Sensitivity Analysis of Risk Assessment with Data-Driven Dependence Modeling

Gabriel Sarazin, Jérôme Morio, Agnès Lagnoux, Mathieu Balesdent, Loic Brevault

Sensitivity Analysis of Risk Assessment with Data-Driven Dependence Modeling

Conference Paper · May 2019

CITATIONS
0

READS
99

5 authors, including:

Gabriel Sarazin
The French Aerospace Lab ONERA
1 PUBLICATION **0** CITATIONS
SEE PROFILE

Jérôme Morio
The French Aerospace Lab ONERA
84 PUBLICATIONS **660** CITATIONS
SEE PROFILE

Agnès Lagnoux
Université Toulouse II - Jean Jaurès
30 PUBLICATIONS **308** CITATIONS
SEE PROFILE

Sensitivity Analysis of Risk Assessment with Data-Driven Dependence Modeling

G. Sarazin & J. Morio

ONERA/DTIS, Université de Toulouse, Toulouse, France. E-mail: Gabriel.Sarazin@onera.fr

A. Lagnoux

IMT, Université de Toulouse Jean Jaurès, Toulouse, France. E-mail: lagnoux@univ-tlse2.fr

M. Balesdent & L. Brevault

ONERA/DTIS, Université de Paris-Saclay, Palaiseau, France. E-mail: Loic.Brevault@onera.fr

The reliability analysis of complex systems often requires dealing with a computationally expensive simulation code. To estimate the failure probability, a frequently used method aims at propagating the input uncertainties through the black-box model. In this paper, as marginal distributions are assumed provided, the lack of knowledge about the joint distribution of input variables is limited to a copula distribution learnt from an industrial dataset obtained during past experiences. To describe complex and polymorphic patterns of dependence, attention has turned to vine copulas whose main advantage rests on their ability to approximate the whole dependence structure with a simple product of judiciously-selected bivariate copulas. The presented approach couples vine copula fitting to model the joint distribution on input variables and importance sampling to estimate the failure probability. For a given training set, the matrix of Kendall's rank correlation coefficients, which collects information about dependence intensities, is deeply involved in the inferential procedure leading to the copula vine specification. In this work, a sensitivity analysis is performed to measure the impact of Kendall's matrix uncertainty due to scarce data on the estimation of the failure probability. As Kendall's coefficients are dependent random variables, sensitivity analysis is achieved with Borgonovo's indices, using bootstrap replications of the available data to have a larger amount of estimations. The ranking of sensitivity indices allows identifying the pair of variables on which one has to acquire new samples in order to reduce variability in risk assessment. This methodology is applied on the buckling of a composite plate.

Keywords: uncertainty quantification, uncertainty management, statistical inference, importance sampling, rank correlation, bootstrap, vine copulas, Borgonovo's indices.

1. Introduction

1.1. *Reliability analysis*

In many industrial fields, while dealing with a highly technical system, reliability cannot be assessed with functional design tools. Instead, the system's behavior is mimicked by a multidisciplinary simulation code that can be seen as a deterministic input-output model $\mathcal{M}$. In such a black-box approach, the level of detail on $\mathcal{M}$ is restricted to the following mathematical description:

$$\begin{array}{rccc} \mathcal{M}: & \Omega \subseteq \mathbb{R}^d & \longrightarrow & \mathbb{R} \\ & \mathbf{x} = (x_1, \ldots, x_d) & \longmapsto & \mathcal{M}(\mathbf{x}). \end{array} \tag{1}$$

As many operating variables are uncertain, a probabilistic framework is adopted and inputs are described as a random vector $\mathbf{X} = (X_1, \ldots, X_d)$ with cumulative distribution function (CDF) $F_{\mathbf{X}}$ and probability density function (PDF) $f_{\mathbf{X}}$. For any given observations $\mathbf{x} \in \Omega$, $\mathcal{M}$ allows to compute the associated scalar output $y = \mathcal{M}(\mathbf{x})$. The uncertainties produced by $\mathbf{X}$ are transferred to Y, which is therefore a random variable as well. The distribution of Y depends on the propagation of those uncertainties and can be sketched out by computing related statistics (especially moments, quantiles and empirical CDF). In our particular case, the quantity of interest is a rare-event probability P_f defined as the risk of exceeding a critical threshold T. The failure domain $\mathcal{D}_f$ refers to the inverse image of the event $\{Y > T\}$ and thus:

$$\begin{aligned} P_f = \mathbb{P}\{Y > T\} &= \mathbb{P}\{\mathcal{M}(\mathbf{X}) > T\} \\ &= \mathbb{P}\{\mathbf{X} \in \mathcal{D}_f\}. \end{aligned} \tag{2}$$

In practice, the input density $f_{\mathbf{X}}$ is often assumed to be known analytically and modeling decisions derive from either engineering expertise or statistical learning. However, in the context of this study, information on $f_{\mathbf{X}}$ is limited to a small-sized dataset $\mathbf{X}_0 \in \mathbb{R}^{N \times d}$. The relatively reduced amount of available observations ($N \approx 500$) in $\mathbf{X}_0$ can be explained by the difficulty to collect usable operational data. In that view, estimating the failure probability P_f may require applying a two-

Proceedings of the 29th European Safety and Reliability Conference.
Edited by Michael Beer and Enrico Zio
 Published by Research Publishing, Singapore
ISBN: 981-973-0000-00-0 :: doi: 10.3850/981-973-0000-00-0_esrel2019-paper

step process. As a first step, the joint distribution of input variables is learnt from $\mathbf{X}_0$. Let us denote by $\hat{f}_0$ the multivariate density obtained after fitting a probabilistic model to $\mathbf{X}_0$. The second step consists in propagating the uncertainties stemming from the density $\hat{f}_0$ through the model $\mathcal{M}$ in order to estimate P_f. In this paper, particular attention has been given to the first step.

1.2. *Probabilistic model estimation*

Given a multivariate sample, learning the underlying joint density remains a major issue in statistics. Inference can be performed using three different approaches: parametric (if considering a family of densities), non-parametric (with kernel density estimation) or hybrid (when combining tools borrowed from the two previous). Another alternative in high dimension seems to be the copula-marginal separation. Indeed, since a major breakthrough due to Sklar (1959), it is well known that any absolutely continuous multivariate distribution function $F_{\mathbf{X}}$ can be written in terms of univariate marginal distributions $F_1, \ldots, F_d$ and a copula $C_{\mathbf{X}}$ which describes the dependence structure existing among the variables:

$$\forall \mathbf{x} = (x_1, \ldots, x_d) \in \mathbb{R}^d,$$
$$F_{\mathbf{X}}(\mathbf{x}) = C_{\mathbf{X}}\Big(F_1(x_1), \ldots, F_d(x_d)\Big). \quad (3)$$

The copula function $C_{\mathbf{X}} : [0\,;1]^d \longrightarrow \mathbb{R}$ is unique and can be interpreted as the joint CDF between the transformed variables $U_i = F_i(X_i)$. The chain rule provides a very simple link between $f_{\mathbf{X}}$ and the copula density $c_{\mathbf{X}}$ computed as the cross partial derivative of $C_{\mathbf{X}}$:

$$f_{\mathbf{X}} = c_{\mathbf{X}}\Big(F_1, \ldots, F_d\Big) \times \prod_{i=1}^{d} f_i\,. \quad (4)$$

In this work, it is assumed that marginal distributions are prescribed and the uncertainties they are expected to introduce under normal conditions are excluded. This situation allows to better focus on the error committed because of the lack of knowledge on the copula function. Moreover, assessing to what extent the presence of epistemic uncertainty on the margins may have an effect on the estimated failure probability has already been deeply investigated, in particular by Der Kiureghian (2008) to cite but one. Under the assumption that the margins are fixed, statistical learning is therefore all about fitting a copula model $\hat{c}_0$ to the rescaled dataset $\mathbf{U}_0 \in [0\,;1]^{N\times d}$. The learnt density can thus be expressed as follows:

$$\hat{f}_0 = \hat{c}_0\Big(F_1, \ldots, F_d\Big) \times \prod_{i=1}^{d} f_i\,. \quad (5)$$

Since adopting a copula-based representation leads to equalize the contributions of all marginal distributions, copula analysis provides information only about intrinsic dependence properties. In the bivariate case, parametric copula families are sufficient to replicate finely any relevant characteristic (correlation, symmetry, orientation, tails...) falling within the description of a dependency. In higher dimension, parametric inference remains tractable provided that dependence modeling incorporates more sophisticated representations, one possibility being the use of regular vine copulas.

It has been proven in Bedford et al. (2002) that any dependence structure arising from a d-copula can be divided, after several conditioning steps, into a small number of pair-copulas. A regular vine copula (R-vine) is an adaptive parametric model, built upon this factorization result, which tries to better understand multivariate patterns. With this method, a very complex arrangement, maybe inextricable at first sight, can be broken down into a simple list of bivariate dependencies that can be treated easily with a parametric approach. Intense research effort over the past few years has resulted in the development of efficient algorithms to use pair-copula constructions. At first, R-vine models were applied in finance engineering to anticipate the joint behavior of asset returns in a context of high portfolio volatility. They have now become a common practice in many sectors (including meteorology, hydrology, insurance and marketing) whenever there is a need to simulate correlated input variables. In reliability analysis, vine copulas have been the subject of particular interest, as evidenced by recent works of Jiang et al. (2015), Benoumechiara et al. (2018), Torre et al. (2018) and Xu et al. (2018).

1.3. *Scope of the paper*

The dependence pattern described by a vine copula can be trivial, for instance if the variables are almost independent, as well as very complicated, if each bivariate copula belongs to a specific family. The ability to recognize and prioritize relationships between pairs of variables determines the quality of the learning process. It should be noted that every pair of variables is driven by a bivariate copula which is involved (maybe indirectly) in the estimated R-vine density $\hat{c}_0$. A natural question that arises is then to identify the pair which has the strongest impact on $\hat{P}_f$ when the joint distribution $\hat{f}_0$ is propagated through $\mathcal{M}$. Targeting the most influential pair might be a valuable indication if one is seeking to increase the knowledge on the input distribution. In this paper, a global sensitivity analysis procedure on the failure probability is proposed with Kendall's rank correlation coefficients used as uncertainty sources. As will be noticed later in Section 2,

the R-vine specification has been designed to give full account of empirical Kendall's matrix $\hat{\tau}_K$. Because of such a close interaction, the uncertainties conveyed by empirical Kendall's coefficients are expected to have an impact on the stability of the R-vine model and, by clear implication, on the whole process of risk estimation.

The paper is organized as follows. In Section 2, a brief introduction to R-vines is provided, with particular emphasis on the inferential procedure. The different stages in the construction of the sensitivity analysis scheme are detailed in Section 3. To illustrate the point, the presented methodology is applied to a physical test case in Section 4 while Section 5 draws some conclusions.

2. Vine Copulas

2.1. *Construction*

For given margins, Joe (1996) introduced a class of multivariate distributions with $d(d-1)/2$ dependence parameters by iteratively mixing conditional distributions. The starting point is the Bayes' theorem for continuous densities which consists in the recursive decomposition of a multivariate density into a product of conditional densities. The key lies in transferring the conditioning variables from densities to copulas. A standard assumption is that conditional pair-copulas depend on their conditioning set only through the conditional margins:

$$\forall i \neq j, \forall A \subset \{1,\dots,d\} \setminus \{i,j\},$$
$$f_{ij|A} = c_{ij|A}\Big(F_{i|A}, F_{j|A}\Big) \times f_{i|A} f_{j|A}\,. \quad (6)$$

It amounts to saying that copulas are conservative with respect to the conditioning process. From Eq. (6), it is straightforward to see that removing any variable X_i from the conditioning set A can be done by incorporating a conditional pair-copula with respect to the reduced set $A_{-i} = A \setminus \{i\}$:

$$\forall A \subset \{1,\dots,d\} \setminus \{j\}, \forall i \in A,$$
$$f_{j|A} = c_{ij|A_{-i}}\Big(F_{i|A_{-i}}, F_{j|A_{-i}}\Big) \times f_{j|A_{-i}}\,. \quad (7)$$

The repeated use of Eq. (7) leads to factorize $c_{\mathbf{X}}$ into a product of $d(d-1)/2$ conditional pair-copulas. As this factorization is not unique, it is important to determine the role played by each variable during the conditioning steps. Since the very start of pair-copula decompositions, a graphical tool has been extensively used in practice for clarity reasons. It has been shown that a sequence of nested trees $\mathcal{V} = \{T_1,\dots,T_{d-1}\}$, each one consisting of nodes N_i and edges E_i, is suited to summarize a copula factorization. The graph $\mathcal{V}$ is called an R-vine tree if it obeys the following construction rules:

- $N_1 = \{1,\dots,d\}$ and $|E_1| = d-1$;
- $\forall i \in \{2,\dots,d-1\}, N_i = E_{i-1}$;
- Two edges in T_i are joined in T_{i+1} if they share a common node in T_i (proximity condition).

For all $i \in \{1,\dots,d\}$, every connecting edge $e \in E_i$ corresponds to a conditional pair-copula $c_{j_e,k_e|D_e}$ with D_e resulting from the successive applications of the proximity condition. Using the same notation, the density corresponding to an R-vine copula may be written as:

$$c_{\mathbf{X}} = \prod_{i=1}^{d-1} \prod_{e \in T_i} c_{j_e,k_e|D_e}\,. \quad (8)$$

Given an R-vine tree specification $\mathcal{V}$, the dependence structure is organized as a cascade of pair-copulas which can be chosen independently among all existing bivariate parametric models. The resulting collection of families is denoted by $\mathcal{F}$ and the associated parameters are grouped together into $\boldsymbol{\theta} \subseteq \mathbb{R}^p$ (with p depending on the family assortment). As a result, full inference for an R-vine distribution should comprise the following steps: (a) selection of the tree structure $\mathcal{V}$, (b) choice of the parametric families $\mathcal{F}$, and (c) estimation of the parameters $\boldsymbol{\theta}$. Steps (a) and (b) constitute what is called model selection because once $\mathcal{V}$ and $\mathcal{F}$ are stated, the copula density $c_{\mathbf{X}}$, as expressed in Eq. (8), becomes fully-parametric and can then be fitted to data with a maximum-likelihood approach. However, the last step implies computing recursively conditional distribution functions involved in conditional pair-copulas. Joe (1996) showed that any conditional distribution function in tree T_{i+1} can be derived from a pair-copula distribution in T_i:

$$\forall i \neq j, \forall A \subset \{1,\dots,d\} \setminus \{i,j\},$$
$$F_{j|A\cup\{i\}} = \frac{\partial C_{ij|A}}{\partial F_{i|A}}\Big(F_{i|A}, F_{j|A}\Big). \quad (9)$$

2.2. *Inference*

A naive method to perform inference should be to maximize the log-likelihood for all possible models $\mathcal{V}$-$\mathcal{F}$ before selecting the best one. As highlighted in Morales-Nápoles (2010), the number of R-vine factorizations increases at a super-exponential rate and scanning all models is therefore intractable in practice. Aas et al. (2009) were the first to introduce the idea of a stepwise tree-by-tree procedure. For a given tree structure $\mathcal{V}$, family selection and parameter estimation are conducted simultaneously, one tree after the other. Within a tree T_i, let us denote $\mathcal{F}_i$ the copula families and $\boldsymbol{\theta}_i$ the copula parameters. Assuming that $\mathcal{V}$ is known, fitting an R-vine copula distribution to a given dataset $\mathbf{U}_0$ may require to proceed as follows:

- Select copula families $\mathcal{F}_1$ in the tree T_1 by maximizing a goodness-of-fit criterion (e.g. Akaike) on each connecting edge.
- Compute copula parameters $\boldsymbol{\theta}_1$ from $\mathbf{U}_0$ by maximum likelihood. If there is a mapping g such that $\boldsymbol{\theta}_1 = g(\tau_K)$, estimation becomes computationally trivial.
- Apply Eq. (9) with $\mathcal{F}_1$ and $\boldsymbol{\theta}_1$ to transform the initial observations $\mathbf{U}_0$ into conditional observations $\mathbf{U}_1$ which will be used in T_2.
- Iterate with T_2 and so on.

At the end of the sequential procedure, $\mathcal{F}$ and $\boldsymbol{\theta}$ are fully determined and it should be noted that all computing tasks have been performed in a bivariate framework. Hence, a final maximum likelihood recalibration on model $\mathcal{V}$-$\mathcal{F}$ is run with $\boldsymbol{\theta}$ as starting value for gradient-descent optimization. Moreover, Dissmann et al. (2013) came up with an heuristic algorithm to improve the tree selection, which had been an open issue until then. Since the tree T_1 has to be chosen among the set $\mathcal{S}_d$ of all spanning trees between nodes $N_1 = \{1, \dots, d\}$, he proposed to select the one which maximizes the sum of absolute empirical Kendall's coefficients computed from $\mathbf{U}_0$:

$$T_1^* = \underset{T \in \mathcal{S}_d}{\arg\max} \sum_{(k,l) \in T} |\hat{\tau}_{kl}| \,. \tag{10}$$

Thus, pairing gives priority to variables which are identified as strongly dependent in terms of Kendall's correlation. It is worth mentionning that, unlike Pearson's correlation coefficient or Spearman's ρ, Kendall's τ is not constructed to quantify linearity, but rather to measure any form of dependence. For the remaining trees $T_2, \dots, T_{d-1}$, an almost identical optimization problem is solved at each step of the tree-by-tree inference procedure. Differences with Eq. (10) include the use of conditional Kendall's correlations and the integration of a constraint related to the proximity condition.

3. Proposed approach

As R-vine distributions seem to be essential in dependence modeling, they are integrated into the learning model:

$$\begin{aligned} \mathcal{L}_N : \mathbb{R}^{N \times d} &\longrightarrow L^2(\mathbb{R}^d) \\ \mathbf{X}_0 &\longmapsto \hat{f}_0(\cdot \mid \mathcal{V}, \mathcal{F}, \boldsymbol{\theta}) \,. \end{aligned} \tag{11}$$

It raises the question of how the sampling variability due to scarce data is likely to affect risk assessment. From one sample $\mathbf{X}_0$ to another, the R-vine specification may be substantially modified, with either the tree arrangement $\mathcal{V}$ or the copula assortment $\mathcal{F}$ being impacted. If a particular kind of model disruption always leads to inaccurate estimates during risk analysis, it is a clear indication that the associated pair of variables deserves to be studied cautiously. In this paper, an attempt is made to quantify how much influence a pair could exert on the failure probability.

3.1. *Uncertainty sources*

The greatest difficulty lies in defining a framework to address this issue. A prerequisite is to find a way to account for the uncertainty inherent to the R-vine specification. For a given pair of variables, it would be ideal to bring into play all the sources of variability which can interfere in the pair-copula representation. Indeed, each time an R-vine distribution is learnt from $\mathbf{X}_0$, one may ask the following questions about the pair $(X_i \,, X_j)$.

- Does the edge (i, j) belong to the tree T_1 ?
- Which candidate family is selected for c_{ij} ?
- How uncertain is the computed parameter θ_{ij} ?

It is not easy to include all elements in a sensitivity analysis procedure with clearly interpretable results. A first idea could be to set an R-vine model $\mathcal{V}$-$\mathcal{F}$ and to focus on the uncertainty propagated by the dependence parameters $\boldsymbol{\theta}$:

$$\begin{aligned} \Psi_{\mathcal{V}\mathcal{F}} : \mathbb{R}^p &\longrightarrow [0\,;1] \\ \boldsymbol{\theta} &\longmapsto \hat{P}_f\big(\hat{f}_0(\boldsymbol{\theta})\big) \,. \end{aligned} \tag{12}$$

Uncertainty on a pair of variables is then restricted to its copula parameter dispersion. Sensitivity analysis on $\Psi_{\mathcal{V}\mathcal{F}}$ can be achieved with maximum likelihood estimators as input variables. In the case of R-vine distributions, Haff et al. (2013) established their asymptotic normality, while Stöber and Schepsmeier (2013) developped a method to routinely estimate their standard errors without using finite difference approximations. Nevertheless, this approach is insufficient because it presupposes to set $\mathcal{V}$ and $\mathcal{F}$, whereas in practice nobody has a clue what a suitable model would look like. To take into account the full variability in automated model selection, changing one's viewpoint becomes necessary.

Empirical Kendall's matrix $\hat{\tau}_K$ may be used as a way of tracing back the pairs of variables which are responsible for variability in the R-vine specification. Each pair of variables can be matched to one and unique Kendall's correlation coefficient (in the lower triangular portion of the matrix). An identification mechanism of this nature was not permitted under $\Psi_{\mathcal{V}\mathcal{F}}$ with parametrization $\boldsymbol{\theta}$ depending only on the pairs involved in T_1. However, it must be clearly understood that the matrix $\hat{\tau}_K$, when separated from $\mathbf{U}_0$, is not sufficient to reconstruct $\hat{f}_0$. As described in Section 2, knowing $\mathbf{U}_0$ is imperative to select copula families and to compute conditional observations. There is a need to characterize the interactions $\mathcal{L}_d$ between

$\hat{\tau}_K$ and $\hat{f}_0$:

$$\begin{array}{rcl} \mathcal{L}_d : [-1\,;1]^{d\times d} & \longrightarrow & L^2(\mathbb{R}^d) \\ \hat{\tau}_K & \longmapsto & \hat{f}_0(\cdot \mid \mathcal{V}, \mathcal{F}, \boldsymbol{\theta})\,. \end{array} \tag{13}$$

One possibility is to consider $\mathcal{L}_d$ as a stochastic model where intrinsic alea affects $\mathcal{F}$ and $\boldsymbol{\theta}$ and originates from the initial sample $\mathbf{U}_0$ used for inference. To observe an input-output match through $\mathcal{L}_d$, both quantities must be derived from a common training set $\mathbf{U}_0$ chosen in advance. Even if $\mathcal{L}_d$ is not a deterministic mapping, this approach has two main benefits:

- As evoked in Eq. (10), $\hat{\tau}_K$ acts as a set of random weights in the maximum spanning tree problem and has a direct impact on $\mathcal{V}$.
- There is an implicit relationship between $\hat{\tau}_K$ and $\boldsymbol{\theta}$ which results in a transfer of uncertainty.

In what follows, sensitivity analysis is then performed on the stochastic model which relates Kendall's matrix and the failure probability:

$$\begin{array}{rcl} \varphi : [-1\,;1]^{d\times d} & \longrightarrow & [0\,;1] \\ \hat{\tau}_K & \longmapsto & \hat{P}_f\,. \end{array} \tag{14}$$

In that regard, to perform sensitivity analysis, one must be able to: (a) simulate input uncertainty on $\hat{\tau}_K$, (b) estimate all failure probabilities with rare-event simulation techniques, and (c) compute sensitivity indices to measure the impact of each Kendall's coefficient on risk assessment.

3.2. *Rare-event probability estimation*

As statistical learning is focused on the inputs $\mathbf{X}$, the distribution of the output Y is still unknown at this point and the Monte Carlo (MC) method must be applied to estimate P_f. Simply using a crude MC procedure proves to be unworkable in practice due to the excessive number of model evaluations needed to enable an accurate estimation. In order to construct a computationally viable estimator, priority is given to rare-event simulation algorithms since they are designed to explore selectively the sample space. Among those techniques, Importance Sampling (IS), which was first introduced by Kahn and Harris (1951), emerged as an almost cure-all solution to alleviate the computational effort. IS brought the innovative idea of generating N_S independant and identically distributed samples $\mathbf{X}^{(k)}$ with a goal-oriented auxiliary density h. The change of sampling density is made possible by the construction of a penalized estimator taking into account the discrepancy between $\hat{f}_0$ and h:

$$\hat{P}_f^{\mathrm{IS}} = \frac{1}{N_S}\sum_{k=1}^{N_S} \mathbb{1}_{\mathcal{D}_f}(\mathbf{X}^{(k)}) \frac{\hat{f}_0(\mathbf{X}^{(k)})}{h(\mathbf{X}^{(k)})}\,. \tag{15}$$

In the equation above (where $\mathbb{1}_{\mathcal{D}_f}$ is the indicator function of the failure domain), the optimal density h_{opt} is the one which minimizes the variance of $\hat{P}_f^{\mathrm{IS}}$. As h_{opt} depends on the quantity of interest P_f, it has to be learnt with an iterative algorithm. In this work, Non-parametric Adaptive Importance Sampling (NAIS) is used to construct an ad-hoc auxiliary distribution, drawing on the initial work of Zhang (1996) and improvements due to Morio (2012). From samples generated with $\hat{f}_0$, a mixture of weighted Gaussian kernels $\hat{h}$ is proposed to learn h_{opt}. At each step of the iterative process, the samples giving rise to the highest outputs are integrated into the auxiliary density which is expected to become closer to $\mathcal{D}_f$. Whether NAIS is a tried-and-tested method for the application described in Section 4, resorting to any other IS technique would not have been a misguided strategy. For instance, Kullback-Liebler's cross-entropy, which is thoroughly examined in Rubinstein and Kroese (2013), would have led to similar results in all what follows.

3.3. *Complete workflow*

As explained in Section 1, the raw data consist of a small amount of observations $\mathbf{X}_0$, transformed into $\mathbf{U}_0$ for inference purposes. From this inital dataset, only a single estimation of P_f can be computed. To be able to observe variability on $\hat{P}_f$, data-driven risk assessment must be performed for other datasets $\tilde{\mathbf{U}}^{(i)}$ stemming from the same underlying distribution than $\mathbf{U}_0$. As the theoretical density $f_{\mathbf{X}}$ is unknown, bootstrapping seems to be the only workable method to replicate the original data. Invented by Efron (1992) and extensively used since then, bootstrap resampling relies upon the empirical distribution function $\hat{F}_N$ to generate new datasets. In this way, N_B bootstrap replications are produced from $\mathbf{U}_0$ and they are expected to give a sufficiently representative idea of how much the dataset may vary from one experiment to the other. For each dataset $\tilde{\mathbf{U}}^{(i)}$, empirical Kendall's matrix $\hat{\tau}_K^{(i)}$ is computed and the learning phase yields the R-vine copula density $\hat{c}^{(i)}$.

To attain a sufficient level of convergence in the estimation of the sensitivity indices, it is often necessary to take large values of N_B. For efficiency reasons, applying the NAIS algorithm to each copula density $\hat{c}^{(i)}$ cannot be contemplated. Indeed, running NAIS turns out to be computationally expensive because the goal-oriented exploration of the sample space Ω implies many evaluations of the Gaussian kernel mixture $\hat{h}$. To couple sensitivity analysis and rare-event simulation, Morio (2011) proposed to complete only a single training of h_{opt} denoted $\hat{h}_0$. After checking

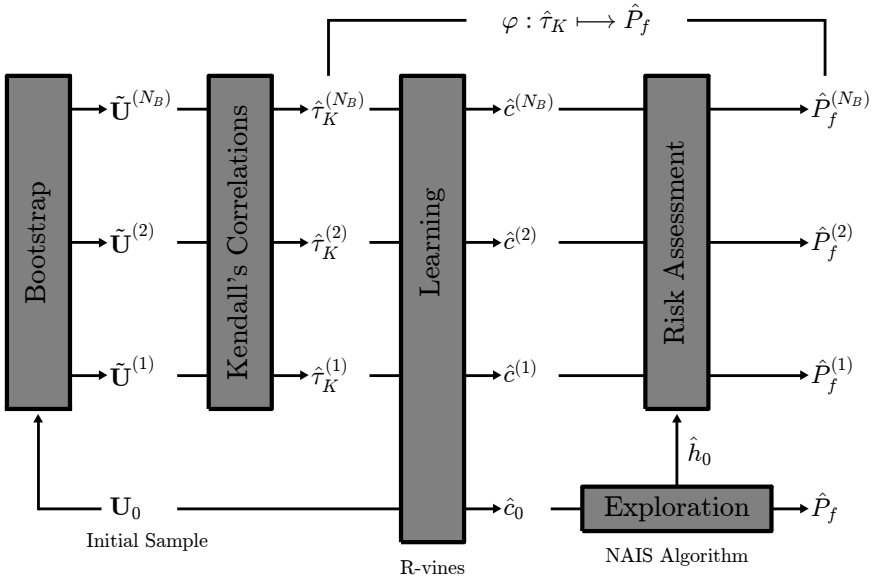

Fig. 1. Simulation scheme to study the sensitivity of risk assessment to pairwise dependence modeling.

that the loss of optimality is not detrimental to estimation accuracy, all other failure probabilities required in sensitivity analysis are computed by plugging $\hat{h}_0$ into a slightly different version of Eq. (15). The samples $\mathbf{X}^{(k)}$ are retrieved from the last iteration of NAIS and used for all estimations:

$$\hat{P}_f^{(i)} = \frac{1}{N_S} \sum_{k=1}^{N_S} \mathbb{1}_{\mathcal{D}_f}(\mathbf{X}^{(k)}) \frac{\hat{f}^{(i)}(\mathbf{X}^{(k)})}{\hat{h}_0(\mathbf{X}^{(k)})} . \quad (16)$$

At the end of the workflow process illustrated on Fig. 1, several input-output samples $(\hat{\tau}_K^{(i)},\ \hat{P}_f^{(i)})$ are available to analyze the stochastic model φ. Special care must be taken to choose appropriate sensitivity indices because Kendall's correlation coefficients are not mutually independent inputs. In this context, a direct computation of Sobol's indices, as advocated when inputs are actually independent, could lead to erroneous interpretations. Over the past few years, as discussed in Iooss and Lemaître (2015), several extensions have been developped to adapt variance-based importance measures defined in the independant case to correlated inputs. Borgonovo's indices are preferred since they encompass the entire distribution of the output rather than only focusing on the second-order moment. They attempt to quantify the impact suffered by the output density when conditioned by one or several input variables. All Borgonovo's indices lie in $[0\,;1]$ but, unlike Sobol's indices, their sum is not equal to one. In this work, an estimation scheme proposed by Derennes et al. (2018) is used to estimate all first-order Borgonovo's indices with a unique joint N_B-sample of $(\hat{\tau}_K^{(i)},\ \hat{P}_f^{(i)})$.

4. Application

4.1. *Test case*

The methodology introduced in Section 3 is now applied to the buckling of a square plate $(a/b = 1)$ under uniaxial compression (see Fig. 2). All four edges are simply supported. The plate is made of a 8-ply carbon/epoxy composite laminate. An acceptable modeling should include two major assumptions: (a) there is no coupling between the membrane behavior and the bending behavior, and (b) the bending behavior is orthotropic. It is assumed that (a) and (b) hold even under the combined effect of the dispersion observed on the elastic properties of the carbon/epoxy-based ply material and the existence of small random errors in the ply orientations ($\pm 2°$). Buckling refers to

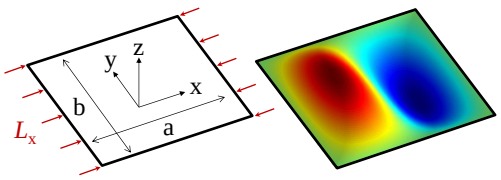

Fig. 2. Buckling of a simply supported plate under uniaxial loading. The buckling mode is $(m^*,\ n^*) = (2\,,1)$. The colormap corresponds to the displacement in the z-direction.

a loss of stability which occurs within the elastic range of the material. When the applied load L_x is increased, the structure deforms into a buckling mode characterized by m half-waves in the x-direction and n half-waves in the y-direction. For any integers $m, n \in \mathbb{N}^*$, the buckling factor $\Lambda_{m,n}$ is a multiplier that gives the magnitude of the load required to produce such a buckling mode.

4.2. *Input-ouput model*

Particular attention is paid to the dependence structure existing among the coefficients of the bending stiffness matrix $\mathbf{D}$. It should be noticed that $\mathbf{D}$ is a 3-by-3 symmetric positive-definite matrix with coefficients D_{13} and D_{23} being neglected to conform to the orthotropy assumption:

$$\mathbf{D} = \begin{pmatrix} D_{11} & D_{12} & 0 \\ D_{12} & D_{22} & 0 \\ 0 & 0 & D_{33} \end{pmatrix} \equiv \begin{pmatrix} D_{11} \\ D_{12} \\ D_{22} \\ D_{33} \end{pmatrix} \in \mathbb{R}^4. \quad (17)$$

Constructing a reliable failure criterion rests on the computation of the buckling factors. As proposed by Berthelot (1999), they can be expressed in terms of the coefficients of the matrix $\mathbf{D}$:

$$\Lambda_{m,n} = \left[D_{11}\left(\frac{m}{a}\right)^4 + D_{22}\left(\frac{n}{b}\right)^4 + 2(D_{12} + 2D_{33})\left(\frac{m}{a}\right)^2\left(\frac{n}{b}\right)^2 \right] \frac{\pi^2}{\left(\frac{m}{a}\right)^2 L_x}. \quad (18)$$

The critical buckling factor Λ is then the smallest buckling factor among all $m, n \in \mathbb{N}^*$ and must be regarded as a deterministic function $\mathcal{M}$ of the bending stiffness matrix:

$$\Lambda = \min_{m,n} \Lambda_{m,n} = \Lambda_{m^*,\, n^*} = \mathcal{M}(\mathbf{D}). \quad (19)$$

The optimal values m^* and n^* give the shape of the buckling mode. Under a given load case L_x, buckling occurs if $\Lambda > 1$ and risk assessment aims at quantifying $P_f = \mathbb{P}\{\mathcal{M}(\mathbf{D}) > 1\}$.

4.3. *Simulation study*

A simulation code $\mathcal{R}$ combining several expert feedbacks within a probabilistic mechanical model allows to generate samples according to the underlying density $f_{\mathbf{D}}$. Compared with only being given $\mathbf{X}_0$, having access to $\mathcal{R}$ brings two major benefits: (a) the marginal distributions can be learnt accurately with kernel density estimation, and (b) the theoretical failure probabilities can be estimated trustfully with a crude MC approach. In view of (a), the marginal distributions are assumed to be known exactly and $\mathbf{X}_0 \in \mathbb{R}^{N\times 4}$ is tranformed neatly into $\mathbf{U}_0 \in [0\,;1]^{N\times 4}$. It was observed numerically that $N_B = 3000$ bootstrap replications are required to obtain low-variance estimators $\hat{\delta}_{\hat{\tau}_{ij}}$ of Borgonovo's indices. During the learning phase, copula modeling is performed with statistical tools provided by the R package `VineCopula`. To shrink as much as possible the intrinsic stochasticity of the NAIS algorithm, $N_P = 3000$ particles are incorporated. The auxiliary density $\hat{h}_0$ is reached after a small number of iterations (two or three according to the rare-event under study).

4.4. *Results*

Sensitivity analysis on risk assessment is performed for two cases. The compressive loads are $L_x^{(1)}$ = 540 N/mm and $L_x^{(2)}$ = 560 N/mm. The failure probabilities are respectively equal to $P_f^{(1)} = 1.82 \times 10^{-2}$ and $P_f^{(2)} = 1.04 \times 10^{-3}$. To simplify notation, the input variables D_{11}, D_{12}, D_{22} and D_{33} play the role of the inputs X_1, X_2, X_3 and X_4. Estimates of all first-order Borgonovo's indices are provided in Table 1. The sensitivity analysis procedure has been run ten times, the resulting estimates have been averaged and the corresponding coefficients of variation $\hat{\gamma}_{ij}$ have been computed. $\hat{\tau}_{12}$, $\hat{\tau}_{23}$ and $\hat{\tau}_{24}$ emerge as the most influential Kendall's coefficients on the estimated failure probability $\hat{P}_f$. Even if the estimators $\hat{\delta}_{\hat{\tau}_{ij}}$ seem to display a large variance, rankings between sensitivity indices are preserved from one estimation to the other.

Table 1. Estimation of Borgonovo's indices.

	$L_x^{(1)}$		$L_x^{(2)}$	
$i-j$	$\hat{\delta}_{\hat{\tau}_{ij}}$	$\hat{\gamma}_{ij}$	$\hat{\delta}_{\hat{\tau}_{ij}}$	$\hat{\gamma}_{ij}$
$1-2$	0.089	16 %	0.038	36 %
$1-3$	0.041	31 %	0.020	36 %
$1-4$	0.028	34 %	0.022	25 %
$2-3$	0.086	21 %	0.044	20 %
$2-4$	0.109	14 %	0.044	25 %
$3-4$	0.025	41 %	0.014	30 %

5. Conclusion

In this paper, a global sensitivity analysis technique has been developped to quantify and to rank the influence of the uncertainties conveyed by each pair-copula modeling on the failure probability. The main idea lies in taking into account the variability related to the R-vine specification through the uncertainties affecting Kendall's matrix. This framework avoids to make any prior assumption on the underlying R-vine model but leads to analyze a stochastic model.

At the end of the sensitivity analysis, a pair of variables (or a small subgroup of pairs) is identified as having a stronger impact on risk assessment. Future works would consist in seeking to intervene upstream in order to facilitate and to improve copula modeling for those pairs. Let us suppose that a pair (X_i, X_j) has been targeted by the procedure. It can be easily imagined solliciting some expert knowledge or aquiring new data in order to improve the way c_{ij} is learnt. Then it raises the obvious question of how this additional knowledge could be integrated smoothly into a predefined R-vine model. If the link (i, j) does not belong to the tree T_1, it is not a trivial issue and it deserves further study.

Acknowledgement

The first author is currently enrolled in a PhD program co-funded by ONERA and the Occitanie region. This financial support is gratefully acknowledged. The authors would also particularly like to thank François-Xavier Irisarri and Antoine Hurmane from ONERA's Materials and Structures Department (DMAS) for proposing this numerical study. Further acknowledgment goes to Pierre Derennes and Sylvain Dubreuil from ONERA's Information Processing and Systems Department (DTIS) for fruitful discussions.

References

Aas, K., C. Czado, A. Frigessi, and H. Bakken (2009). Pair-copula constructions of multiple dependence. *Insurance: Mathematics and economics 44*(2), 182–198.

Bedford, T., R. M. Cooke, et al. (2002). Vines–a new graphical model for dependent random variables. *The Annals of Statistics 30*(4), 1031–1068.

Benoumechiara, N., B. Michel, P. Saint-Pierre, and N. Bousquet (2018). Detecting and modeling worst-case dependence structures between random inputs of computational reliability models. *arXiv preprint arXiv:1804.10527*.

Berthelot, J.-M. (1999). *Composite materials: mechanical behavior and structural analysis*. Mechanical engineering series. New York, NY: Springer. OCLC: 845215936.

Der Kiureghian, A. (2008). Analysis of structural reliability under parameter uncertainties. *Probabilistic engineering mechanics 23*(4), 351–358.

Derennes, P., J. Morio, and F. Simatos (2018). Estimation of moment independent importance measures using a copula and maximum entropy framework. In *2018 Winter Simulation Conference (WSC)*, pp. 1623–1634. IEEE.

Dissmann, J., E. C. Brechmann, C. Czado, and D. Kurowicka (2013). Selecting and estimating regular vine copulae and application to financial returns. *Computational Statistics & Data Analysis 59*, 52–69.

Efron, B. (1992). Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics*, pp. 569–593. Springer.

Haff, I. H. et al. (2013). Parameter estimation for pair-copula constructions. *Bernoulli 19*(2), 462–491.

Iooss, B. and P. Lemaître (2015). A review on global sensitivity analysis methods. In *Uncertainty management in simulation-optimization of complex systems*, pp. 101–122. Springer.

Jiang, C., W. Zhang, X. Han, B. Ni, and L. Song (2015). A vine-copula-based reliability analysis method for structures with multidimensional correlation. *Journal of Mechanical Design 137*(6), 061405.

Joe, H. (1996). Families of m-variate distributions with given margins and m(m-1)/2 bivariate dependence parameters. *Lecture Notes-Monograph Series*, 120–141.

Kahn, H. and T. E. Harris (1951). Estimation of particle transmission by random sampling. *National Bureau of Standards applied mathematics series 12*, 27–30.

Morales-Nápoles, O. (2010). Counting vines. In *Dependence modeling: Vine copula handbook*, pp. 189–218. World Scientific.

Morio, J. (2011). Influence of input pdf parameters of a model on a failure probability estimation. *Simulation Modelling Practice and Theory 19*(10), 2244–2255.

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

Rubinstein, R. Y. and D. P. Kroese (2013). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Sklar, M. (1959). Fonctions de répartition à n dimensions et leurs marges. *Publ. inst. statist. univ. Paris 8*, 229–231.

Stöber, J. and U. Schepsmeier (2013). Estimating standard errors in regular vine copula models. *Computational Statistics 28*(6), 2679–2707.

Torre, E., S. Marelli, P. Embrechts, and B. Sudret (2018). A general framework for data-driven uncertainty quantification under complex input dependencies using vine copulas. *Probabilistic Engineering Mechanics*.

Xu, D., M. Xing, Q. Wei, Y. Qin, J. Xu, Y. Chen, and R. Kang (2018). Failure behavior modeling and reliability estimation of product based on vine-copula and accelerated degradation data. *Mechanical Systems and Signal Processing 113*, 50–64.

Zhang, P. (1996). Nonparametric importance sampling. *Journal of the American Statistical Association 91*(435), 1245–1253.