Computer Physics Communications 183 (2012) 937–946

Contents lists available at SciVerse ScienceDirect

Computer Physics Communications

www.elsevier.com/locate/cpc

Estimation of global sensitivity indices for models with dependent variables

S. Kucherenko [a,b,*], S. Tarantola [a], P. Annoni [a]

[a] Joint Research Centre of the European Commission, TP 361, Via E. Fermi, 2749, 21027 Ispra (VA), Italy
[b] Imperial College London, London, SW7 2AZ, UK

ARTICLE INFO

Article history:
Received 15 March 2011
Received in revised form 28 November 2011
Accepted 23 December 2011
Available online 29 December 2011

Keywords:
Global sensitivity analysis
Correlated inputs
Gaussian copula
Quasi Monte Carlo methods
Sobol' sensitivity indices
Sobol' sequences

ABSTRACT

A novel approach for estimation variance-based sensitivity indices for models with dependent variables is presented. Both the first order and total sensitivity indices are derived as generalizations of Sobol' sensitivity indices. Formulas and Monte Carlo numerical estimates similar to Sobol' formulas are derived. A copula-based approach is proposed for sampling from arbitrary multivariate probability distributions. A good agreement between analytical and numerical values of the first order and total indices for considered test cases is obtained. The behavior of sensitivity indices depends on the relative predominance of interactions and correlations. The method is shown to be efficient and general.

1. Introduction

High complexity of models in physics, chemistry and other fields often results in the increased uncertainty in model parameters. In such cases good modelling practices require the use of global sensitivity analysis (SA) [1–3]. Over the last decade ANOVA-based SA has gained acceptance among practitioners in the process of model development and corroboration. The ANOVA decomposition is unique only if the input variables are independent, and methods like that of Sobol' [4] have been developed to deal with independent inputs. However, in many fields such as nuclear engineering [5], mechanical engineering [6], medical decision making [7], extinction risks assessments in biology [8], just to name a few input variables are dependent. The dependency can arise as the result of constraints in the inputs space due to properties of inputs (i.e. composition constraints in material science) or because of the complex input structure where inputs may be themselves the output of some other model or experiment [9]. In these cases the simple description of input uncertainty through independent marginal distribution functions is not adequate. Correct procedures require sampling from the joint and conditional distribution functions of inputs.

Over the past ten years or so many techniques have been proposed to generalize variance-based SA for the case of dependent variables. Tim Bedford [10] applied a Gram–Schmidt orthogonalization of the inputs, followed by the classical decomposition of variance of the new orthogonal set of variables. The proposed decomposition is not unique as it depends on the ordering in which the original variables are included in the orthogonalization procedure resulting in interpretation difficulties. Later, Saltelli and Tarantola [11] proposed an algorithm to identify most important inputs in the case of dependent variables. Each sensitivity index of the first (or higher) order is estimated using its straightforward definition, which numerically results in the so-called brute force double loop Monte Carlo (MC) estimate. The major drawback of this approach is the high computational cost usually required and not always affordable. Recently, Xu and Gertner [12] developed an approach by splitting the contribution of an individual input to the uncertainty of the model output into two components: the correlated contribution and the uncorrelated one. They propose a regression-based method for estimating the correlated and uncorrelated contribution of the inputs. The method relies on the assumption that the relationship between response and parameters is approximately linear. In [13], Li et al. introduced a unified framework which generalizes the ANOVA-HDMR including covariances in the decomposition of the model output variance. Similarly to Xu and Gertner [12], they distinguish between structural and correlative contribution of a given input. The method consists of two steps: (i) an approximation of the model obtained by estimating a (reduced) functional decomposition via a metamodel; (ii) computation of variances and covariances of the model function components with the model output. These yield two sensitivity indices for each input, the structural and correlative part. Although this method represents an improvement in the sensitivity analysis with correlated inputs, it also presents some critical issues such as non-uniqueness of the functional decomposition (how

* Corresponding author at: Imperial College London, London, SW7 2AZ, UK.
E-mail address: s.kucherenko@ic.ac.uk (S. Kucherenko).

0010-4655/$ – see front matter
doi:10.1016/j.cpc.2011.12.020

many components should be included in the decomposition, the choice of the appropriate surrogate model, etc.).

In this paper we propose a generalization of the variance-based sensitivity indices for the case of dependent variables. The generalization does not involve the use of surrogate models, data-fitting procedures or orthogonalization of the input factor space. We derive formulas and numerical estimates for sensitivity indices, analogous to Sobol' formulas for the case of independent variables [4,14]. The values of the sensitivity indices are unique and their interpretation is not ambiguous. A priori knowledge of probability distribution functions is required. In the case of normal distributions an application of formulas is straightforward. In the general case we propose to use a Gaussian copula which is determined by marginal distributions of inputs and their correlation structure.

This paper is organized as follows: the next section presents formulas for the first order effect and total indices for models with dependent variables. Section 3 contains general results for the case of a correlated multivariate normal distribution. It also considers the particular case of a linear model with two correlated bivariate normal variables. In Section 4 the general case of correlated variables each of which represented by its own marginal distribution and a correlation structure is presented. Application of the Gaussian copulas is proposed to reduce the arbitrary case of correlated variables to the case of correlated multivariate normal distribution. Generation of correlated variables is considered in Section 4. MC estimates of the derived formulas are presented in Section 5. Section 6 introduces an MC sampling procedure for calculation of MC estimates. Section 6 presents three test cases. It compares the numerical and analytical results and discusses the convergence rates of the MC and Quasi Monte Carlo (QMC) methods for the derived formulas and for the brute force approach. Finally, conclusions are presented in the last section.

2. Formulas for the first order effect and total indices

We consider a model function $f(x_1,\dots,x_n)$ defined in R^n with an input vector $x=(x_1,\dots,x_n)$. Here x is a real-valued random variable with a continuous distribution function $p(x_1,\dots,x_n)$. It is assumed that $f(x_1,\dots,x_n)$ has a finite variance. Consider an arbitrary subset of the variables $y=(x_{i_1},\dots,x_{i_s})$, $1\leqslant s<n$, and a complementary subset $z=(x_{i_1},\dots,x_{i_{n-s}})$, so that $x=(y,z)$.

The total variance of $f(x_1,\dots,x_n)$ can be decomposed as

$$D=D_y\big[E_z\big(f(y,\bar z)\big)\big]+E_y\big[D_z\big(f(y,\bar z)\big)\big].$$

Here

$$E_z\big(f(y,\bar z)\big)=\int_{R^{n-s}} f(y,\bar z)p(y,\bar z|y)\,d\bar z,$$

$$D_y\big[E_z\big(f(y,\bar z)\big)\big]=\int_{R^s}\big[E_z\big(f(y,\bar z)\big)\big]^2p(y)\,dy-f_0^2,$$

$$D_z\big(f(y,\bar z)\big)=\int_{R^s}\big(f(y,\bar z)\big)^2p(y,\bar z|y)\,d\bar z-f_0^2,$$

$$E_y\big[D_z\big(f(y,\bar z)\big)\big]=\int_{R^s}\big[D_z\big(f(y,\bar z)\big)\big]^2p(y)\,dy.$$

We use notations z and $\bar z$ to distinguish a random vector z generated from a joint probability density function $p(y,z)$ and a random vector $\bar z$ generated from a conditional distribution $p(y,\bar z|y)$. Normalized by the total variance, this expression leads to the equality

$$1=\frac{D_y[E_z(f(y,\bar z))]}{D}+\frac{E_y[D_z(f(y,\bar z))]}{D}. \tag{2.1}$$

2.1. The first order effect index

The ratio

$$S_y=\frac{D_y[E_z(f(y,\bar z))]}{D} \tag{2.2}$$

is known as the first order effect index of the subset y, while the second term in (2.1)

$$S_z^T=\frac{E_y[D_z(f(y,\bar z))]}{D}$$

is known as the total effect of the subset z. For consistency we also present the total effect of the subset y

$$S_y^T=\frac{E_z[D_y(f(\bar y,z))]}{D}. \tag{2.3}$$

Collectively S_y, S_y^T in the case of independent variables are known as Sobol' indices [3,14].

The full expression for the first order effect index is given by the following formula:

$$S_y=\frac{1}{D}\left[\int_{R^s}p(y)\,dy\left[\int_{R^{n-s}}f(y,\bar z)p(y,\bar z|y)\,d\bar z\right]^2-f_0^2\right]. \tag{2.4}$$

Here $f_0=E(f(x))$, $p(y,\bar z|y)$ is a conditional probability density function, $p(y)$ is a marginal distribution function. For majority of practical problems an analytical calculation of S_y is not feasible. In the MC framework a direct application of formula (2.2) requires the so-called double loop of computations also known as the brute force approach. It results in a very high computational cost and hence it is not applicable for practical applications [11].

Expression (2.4) can be written as

$$S_y=\frac{1}{D}\left[\int_{R^s}p(y)\,dy\left[\int_{R^{n-s}}f(y,\bar z)p(y,\bar z|y)\,d\bar z\times\int_{R^{n-s}}f\big(y,\bar z'\big)p\big(y,\bar z'\big|y\big)\,d\bar z'\right]-f_0^2\right]. \tag{2.5}$$

Formula (2.5) allows for overcoming the limitation of the double loop approach although the dimensionality of the integral is increased from n to $2n-s$. Application of formula (2.5) requires generating (i) a random vector (y,z) from the joint distribution $p(y,z)$; (ii) a random vector $(y,\bar z')$ from the conditional distribution $p(y,\bar z'|y)$. The details of the MC scheme are given in the following sections. For the uncorrelated case conditional distributions are reduced to marginal distributions: $p(y,\bar z|y)=p(\bar z)$, $p(y,\bar z'|y)=p(\bar z')$ and formula (2.5) becomes equivalent to the standard Sobol' formula for the first order effect [4,14].

Modified formula: Formula (2.5) can be modified by noticing that

$$f_0^2=\left[\int_{R^n}f(y,z)p(y,z)\,dy\,dz\right]^2=\left[\int_{R^n}f(y,z)p(y,z)\,dy\,dz\int_{R^n}f\big(y',z'\big)p\big(y',z'\big)\,dy'\,dz'\right]. \tag{2.6}$$

Here (y,z) and (y',z') are two different random vectors generated from the joint distribution $p(y,z)$. Combining (2.5) and (2.6) we obtain

$$S_y = \frac{1}{D}\Bigg[\int_{R^s} p(y)\,dy\Bigg[\int_{R^{n-s}} f(y,\bar{z})p(y,\bar{z}|y)\,d\bar{z} \times \int_{R^{n-s}} f\left(y,\bar{z}'\right)p\left(y,\bar{z}'|y\right)d\bar{z}'\Bigg] - \Bigg[\int_{R^n} f(y,z)p(y,z)\,dy\,dz \int_{R^n} f\left(y',z'\right)p\left(y',z'\right)dy'\,dz'\Bigg]\Bigg]$$

which can be further simplified by using the Bayes' formula $p(y,\bar{z}|y)p(y) = p(y,z)$ and taking the common multiplier $\int_{R^n} f(y,z)p(y,z)\,dy\,dz$ out:

$$S_y = \frac{1}{D}\Bigg[\int_{R^n} f(y,z)p(y,z)\,dy\,dz\Bigg[\int_{R^{n-s}} f\left(y,\bar{z}'\right)p\left(y,\bar{z}'|y\right)d\bar{z}' - \int_{R^n} f\left(y',z'\right)p\left(y',z'\right)dy'\,dz'\Bigg]\Bigg]. \tag{2.7}$$

This formula significantly improves the convergence of the numerical estimates for the case of small indices [15,16].

Alternative formula: We can also write (2.4) with respect to y'

$$S_y = \frac{1}{D}\Bigg[\int_{R^s} p\left(y'\right)dy'\Bigg[\int_{R^{n-s}} f\left(y',\hat{z}\right)p\left(y',\hat{z}|y'\right)d\hat{z}\Bigg]^2 - f_0^2\Bigg]$$

in which case (2.7) can be written as

$$S_y = \frac{1}{D}\Bigg[\int_{R^n} f\left(y',z'\right)p\left(y',z'\right)dy'\,dz'\Bigg[\int_{R^{n-s}} f\left(y',\hat{z}\right)p\left(y',\hat{z}|y'\right)d\hat{z} - \int_{R^n} f(y,z)p(y,z)\,dy\,dz\Bigg]\Bigg]. \tag{2.8}$$

Here we used notations z, z', $\hat{z}$ to distinguish random vectors z and z' generated from the joint probability distribution $p(y,z)$ and a random vector $\hat{z}$ generated from a conditional probability distribution function $p(y',\hat{z}|y')$.

In the case of independent inputs an alternative formula (2.8) is reduced to

$$S_y = \frac{1}{D}\Bigg[\int_{R^n} f\left(y',z'\right)p\left(y'\right)p\left(z'\right)dy'\,dz'\Bigg[\int_{R^{n-s}} f\left(y',z\right)p(z)\,dz - \int_{R^n} f(y,z)p(y)p(z)\,dy\,dz\Bigg]\Bigg]. \tag{2.9}$$

The alternative formula allows to reduce the number of function evaluations in the case of independent inputs [17]. For the case of dependent inputs an application of the alternative formula (2.8) requires the same number of function evaluations as formula (2.7). Nevertheless, MC estimates of formula (2.8) can be more efficient than formula (2.5) or (2.7) in conjunction with an application of QMC methods. The details are given in Section 6.

2.2. Total effects

The definition for the total effect index (2.3) is derived from the following relationship:

$$D_y^T = D - D_z\left[E_y\left(f(\bar{y},z)\right)\right]. \tag{2.10}$$

Here

$$E_y\left(f(\bar{y},z)\right) = \int_{R^s} f(\bar{y},z)p(\bar{y},z|z)\,d\bar{y},$$

$$D_z\left[E_y\left(f(\bar{y},z)\right)\right] = \int_{R^{n-s}} \left[E_y\left(f(\bar{y},z)\right)\right]^2 p(z)\,dz - f_0^2.$$

Using (2.5) formula (2.10) can be transformed to the explicit formula for S_y^T

$$S_y^T = \frac{1}{2D}\int_{R^{n+s}} \left[f(y,z) - f\left(\bar{y}',z\right)\right]^2 p(y,z)p\left(\bar{y}',z|z\right)dy\,d\bar{y}'\,dz. \tag{2.11}$$

The details of derivation are given in Appendix A. This formula is the generalization of the Jensen–Sobol' formula [14,18] for the case of dependent variables.

For generating input vectors (y,z) and $(\bar{y}',z)$ a joint probability distribution function $p(y,z)$ and a conditional distribution $p(\bar{y}',z|z)$ are used correspondingly.

In the case of independent variables formula (2.11) is reduced to the well-known Jensen–Sobol' formula [14,18]:

$$S_y^T = \frac{1}{2D}\int_{R^{n+s}} \left[f(y,z) - f\left(y',z\right)\right]^2 p(y)p\left(y'\right)p(z)\,dy\,dy'\,dz. \tag{2.12}$$

3. Correlated normal distributions

In this section we consider models with input vectors which follow normal distributions. In this case an application of the derived formulas is rather straightforward.

3.1. n-Dimensional normal distribution

Consider n-dimensional multivariate normal distribution with mean vector μ and covariance matrix Σ:

$$\Phi_n(x) = \frac{1}{(2\pi)^{n/2}\sqrt{|\Sigma|}}e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}. \tag{3.1}$$

Here $|\Sigma|$ denotes the determinant of matrix Σ. We use a different notation for a multivariate normal distribution function $\Phi_n(x)$ to distinguish it from other distributions which we denote as $p(x)$. The components y, z of the vector x are also normally distributed with mean vectors μ_y, μ_z and covariance matrices Σ_y, Σ_z correspondingly. The mean vector μ and the covariance matrix Σ can be partitioned as

$$\mu = \begin{bmatrix}\mu_y\\ \mu_z\end{bmatrix}, \qquad \Sigma = \begin{bmatrix}\Sigma_y & \Sigma_{yz}\\ \Sigma_{zy} & \Sigma_z\end{bmatrix}. \tag{3.2}$$

The conditional distribution of $\Phi_{n-s}(y,\bar{z}|y)$ is also a normal distribution:

$$\Phi_{n-s}(y,\bar{z}|y) = \frac{1}{(2\pi)^{(n-s)/2}\sqrt{|\Sigma_{zc}|}}e^{-\frac{1}{2}(\bar{z}-\mu_{zc})^T\Sigma_{zc}^{-1}(\bar{z}-\mu_{zc})} \tag{3.3}$$

with mean vector μ_{zc}:

$$\mu_{zc} = \mu_z + \Sigma_{yz}\Sigma_y^{-1}[y-\mu_y] \tag{3.4}$$

and covariance matrix Σ_{zc}:

$$\Sigma_{zc} = \Sigma_z - \Sigma_{zy}\Sigma_y^{-1}\Sigma_{yz}. \tag{3.5}$$

A marginal distribution $p(y)$ has a form:

$$\Phi_s(y) = \frac{1}{(2\pi)^{s/2}\sqrt{|\Sigma_y|}}e^{-\frac{1}{2}(y-\mu_y)^T\Sigma_y^{-1}(y-\mu_y)}. \tag{3.6}$$

Formulas for the first order effect (2.5) and the total indices (2.11) in the case of a model with inputs following multivariate normal distribution are

$$S_y = \frac{1}{D}\Bigg[\int\limits_{R^s} \Phi_s(y)\,dy \Bigg[\int\limits_{R^{n-s}} f(y,\bar{z})\Phi_{n-s}(y,\bar{z}|y)\,d\bar{z} \times \int\limits_{R^{n-s}} f\big(y,\bar{z}'\big)\Phi_{n-s}\big(y,\bar{z}'\big|y\big)\,d\bar{z}'\Bigg] - f_0^2\Bigg],$$

$$S_y^T = \frac{1}{2D}\int\limits_{R^{n+s}} \big[f(y,z) - f\big(\bar{y}',z\big)\big]^2 \Phi_n(y,z)\Phi_s\big(\bar{y}',z\big|z\big)\,dy\,d\bar{y}'\,dz. \tag{3.7}$$

3.2. Bivariate normal distribution

To illustrate the developed technique we consider a simple case of the bivariate normal distribution:

$$\Phi(y,z) = \frac{1}{\sigma_y\sigma_z\sqrt{2\pi(1-\rho^2)}} \times e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_y)^2}{\sigma_y^2} + \frac{(z-\mu_z)^2}{\sigma_z^2} - \frac{2\rho(y-\mu_y)(z-\mu_z)}{\sigma_z\sigma_y}\right]}, \tag{3.8}$$

where ρ is the correlation coefficient between y and z. Here y, z are two elements of the input vector x. In this case moments (3.2) have a simple form:

$$\mu = \begin{bmatrix} \mu_y \\ \mu_z \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} \sigma_y^2 & \rho\sigma_z\sigma_y \\ \rho\sigma_z\sigma_y & \sigma_z^2 \end{bmatrix},$$

and the conditional distribution (3.3) simplifies to

$$\Phi(y,\bar{z}|y) = \frac{1}{\sigma_z\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(\bar{z}-\mu_{zc})^2}{2(1-\rho^2)\sigma_z^2}},$$

where $\mu_{zc} = \mu_z + \frac{\rho\sigma_z}{\sigma_y}[y-\mu_y]$.

The formula for the first order effect index (3.7) has a form:

$$S_y = \frac{1}{D}\Bigg[\int\limits_R \frac{1}{\sqrt{2\pi\sigma_y^2}} e^{-\frac{(y-\mu_y)^2}{2\sigma_y^2}}\,dy \times \Bigg[\int\limits_R f(y,\bar{z}) \frac{1}{\sigma_z\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(\bar{z}-\mu_{zc})^2}{2(1-\rho^2)\sigma_z^2}}\,d\bar{z} \times \int\limits_R f\big(y,\bar{z}'\big) \frac{1}{\sigma_z\sqrt{2\pi(1-\rho^2)}} e^{-\frac{(\bar{z}'-\mu_{zc})^2}{2(1-\rho^2)\sigma_z^2}}\,d\bar{z}'\Bigg] - f_0^2\Bigg]. \tag{3.9}$$

A formula for the total index can be written similarly. However, in this particular case it may be simpler to use definition (2.10) for calculating S_y^T once the first order effect index is calculated.

3.3. Bivariate normal distribution. Linear additive model

Consider an additive model:

$$f(y,z) = a_1 y + a_2 z.$$

Using (3.9) we obtain

$$D_y = (a_1\sigma_y + \rho a_2\sigma_z)^2.$$

In the limiting case of no correlation $\rho = 0$ this expression is reduced to:

$$D_y = a_1^2\sigma_y^2.$$

Noticing that

$$D = a_1^2\sigma_y^2 + a_2^2\sigma_z^2 + 2\rho a_1 a_2\sigma_y\sigma_z$$

it is easy to obtain the value for D_y^T:

$$D_y^T = a_1^2\sigma_y^2(1-\rho^2).$$

Condition $D_y^T \geqslant D_y$ is satisfied only if $\rho \leqslant -\frac{2a_1a_2\sigma_y\sigma_z}{a_1^2\sigma_y^2+a_2^2\sigma_z^2}$ for $\rho > 0$ or $\rho \geqslant -\frac{2a_1a_2\sigma_y\sigma_z}{a_1^2\sigma_y^2+a_2^2\sigma_z^2}$ for $\rho < 0$. If both coefficients a_1, a_2 are positive or negative, then $D_y^T < D_y$ for positive values of ρ. For negative values of ρ $D_y^T < D_y$ if coefficients a_1, a_2 have opposite signs. Such a situation is not possible in the case of independent variables, where it is always $D_y^T \geqslant D_y$.

4. The general case of correlated variables

Although normal distributions are often encountered in practice, there is a need to account for different types of dependences among input variables which are often present in practical applications. In this section we consider the general case of pairwise correlated variables with each variable being represented by its own marginal distribution. We propose to use a Gaussian copula which allows us to reduce the problem of evaluation of main and total indices to the case of the correlated normal distribution. Other copulas can also be used (f.e. Archimedean copulas can be used to account for tail dependence of input variables). Copulas are widely used in applications of financial risk management and actuarial analysis [19,20]. We also note that the presented method is general and it allows using any form of dependency between inputs provided that the joint and conditional probability distributions are known.

4.1. Uniformly distributed random variables

Let $u = u_1,\ldots,u_n$, $u_i \in [0,1]$, $i = 1,\ldots,n$, be uniformly distributed random variables with correlation matrix Σ_u.

Definition. Let $F_n(x)$ be the n-variate cumulative normal distribution function and $F(x_i)$ be the univariate cumulative normal distribution function. The n-variate Gaussian copula function is

$$C(u_1,\ldots,u_n;\Sigma_u) = F_n\big(F^{-1}(u_1),\ldots,F^{-1}(u_n);\Sigma\big). \tag{4.1}$$

Here F^{-1} is the inverse normal cumulative distribution function and $x_i = F^{-1}(u_i)$ [21].

From this definition the joint distribution function for correlated uniformly distributed random variables $p(u_1,\ldots,u_n)$ can be written explicitly:

$$p(u_1,\ldots,u_n) = \frac{\partial^n C(u_1,\ldots,u_n)}{\partial u_1\cdots\partial u_n} = \frac{\Phi_n(x_1,\ldots,x_n;\Sigma)}{\prod_{i=1}^n \Phi(x_i)}. \tag{4.2}$$

Here $\Phi_n(x_1,\ldots,x_n;\Sigma)$ is the multivariate joint normal distribution function (3.1), $\Phi(x_i)$ is the univariate normal distribution function with zero mean and variance σ_i^2.

An application of the Gaussian copula function requires mapping of an original correlation matrix Σ_u to the correlation matrix Σ of the corresponding joint normal distribution function. The Pearson correlation coefficients for uniformly distributed random variables $r(u_i,u_j)$ are equal to

$$r(u_i,u_j) = 12E(u_iu_j) - 3,$$

where $E(u_i u_j) = \iint_{H^2} u_i u_j p(u_i, u_j)\,du_i\,du_j$. Using a relationship $p(u_i, u_j)\,du_i\,du_j = \Phi_2(x_i, x_j; \rho)\,dx_i\,dx_j$, $E(u_i, u_j)$ can be presented as

$$E(u_i u_j) = \iint_{R^2} F(x_i)F(x_j)\Phi_2(x_i, x_j; \rho)\,dx_i\,dx_j,$$

where

$$\Phi_2(x_i, x_j; \rho_{ij}) = \frac{1}{2\pi\sqrt{1-\rho_{ij}^2}} e^{-\frac{x_i^2+x_j^2-2\rho_{ij}xy}{2(1-\rho_{ij}^2)}}.$$

Hence the correlation coefficients for uniformly distributed random variables $r(u_i, u_j)$ and the corresponding correlation coefficients ρ_{ij} of the normal distribution are linked by the following formula:

$$r(u_i, u_j) = 12E\big(F(x_i)F(x_j)\big) - 3. \tag{4.3}$$

Components of the uniformly distributed random vector u are obtained by the transformation $u_i = F(x_i)$, $i = 1, \ldots, n$.

Applying the transformation law of probabilities (4.2) formulas for the first order effect (2.5) and the total indices (2.11) can be presented in the following form:

$$\begin{aligned}
S_y &= \frac{1}{D}\Bigg[\int_{R^s} \Phi_s(y)\,dy \Bigg[\int_{R^{n-s}} f\big(\bar{F}_s(y), \bar{F}_{n-s}(\bar{z})\big)\Phi_{n-s}(y, \bar{z}|y)\,d\bar{z} \\
&\quad \times \int_{R^{n-s}} f\big(\bar{F}_s(y), \bar{F}_{n-s}(\bar{z}')\big)\Phi_{n-s}\big(y, \bar{z}'\big|y\big)\,d\bar{z}'\Bigg] - f_0^2\Bigg], \\
S_y^T &= \frac{1}{2D}\int_{R^{n+s}} \big[f\big(\bar{F}_s(y), \bar{F}_{n-s}(z)\big) - f\big(\bar{F}_s(\bar{y}'), \bar{F}_{n-s}(z)\big)\big]^2 \\
&\quad \times \Phi_n(y, z)\Phi_s\big(\bar{y}', z\big|z\big)\,dy\,d\bar{y}'\,dz.
\end{aligned} \tag{4.4}$$

Here $\bar{F}_s(y)$ is a vector of univariate marginal distribution functions: $\bar{F}_s(y) = (F(x_1), \ldots, F(x_s))$. Similarly can be written all other formulas derived in Section 1.

In comparison with the case of normally distributed variables an application of formulas (4.4) requires two additional steps:

(1) mapping of the original correlation matrix Σ_u to the correlation matrix Σ;
(2) an application of the univariate cumulative normal distribution function $F(x_i)$ to each corresponding normal component x_i for construction the input factor u_i.

4.2. Arbitrary distributions

In the general case of arbitrarily distributed variables $\xi_1, \ldots, \xi_n$ with correlation matrix Σ_ξ and a marginal univariate cumulative distribution function G the n-variate Gaussian copula function has a form:

$$\begin{aligned}
&C\big(G_1(\xi_1), \ldots, G_n(\xi_n); \Sigma_\xi\big) \\
&\quad = F_n\big(F^{-1}\big(G_1(\xi_1)\big), \ldots, F^{-1}\big(G_n(\xi_n)\big); \Sigma\big).
\end{aligned}$$

As in the previous case of uniformly distributed random variables we need to map Σ_ξ to the correlation matrix Σ using the definition:

$$r(\xi_i, \xi_j) = \frac{E[(\xi_i - \bar{\xi}_i)(\xi_j - \bar{\xi}_j)]}{[D(\xi_i)D(\xi_j)]^{1/2}} = \frac{E(\xi_i\xi_j) - (\bar{\xi}_i\bar{\xi}_j)^2}{[D(\xi_i)D(\xi_j)]^{1/2}}, \tag{4.5}$$

where $E(\xi_i, \xi_j)$ is calculated as

$$E(\xi_i\xi_j) = \iint G_i^{-1}\big(F(x_i)\big)G_j^{-1}\big(F(x_j)\big)\Phi_2(x_i, x_j; \rho_{ij})\,dx_i\,dx_j.$$

Here G^{-1} is the inverse cumulative distribution function.

It was shown in [22] that a relationship between ρ_{ij} and $r(\xi_i, \xi_j)$ can be presented as $\rho_{ij} = r(\xi_i, \xi_j)\psi$. Here ψ, satisfying $\psi \geqslant 1$ is a function of $r(\xi_i, \xi_j)$ and marginal distributions $G_i(\xi_i)$ and $G_j(\xi_j)$. For commonly used distributions values of ψ are given in [22]. For a uniform distribution ψ is almost a constant: $\psi \approx 1$. A similar but less general approach was developed in [23].

Components of random vector ξ are obtained by the transformation

$$\xi_i = G_i^{-1}\big(F(x_i)\big), \quad i = 1, \ldots, n. \tag{4.6}$$

Formulas for the first order effect and the total indices have the following form:

$$\begin{aligned}
S_y &= \frac{1}{D}\Bigg[\int_{R^s} \Phi_s(y)\,dy \\
&\quad \times \Bigg[\int_{R^{n-s}} f\big(\bar{G}_s^{-1}\big(\bar{F}_s(y)\big), \bar{G}_{n-s}^{-1}\big(\bar{F}_{n-s}(\bar{z})\big)\big)\Phi_{n-s}(y, \bar{z}|y)\,d\bar{z} \\
&\quad \times \int_{R^{n-s}} f\big(\bar{G}_s^{-1}\big(\bar{F}_s(y)\big), \bar{G}_{n-s}^{-1}\big(\bar{F}_{n-s}(\bar{z}')\big)\big) \\
&\quad \times \Phi_{n-s}\big(y, \bar{z}'\big|y\big)\,d\bar{z}'\Bigg] - f_0^2\Bigg], \\
S_y^T &= \frac{1}{2D}\int_{R^{n+s}} \big[f\big(\bar{G}_s^{-1}\big(\bar{F}_s(y)\big), \bar{G}_{n-s}^{-1}\big(\bar{F}_{n-s}(z)\big)\big) \\
&\quad - f\big(\bar{G}_s^{-1}\big(\bar{F}_s(\bar{y}')\big), \bar{G}_{n-s}^{-1}\big(\bar{F}_{n-s}(z)\big)\big)\big]^2 \\
&\quad \times \Phi_n(y, z)\Phi_s\big(\bar{y}', z\big|z\big)\,dy\,d\bar{y}'\,dz.
\end{aligned} \tag{4.7}$$

Here we use the following notation: $\bar{G}_s^{-1}(\bar{F}_s(y)) = (G_1^{-1}(F(x_1)), \ldots, G_s^{-1}(F(x_s)))$.

There are three additional steps in comparison to the case of normally distributed variables:

(1) mapping of the original correlation matrix Σ_ξ to the correlation matrix Σ;
(2) an application of the univariate cumulative normal distribution function $F(x_i)$ to each corresponding normal component x_i of the input factor ξ_i;
(3) an application of the inverse cumulative distribution G_i^{-1} for construction the input variables ξ_i using transformation (4.6).

5. Monte Carlo estimates

In this section we construct an MC algorithm for numerical estimation of S_y and S_y^T. An MC estimator of the original formula (2.4) has a form:

$$S_y = \frac{\frac{1}{N_y}\sum_{j=1}^{N_y}\big(\frac{1}{N_z}\sum_{k=1}^{N_z} f(y_j, \bar{z}_k)\big)^2 - f_0^2}{D}. \tag{5.1}$$

It requires (i) double loop calculations consisting of sampling N_y points of vector y distributed according to $p(y)$ and N_z points of vector $\bar{z}$ distributed according to a conditional distribution $p(y, \bar{z}|y)$ to calculate the first term in the nominator; (ii) generation of N_{yz} points of a random vector (y, z) according to a joint probability density function $p(y, z)$ to evaluate f_0 and D:

$$f_0 = \frac{1}{N_{yz}}\sum_{l=1}^{N_{yz}} f(y_l, z_l),$$

$$D = \frac{1}{N_{yz}} \sum_{l=1}^{N_{yz}} \big(f(y_l, z_l)\big)^2 - f_0^2. \tag{5.2}$$

The total number of sampled points and function evaluations per one set (y, z) in this case is equal to $N = N_y N_z + N_{yz}$. To achieve a good convergence both N_y and N_z should be large. This algorithm which was considered in [11] requires high computational efforts.

Formula (2.5) allows for using a much more efficient MC estimator similar to the one originally suggested by Sobol' in [4,14] for the case of uncorrelated variables:

$$S_y = \frac{\frac{1}{N}\sum_{j=1}^{N} (f(y_j, z_j) f(y_j, \bar{z}'_j)) - \left(\frac{1}{N}\sum_{j=1}^{N} f(y_j, z_j)\right)^2}{D}. \tag{5.3}$$

Here estimators (5.2) are used for the computation of f_0 and D with the number of sampled points equal to $N_{yz} = N$.

An estimator for the total effect index (2.11) has a form:

$$S_y^T = \frac{\frac{1}{N}\sum_{j=1}^{N} (f(y_j, z_j) - f(\bar{y}'_j, z_j))^2}{2D}. \tag{5.4}$$

For estimation S_y and S_y^T each MC trial requires three function evaluations: $f(y_j, z_j)$, $f(y_j, \bar{z}'_j)$ and $f(\bar{y}'_j, z_j)$. Hence the total number of function evaluations for a set (S_i, S_i^T), $i = 1, \ldots, n$, is equal to $N_F = N(2n + 1)$.

The modified formula (2.7) has the following MC estimator:

$$S_y = \frac{\frac{1}{N}\sum_{j=1}^{N} f(y_j, z_j)(f(y_j, \bar{z}'_j) - f(y'_j, z'_j))}{D}. \tag{5.5}$$

Although it has improved convergence properties, for estimating S_y and S_y^T each MC trial requires four function evaluations: $f(y_j, z_j)$, $f(y'_j, z'_j)$, $f(y_j, \bar{z}'_j)$ and $f(\bar{y}'_j, z_j)$. The total number of function evaluations for a set (S_i, S_i^T), $i = 1, \ldots, n$, in this case is equal to $N_F = N(2n + 2)$.

The estimator for the alternative formula (2.8) for S_y has a form:

$$S_y = \frac{\frac{1}{N}\sum_{j=1}^{N} (f(y'_j, z'_j)(f(y'_j, \hat{z}_j) - f(y_j, z_j)))}{D}. \tag{5.6}$$

Four function evaluations $f(y_j, z_j)$, $f(y'_j, z'_j)$, $f(y'_j, \hat{z}_j)$ and $f(\bar{y}'_j, z_j)$ are required at each MC trial for estimation S_y and S_y^T. The total number of function evaluations for a full set (S_i, S_i^T), $i = 1, \ldots, n$, is equal to $N_F = N(2n + 2)$. In the case of independent variables $f(y'_j, \hat{z}_j) = f(\bar{y}'_j, z_j) = f(y'_j, z_j)$ and N_F is reduced to $N(n + 2)$ [17].

6. Generation of correlated variables

MC estimates presented in the previous section require MC sampling algorithms. Firstly, we present an algorithm for generating n-dimensional random normal variables. The algorithm consists of the following steps:

(a) generate n-dimensional random vector u uniformly distributed between 0 and 1 using random numbers or quasi-random sequences;
(b) transform every element of u_i into a standard normal vector $\tilde{x}_i$ with zero mean and unit variance using the inverse normal cumulative distribution function: $\tilde{x}_i = F^{-1}(u_i)$;
(c) find Cholesky or other suitable decomposition of correlation matrix Σ:

$$AA^T = \Sigma;$$

(d) compute vector x which follows n-dimensional normal distribution (3.1):

$$x = A\tilde{x} + \mu.$$

Secondly, we consider construction of the sets $(y, \bar{z}')$ and $(\bar{y}', z)$ of normal random variables required to calculate (S_y, S_y^T) using standard formulas (2.7), (2.11) or $(y', \hat{z})$ and $(\bar{y}', z)$ if an alternative formula (2.8) is used. Here we only consider the construction of the set $(y, \bar{z}')$ (other sets can be constructed similarly) by following these steps:

(a) generate two independent n-dimensional random vectors u and u' uniformly distributed between 0 and 1. In the case of QMC sampling a vector of the total dimension $2n$ needs to be generated and then split into two vectors u and u'. The initial $(1:n)$ low-dimensional coordinates of the original vector are used to create a vector u, while the later high-dimensional coordinates $(n+1:2n)$ are used to create a vector u';
(b) split u' into two subsets of variables $v' = (u'_{i_1}, \ldots, u'_{i_s})$, $1 \leqslant s < n$, and a complementary subset w' of $n - s$ variables, so that $u' = (v', w')$;
(c) transform vector u into a normal vector x with mean μ and covariance matrix Σ as described above;
(d) split vector x into two subsets (y, z);
(e) transform every element u'_i of the complementary subset w' into a standard normal variate $\tilde{z}_i$ with zero mean and unit variance using the inverse normal cumulative distribution function;
(f) compute μ_{zc} given in (3.4) using vector y;
(g) find Cholesky or other suitable decomposition of correlation matrix Σ_{zc} given by (3.5):

$$A_{zc} A_{zc}^T = \Sigma_{zc};$$

(h) compute vector $\bar{z}'$ which follows the $(n - s)$-dimensional conditional distribution given by (3.3):

$$\bar{z}' = A_{zc}\tilde{z}' + \mu_{zc};$$

(i) combine vectors y and $\bar{z}'$ to obtain $(y, \bar{z}')$.

For the case of arbitrary distributions one can apply a copula-based approach as described in Section 4.

An application of the alternative formula (2.8) requires construction of a set $(y', \hat{z})$ which is formed similarly to a set $(y, \bar{z}')$. The only major difference between these two sets is that a set w' of the original uniform vector u' is used for construction of a vector $\bar{z}'$, while for a vector $\hat{z}$ the w part of the original uniform vector u is used. This difference does not affect the performance of MC estimators (5.5) and (5.6) when random numbers are used for sampling. However, when QMC sampling is used the performance of the alternative estimator (5.6) can be higher than that of the estimator (5.5) due to the specifics of the low discrepancy sequences [24,25]. It is well known that the initial low-dimensional coordinates of the low discrepancy sequences such as Sobol' sequences are more uniformly distributed than the later high-dimensional coordinates [24–26]. Initial low-dimensional coordinates are used by a vector w, while the later high-dimensional coordinates are used by a vector w'. In other words, an application of the alternative formula (2.8) reduces an effective dimension of an integrand. It results in the increased efficiency of QMC sampling [16] and higher performance of the estimator (5.6).

7. Numerical tests

In this section the developed approach is illustrated by three different models with correlated input variables. For first two

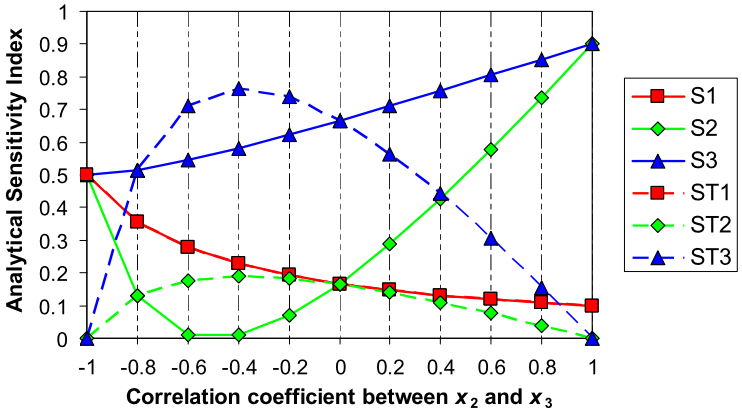

Fig. 1. Test case 1: Evolution of the first and total order indices at different values of ρ estimated at the sample size $N = 2^{13}$. Solid lines refer to S_i; broken lines refer to S_i^T.

models with normally distributed variables analytical solutions are computed and compared with the results of numerical tests. In the third model variables are uniformly distributed and the analytical solution is available only for the case of independent variables. MC estimates (5.5) for S_i and (5.4) for S_i^T were used in all tests. QMC sampling is performed using Sobol' sequences [24,26].

For the first test case we also compared the efficiency of the proposed estimators with that of a modified brute force approach. For the brute force estimates N points $x^{(j)}$, $j = 1, \ldots, N$, are generated from the joint probability distribution. For each random variable $y = x_i$ the sample set $x^{(j)}$, $j = 1, \ldots, N$, is sorted and subdivided in M equally populated partitions (bins) with $N_m = N/M$ points each ($M < N$). Within each bin we calculate the local mean value $E_z(f(y, \bar{z})|y) \approx \sum_{k=1}^{N_m} f(y_k, z_k)$. Note, that $f(y_k, z_k)$ is only an approximation to $f(y_k, \bar{z}_k)$. Finally, the variance of all conditional averages is computed as

$$D_y = \frac{1}{M}\sum_{j=1}^{M}\left(\frac{1}{N_m^j}\sum_{k=1}^{N_m^j} f(y_k, z_k)\right)^2 - f_0^2.$$

The subdivision in bins is done in the same way for all inputs using the same set of sample points.

For the three-dimensional problem it is possible to apply a similar modified brute force approach to calculate the total sensitivity index. However, its extension to higher dimension is not practical. The same set of points can be used for both the first order and total sensitivity indices. To calculate S_1^T we take $y = x_1$, $z = (x_2, x_3)$ and divide the two-dimensional space of z into cells which contain approximately the same number of points. Within the jth cell we estimate $E_j(f(\bar{y}, z)|z) \approx \frac{1}{N_j}\sum_{k=1}^{N_j} f(y_k, z_k)$ where N_j is the number of sample points in the cell j. Then we compute the sample variance over all these conditional averages to obtain S_{23} and apply formula (2.10): in the considered case $S_1^T = 1 - S_{23}$. The problem with this approach is that the correlation between inputs may leave some cells with very few sample points which can cause numerical problems.

7.1. Test case 1. Linear model

Consider the linear model

$$f(x_1, x_2, x_3) = x_1 + x_2 + x_3,$$

where all input variables are normally distributed with zero mean and covariance matrix:

$$\Sigma = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & \rho\sigma \\ 0 & \rho\sigma & \sigma^2 \end{pmatrix}.$$

Table 1
Numerical estimates and analytical values of the first and total order indices for the linear additive model. Notations: the subscript (A) stands for "analytical".

Correlation	S_i	$S_i^{(A)}$	S_i^T	$S_i^{T(A)}$
$\rho = 0.0$				
X_1	0.167	0.167	0.167	0.167
X_2	0.168	0.167	0.167	0.167
X_3	0.669	0.667	0.669	0.667
$\rho = 0.5$				
X_1	0.126	0.125	0.125	0.125
X_2	0.502	0.5	0.094	0.094
X_3	0.784	0.781	0.376	0.375
$\rho = -0.5$				
X_1	0.251	0.250	0.251	0.250
X_2	0.000	0.000	0.188	0.188
X_3	0.564	0.563	0.752	0.750
$\rho = 0.8$				
X_1	0.109	0.109	0.109	0.109
X_2	0.737	0.735	0.039	0.039
X_3	0.855	0.852	0.157	0.157
$\rho = -0.8$				
X_1	0.357	0.357	0.358	0.357
X_2	0.129	0.129	0.129	0.129
X_3	0.515	0.514	0.515	0.514

Analytical values for the first order indices were given in [27], while we calculated the total order indices:

$$S_1 = \frac{1}{2+\sigma^2+2\rho\sigma}, \qquad S_1^T = \frac{1}{2+\sigma^2+2\rho\sigma};$$

$$S_2 = \frac{(1+\rho\sigma)^2}{2+\sigma^2+2\rho\sigma}, \qquad S_2^T = \frac{1-\rho^2}{2+\sigma^2+2\rho\sigma};$$

$$S_3 = \frac{(\sigma+\rho)^2}{2+\sigma^2+2\rho\sigma}, \qquad S_3^T = \frac{\sigma^2(1-\rho^2)}{2+\sigma^2+2\rho\sigma}.$$

Analysis shows that the total indices for the correlated inputs S_2^T and S_3^T tend to zero as $|\rho| \to 1$. It is also interesting to note that $S_i^T \leqslant S_i$, $i = 2, 3$, if $\rho \geqslant 0$ or $\rho \leqslant -\frac{2\sigma}{\sigma^2+1}$.

In all further tests σ was taken to be equal 2. Fig. 1 shows the analytical values of the first and total order sensitivity indices versus ρ. For the chosen σ $S_i^T \leqslant S_i$, $i = 2, 3$, if $\rho \geqslant 0$ or $\rho \leqslant -0.8$, while for variable x_1 $S_1^T = S_1$ for all ρ and both indices never reach zero.

Table 1 shows that the numerical results are in perfect agreement with the analytical values. In all numerical tests the number of sampled points were chosen to be $N = 2^{13}$, i.e. sufficiently large to reach convergence (see also Fig. 2). The computational cost

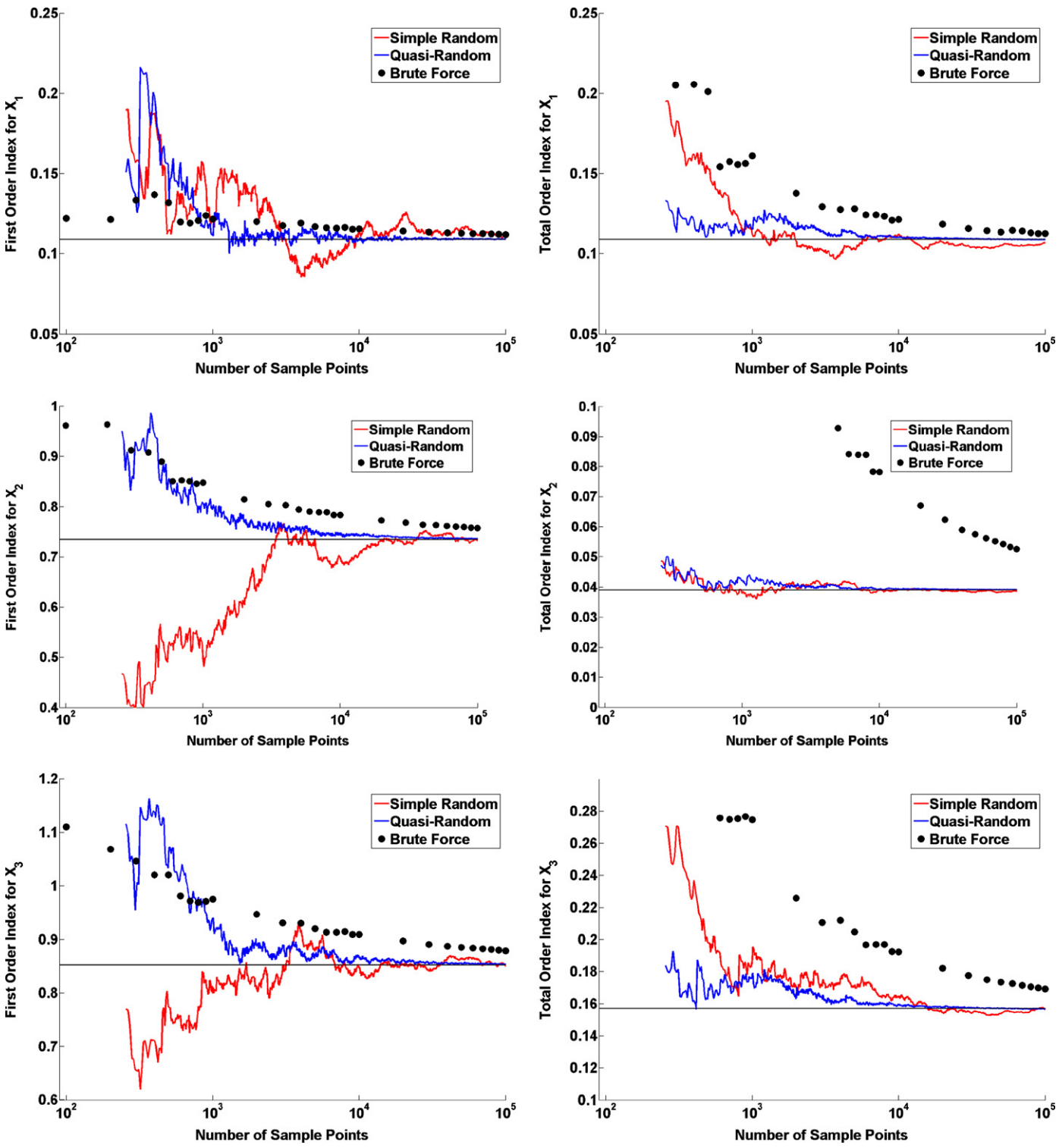

Fig. 2. Test case 1: Convergence plots of the first order and total indices for three variables at $\rho = 0.8$. The red solid line refers to random sampling; the blue dotted line refers to quasi-random sampling; black circles refer to the brute force method. Horizontal lines show analytic values given in Table 1. (For interpretation of colors in this figure, the reader is referred to the web version of this article.)

(i.e. number of model evaluations) required to obtain both the first and total order indices for all three inputs was $N(2n+2) = 2^{16}$.

Fig. 2 shows the numerical convergence of the first and total indices at $\rho = 0.8$. A comparison between the MC and QMC methods shows that the QMC method provides much faster convergence. Among the three methods the brute force approach is by far the least efficient. For monitoring convergence the number of sampled points should be increased incrementally which is not possible for the brute force approach. This approach requires fixing the number of points in advance.

7.2. Test case 2. Portfolio model

The second test case is:

$$f(x_1, x_2, x_3, x_4) = x_1 x_3 + x_2 x_4,$$

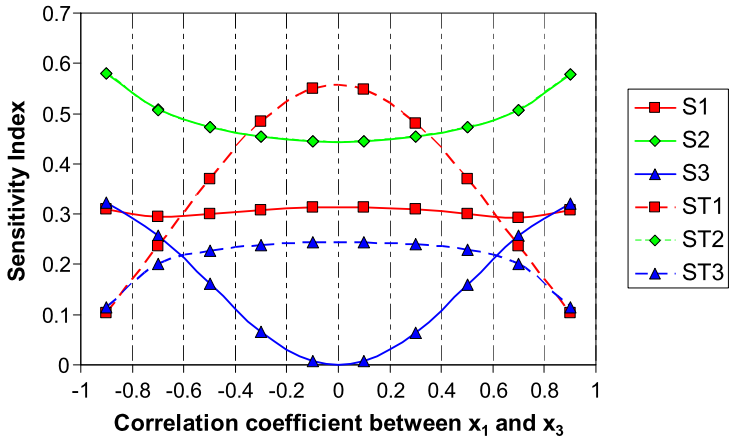

Fig. 3. Evolution of the first order and total order sensitivity indices of the Ishigami function at different correlation values ρ_{13} estimated at the sample size $N = 8192$. Solid lines refer to S_i; broken lines refer to S_i^T.

where $(x_1, x_2, x_3, x_4) \sim N(\mu, \Sigma)$, with $\mu \equiv (0, 0, \mu_3, \mu_4)$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & 0 & 0 \\ \sigma_{21} & \sigma_2^2 & 0 & 0 \\ 0 & 0 & \sigma_3^2 & \sigma_{34} \\ 0 & 0 & \sigma_{43} & \sigma_4^2 \end{pmatrix}.$$

This case is a four variable version of the portfolio model discussed in [28] which includes 6 variables. Analytical values of first and total order indices are:

$$S_1 = \frac{\sigma_1^2(\mu_3 + \mu_4\rho_{12}\frac{\sigma_2}{\sigma_1})^2}{D}, \qquad S_1^T = \frac{\sigma_1^2(1-\rho_{12}^2)(\sigma_3^2+\mu_3^2)}{D};$$

$$S_2 = \frac{\sigma_2^2(\mu_4 + \mu_3\rho_{12}\frac{\sigma_1}{\sigma_2})^2}{D}, \qquad S_2^T = \frac{\sigma_2^2(1-\rho_{12}^2)(\sigma_4^2+\mu_4^2)}{D};$$

$$S_3 = 0, \qquad S_3^T = \frac{\sigma_1^2\sigma_3^2(1-\rho_{34}^2)}{D};$$

$$S_4 = 0, \qquad S_4^T = \frac{\sigma_2^2\sigma_4^2(1-\rho_{34}^2)}{D},$$

where $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i\sigma_j}$ and $D = \sigma_1^2(\sigma_3^2+\mu_3^2)+\sigma_2^2(\sigma_4^2+\mu_4^2)+2\sigma_{12}(\sigma_{34}+\mu_3\mu_4)$. Similarly to the test case 1 values of total order indices $S_i^T \to 0$ as $|\rho_{12}| \to 1$ or $|\rho_{34}| \to 1$.

For numerical tests we used the following values of parameters:

$$\mu = (0, 0, 250, 400),$$

$$\Sigma = \begin{pmatrix} 16 & 2.4 & 0 & 0 \\ 2.4 & 4 & 0 & 0 \\ 0 & 0 & 4\cdot 10^4 & -1.8\cdot 10^4 \\ 0 & 0 & -1.8\cdot 10^4 & 9\cdot 10^4 \end{pmatrix}.$$

Table 2 shows that both analytical and numerical values are in a very good agreement. All numerical results were obtained with using $N = 1500$ sampled points. Given that $n = 4$, the total cost of the numerical estimates is $N(2n+2) = 15\,000$ model runs. Comparison shows that the QMC method is much more efficient than the MC one (plots not presented in this paper).

7.3. Test case 3. Ishigami function

The last test case is the Ishigami function

$$f(x_1, x_2, x_3) = \sin(x_1) + 7\sin^2(x_2) + 0.1x_3^4\sin(x_1)$$

with input variables being uniformly distributed: $-\pi \leqslant x_i \leqslant \pi$, $i = 1, 2, 3$ [29]. This model is widely used as a benchmark for sensitivity tests as it features strong non-linearity and non-monotonicity.

Table 2
Numerical estimates and analytical values of the first and total order indices for the portfolio model. Notations: the subscript (A) stands for "analytical".

	S_i	$S_i^{(A)}$	S_i^T	$S_i^{T(A)}$
X_1	0.503	0.507	0.495	0.492
X_2	0.400	0.399	0.298	0.300
X_3	0.009	0.000	0.185	0.192
X_4	0.007	0.000	0.108	0.108

For the Ishigami function analytical values of first and total order indices are available only for independent inputs.

In this test case we assume the presence of correlation between variables x_1 and x_3 and explore the variation of sensitivity estimates at different values of correlation ranging from $\rho_{13} = -0.9$ to $\rho_{13} = 0.9$. Numerical values of sensitivity indices are obtained using the copula approach presented in Section 4.1. Fig. 3 shows the first order and total order sensitivity indices at different values of ρ_{13}. The first order index for x_1 is only weakly influenced by the correlation while first order indices for x_2 and x_3 increase with correlation. $S_2^T = S_2$ as variable x_2 is not involved neither in interaction nor in correlation with other variables. The total indices for x_1 and x_3 go to zero as $|\rho_{13}| \to 1$. Note that $S_1^T > S_1$, $S_3^T > S_3$ at $|\rho_{13}| \leqslant 0.6$ due to the presence of interaction between x_1 and x_3 but at $|\rho_{13}| > 0.6$ the relationship between the first order and total indices changes: $S_1^T < S_1$, $S_3^T < S_3$ due to the dominant effect of correlation.

8. Conclusions

A novel approach for estimating global sensitivity indices for models with dependent input variables is presented. The method can be seen as a generalization of Sobol' sensitivity indices which have been originally designed for models with independent variables. Formulas for dependent variables are derived for both the first order and total sensitivity indices. The dependency can be in a form of constraints or correlations between variables. A priori knowledge of probability distribution functions is required. In the case of normal distributions the application of the proposed formulas is straightforward. In the general case we propose to use a copula-based approach.

Three different test functions were used for testing and comparison. For two of them we obtained analytical solutions and showed a good agreement between numerical and analytical results. Numerical tests also demonstrated that the rate of convergence of the proposed approach is much higher than that of the brute force method.

Results show an interesting effect of correlation on the sensitivity indices. The first order indices can be higher than the total indices depending on the level of correlation. Total indices always tend to zero as correlation $|\rho| \to 1$.

Acknowledgement

The authors would like to thank A. Saltelli for helpful discussions and his support of his work.

Appendix A

The subsets total variance D_y^T by definition is equal to

$$D_y^T = D - D_z\left[E_y\left(f(\bar{y},z)\right)\right]. \tag{A.1}$$

We notice that D can be written as

$$\begin{aligned} D &= E\left(f^2(y,z)\right) - f_0^2 \\ &= \frac{1}{2}\int_{R^n} f^2(y,z)p(y,z)\,dy\,dz \\ &\quad + \frac{1}{2}\int_{R^n} f^2\left(\bar{y}',z\right)p\left(\bar{y}',z\right)d\bar{y}'\,dz - f_0^2. \end{aligned} \tag{A.2}$$

The first term on the right can be presented in an equivalent form using the Bayes' formula $p(y,\bar{z}|y)p(y) = p(y,z)$:

$$\frac{1}{2}\int_{R^n} f^2(y,z)p(y,z)\,dy\,dz = \frac{1}{2}\int_{R^n} f^2(\bar{y},z)p(z)p(\bar{y},z|z)\,d\bar{y}\,dz. \tag{A.3}$$

We use formula (2.5) written for $D_z[E_y(f(\bar{y},z))]$ and (A.1)–(A.3) to present D_y^T as

$$\begin{aligned} D_y^T &= \frac{1}{2}\int_{R^n} f^2(\bar{y},z)p(z)p(\bar{y},z|z)\,d\bar{y}\,dz \\ &\quad + \frac{1}{2}\int_{R^n} f^2\left(\bar{y}',z\right)p(z)p\left(\bar{y}',z\middle|z\right)d\bar{y}'\,dz - f_0^2 \\ &\quad - \left[\int_{R^{n-s}} p(z)\,dz\left[\int_{R^s} f(\bar{y},z)p(\bar{y},z|z)\,d\bar{y}\right.\right. \\ &\quad \left.\left.\times \int_{R^s} f\left(\bar{y}',z\right)p\left(\bar{y}',z\middle|z\right)d\bar{y}'\right] - f_0^2\right]. \end{aligned}$$

It can be further transformed as

$$\begin{aligned} D_y^T &= \frac{1}{2}\int_{R^n} f^2(\bar{y},z)p(z)p(\bar{y},z|z)\,d\bar{y}\,dz \\ &\quad - \int_{R^{n-s}} p(z)\,dz\left[\int_{R^s} f(\bar{y},z)p(\bar{y},z|z)\,d\bar{y}\right. \\ &\quad \left.\times \int_{R^s} f\left(\bar{y}',z\right)p\left(\bar{y}',z\middle|z\right)d\bar{y}'\right] \\ &\quad + \int_{R^n} f^2\left(\bar{y}',z\right)p(z)p\left(\bar{y}',z\middle|z\right)d\bar{y}'dz. \end{aligned} \tag{A.4}$$

(A.4) is equivalent to

$$D_y^T = \frac{1}{2}\int_{R^{n+s}} \left[f(y,z) - f\left(\bar{y}',z\right)\right]^2 p(y,z)p\left(\bar{y}',z\middle|z\right)dy\,d\bar{y}'\,dz.$$

Hence the final formula for S_y^T has the form:

$$S_y^T = \frac{1}{2D}\int_{R^{n+s}} \left[f(y,z) - f\left(\bar{y}',z\right)\right]^2 p(y,z)p\left(\bar{y}',z\middle|z\right)dy\,d\bar{y}'\,dz.$$

References

[1] A. Saltelli, M. Ratto, S. Tarantola, F. Campolongo, Chem. Rev. 105 (2005) 2811.
[2] I. Sobol', S. Kucherenko, Wilmott (2005) 56.
[3] A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, S. Tarantola, Comput. Phys. Commun. 181 (2010) 259.
[4] I. Sobol', Mat. Model. 2 (1990) 112 (in Russian); translated in: I.M. Sobol', Mat. Model. Comp. Exper. 26 (1993) 407.
[5] J. Jacques, C. Lavergne, N. Devictor, Rel. Eng. Syst. Safety 91 (2006) 1126.
[6] H. Keitel, A. Dimmig-Osburg, Eng. Struct. 32 (2010) 3758.
[7] A.E. Ades, K. Claxton, M. Sculpher, Health Econ. 15 (2006) 373.
[8] S. Ferson, M. Burgman, Biol. Conserv. 73 (1995) 101.
[9] G. Brell, G. Li, H. Rabitz, J. Chem. Phys. 132 (2010) 174103.
[10] T. Bedford, in: Proceedings of the Second International Symposium on Sensitivity Analysis of Model Output, Venice, 1998.
[11] A. Saltelli, S. Tarantola, J. Amer. Statist. Assoc. 97 (2002) 702.
[12] C. Xu, G.Z. Gertner, Rel. Eng. Syst. Safety 93 (2008) 1563.
[13] G. Li, H. Rabitz, P. Yelvington, O. Oluwole, F. Bacon, C. Kolb, J. Schoendorf, J. Phys. Chem. A 114 (2010) 6022.
[14] I. Sobol', Math. Comput. Simulation 55 (2001) 271.
[15] I. Sobol', S. Tarantola, D. Gatelli, S. Kucherenko, W. Mauntz, Rel. Eng. Syst. Safety 92 (2007) 957.
[16] S. Kucherenko, B. Feil, N. Shah, W. Mauntz, Rel. Eng. Syst. Safety 96 (2011) 440.
[17] A. Saltelli, Comput. Phys. Commun. 145 (2002) 280.
[18] M.J.W. Jansen, W.A.H. Rossing, R.A. Daamen, in: J. Grasman, G. Van Straten (Eds.), Predictability and Nonlinear Modelling in Natural Sciences and Economics, Kluwer Academic Publishers, Dordrecht, 1994, p. 334.
[19] U. Cherubini, E. Luciano, W. Vecchiato, Copula Methods in Finance, Wiley, 2004.
[20] A.J. McNeil, R. Frey, P. Embrechts, Quantitative Risk Management: Concepts, Techniques, and Tools, Princeton Series in Finance, 2005.
[21] A. Sklar, Publ. Inst. Statist. Univ. Paris 8 (1959) 229.
[22] P.L. Liu, A. Der Kiureghian, Probab. Eng. Mech. 1 (1986) 105.
[23] I. Sobol, E. Myshetskaya, Monte Carlo Methods Appl. 1 (2003) 67.
[24] P. Bratley, B. Fox, ACM Trans. Math. Software 14 (1988) 88.
[25] H. Niederreiter, Random Number Generation and Quasi-Monte Carlo Methods, SIAM, 1992.
[26] I. Sobol', Math. Comput. Simulation 47 (1998) 103.
[27] S. Da Veiga, F. Whal, F. Gamboa, Technometrics 51 (2009) 452.
[28] A. Saltelli, M. Ratto, S. Tarantola, F. Campolongo, Global Sensitivity Analysis, Wiley, 2004.
[29] T. Ishigami, T. Homma, in: Proceedings of ISUMA, First International Symposium on Uncertainty Modelling and Analysis, 1990, p. 398.