = Z_v + h. \tag{52}$$

The six inputs' probabilistic structure is described in Table 1. The problem is of dimension $d = 6$. Under the TSA paradigm, the variable of interest is $\mathbb{1}_{\{G(X)>t\}}(X)$ with t representing the dyke's height, fixed to $t = 54.5$ m. The reference failure probability (see, Eq. (16)), computed here with a Monte Carlo sample of large size (here 10^7 samples) is equal to $p_t^Y = 4.5 \times 10^{-3}$.

Input	Description	Unit	Distribution
Q	maximal annual flow rate	$\mathrm{m}^3.\mathrm{s}^{-1}$	Gumbel$(1013, 558)$ truncated to $[500, 3000]$
K_s	Strickler friction coefficient	-	Normal$(30, 7)$ truncated to $[15, +\infty)$
Z_v	river downstream level	m	Triangular$(49, 50, 51)$
Z_m	river upstream level	m	Triangular$(54, 55, 56)$
L	length of the river stretch	m	Triangular$(4990, 5000, 5010)$
B	river width	m	Triangular$(295, 300, 305)$
t	dyke height (threshold)	m	Fixed to 54.5

Table 1: Input variables and distributions for the flood model.

In the same fashion as in Chastaing et al. (2012), three pairs of inputs are assumed to be linearly dependent: Q and K_s with $\rho(Q, K_s) = 0.5$, Z_v and Z_m with $\rho(Z_v, Z_m) = 0.3$, L and B with $\rho(L, B) = 0.3$. The aim of this use-case is to assess the relevance of the target Shapley effects in a complex environment. In Chastaing et al. (2012), it is shown that, from a GSA standpoint (using a generalized variance decomposition for dependent variables), the two most influential inputs on the annual water level are Q, the maximal annual flow rate, and Z_v, the river downstream level.

An i.i.d. sample of $N = 2 \times 10^5$ input realizations is drawn (note that the linear correlations are injected following the algorithm proposed by Schumann (2009)) which leads to N model evaluations. Figure 7 presents the estimated target Shapley effects on this i.i.d. sample, using the nearest-neighbor procedure depicted in Subsection 4.2 with an arbitrary number of neighbors set at $N_s = 2$. 300 repetitions of the simulation and the estimation procedure allow for the assessment of the estimation procedure's variance (represented by boxplots in Figure 7). One can notice that Q is granted an influence of 24.3% ($\pm$1.3%), K_s has 22.6% ($\pm$1.3%) and Z_v around 16.7% ($\pm$1%). The other inputs are attributed a share of around 12%. Compared to results obtained by from a GSA standpoint, without correlations (Iooss and Lemaître, 2015) and with correlations (Chastaing et al., 2012), these TSA results allow for granting a much larger share to K_s and non-negligible effects to Z_m, L and B. This was expected due to the interactions induced by the considered TSA variable of interest. This example illustrates the ability of the target Shapley effects to quantify the importance of input variables in a use-case involving input correlation.

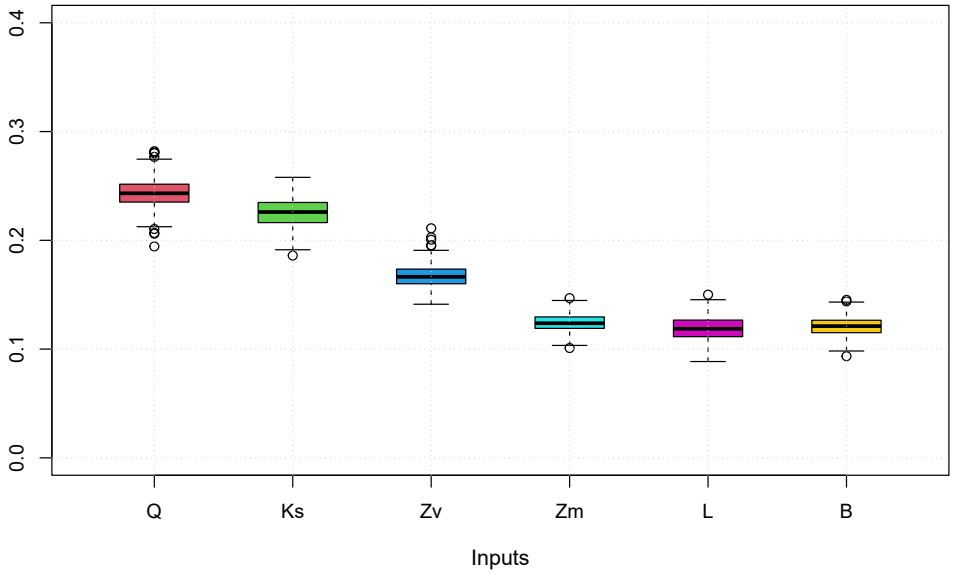

Figure 7: Estimated target Shapley effects for the flood case.

6.2. A COVID-19 epidemiological model

In 2020, the COVID-19 pandemic has raised important questions on the usefulness of epidemic modeling, especially on their ability to produce relevant insights to public policy decision makers. Saltelli et al. (2020) have taken this example to insist on the essential use of GSA on such models, which claim to predict the potential consequences of intervention policies. A first study has been proposed by Lu and Borgonovo (2020), in the context of COVID-19 in Italy, to assess the sensitivity of important epidemiological model outcomes, such as the number of people being either quarantined, recovering, or dead due to COVID-19. Another GSA has been performed in Da Veiga et al. (2021) in the French context of the first COVID-19 outbreak. By using data coming from this last analysis (thanks to the authors' agreement), the goal of this section is to demonstrate how TSA can help to characterize the influence of various uncertain parameters on a real-scale model.

6.2.1. Description of the model and its inputs

The deterministic compartmental model developed in Da Veiga et al. (2021) (also presented in Da Veiga (2020)) is representative of the COVID-19 French epidemic (from March to May) by taking into account the asymptomatic individuals, the testing strategies, the hospitalized individuals, and people admitted to Intensive Care Unit (ICU). Using several assumptions, it is based on a system of 10 ordinary differential equations that can be fully retrieved in references Da Veiga (2020) and Da Veiga et al. (2021). Each equation models path of individuals between different compartments (corresponding to their infectious and illness states), as shown in Figure 8. These equations involve many input parameters and model the dynamic between the different compartments. Table 2 presents the 20 continuous input parameters with their prior distribution (chosen from literature studies), which form the inputs X, assumed to be independent between each other.

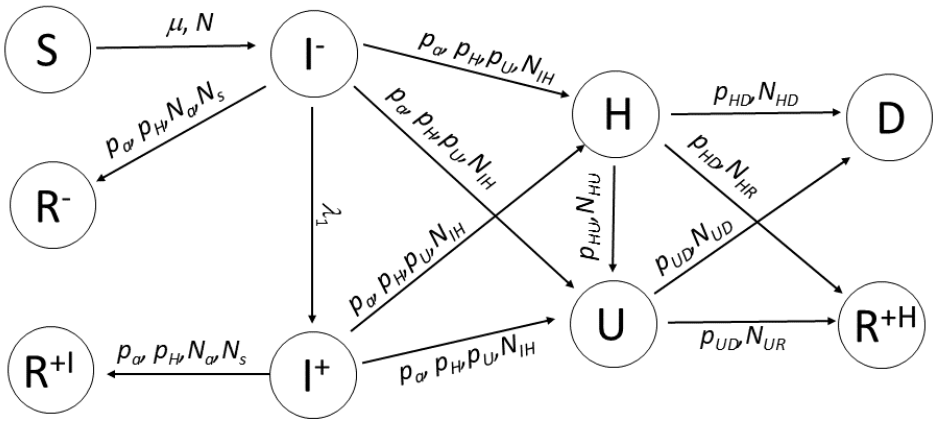

Figure 8: Overview of the COVID-19 compartmental model. The different compartments contain the following individuals: susceptible (S), undetected infected (I^-), unhospitalized detected infected (I^+), recovered undetected (R^-), recovered unhospitalized detected (R^{+I}), hospitalized in severe illness (H), hospitalized in intensive care (U), recovered hospitalized (R^{+H}) and dead (D).

Input	Description	Prior distribution
p_a	Conditioned on being infected, the probability of having mild symptoms or no symptoms	$\mathcal{U}(0.5, 0.9)$
p_H	Conditioned on being mild/severely ill, the probability of needing hospitalization (H or U)	$\mathcal{U}(0.15, 0.2)$
p_U	Conditioned on hospitalisation, the probability of being admitted to ICU	$\mathcal{U}(0.15, 0.2)$
p_{HD}	Conditioned on being hospitalized but not in ICU, the probability of dying	$\mathcal{U}(0.15, 0.25)$
p_{UD}	Conditioned on being admitted to ICU, the probability of dying	$\mathcal{U}(0.2, 0.3)$
N_a	If asymptomatic, number of days until recovery	$\mathcal{U}(8, 12)$
N_s	If symptomatic, number of days until recovery without hospitalisation	$\mathcal{U}(8, 12)$
N_{IH}	If severe symptomatic, number of days until hospitalization	$\mathcal{U}(8, 12)$
N_{HD}	If in H, number of days until death	$\mathcal{U}(15, 20)$
N_{UD}	If in ICU, number of days until death	$\mathcal{U}(8, 12)$
N_{HR}	If hospitalized but not in ICU, the number of days until recovery	$\mathcal{U}(15, 25)$
N_{UR}	If in ICU, number of days until recovery	$\mathcal{U}(15, 25)$
R_0	Basic reproduction number	$\mathcal{U}(3, 3.5)$
t_0	Starting date of epidemics (in 2020)	$\mathcal{U}(01/25, 02/24)$
μ	Decaying rate for transmission (after social distanciation and lockdown)	$\mathcal{U}(0.03, 0.08)$
N	Date of effect of social distanciation and lockdown	$\mathcal{U}(20, 50)$
λ_1	Type-1 testing rate	$\mathcal{U}(1e-4, 1e-3)$
p_{HU}	Conditioned on being hospitalized in H, the probability of being admitted to ICU	$\mathcal{U}(0.15, 0.2)$
N_{HU}	If in H, number of days until ICU	$\mathcal{U}(1, 10)$
I_0^-	Number of infected undetected at the start of epidemics	$\mathcal{U}(1, 100)$

Table 2: Model inputs and their prior distribution. H is the number of hospitalized individuals with severe symptoms. U is the number of hospitalized individuals in ICU.

For the present study, our variable of interest, which is a particular model output, then writes

$$U_{\max}^{\mathrm{p}} = \max_{v \in \text{time range}} \Big\{ U_v(X) \Big\} \tag{53}$$

where U_v is the the number of hospitalized patients in ICU at time v. Note that the "p" in $U_{\max}^{\mathrm{p}}$ stands for "prior" as this quantity corresponds to the variable of interest before any calibration w.r.t. the available data.

In Da Veiga et al. (2021), after a first screening step which allows for suppressing non-influential inputs, the model is calibrated on real data by using a Bayesian calibration technique. After the analysis of this step, the selected remaining inputs are

$$X_{\mathrm{sel}} = (p_a, N_a, N_s, R_0, t_0, \mu, N, I_0^-)^\top \tag{54}$$

and their distributions are obtained from a sample given by the calibration process. The non-influential inputs are fixed to their nominal values and the posterior variable of interest becomes

$$U_{\max} = \max_{v \in \text{time range}} \Big\{ U_v(X_{\mathrm{sel}}) \Big\} \tag{55}$$

with $U_{\max}$ being the maximum number of hospitalized people in ICU who need special medical care on the studied temporal range, and U_v is the number of hospitalized patients in ICU at time v.

6.2.2. Input importance for ICU bed shortage

The central question of this study would be to determine which inputs influence the event of a country experiencing a shortage of ICU bed capacity during the time period. For that purpose, one can introduce a threshold k, which represents the total number of ICU beds in the country, which is assumed to be constant during the studied time period. The new variable of interest would then be $\mathbb{1}_{\{U_{\max}^{\mathrm{p}} > k\}}(X)$ for the full compartmental model (preliminary study) and $\mathbb{1}_{\{U_{\max} > k\}}(X_{\mathrm{sel}})$ for the model with selected inputs (post-calibration study). Two input-output samples of size $n = 5000$ are available. The first one (preliminary study) includes all the inputs following their prior distribution (see Table 2) and the corresponding output $U_{\max}^P$ of the compartmental model. The second one (post-calibration study) is composed of a sample of X_{sel} after the Bayesian calibration, and the corresponding output $U_{\max}$ of the compartmental model with the non-selected inputs fixed to their nominal values.

Five different thresholds are studied on $U_{\max}^P$: $5 \cdot 10^3$, 10^4, $5 \cdot 10^4$, 10^5 and $2 \cdot 10^5$, with respectively 58.1%, 47.7%, 22%, 10.1% and 2.2% of the total output samples being in a failure state. This illustrates the behavior of the target Shapley effects when the failure probability decreases. The threshold of 6300 has been chosen for $U_{\max}$, with 10.9% of the total output samples being above this threshold. Figure 9 illustrates two different thresholds, and the corresponding estimated failure probability on the histogram of both outputs.

The target Shapley effects have been estimated using a variant of the estimation scheme presented in Subsection 4.2, with a fixed number of random permutations of 10^3, and with a number of neighbors set to 3, following the rule of thumb guideline of Broto et al. (2020), due to the sheer complexity of this estimation algorithm. Since the compartmental model is deterministic, the target Shapley effects have been forced to sum up to one. Figure 10 presents the main results for $U_{\max}^P$, with the red dotted

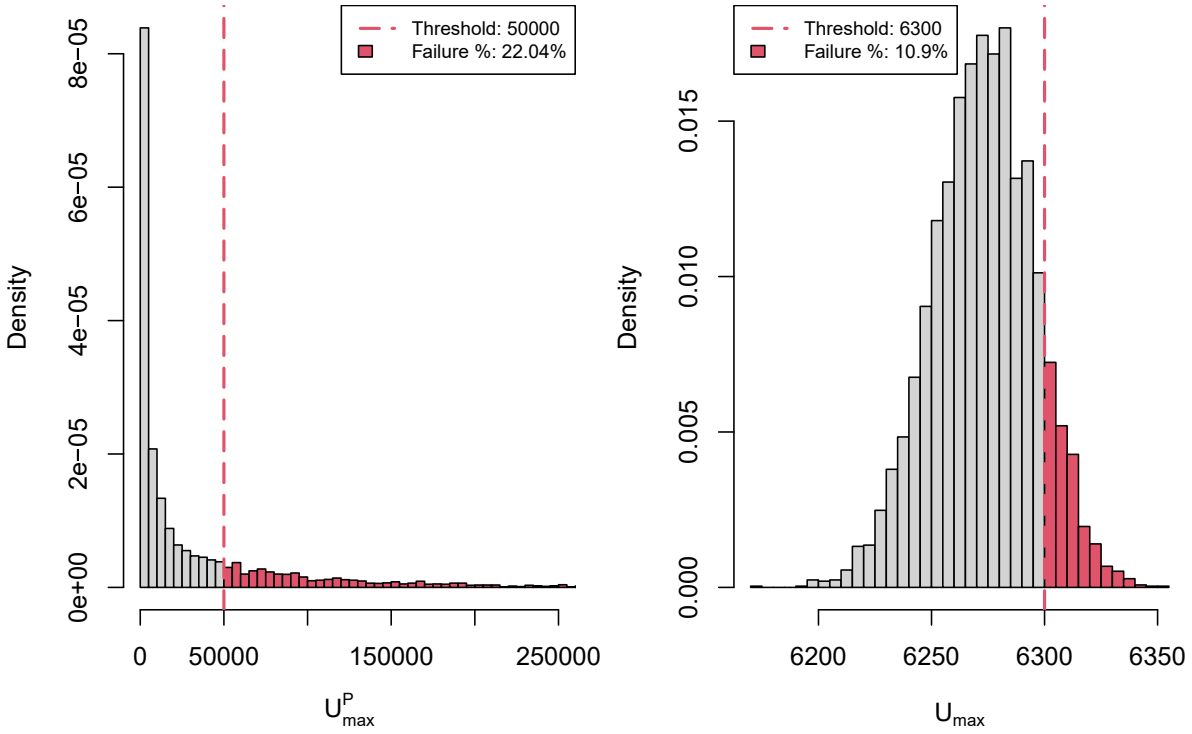

Figure 9: Illustration of thresholds on the histograms of $U^P_{\max}$ (left) and $U_{\max}$ (right).

line being the average influence of an input, in the case of similar importance (i.e., $\frac{1}{20}$). One can remark that for less restrictive thresholds (i.e., threshold for which the failure probability is high), the input N, the effective date of lockdown/social distanciation measures, seem to be the most influential, reaching more than 50% of the TSA variable of interest's variance. However, as soon as the threshold becomes more and more restrictive (i.e., the failure probability decomes lower and lower), the effect of N decreases, and the effects of the other inputs increase accordingly, in order to reach what seem to be an equilibrium at the value $\frac{1}{20}$. This behavior can be explained by two main reasons:

- As outlined in Subsection 3.1, the nature of a restrictive TSA variable of interest induces high interaction between the inputs;
- The Shapley allocation system, when applied to variance as a production value, redistributes the interaction effects equally between all inputs (there is no correlation between inputs in this prior study).

One can argue that, as soon as k becomes very restrictive, the combined interaction effects outweighs the effect of N itself, and since these effects are equally distributed among all the inputs, their share will tend to go towards $\frac{1}{20}$.

For the post-calibration study, some selected inputs X_{sel} are linearly correlated (see Figure 11 - top). This is typically the case for N and μ, with an estimated correlation coefficient $\hat{\rho}(N, \mu) = 0.69$, and for R_0 and N with an estimated correlation coefficient of $\hat{\rho}(N, R_0) = -0.66$. This correlation structure does not allow for interpretable Sobol' indices, as outlined in Section 2, which encourages

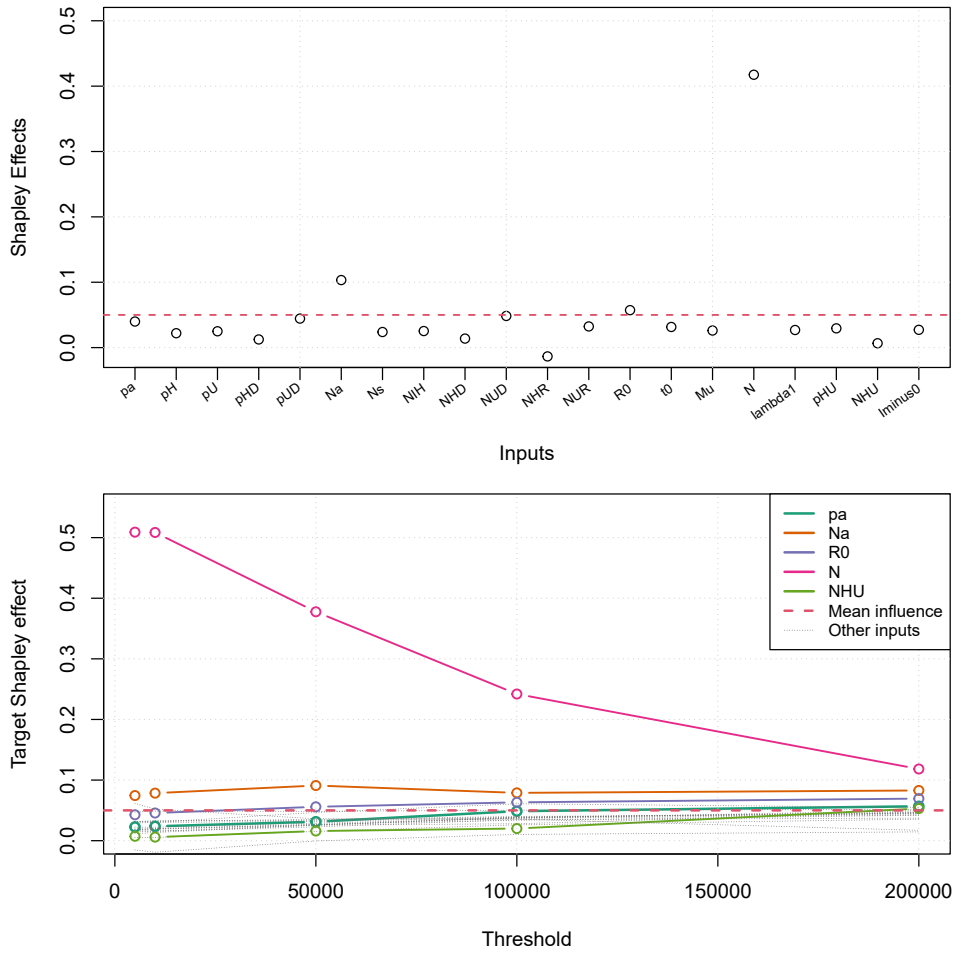

Figure 10: Shapley effects (top) and target Shapley effects for different thresholds (bottom) for $U^P_{\max}$.

the use of Shapley-inspired indices. The Shapley effects and the target Shapley effects of X_{sel} for $U_{\max}$ have been computed using the nearest-neighbor procedure, with a fixed number of neighbors of 3, and forced to sum to one because of the deterministic nature of the model.

In Figure 11 (bottom), one can remark that N_a, the number of days until recovery, seem to be the most important input in explaining the number maximum number of ICU patients on the studied time range, with a Shapley effect of around 35% of the output variance. The inputs p_a, N_s, R_0 and N seem to present average effects, that is around $\frac{1}{8}$, while t_0, μ and I_0^- seem to be less influential, with around 5% of explained variance each.

However, focusing on the occurrence of a ICU bed shortage, one can remark that the target Shapley effect of N_a is lower (around 22%), with the influence of N being higher (around 15%) than their Shapley effects. Moreover, t_0, μ and I_0^- present higher TSA effects, i.e., slightly under 10%, due to the interaction induced by the indicator function. One can also remark that the influence of N_s is higher than that of R_0 in the TSA setting, which was the inverse for the Shapley effects. This would indicate that N_s, the number of days until recovery for a symptomatic patient without hospitalization, has more influence on the event of a bed shortage than the basic reproducing number of the virus, R_0.

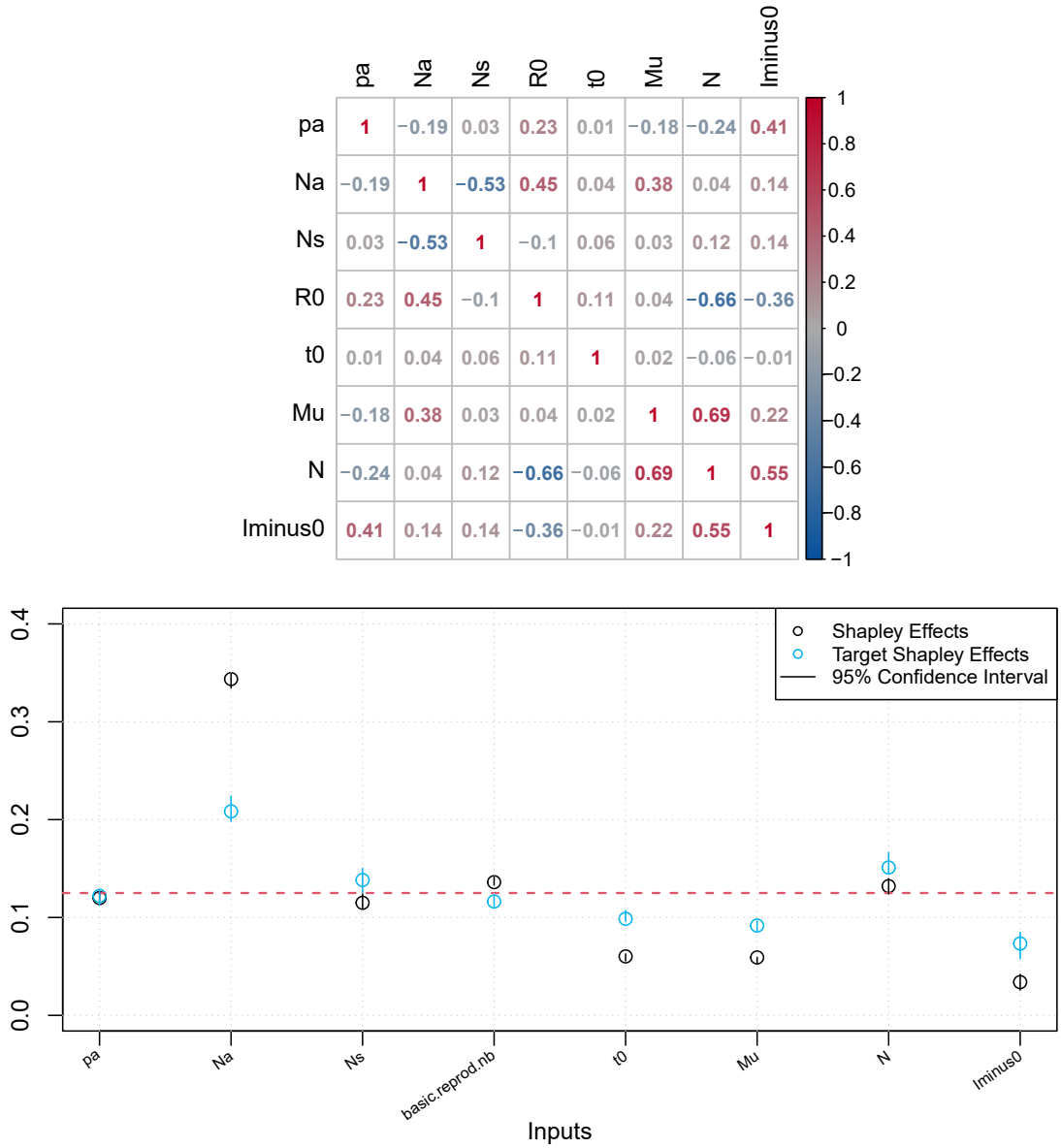

Figure 11: Input correlation matrix (top), Shapley effects for $U_{\max}$ and target Shapley effects (bottom) for $\mathbb{1}_{\{U_{\max}>t\}}(X_{\text{sel}})$. The 95% confidence intervals have been computed by uniformly selecting 80% of the observations, for 100 repetitions, without replacement.

7. Conclusion

This paper proposes a set of novel indices adapted to target sensitivity analysis while being able to handle correlated inputs. The objective is to quantify the importance of inputs on the occurrence of a critical failure event of the system under study. The proposed indices are based on a cooperative Shapley procedure which aims at allocating the effects of the interaction and correlation equally between all the inputs in the same manner as the Shapley effects in global sensitivity analysis. Thus, a general class of distance-based indices is proposed, namely the $(\mathcal{D})$-target Shapley effects and some

relevant properties are highlighted. Depending on the choice of the distance $\mathcal{D}$, well-known preexisting indices can be used as cost functions in the Shapley formulation. Therefore, these indices allow for the allocation, among the different inputs, of shares of several dispersion statistics (e.g., mean absolute deviation for the ℓ^1 case, variance for the ℓ^2 case). These indices are easily usable in practice, as they can be interpreted as percentages of the dispersion statistic, allocated to each input. This versatile procedure produces input importance measures according to a specific metric, driven by the choice of the distance.

In particular, the (ℓ^2)-target Shapley effects (called target Shapley effects to simplify), which represents percentages of variance, have been studied more extensively and two dedicated estimation methods have been proposed. These particular indices have then been applied, analyzed and discussed through simple Gaussian toy-cases. Finally, two real-world use-cases have been studied: the modeling of a river flood and the ICU bed shortage during the COVID-19 pandemic. These indices are revealed to be able to detect influential inputs in the context of correlated inputs. For target sensitivity analysis, such a tool is valuable and can be used as a complement of more standard procedures. The clear advantage of this method is that only one set of indices is needed in order to produce easily interpretable and meaningful insights regarding the studied phenomenon. Moreover, the proposed indices can be estimated in a given-data context which can be adapted to applications for which no computer model is available. However, the major limitations of the approach are primarily related to the target aspect of the analysis. Indeed, as soon as the event becomes increasingly rare, all the inputs tend to be influential and making a clear distinction between interactions and correlation effects becomes difficult.

To overcome these limitations, a first approach could be to improve the estimation strategies. The sampling-based method could benefit from a better sampling scheme, such as importance sampling, as described by Rubinstein and Kroese (2008), which could reduce the estimator's variance. Recent results from Sarazin et al. (2020) using copulas are also promising in the extent to which they show efficient estimations of the Shapley effects. Moreover, adapting recent results from Spagnol (2020), with a link between the target Sobol' indices and the Squared Mutual Information Sugiyama (2012), should allow for other possibilities of given-data estimation methods. Another method based on a random forest given-data procedure, explored by Elie-Dit-Cosaque (2020) in the context of quantile-oriented importance measure estimation, could also yield promising results if transposed to a reliability-oriented setting.

Even if the Shapley attribution system is a solution when dealing with input statistical dependencies, it lacks a finer decomposition allowing to quantify the origin of each effect (e.g., statistical dependence and interaction). Future work could use the recent developments in Rabitti and Borgonovo (2019) in order to quantify interaction effects, by transposition of these results to the target sensitivity analysis setting.

Finally, it has been shown in Soofi et al. (2000) that the Shapley attribution system is equivalent to a maximum entropy distribution (e.g., uniform) over all possible orderings of inputs (the Shapley weights). Developments towards other forms of data-driven allocation systems could also open a path for further improvements.

Acknowledgments

We are grateful to the three anonymous referees as well as Jérôme Morio, François Bachoc and Julien Demange-Chryst for their helpful remarks. We also thank Sébastien Da Veiga, Clémentine Prieur and Fabrice Gamboa for interesting discussions and for having provided the dataset on the COVID-19 model. Finally, we would like to thank Victoria Stanford for her help in proofreading this work.

References

E. De Rocquigny, N. Devictor, S. Tarantola, Uncertainty in industrial practice: a guide to quantitative uncertainty management, Wiley, 2008.

K. Beven, Environmental Modelling: An Uncertain Future?, CRC Press, 2008.

F. Pianosi, K. Beven, J. Freer, J. Hall, J. Rougier, D. Stephenson, T. Wagener, Sensitivity analysis of environmental models: A systematic review with practical workflow, Environmental Modelling & Software 79 (2016) 214–232.

S. Razavi, A. Jakeman, A. Saltelli, C. Prieur, B. Iooss, E. Borgonovo, E. Plischke, S. Lo Piano, T. Iwanaga, W. Becker, S. Tarantola, J. Guillaume, J. Jakeman, H. Gupta, N. Melillo, G. Rabiti, V. Chabridon, Q. Duan, X. Sun, S. Smith, R. Sheikholeslami, N. Hosseini, M. Asadzadeh, A. Puy, S. Kucherenko, H. Maier, The future of sensitivity analysis: An essential discipline for systems modelling and policy making, Environmental Modelling and Software 137 (2021) 104954.

A. Saltelli, M. Ratto, T. Andres, F. Campolongo, J. Cariboni, D. Gatelli, M. Saisana, S. Tarantola, Global Sensitivity Analysis. The Primer, Wiley, 2008.

B. Iooss, P. Lemaître, A review on global sensitivity analysis methods, in: C. Meloni, G. Dellino (Eds.), Uncertainty management in Simulation-Optimization of Complex Systems: Algorithms and Applications, Springer, 2015, pp. 101–122.

M. Lemaire, A. Chateauneuf, J.-C. Mitteau, Structural Reliability, ISTE Ltd & John Wiley & Sons, Inc., 2009.

Y. Richet, V. Bacchi, Inversion algorithm for civil flood defense optimization: Application to two-dimensional numerical model of the garonne river in france, Frontiers in Environmental Science 7 (2019) 160.

R. T. Rockafellar, J. O. Royset, Engineering Decisions under Risk Averseness, ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering 1 (2015) 1–12.

J. Morio, M. Balesdent, Estimation of Rare Event Probabilities in Complex Aerospace and Other Systems: A Practical Approach, Woodhead Publishing, Elsevier, 2015.

Y.-T. Wu, Computational Methods for Efficient Structural Reliability and Reliability Sensitivity Analysis, AIAA Journal 32 (1994) 1717–1723.

S. Song, Z. Lu, H. Qiao, Subset simulation for structural reliability sensitivity analysis, Reliability Engineering and System Safety 94 (2009) 658–665.

P. Wei, Z. Lu, W. Hao, J. Feng, B. Wang, Efficient sampling methods for global reliability sensitivity analysis, Computer Physics Communications 183 (2012) 1728–1743.

V. Chabridon, Reliability-oriented sensitivity analysis under probabilistic model uncertainty – Application to aerospace systems, Ph.D. thesis, Université Clermont Auvergne, 2018.

G. Perrin, G. Defaux, Efficient Evaluation of Reliability-Oriented Sensitivity Indices, Journal of Scientific Computing 79 (2019) 1433–1455.

I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, Mathematical Modelling and Computational Experiments 1 (1993) 407–414.

E. Borgonovo, A new uncertainty importance measure, Reliability Engineering & System Safety 92 (2007) 771–784.

H. Raguet, A. Marrel, Target and conditional sensitivity analysis with emphasis on dependence measures, Working paper (2018). URL: `https://arxiv.org/abs/1801.10047`.

L. Li, Z. Lu, F. Jun, W. Bintuan, Moment-independent importance measure of basic variable and its state dependent parameter solution, Structural Safety 38 (2012) 40–47.

A. Marrel, V. Chabridon, Statistical developments for target and conditional sensitivity analysis: application on safety studies for nuclear reactor, Reliability Engineering and System Safety, in press (2021). URL: `https://hal.archives-ouvertes.fr/hal-02541142v2`.

W. Hoeffding, A class of statistics with asymptotically normal distribution, The Annals of Mathematical Statistics 19 (1948) 293–325.

J. Jacques, C. Lavergne, N. Devictor, Sensitivity analysis in presence of model uncertainty and correlated inputs, Reliability Engineering & System Safety 91 (2006) 1126–1134.

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

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

C. Xu, G. Z. Gertner, Uncertainty and sensitivity analysis for models with correlated parameters, Reliability Engineering & System Safety 93 (2008) 1563–1573.

T. Mara, S. Tarantola, Variance-based sensitivity indices for models with dependent inputs, Reliability Engineering & System Safety 107 (2012) 115–121.

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

N. Benoumechiara, K. Elie-Dit-Cosaque, Shapley effects for sensitivity analysis with dependent inputs: bootstrap and kriging-based algorithms, ESAIM: Proceedings and Surveys 65 (2019) 266–293.

N. Do, S. Razavi, Correlation effects? A major but often neglected component in sensitivity and uncertainty analysis, Water Resources Research 56 (2020) e2019WR025436.

L. S. Shapley, A value for n-person games, in: H. Kuhn, A. W. Tucker (Eds.), Contributions to the Theory of Games, Volume II, Annals of Mathematics Studies, Princeton University Press, Princeton, NJ, 1953, pp. 307–317.

M. Osborne, A. Rubinstein, A Course in Game Theory, MIT Press, 1994.

A. B. Owen, Sobol' indices and Shapley value, SIAM/ASA Journal of Uncertainty Quantification 2 (2014) 245–251.

A. B. Owen, C. Prieur, On Shapley value for measuring importance of dependent inputs, SIAM/ASA Journal of Uncertainty Quantification 5 (2017) 986–1002.

B. Iooss, C. Prieur, Shapley effects for Sensitivity Analysis with correlated inputs : Comparisons with Sobol' Indices, Numerical Estimation and Applications, International Journal for Uncertainty Quantification 9 (2019) 493–514.

A. Spagnol, Kernel-based sensitivity indices for high-dimensional optimization problems, Ph.D. thesis, Ecole des Mines de Saint-Etienne, 2020.

B. Broto, F. Bachoc, M. Depecker, Variance reduction for estimation of shapley effects and adaptation to unknown input distribution, SIAM/ASA Journal on Uncertainty Quantification 8 (2020) 693–716.

R. Christensen, Linear models for multivariate, time series and spatial data, Springer-Verlag, 1990.

T. Hastie, R. Tibshirani, J. Friedman, The elements of statistical learning, Springer, 2002.

J. Helton, J. Johnson, C. Salaberry, C. Storlie, Survey of sampling-based methods for uncertainty and sensitivity analysis, Reliability Engineering & System Safety 91 (2006) 1175–1209.

J. Johnson, J. LeBreton, History and use of relative importance indices in organizational research, Organizational Research Methods 7 (2004) 238–257.

L. Clouvel, Uncertainty quantification of the fast flux calculation for a PWR vessel, Ph.D. thesis, Université Paris-Saclay, 2019.

R. H. Lindeman, P. F. Merenda, R. Z. Gold, Introduction to bivariate and multivariate analysis, Scott Foresman and Company, Glenview, IL, 1980.

U. Grömping, Relative importance for linear regression in R: the Package relaimpo, Journal of Statistical Software 17 (2006) 1–27.

J. Nossent, P. Elsen, W. Bauwens, Sobol' sensitivity analysis of a complex environmental model, Environmental Modelling & Software 26 (2011) 1515 – 1525.

E. Song, B. Nelson, J. Staum, Shapley effects for global sensitivity analysis: Theory and computation, SIAM/ASA Journal on Uncertainty Quantification 4 (2016) 1060–1083.

A. Brandenburger, Cooperative Game Theory: Characteristic Functions, Allocations, Marginal Contribution, 2007. URL: https://web.archive.org/web/20170829122711/http://www.uib.cat/depart/deeweb/pdi/hde

P. Derennes, J. Morio, F. Simatos, Simultaneous estimation of complementary moment independent and reliability-oriented sensitivity measures, Mathematics and Computers in Simulation 182 (2021) 721–737.

J. Morio, Extreme quantile estimation with nonparametric adaptive importance sampling, Simulation Modelling Practice and Theory 27 (2012) 76–89.

V. Chabridon, M. Balesdent, G. Perrin, J. Morio, J.-M. Bourinet, N. Gayton, Mechanical Engineering Under Uncertainties, Wiley - ISTE Ltd, 2020, pp. 1–43.

L. Cui, Z. Lu, X. Zhao, Moment-independent importance measure of basic random variable and its probability density evolution solution, Science China Technical Sciences 53 (2010) 1138–1145.

J.-C. Fort, T. Klein, N. Rachdi, New sensitivity analysis subordinated to a contrast, Communications in Statistics - Theory and Methods 45 (2016) 4349–4364.

T. Browne, J.-C. Fort, B. Iooss, L. Le Gratiet, Estimate of quantile-oriented sensitivity indices, HAL, hal-01450891, version 1 (2017).

V. Maume-Deschamps, I. Niang, Estimation of quantile oriented sensitivity indices, Statistics and Probability Letters 134 (2018) 122–127.

S. Kucherenko, S. Song, L. Wang, Quantile based global sensitivity measures, Reliability Engineering and System Safety 185 (2019) 35–48.

L. Li, Z. Lu, C. Chen, Moment-independent importance measure of correlated input variable and its state dependent parameter solution, Aerospace Science and Technology 48 (2016) 281–290.

B. Iooss, S. Da Veiga, A. Janon, G. Pujol, sensitivity: Global Sensitivity Analysis of Model Outputs, 2021. URL: https://CRAN.R-project.org/package=sensitivity, R package version 1.25.0.

P. Lemaitre, Analyse de sensibilité en fiabilité des structures, Ph.D. thesis, Université de Bordeaux, 2014.

J. Fox, G. Monette, Generalized collinearity diagnostics, Journal of the American Statistical Association 87 (1992) 178–183.

K. M. Ortmann, The proportional value for positive cooperative games, Mathematical Methods of Operations Research (ZOR) 51 (2000) 235–248. URL: http://link.springer.com/10.1007/s001860050086. doi:10.1007/s001860050086.

B. E. Feldman, Relative Importance and Value, SSRN Electronic Journal (2005). URL: http://www.ssrn.com/abstract=2255827. doi:10.2139/ssrn.2255827.

E. Schumann, Generating correlated uniform variates. (2009). URL: http://comisef.wikidot.com/tutorial:correlateduniformvariates.

A. Saltelli, G. Bammer, I. Bruno, E. Charters, M. D. Fiore, et al., Five ways to ensure that models serve society: a manifesto (short comments), Nature 582 (2020) 482–484.

X. Lu, E. Borgonovo, Is time to intervention in the COVID-19 outbreak really important? A global sensitivity analysis approach, Preprint (2020). ArXiv:2005.01833.

S. Da Veiga, F. Gamboa, B. Iooss, C. Prieur, Basics and trends in sensitivity analysis, SIAM, In press, 2021.

S. Da Veiga, Calibration and sensitivity analysis of a COVID-19 epidemics model, Meeting AppliBUGS (Applications du Bayesian Unified Group of Statisticians), December 2020. URL: genome.jouy.inra.fr/applibugs/Daveiga_AppliBUGSDec2020.pdf.

R. Y. Rubinstein, D. P. Kroese, Simulation and the Monte Carlo method, Second ed., Wiley, 2008.

G. Sarazin, P. Derennes, J. Morio, Estimation of high-order moment-independent importance measures for Shapley value analysis, Applied Mathematical Modelling 88 (2020) 396–417.

M. Sugiyama, Machine learning with squared-loss mutual information, Entropy 15 (2012) 80–112.

K. Elie-Dit-Cosaque, Développement de mesures d'incertitudes pour le risque de modèle dans des contextes incluant de la dépendance stochastique, Ph.D. thesis, Université Claude Bernard - Lyon 1, 2020.

G. Rabitti, E. Borgonovo, A Shapley–owen index for interaction quantification, SIAM/ASA Journal on Uncertainty Quantification 7 (2019) 1060–1075.

E. S. Soofi, J. J. Retzer, M. Yasai-Ardekani, A Framework for Measuring the Importance of Variables with Applications to Management Research and Decision Models*, Decision Sciences 31 (2000) 595 – 625.

Appendix A. ANOVA and Sobol' indices

In the general non-linear case, as for the ANOVA of the linear model case (see Subsection 2.1), the idea is to find a general decomposition of the output variance. This can be done through the decomposition of a function with finite variance (L^2 mathematical property), called the *Hoeffding decomposition* Hoeffding (1948), which allows to rewrite $G(X)$ as a sum of centered components related to each possible subset of inputs. For example, in the case of a model with three inputs $X = (X_1, X_2, X_3)$, $G(X)$ can be decomposed into four components:

$$\begin{aligned} G(X) = & \; G_{\emptyset} && \text{(Mean behavior)} \\ & + G_1(X_1) + G_2(X_2) + G_3(X_3) && \text{(First-order)} \\ & + G_{\{1,2\}}(X_1, X_2) + G_{\{1,3\}}(X_1, X_3) + G_{\{2,3\}}(X_2, X_3) && \text{(Second-order)} \\ & + G_{\{1,2,3\}}(X) \; . && \text{(Third-order)} \end{aligned}$$

Moreover, if the inputs are assumed to be independent, each term is orthogonal to one another and writes

$$G_A(x_A) = \sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{E}\left[G(X)|X_B = x_B\right] \tag{A.1}$$

where $A \in \mathcal{P}_d$ is a subset of indices and $\mathcal{P}_d$ the set of all possible subsets of $\{1, \ldots, d\}$, $|A|$ is the cardinal of A and X_A denotes the subset of inputs, selected by the indices in A ($X_A = (X_i)_{i \in A}$). Then, the Hoeffding decomposition is unique and leads to a variance decomposition called "functional ANOVA":

$$\mathbb{V}[G(X)] = \sum_{A \in \mathcal{P}_d, A \neq 0} \mathbb{V}[G_A(x_A)] \, . \tag{A.2}$$

This leads to the definition of the Sobol' indices Sobol (1993):

$$S_A = \frac{\mathbb{V}[G_A(X_A)]}{\mathbb{V}[G(X)]} = \frac{\sum_{B \subset A} (-1)^{|A|-|B|} \mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_B\right]\right)}{\mathbb{V}[G(X)]} \, . \tag{A.3}$$

The sum of the Sobol' indices over all subset on inputs $A \in \mathcal{P}_d$ being equal to one, they can be directly interpreted as the percentage of the output variance due to each subset of input Sobol (1993); Saltelli et al. (2008). The Sobol' indices of higher orders than one can be interpreted as a means of quantifying the share of variance due to the interaction effects induced by the structure of the model $G(\cdot)$ between the selected subset of inputs.

Another useful sensitivity index is the closed Sobol' index Sobol (1993) which writes

$$S_A^{\text{clos}} = \sum_{B \subset A} S_B = \frac{\mathbb{V}\left(\mathbb{E}\left[G(X) \mid X_A\right]\right)}{\mathbb{V}[G(X)]} \tag{A.4}$$

In the independent setting, it can be interpreted as the percentage of variability induced by all the variables in a selected subset and their interactions. Figure Appendix A.1 provides an illustration of the Sobol' indices and the closed Sobol' indices for a model with three inputs. Each Venn diagram represents the variance of the output, with the representation of each of the two Sobol' indices presented above. While this representation is useful in the GSA context, it relies on the assumption of independence between the inputs.

Appendix B. Axioms of Shapley values

Consider a game with d players, and let $\text{val}(A) \in \mathbb{R}$ be the cost function quantifying the production value of a coalition (i.e., set of players) $A \in \mathcal{P}_d$, under the assumption that $\text{val}(\emptyset) = 0$. The Shapley value $\phi_j = \phi_j(\text{val}), j = 1, \ldots, d$ attributed to each player can be defined by the following set of axioms:

1. (Efficiency) $\sum_{j=1}^{d} \phi_j = \text{val}(\{1 \ldots, d\})$, meaning that the sum of the allocated values have to be equal to the value produced by the cooperation of all the players.
2. (Symmetry) If $\text{val}(A \cup \{i\}) = \text{val}(A \cup \{j\})$ for all $A \in \mathcal{P}_d$, then $\phi_i = \phi_j$, meaning that if two players allow for the same contribution to every coalition, their attribution should be the same.
3. (Dummy) If $\text{val}(A \cup \{i\}) = \text{val}(A)$ for all $A \in \mathcal{P}_d$, then $\phi_i = 0$, meaning that if a player does not contribute the the production of resources for all coalition, he should not be attributed any

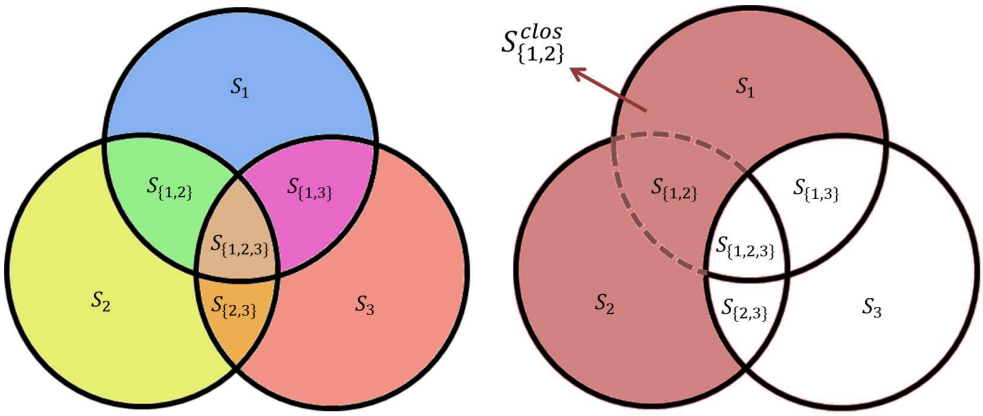

Figure Appendix A.1: Sobol' indices (left) and closed Sobol' indices (right).

resources.

4. (Additivity) If val and val$'$ have Shapley Values ϕ and ϕ' respectively, then the game with cost function val + val$'$ has Shapley values $\phi_j + \phi'_j$ for $j \in \{1, \dots, d\}$.

These four axioms guarantee a cooperative allocation of val$(\{1, \dots, d\})$. The unique attribution method that satisfies these four axioms are the Shapley values Osborne and Rubinstein (1994), defined by:

$$\phi_j = \frac{1}{d} \sum_{A \subset -j} \binom{d-1}{|A|}^{-1} (\text{val}(A \cup \{j\}) - \text{val}(A)), \quad j = 1, \dots, d \tag{B.1}$$

where $\{-j\} = \{1, \dots, d\} \backslash j$. One can additionally remark that $\phi_j(\text{val})$ is a linear operator, meaning that for some constant $c \in \mathbb{R}$, $\phi_j(c \times \text{val}) = c \times \phi_j(\text{val})$.

Appendix C. Mathematical proofs

Appendix C.1. Positivity of the (ℓ^1)-target Shapley effects

Let $A \subseteq \{1, \dots, d\} \setminus \{j\}$, for $j \in \{1, \dots, d\}$. In order to show that the (ℓ^1)-target Shapley effects are positive, one needs to prove that:

$$\text{T-S}^{\ell^1}_{A \cup \{j\}} \geq \text{T-S}^{\ell^1}_{A}. \tag{C.1}$$

In Cui et al. (2010), it was shown that the following property holds:

$$\eta_{A \cup \{j\}} \geq \eta_A \tag{C.2}$$

with η_A being defined in Eq. (19). From the definition of $\text{T-S}^{\ell^1}_A$,

$$\text{T-S}^{\ell^1}_A = \frac{2}{\mathbb{E}\Big[\big|\mathbb{1}_{\mathcal{F}_t}(X) - \mathbb{E}\left[\mathbb{1}_{\mathcal{F}_t}(X)\right]\big|\Big]} \eta_A, \tag{C.3}$$

one gets immediately the property C.1.

Appendix C.2. Positivity of the (ℓ^2)-target Shapley effects

Let $X = (X_1, \ldots, X_d) \in \mathbb{R}^d$, be a real-valued random vector admitting a probability measure P_X on the usual real measurable space. Let $L^2(P_X)$ be the functional space such that, for a measurable function f, $\|f\|_{L^2} \overset{\text{def}}{=} \int_{\mathbb{R}^d} f^2(x) dP_X(x) < +\infty$. Let $G(\cdot) \in L^2$ be the studied numerical model, and denote the random variable $Y = G(X)$ be the model output (or $Y = \mathbb{1}_{G(X)>t}(X)$ the TSA variable of interest, without loss of generality). Let $A \subseteq \{1, \ldots, d\} \setminus \{j\}$ be the indices of the subset of inputs X_A and $j \in \{1, \ldots, d\}$. In order to show that T-Sh$_j \geq 0$, one needs to prove that:

$$\text{T-S}_{A\cup\{j\}} - \text{T-S}_A \geq 0 \tag{C.4}$$

which is equivalent to

$$\mathbb{V}\Big(\mathbb{E}\left[Y|X_A\right]\Big) \leq \mathbb{V}\Big(\mathbb{E}\left[Y|X_{A\cup\{j\}}\right]\Big). \tag{C.5}$$

From the Pythagorean theorem, one has:

$$\|Y\|_{L^2} = \big\|\mathbb{E}\big[Y \mid X_A\big]\big\|_{L^2} + \big\|Y - \mathbb{E}\big[Y \mid X_A\big]\big\|_{L^2}, \tag{C.6}$$

which is equivalent to

$$\mathbb{E}\big[Y^2\big] = \mathbb{E}\Big[\big(\mathbb{E}\big[Y \mid X_A\big]\big)^2\Big] + \mathbb{E}\Big[\big(Y - \mathbb{E}\big[Y \mid X_A\big]\big)^2\Big]. \tag{C.7}$$

By removing $\big(\mathbb{E}\big[Y\big]\big)^2$ to both sides of the equality, one obtains:

$$\mathbb{V}\Big(\mathbb{E}\left[Y|X_A\right]\Big) = \mathbb{V}\left(Y\right) - \left\|Y - \mathbb{E}\left[Y|X_A\right]\right\|_{L^2}. \tag{C.8}$$

By using the formula $\mathbb{E}\big[Y \mid X_A\big] = \underset{Z\in\sigma(X_A)}{\operatorname{argmin}} \|Y - Z\|_{L^2}$, with $\sigma(X_A)$ being the span of X_A, we deduce that $\mathbb{E}\big[Y \mid X_A\big] \leq \mathbb{E}\big[Y \mid X_{A\cup\{j\}}\big]$ since $\sigma\big(X_A\big) \subseteq \sigma\big(X_{A\cup\{j\}}\big)$. This leads to

$$\mathbb{V}\left(Y\right) - \big\|Y - \mathbb{E}\big[Y \mid X_A\big]\big\|_{L^2} \leq \mathbb{V}\left(Y\right) - \big\|Y - \mathbb{E}\big[Y \mid X_{A\cup\{j\}}\big]\big\|_{L^2}. \tag{C.9}$$

Finally, from Eq. (C.8) and Eq. (C.9), we obtain

$$\mathbb{V}\Big(\mathbb{E}\big[Y \mid X_A\big]\Big) \leq \mathbb{V}\Big(\mathbb{E}\big[Y \mid X_{A\cup\{j\}}\big]\Big) \tag{C.10}$$

which concludes the proof.

Appendix D. Minimal R code examples for the estimation methods

Appendix D.1. Monte Carlo sampling estimator

```
#Packages
library(sensitivity)
library(mvtnorm)
library(condMVNorm)
```

```r
#Model definition
model.linear <- function(X) as.numeric(apply(X,1,sum)>0)

#Parameters
d <- 3
mu <- rep(0,d)
sig <- c(1,1,2)
ro <- 0.9
Cormat <- matrix(c(1,0,0,0,1,ro,0,ro,1),d,d)
Covmat <- ( sig %*% t(sig) ) * Cormat

#Total and marginal simulation function
Xall <- function(n) mvtnorm::rmvnorm(n,mu,Covmat)

#Conditional simulation function
Xset <- function(n, Sj, Sjc, xjc){
  if (is.null(Sjc)){
    if (length(Sj) == 1){ rnorm(n,mu[Sj],sqrt(Covmat[Sj,Sj]))
    }else{
    mvtnorm::rmvnorm(n,mu[Sj],Covmat[Sj,Sj])
    }
  }else{
  condMVNorm::rcmvnorm(n,
                    mu,
                    Covmat,
                    dependent.ind=Sj,
                    given.ind=Sjc,
                    X.given=xjc)
    }
}

#(l2)-target Shapley effects estimation
l2_tse.mc <- shapleyPermEx(model = modlin,
                           Xall=Xall,
                           Xset=Xset,
                           d=d,
                           Nv=1e4,
                           No = 1e3,
                           Ni = 3)
#Plot the results
print(l2_tse.mc)

#(l2)-target Shapley effects estimation with random permutations
l2_tse.mc.randperm<-shapleyPermRand(model = modlin,
```

```
                              Xall=Xall,
                              Xset=Xset,
                              d=d,
                              Nv=1e4,
                              No = 1e3,
                              Ni = 3,
                              m=5)
#Plot the results
plot(l2_tse.mc.randperm)
```

Listing 1: Minimal R code example for the Monte Carlo estimation.

Appendix D.2. Nearest-neighbor estimator

```
#Packages
library(sensitivity)
library(mvtnorm)

#Random sample of inputs-output
X<-rmvnorm(2000, rep(0,3), diag(3))
Y<-rbinom(2000, 1, 0.7)

#(l2)-target Shapley effects estimation
l2_tse.knn<-sobolshap_knn(model=NULL,
              X=X)
tell(l2_tse.knn, Y)

#Plot the results
plot(l2_tse.knn)

#(l2)-target Shapley effects estimation with random permutations
l2_tse.knn.randperm<-sobolshap_knn(model=NULL,
              X=X,
              rand.perm=T,
              n.perm=5)
tell(l2_tse.knn.randperm, Y)

#Plot the results
plot(l2_tse.knn.randperm)
```

Listing 2: Minimal R code example for the nearest-neighbor estimation.

Appendix E. Empirical convergence rate of the estimation scheme

Appendix E.1. Empirical convergence of the Monte Carlo estimation procedure

In order to illustrate the Monte Carlo estimation procedure of the target Shapley effects (see, Section 4.1), the following model is considered:

$$X = \begin{pmatrix} X_1 \\ X_2 \\ X_3 \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0.6 \\ 0 & 0.6 & 1 \end{pmatrix} \right), \quad Y = \sum_{i=1}^{3} X_i \tag{E.1}$$

where the considered TSA variable of interest is $\mathbb{1}_{Y>3}(X)$. This represents a failure probability $p_t^Y \simeq 0.071$. The empirical rate of convergence is studied on 100 repetitions, with respect to several values of N_v, with fixed sample sizes $N = 10^5$ and $N_p = 3$. The empirical convergence results are illustrated in Figure Appendix E.1.

Appendix E.2. Empirical convergence of the nearest-neighbor estimation procedure

The empirical convergence rate of the nearest-neighbor estimator of the target Shapley effects (see, Section 4.2) is illustrated on the same test-case as in Appendix E.1, with the same TSA variable of interest is $\mathbb{1}_{Y>3}(X)$. The proposed indices have been estimated on 100 generated samples of X, for several sample sizes. The results are presented in Figure Appendix E.2.

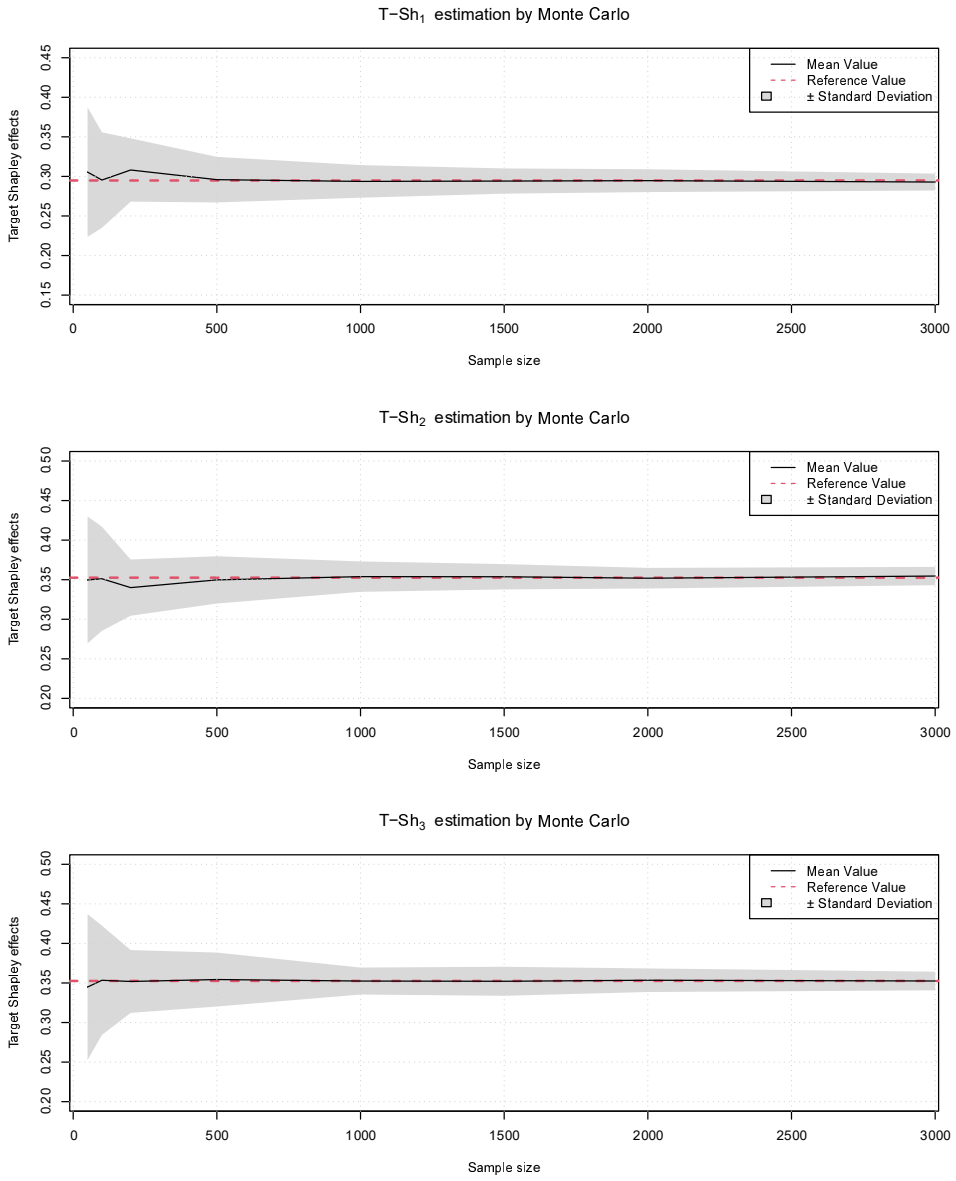

Figure Appendix E.1: Empirical convergence of the Monte Carlo estimation procedure with respect to the sample size.

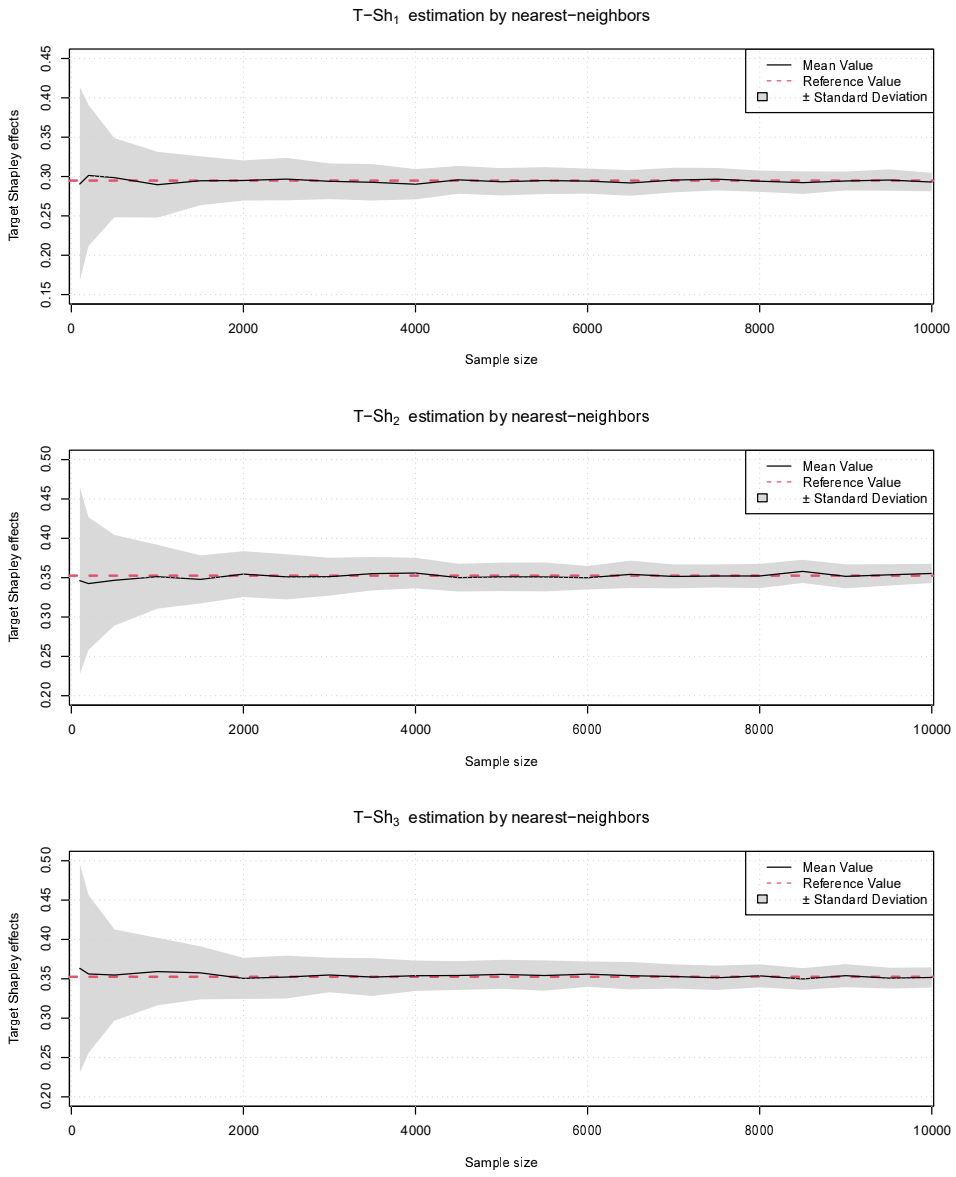

Figure Appendix E.2: Empirical convergence of the nearest-neighbor estimation with respect to the sample size.