Mutual information for global sensitivity analysis and adaptive-learning surrogate modelling

Umberto Alibrandi
Department of Civil and Architectural Engineering, AArhus University, Denmark, e-mail: ua@cae.au.dk

Lars V. Andersen
Department of Civil and Architectural Engineering, AArhus University, Denmark, e-mail: lva@cae.au.dk

Enrico Zio
MINES ParisTech / PSL Université Paris, Centre de Recherche sur les Risques et les Crises (CRC), Sophia Antipolis, France
Energy Department, Politecnico di Milano, Via La Masa 34, Milano, 20156, Italy
e-mail: enrico.zio@polimi.it

Abstract. In this paper tools of information theory are applied for probabilistic sensitivity analysis and surrogate modelling with adaptive sampling. One of the authors has recently proposed the adoption of the informational coefficient of correlation as a measure of dependence between the random variables, instead of the largely adopted linear coefficient of correlation. First, it is shown that it can be used for probabilistic sensitivity analysis. Moreover, two novel learning functions for adaptive sampling are proposed. The first, called $\mathcal{H}$- function, gives rise to the method AK-H (Adaptive Kriging - Entropy), and it describes the epistemic uncertainty through the entropy. The second, called $\mathcal{MI}$- function, gives rise to the method AL-MI (Active Learning - Mutual Information), which describes the model error through the Mutual Information. The second has the peculiarity that allows the implementation of adaptive learning in any kind of surrogate modelling, even different from Kriging. A simple numerical example shows the features and the potential of the proposed approach.

Keywords: Information theory, informational coefficient of correlation, global sensitivity Analysis, Active Learning, Kriging

1. Introduction

According to structural reliability theory (Ditlevsen and Madsen, 1996; Madsen et al., 1986; Melchers, 1999), the failure probability with respect to an assigned limit state is defined as

$$P_f = \int_{G(\mathbf{x})\leq 0} f_{\mathbf{X}}(\mathbf{x})d\mathbf{x} \tag{1}$$

where $\mathbf{x}$ is an n-dimensional vector collecting the basic random variables $x_1, x_2, ..., x_n$, $y = G(\mathbf{x})$ is the limit state function, $G(\mathbf{x}) = 0$ is the limit state surface separating the failure set $G(\mathbf{x}) \leq 0$ from the safe set $G(\mathbf{x}) > 0$ and $f_{\mathbf{X}}(\mathbf{x})$ is the joint probability density function of the random variables $\mathbf{x}$.

The failure probability can be evaluated by using analytical methods, like the First- and Second-Order Reliability Methods (FORM/SORM) FORM has gained wide popularity because of its simplicity and computational cost; however it presents two known main drawbacks: (1) its approximation may not be adequate for limit state surfaces departing significantly from linearity around the region of probabilistic interest, or in the presence of multiple design points (Der Kiureghian and Dakessian, 1998), and (2) it does not provide information about the degree of accuracy achieved.

Monte Carlo Simulation (MCS) is the most robust procedure but requires excessive computational effort for the evaluation of very small failure probabilities, in particular in its crude form. Smart sampling techniques have been proposed, e.g. Importance Sampling (IS) (Melchers, 1999; Rubinstein and Krose, 2008), Subset Simulation (SS) (Au and Beck, 2001), Line Sampling (De Angelis et al., 2015). These approaches are robust, but they typically remain computationally demanding for industrial applications.

To reduce the computational cost, an alternative strategy is given by the Response Surface Methodology (RSM), which builds a surrogate model of the target limit state function, defined in a simple and explicit mathematical form (Faravelli, 1989). FORM/SORM can, then be considered are particular kinds of response surfaces, which approximate the limit state with linear/quadratic response surfaces passing through the design point $\mathbf{u}^*$. Response surface models can be built to find the design point with reduced computational cost (Bucher and Burgound, 1990; Rajashekhar and Ellingwood, 1993; Zheng and Das, 2000). The approximation provided by these formulations however, shares the same drawbacks of FORM/SORM, and response surfaces aiming at modelling the whole limit state surface have been proposed, like the Moving Least-Square Regression (Bucher and Most, 2008), High-Dimensional Model Representation (HDMR) (Li et al., 2001; Chowdury et al., 2005; Alibrandi, 2014) or the Polynomial Chaos Expansion (PCE) (Sudret and Der Kiureghian, 2002; Choi et al., 2004; Berveiller et al., 2006; Blatman and Sudret, 2010).

Recently, increasing popularity has been gained by the Gaussian process models, also known as Kriging interpolation models (Rasmussen and Williams, 2006; Forrester et al., 2008), which are attractive because of their capability to: (i) interpolate data at chosen training samples and, (ii) to provide measures of local uncertainty for the model predictions. In (Kaymaz, 2005), Kriging is applied for structural reliability analysis, and it is shown that the Gaussian function is suitable for limit states nonlinear around the design point. In (Gaspar et al., 2014), it is shown that a polynomial of order zero (constant) can be considered as a basis function for practical applications, and that Kriging is well tailored for adaptive experimental designs. In Adaptive Kriging (AK), a *learning function* allows selecting the suitable new points of the experimental design. AK has been used jointly with different strategies of sampling: Monte Carlo (AK-MCS) (Echard et al., 2011), Importance Sampling (AK-IS) (Echard et al., 2013; Duborg et al., 2013), Subset Simulation (AK-SS) (Huang et al., 2016). A key factor for the efficiency of AK is represented by the choice of the learning function. The most commonly used are the Expected Feasibility Function (EFF) (Bichon et al., 2008), and the $\mathcal{U}$ learning function (Echard et al., 2011). Some alternative learning functions have been proposed recently (LV et al., 2015; Zhang et al., 2019). This is also the direction pursued in this paper.

In (Alibrandi and Mosalam, 2019; Alibrandi and Mosalam, 2020; Alibrandi and Mosalam, 2021) a framework of data-driven uncertainty quantification and risk analysis based on information theory is proposed; moreover the authors propose the adoption of the *informational coefficient of correlation* as a measure of dependence. This is based on the concept of Mutual Information (MI) and it is capable to describe the full dependencies between random variables. Starting from these findings, in this paper it is proposed an alternative approach for AK based on the Mutual Information (MI), giving rise to AK-MI. First, MI is used for probabilistic sensitivity analysis in order to obtain an initial ranking of the most important variables to be included in the surrogate model. Second, Entropy and MI are adopted to define a learning function for adaptive learning. A simple numerical example shows the potential of the proposed approach.

2. Adaptive Kriging (AK) for structural reliability analysis

At first, as typically done in reliability analysis, a isoprobabilistic transformation toward the normal standard space is pursued, so that the faiure probability is

$$P_f = \int_{G(\mathbf{u})\leq 0} \varphi_n(\mathbf{u})d\mathbf{u} \tag{2}$$

where $\varphi_n(\mathbf{u})$ is the multivariate normal standard distribution, whereas $G(\mathbf{u})$ is the limit state function defined in the $\mathbf{u}$-space.

2.1. Kriging Metamodel

.

Consider a set of m observations collected in the matrix $\mathcal{D}^{(m)} \equiv [\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, ..., \mathbf{d}^{(m)}]$ of order $(n+1)\times m$, where $\mathbf{d}^{(k)} = \{\mathbf{u}^{(k)}, y_k\}$, $k = 1, 2, ..., m$ with $y_k = G\left(\mathbf{u}^{(k)}\right)$. The idea behind Kriging is to approximate the target limit state function $G(\mathbf{u})$ through a Gaussian stochastic process $\epsilon\left(\mathbf{u}\right)$

$$\begin{aligned} \hat{G}\left(\mathbf{u}\right) &= \mu\left(\mathbf{u}\right) + \epsilon\left(\mathbf{u}\right) \\ &= \mathbf{h}^T\left(\mathbf{u}\right)\mathbf{w} + \epsilon\left(\mathbf{u}\right), \end{aligned} \tag{3}$$

where $\mu\left(\mathbf{u}\right)$ is the mean function approximating the output, being $h\left(\mathbf{u}\right)$ a vector collecting a set of p basis functions $h_1(u), h_2(u), ..., h_p(u)$, whereas $\mathbf{w} \equiv \{w_1, w_2, ..., w_p\}$ collects the parameters of the regression. In eq.(3) the contribution $\epsilon\left(\mathbf{u}\right)$ describes the uncertain departure of $\hat{G}\left(\mathbf{u}\right)$ from its mean value and it is modelled through a stationary zero-mean Gaussian process, whose autocovariance function is

$$Cov\left(\mathbf{u}, \mathbf{u}'\right) = \sigma_\epsilon^2 R_\epsilon\left(\mathbf{u}, \mathbf{u}'; \theta_\epsilon\right), \tag{4}$$

where $R_\epsilon\left(\cdot,\cdot\right)$ is a suitable autocorrelation function providing the dependence structure, σ_ϵ^2 is the variance of the process, and θ_ϵ collects all the parameters of $R_\epsilon\left(\cdot,\cdot\right)$. Several correlation functions exist in literature; in practice, square exponential models are often adopted

$$R_\epsilon\left(\mathbf{u},\mathbf{u}';\theta_\epsilon\right)=\exp\left(-\sum_{k=1}^{n}\theta_{\epsilon,k}|\ u-u_k'|^2\right). \tag{5}$$

2.2. Estimation of the Kriging parameters

.

For the given set of m points $\{\mathbf{u}^{(1)},\mathbf{u}^{(2)},...,\mathbf{u}^{(m)}\}$, eq.(3)becomes

$$\mathbf{y}=\mathbf{H}\mathbf{w}+\mathbf{E} \tag{6}$$

where

$$\mathbf{y}=\begin{Bmatrix}G\left(\mathbf{u}^{(1)}\right)\\G\left(\mathbf{u}^{(2)}\right)\\...\\G\left(\mathbf{u}^{(m)}\right)\end{Bmatrix},\qquad \mathbf{H}=\begin{bmatrix}h_1\left(\mathbf{u}^{(1)}\right) & ... & h_p\left(\mathbf{u}^{(1)}\right)\\h_1\left(\mathbf{u}^{(2}\right) & ... & h_p\left(\mathbf{u}^{(2}\right)\\ & ... & \\h_1\left(\mathbf{u}^{(m}\right) & ... & h_p\left(\mathbf{u}^{(m}\right)\end{bmatrix},\qquad \mathbf{E}=\begin{Bmatrix}\epsilon\left(\mathbf{u}^{(1)}\right)\\\epsilon\left(\mathbf{u}^{(2)}\right)\\...\\\epsilon\left(\mathbf{u}^{(m)}\right)\end{Bmatrix}. \tag{7}$$

By assuming that the vector of model error E is Gaussian, the parameters σ_ϵ^2 and w can be determined through the Maximum Likelihood Estimation (MLE):

$$\begin{aligned}\left(\hat{\mathbf{w}},\hat{\sigma}_\epsilon^2\right)&=\underset{\mathbf{w},\sigma_\epsilon^2}{\operatorname{argmax}}\,L\left(\mathbf{w},\sigma_\epsilon^2|\mathbf{y}\right)\\&=\frac{1}{\left(2\pi\sigma_\epsilon^2\right)^{\frac{n}{2}}\sqrt{|\mathbf{R}_\epsilon|}}\exp\left[-\frac{1}{2\sigma_\epsilon^2}\left(\mathbf{y}-\mathbf{HW}\right)^T\mathbf{R}_\epsilon^{-1}\left(\mathbf{y}-\mathbf{HW}\right)\right],\end{aligned} \tag{8}$$

where $\mathbf{R}_\epsilon\left(\theta_\epsilon\right)$ is the matrix of correlation between the training points

$$\mathbf{R}_\epsilon\left(\theta_\epsilon\right)=\begin{bmatrix}R_\epsilon\left(\mathbf{u}^{(1)},\mathbf{u}^{(1)}\right) & ... & R_\epsilon\left(\mathbf{u}^{(1)},\mathbf{u}^{(m)}\right)\\R_\epsilon\left(\mathbf{u}^{(2)},\mathbf{u}^{(1)}\right) & ... & R_\epsilon\left(\mathbf{u}^{(2)},\mathbf{u}^{(m)}\right)\\ & ... & \\R_\epsilon\left(\mathbf{u}^{(m)},\mathbf{u}^{(1)}\right) & ... & R_\epsilon\left(\mathbf{u}^{(m)},\mathbf{u}^{(m)}\right).\end{bmatrix} \tag{9}$$

The solutions of eqs.(8) and (9) are

$$\begin{aligned}\hat{\mathbf{w}}&=\left(\mathbf{H}^T\mathbf{R}_\epsilon^{-1}\mathbf{H}\right)^{-1}\mathbf{H}^T\mathbf{R}_\epsilon^{-1}\mathbf{y},\\\hat{\sigma}_\epsilon^2&=\frac{1}{m}\left(\mathbf{y}-\mathbf{HW}\right)^T\mathbf{R}_\epsilon^{-1}\left(\mathbf{y}-\mathbf{HW}\right).\end{aligned} \tag{10}$$

2.3. Kriging prediction

The Kriging estimator at a new point is by definition a Gaussian random variate $\hat{G}(\mathbf{u})$ obtained as the Best Linear Unbiased Estimator (BLUE) of $G(\mathbf{u})$, conditioned to the observed samples. It is a linear combination of the samples and it the unbiased estimator with minimum variance. Given the set of m training points $\mathbf{d}^{(k)} = \{\mathbf{u}^{(k)}, y_k\}$, $k = 1, 2, ..., m$, the definition of Gaussian process implies that $\{\mathbf{y}, \hat{G}(\mathbf{u})\}$ are jointly Gaussian

$$\begin{Bmatrix} \mathbf{y} \\ \hat{G} \end{Bmatrix} \sim \mathcal{N}\left(\begin{Bmatrix} \mu(\mathbf{u}) \\ \hat{\mu} \end{Bmatrix}, \quad \sigma_\epsilon^2 \begin{bmatrix} \mathbf{R} & \mathbf{r}(\mathbf{u}) \\ \mathbf{r}^T(\mathbf{u}) & 1 \end{bmatrix}\right) \tag{11}$$

where $\mu(\mathbf{u}) = \mathbf{h}^T(\mathbf{u})\,\hat{\mathbf{w}}$ and $\mathbf{r}(\mathbf{u})$ is a vector of order m which collects the correlations between the point $\mathbf{u}$ and the m training points, i.e. $r_i(\mathbf{u}) = R_\epsilon\left(\mathbf{u}, \mathbf{u}^{(i)}\right)$. Using eq.(11), the posterior distribution is obtained

$$f\left(\hat{G}|\mathbf{y}\right) = \mathcal{N}\left(\hat{G}|\mu_{\hat{g}}(\mathbf{u}), \sigma_{\hat{g}}^2(\mathbf{u})\right) \tag{12}$$

where the posterior Kriging mean and variance are

$$\begin{aligned} \mu_{\hat{g}}(\mathbf{u}) &= \mathbf{h}^T(\mathbf{u})\,\hat{\mathbf{w}} + \mathbf{r}^T(\mathbf{u})\,\mathbf{R}_\epsilon^{-1}(\mathbf{y} - \mathbf{H}\hat{\mathbf{w}}), \\ \sigma_{\hat{g}}(\mathbf{u}) &= \sigma_\epsilon^2\left[1 - \mathbf{r}^T(\mathbf{u})\,\mathbf{R}_\epsilon^{-1}\mathbf{r}(\mathbf{u}) + \mathbf{q}^T(\mathbf{u})\left(\mathbf{H}^T\mathbf{R}_\epsilon^{-1}\mathbf{H}\right)^{-1}\mathbf{q}(\mathbf{u})\right]. \end{aligned} \tag{13}$$

with $\mathbf{q}(\mathbf{u}) = \mathbf{H}\mathbf{R}_\epsilon^{-1}\mathbf{r}(\mathbf{u}) - \mathbf{h}(\mathbf{u})$.

The Kriging surrogate has several interesting features. First, at the training points the prediction is exact and the corresponding variance is zero. Second, when the number of training points increases, the overall variance of the process decreases (i.e. $\sigma_\epsilon \to 0$ when $m \to \infty$). Since the prediction $\hat{G}(\mathbf{u})$ is a Gaussian random variable of mean $\mu_{\hat{g}}(\mathbf{u})$ and variance $\sigma_{\hat{g}}^2(\mathbf{u})$, it is possible to derive confidence bounds over the prediction. The variance process $\sigma_{\hat{g}}^2(\mathbf{u})$ is typically used as an error measure of the epistemic uncertainty. If the prediction of $\hat{G}(\mathbf{u})$ in a point $\mathbf{u}_{new}$ (different from the previously selected m training points) shows high variance $\sigma_{\hat{g}}^2(\mathbf{u}_{new})$, this implies non-negligible epistemic uncertainty of the model prediction $\mu_{\hat{g}}(\mathbf{u}_{new})$. Therefore, to improve the global accuracy of the surrogate, it can be convenient choose a new training point $\mathbf{u}^{(m+1)} \equiv \mathbf{u}_{new}$. This last feature allows to defining adaptive sampling plans for the modelling of the response surface.

2.4. Adaptive Kriging (AK)

To address AK, in literature, is defined a *learning function* capable of detecting a new training point $\mathbf{u}^{(m+1)}$ in a way to improve the overall accuracy of the Kriging metamodel. Typically, in AK the following steps are followed:

1. Initialization
   a) Generate a large set of N_c training candidates $\mathcal{P}_c \equiv \{\mathbf{u}_c^{(1)}, \mathbf{u}_c^{(2)}, ..., \mathbf{u}_c^{(Nc)}\}$
   b) Generate an initial training set of m points $\mathcal{P}^{(m)} \equiv \{\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, ..., \mathbf{u}^{(m)}\}$, with $\mathbf{u}^{(k)} \in \mathcal{P}_c$, $m \ll N_c$
   c) Evaluate the limit state for the m training points, $y_k = G\left(\mathbf{u}^{(k)}\right)$
   d) Collect the dataset in $\mathcal{D}^{(m)} \equiv \{\mathbf{d}^{(1)}, \mathbf{d}^{(2)}, ..., \mathbf{d}^{(m)}\}$, with $\mathbf{d}^{(k)} \equiv \{\mathbf{u}^{(k)}, y_k\}$

2. Build a first Kriging model $\hat{G}_{(0)}\left(\mathbf{u}|\mathbf{w},\sigma_\epsilon^2,\theta_\epsilon\right)$ trained over $\mathcal{D}^{(m)}$
3. Evaluate the failure probability $P_{f,(0)}$ using the first Kriging model $\hat{G}_{(0)}$
4. Choose the learning function $\mathcal{L}(\mathbf{u})$
5. Adaptive sampling
   a) $r=0$
   b) $r=r+1$
   c) Through the selected learning function $\mathcal{L}(\mathbf{u})$, pick from the set $\mathcal{P}_c$ the best "new point" of the sampling plan $\mathbf{u}^{(m+r)}$; no evaluation of the limit state is needed in this stage
   d) Evaluate the limit state for the chosen point, $y_{m+r}=G\left(\mathbf{u}^{(m+r)}\right)$
   e) Collect the dataset in $\mathcal{D}^{(m+r)}\equiv\{\mathbf{d}^{(1)},\mathbf{d}^{(2)},...,\mathbf{d}^{(m+r)}\}$
   f) Build a Gaussian model $\hat{G}_{(r)}\left(\mathbf{u}|\mathbf{w},\sigma_\epsilon^2,\theta_\epsilon\right)$ trained over $\mathcal{D}^{(m+r)}$
   g) Evaluate the failure probability $P_{f,(r)}$ using the updated Kriging model $\hat{G}_{(r)}$
   h) Check convergence: if satisfied, stop, otherwise go to step 4(b).

A largely adopted learning function is the so-called $\mathcal{U}$-function (Echard et al., 2011)

$$\mathcal{U}(\mathbf{u})=\frac{|\mu_{\hat{G}}(\mathbf{u})|}{\sigma_{\hat{G}}(\mathbf{u})} \tag{14}$$

which is adopted in the approach AK-MCS (Adaptive Kriging + Monte Carlo Simulation). A small value of $\mathcal{U}(\mathbf{u})$ implies that the Kriging model is close to the limit state surface (i.e. $\mu_{\hat{G}}(\mathbf{u})=0$) or the uncertainty epistemic $\sigma_{\hat{G}}(\mathbf{u})$ is large, or both. The selected new point of the adaptive procedure has the minimum value of $\mathcal{U}(\mathbf{u})$

$$\mathbf{u}^{(m+1)}=\underset{\mathbf{u}\in\mathcal{P}_c}{\operatorname{argmin}}\left[\mathcal{U}\left(\mathbf{u}_c^{(1)}\right),\mathcal{U}\left(\mathbf{u}_c^{(2)}\right),...,\mathcal{U}\left(\mathbf{u}_c^{(N_c)}\right)\right]. \tag{15}$$

3. Information Theory for Uncertainty Quantification and Structural Reliability Analysis

Recently, one of the authors has underlined the relationships between Information Theory with Uncertainty Quantification and Risk Analysis (Alibrandi and Mosalam, 2019; Alibrandi and Mosalam, 2020; Alibrandi and Mosalam, 2021). In the following, some findings that are relevant for active-learning surrogate modelling design are discussed.

3.1. Copula

In the basic problem of structural reliability, see eq.(1), the set of basic random variables $x_1, x_2, ..., x_n$ is given, whose marginal Probability Density Functions (PDF) and Cumulative Distribution Function (CDF) are $f_i(x_i)$ and $F_i(x_i)$, respectively. Consider the transformation $v_i = F_i(x_i)$, $i = 1, 2, ..., n$. The vector $\mathbf{v} = \{v_1, v_2, ..., v_n\}$ has uniformly distributed marginals, whereas the copula is the joint CDF $F(v_1, v_2, ..., v_n)$. The copula describes the *stochastic dependence* between the random variables and it is unique. The theorem of Sklar allows to describe any multivariate distribution as the product of its marginal PDFs $f_i(x_i)$ and a copula density $c(v_1, v_2, ..., v_n)$,

$$f\left(x_1, x_2, ..., x_n\right) = c\left(v_1, v_2, ..., v_n\right) f_1\left(x_1\right) f_2\left(x_2\right) ... f_n\left(x_n\right). \tag{16}$$

Many copulas have been proposed to model the dependence between random variables, and each one of them imposes a different dependence structure on the data. One commonly applied parametric copula is the Gaussian, whose copula density is

$$c_G\left(\mathbf{v}\right) = \frac{1}{\sqrt{|\mathbf{R}_c|}} \exp\left[-\frac{1}{2}\mathbf{u}^T\left(\mathbf{R}_c^{-1} - \mathbf{I}\right)\mathbf{u}\right], \tag{17}$$

where $\mathbf{u} = \{u_1, u_2, ..., u_n\}$ collects a set of normal standard random variables mapped from $\mathbf{v}$, i.e. $u_i = \Phi^{-1}\left(v_i\right)$, with $\Phi(\cdot)$ the normal standard CDF, and $\mathbf{R}_c$ the matrix of correlation of the copula (different from the matrix of correlation $\mathbf{R}_\epsilon$ of the model error of the Kriging model), whose elements may be defined through linear correlations $R_{ij} = \rho\left(v_i, v_j\right)$. For further details about copula, see (Embrechts et al., 2000; Dutfoy and Lebrun, 2009; Alibrandi and Mosalam, 2020; Alibrandi and Mosalam, 2021).

3.2. Entropy

In information theory, the entropy of a random variable is a *measure of uncertainty* and it can be interpreted as the degree of information that the observation of the variable gives. The (differential) entropy of a continuous-valued random variable X with probability density function $f(x)$ is

$$H\left(x\right) = -\int f\left(x\right)\log f\left(x\right)\ dx. \tag{18}$$

For a Gaussian random variable of standard deviation σ the entropy is

$$H_G\left(x\right) = \frac{1}{2}\left(1 + \log 2\pi\right) + \log\sigma. \tag{19}$$

In the case of Gaussian distribution, the entropy (i.e. degree of uncertainty) is proportional to the standard deviation; this is expected, since the uncertainty of a Gaussian distribution is described uniquely by σ. Given the limit state function $y = G\left(x_1, x_2, ..., x_n\right)$, the entropy $H\left(G\right)$ provides the uncertainty over G

$$H\left(G\right) = -\int f\left(g\right)\log f\left(g\right)\ dg. \tag{20}$$

If in particular, the limit state function $G(\mathbf{u})$ follows a Gaussian distribution or it is well approximated by a parametric distribution, then its entropy can be determined in closed form as a function of its parameters.

3.3. Mutual Information

The mutual information $I(x_1, x_2)$ of two random variables X_1 and X_2 measures the information that they share. Thus, it is a measure of their dependence and it is defined as

$$MI(x_1, x_2) = \iint f(x_1, x_2) \log \frac{f(x_1, x_2)}{f_1(x_1) f_2(x_2)} \, dx_1 \, dx_2, \tag{21}$$

Different from the coefficient of correlation ρ, the MI is able to describe full dependence between variables, not just linear correlations (i.e. second-order dependencies). While the attractive properties of the mutual information as a measure of dependence are well known since long time, it was not readily adopted in statistics and uncertainty quantification because it was originally applied mostly to discrete random variables, and its evaluation is in such cases quite involved. This gap of knowledge has been analyzed in (Alibrandi and Mosalam, 2020; Alibrandi and Mosalam, 2021), where connections with copulas have been underlined. By replacing eq.(16) into eq.(21) one has

$$MI(x_1, x_2) = \iint f(x_1, x_2) \log c(v_1, v_2) \, dv_1 \, dv_2 = E[c(v_1, v_2)], \tag{22}$$

and the MI is evaluated as the expected value of the logarithm of the copula density. In the case of Gaussian copula, see eqs. (17) and (22), one has

$$MI_G(x_1, x_2) = -\frac{1}{2} \log |\mathbf{R}_c|. \tag{23}$$

3.4. Informational Coefficient of Correlation

The informational coefficient of correlation ρ_{MI} is defined as

$$\rho_{MI}(x_1, x_2) = \sqrt{1 - \exp[-2MI(x_1, x_2)]} \tag{24}$$

and it is a *proper measure of dependence*. In particular $\rho_{MI} = 0$ for independent (not just uncorrelated) random variables, whereas for perfect dependence $\rho_{MI} = 1$. Also, ρ_{MI} does not suffer any limitations of the linear coefficient of correlation, and they coincide in the case of joint Gaussian distribution. For further details see (Linfoot, 1957; Alibrandi and Mosalam, 2020; Alibrandi and Mosalam, 2021)

4. Global Sensitivity Analysis

Given the limit state function $y = G(u_1, u_2, ..., u_n)$, the Mutual Information $MI(G, u_k)$ can be interpreted as the *reduction in uncertainty* about G after observing u_k:

$$MI\left(G,u_k\right)=H\left(G\right)-H\left(G|u_k\right), \tag{25}$$

where $H(G)$ is the entropy (i.e. uncertainty) of G, whereas $H(G|u_k)$ is the conditional entropy of G given u_k. This shows, also, that on average conditioning reduces entropy. In Fig.1 a Diagram of Venn is shown, where the area of the sets represents their uncertainty, whereas their intersection is the mutual information.

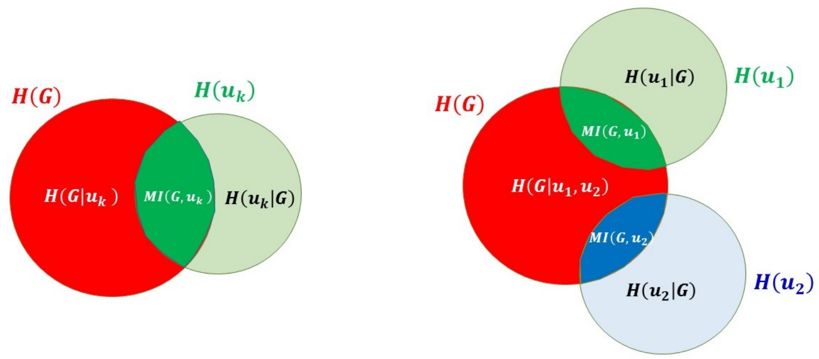

Figure 1. Diagrams of Venn for Mutual Information

Since the normal standard random variables are independent, their mutual informations are zero, i.e. $MI(u_i,u_j)=0$; thus, since there are no interactions between the variables $u_1,u_2,...,u_n$, the informational coefficient of correlation $\rho_{MI}\left(G,u_k\right)$ between G and u_k provides a proper measure of the importance of the variables.

$$\delta_k=\rho_{MI}\left(G,u_k\right) \tag{26}$$

where δ_k are the *informational sensitivity indices.*

5. Adaptive Kriging based on the Mutual Information (AK-MI)

5.1. Adaptive Kriging based on the Entropy: AK-H

Since the entropy can be considered as a generalized measure of uncertainty, which includes the standard deviation in the case of Gaussian random variables, following the idea of the $\mathcal{U}$-function, see eq.(14) (Echard et al., 2011), a learning function based on the entropy of $\hat{G}\left(\mathbf{u}\right)$ is proposed:

$$\mathcal{H}\left(\mathbf{u}\right)=\frac{|\mu_{\hat{G}}\left(\mathbf{u}\right)|}{H_{\hat{G}}\left(\mathbf{u}\right)}=\frac{|\mu_{\hat{G}}\left(\mathbf{u}\right)|}{\frac{1}{2}\left(1+\log 2\pi\right)+\log\sigma_{\hat{G}}\left(\mathbf{u}\right)}. \tag{27}$$

A small value of $\mathcal{H}\left(\mathbf{u}\right)$ implies that the Kriging model is close to the limit state surface (i.e. $\mu_{\hat{G}}\left(\mathbf{u}\right)=0$) or the epistemic uncertainty $H_{\hat{G}}\left(\mathbf{u}\right)$ is large, or both. The selected new point of the adaptive procedure has the minimum value of $\mathcal{H}\left(\mathbf{u}\right)$:

$$\mathbf{u}^{(m+1)} = \underset{\mathbf{u}\in\mathcal{P}_c}{\operatorname{argmin}} \left[\mathcal{H}\left(\mathbf{u}_c^{(1)}\right), \mathcal{H}\left(\mathbf{u}_c^{(2)}\right), ..., \mathcal{H}\left(\mathbf{u}_c^{(N_c)}\right) \right]. \tag{28}$$

5.2. Adaptive Learning based on the Mutual Information: AL-MI

Typically, adaptive learning is implemented together with Kriging for its capabilities to describe the error of the epistemic uncertainty. However many well-performing surrogate models do not manage this feature: Thus, they are not used inside a procedure of Adaptive Learning. Alternatively, variants of these methods have been proposed in order to describe the model uncertainty.

Having in mind the adoption of Adaptive Learning with surrogate models different from Kriging, a second learning function is here proposed. Given the limit state function $G = G(u_1, u_2, ..., u_n)$, the total uncertainty is

$$\begin{aligned} H(G, u_1, u_2, ..., u_n) &= H(G|u_1, u_2, ..., u_n) + H(u_1, u_2, ..., u_n) \\ &= H(G|u_1, u_2, ..., u_n) + \sum_{k=1}^{n} H(u_i) = \\ &= H(G|u_1, u_2, ..., u_n) + \frac{n}{2}(1 + \log 2\pi), \end{aligned} \tag{29}$$

and the total uncertainty is reduced to the conditional entropy of G. From Fig.1 it is seen that the maximum uncertainty over G is equivalent to the minimimum of the sum of the mutual information between G and the random variables $u_1, u_2, ..., u_n$, that is

$$\max_{\mathbf{u}} H(G|u_1, u_2, ..., u_n) = \min_{\mathbf{u}} \sum_{k=1}^{n} MI(G, u_k). \tag{30}$$

Since the informational coefficient of correlation is bounded between zero and one, $0 \le \rho_{MI} \le 1$, a proxy in eq.(27) for the epistemic uncertainty is $1 - \rho_{MI}(\mathbf{u})$. The proposed learning function based on the mutual information is, then,

$$\mathcal{MI}(\mathbf{u}) = \frac{|\mu_{\hat{G}}(\mathbf{u})|}{1 - \rho_{MI}(\mathbf{u})}. \tag{31}$$

The selected new point of the adaptive procedure has the minimum value of $\mathcal{MI}(\mathbf{u})$

$$\mathbf{u}^{(m+1)} = \underset{\mathbf{u}\in\mathcal{P}_c}{\operatorname{argmin}} \left[\mathcal{MI}\left(\mathbf{u}_c^{(1)}\right), \mathcal{MI}\left(\mathbf{u}_c^{(2)}\right), ..., \mathcal{MI}\left(\mathbf{u}_c^{(N_c)}\right) \right]. \tag{32}$$

6. Numerical Application

Let us consider a limit state function with multiple design points, modified from (Der Kiureghian and Dakessian, 1998)

$$g(u_1, u_2) = 3 - u_2 - 0.3u_1^2, \tag{33}$$

where u_1 and u_2 are normal standard random variables. All the reliability computations have been developed through OpenAIUQ software of data-driven uncertainty quantification and risk-analysis (Alibrandi et al., 2021).

6.1. Probabilistic sensitivity analysis

First, a sensitivity analysis is pursued. The informational sensitivity indices are $\delta_1 \equiv \rho_{MI}(G, u_1) = 0.33$ and $\delta_2 \equiv \rho_{MI}(G, u_2) = 0.92$, see eq.(26).The results show that both variables are important, and that to reduce the uncertainty over G it is more interesting to gain information about u_2. Note that the informational coefficient of correlation ρ_{MI} may detect also nonlinear correlations, which is not so with the linear coefficient of correlation ρ. For sake of illustration, in Fig.2 it is shown the limit state function $y = G(u_1, u_2)$, and the scatter plots of the points (u_1, G) and (u_2, G). The application of the linear coefficient of correlation ρ provides $\rho(G, u_1) = 0.0$ and $\rho(G, u_2) = 0.92$. As expected, only linear correlation may be well estimated by ρ.

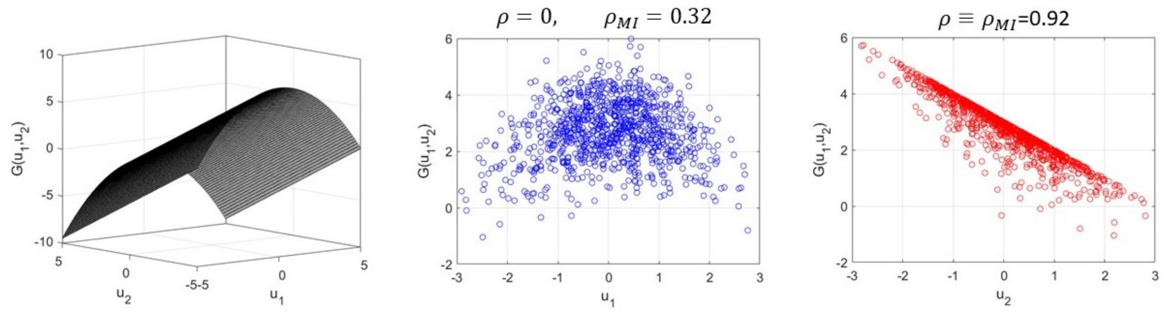

Figure 2. Sensitivity of the limit state function

6.2. Adaptive Kriging

The "reference" failure probability $P_f = 9.737 \times 10^{-3}$ has been evaluated as the average values of 50 MCS runs, each one with a coefficient of variation $\nu = 1\%$. The design point is found to be $\mathbf{u}^* = \{2.11, 1.66\}$; the FORM solution provides $P_{f,FORM} = 3.60 \times 10^{-3}$, with a relative error $e_{FORM} = 63.02\%$. For this toy example, aimed at benchmarking the procedure, the set of the candidate points $\mathcal{P}_c$ includes a uniform grid $u_1 \in [-5, 5]$, $u_2 \in [-5, 5]$, with a step $\Delta u = 0.10$.

For the Kriging model, an initial sampling plan of $m = 9$ points is chosen, distributed along an uniform grid $u_1 \in [-1, 1]$, $u_2 \in [-1, 1]$ with a step $\Delta u = 1$, see Fig.3. To analyze the capabilities of convergence of Kriging, the design point is not included in the initial training set $\mathcal{D}^{(9)}$. A first Kriging model $\hat{G}_{(0)}$ is evaluated and represented in Fig.3 (blue line), together with the target limit state function (black line) and the initial training points (blue points). In the right panel of Fig.3

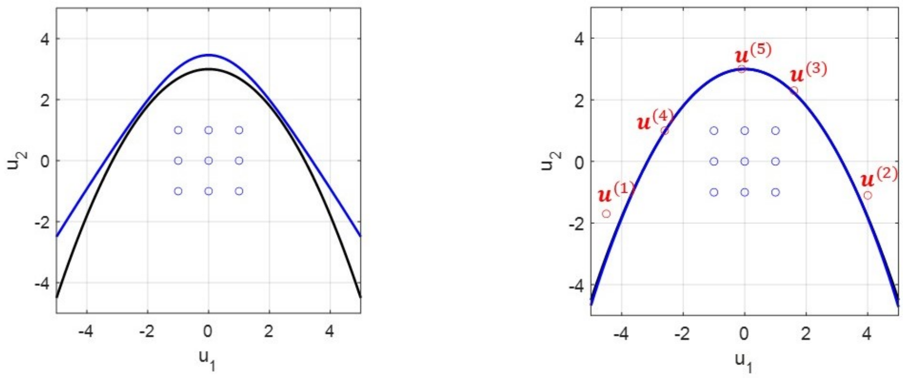

Figure 3. Adaptive Kriging, limit state

the first five points of the adaptive plan are shown (red dots), together with the corresponding Kriging model $\hat{G}_{(5)}$ (blue line).

To check the convergence properties, we represent in Fig.4 (on the left) the loss function of AK-MI, evaluated with respect to all the candidate points of the set $\mathcal{P}_c$. On the right panel of the Figure, it is represented the relative error of FORM (magenta line) and AK (blue line) with respect to the target failure probability P_f. The failure probabilities with Kriging model are estimated through MCS with a coefficient of variation $\nu = 5\%$. It is seen that after only two points of the adaptive sampling there is an error of approximately 10%, which is inside the range of the chosen coefficient of variation. The red line represents the average value of the probabilities estimated by the Kriging models $\hat{G}_{(k)}$ when more training points are added, which shows the convergence toward the target failure probability P_f.

7. Conclusions

In this paper, we have shown the application of information theory for probabilistic sensitivity analysis and surrogate modelling with adaptive sampling. One of the authors has recently proposed the adoption of the informational coefficient of correlation as a measure of dependence between the random variables, instead of the largely adopted linear coefficient of correlation.

In this paper two main contributions are introduced. First, it is shown that the informational coefficient of correlation ρ_{MI} can be used for probabilistic sensitivity analysis. One application shows how it outperforms the classical coefficient of correlation ρ. With reference to surrogate modelling with adaptive sampling, typically the epistemic uncertainty is modelled through the variance of the Gaussian process. Two novel learning functions for adaptive sampling are proposed. The first, called the $\mathcal{H}$- function, giving rise to the method AK-H (Adaptive Kriging - Entropy), describes the model error through the entropy; the second, called the $\mathcal{MI}$- function, giving rise to the method AL-MI (Active Learning - Mutual Information), describes the model error through the Mutual Information.

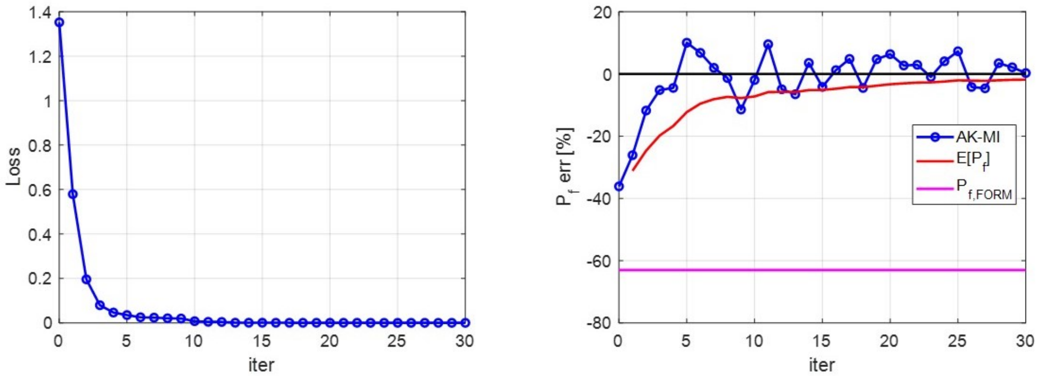

Figure 4. Loss function and error of the Kriging model

The $\mathcal{MI}$-function has the peculiarity that it allows the implementation of adaptive learning in any kind of surrogate modelling, which can differ from Kriging. This capability is attractive, given the high number of surrogate models which do not provide estimates of errors over their predictions. Further research has to be developed to check the computational performances of the proposed approaches, also in comparison with the other learning functions existing in the literature.

References

Alibrandi, U. A response surface method for stochastic dynamic analysis. *Reliab. Eng. Syst. Saf.*, 26:44–53, 2014.

Alibrandi, U. and L. V. Anderssen and K. M. Mosalam OpenAIUQ: data driven uncertainty quantification, risk analysis for risk-informed digital twins. In: *31st European Safety and Reliability Conference*, Angers, France, 2021

Alibrandi, U. and K. M. Mosalam Distribution with Independent Components for Uncertainty Quantification and Structural Reliability Analysis. In: *P13th International Conference on Applications of Statistics and Probability in Civil Engineering, ICASP13*, Seoul, South Korea, 2019

Alibrandi, U. and K. M. Mosalam Information Theory for data-driven Risk Analysis. In: *29th European Safety and Reliability Conference*, Hannover, Germany, 2020

Alibrandi, U. and K. M. Mosalam Information Theory and Machine Learning: Uncertainty Quantification and Risk Analysis revisited. *Reliability Engineering and System Safety*, (submitted)

Au, S. K. and J. L. Beck Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001.

Berveiller, M. and B. Sudret and M. Lemaire Stochastic finite elements: a non-intrusive approach by regression. *Eur. J. Comput. Mech.*, 15(1-3):81–92, 2006.

Bichon, B. J. and M. S. Eldred and L. P. Swiler and S. Mahadevan and J. M. MaFarland Efficient global reliability analysis for nonlinear implicit performance functions. *AIAA J*, 46(10):2459–2468, 2008.

Blatman, G. and B. Sudret An adaptive algorithm to build-up sparse polynomial chaos expansion for stochastic finite elements. *Prob. Eng. Mech.*, 25(2):183–197, 2010.

Bucher, C. and U. Burgound A fast and efficient response surface approach for structural reliability problems. *Structural Safety*, 7:57–66, 1990.

Bucher, C. and T. Most A comparison of approximate response functions in structural reliability analysis. *Probab. Eng. Mech.*, 23:154–163, 2008.

Choi, S. and R. Grandhi and R. Canfield Structural Reliability under non-Gaussian stochastic behavior. *Computers & Structures*, 82:1113–1121, 2004.

Chowdury, R. and B. N. Rao and A. M. Prasad High-dimensional model representation for structural reliability analysis. *J. Commun. Numer. Methods Eng.*, 25:301–307, 2005.

De Angelis, M. and E. Patelli and M. Beer Advanced line sampling for efficient robust reliability analysis. *Structural Safety*, 52:170–182, 2015.

Der Kiureghian, A. Analysis of structural reliability under parameter uncertainties. *Probabilistic Engineering Mechanics*, 23:351–358, 2008.

Der Kiureghian, A. and T. Dakessian Multiple design points in first- and second-order reliability. *Structural Safety*, 20:37–49, 1998.

Ditlevsen, O. and H. O. Madsen. Structural Reliability Methods. Wiley, 1996.

Duborg, V. and B. Sudret and F. Deheeger Metamodel-based importance sampling for structural reliability analysis. *Probab Eng Mech*, 33:47–57, 2013.

Dutfoy, A. and R. Lebrun Practical approach to dependence modelling using copulas. *Proc. IMech. E., Part O: J. Risk and Reliability*, 223:347–361, 2009.

Echard, B. and N. Gayton and M. Lemaire AK-MCS: an active learning reliability method combining Kriging and Monte Carlo simulation. *Struct Saf*, 33(2):145–154, 2011.

Echard, B. and N. Gayton and M. Lemaire and N. Relun A combined Importance Sampling and Kriging reliability method for small failure probabilities with time-demanding numerical models. *Reliab Eng Syst Saf*, 111:232–240, 2013.

Embrechts, A. and A. McNeil and D. Straumann Correlation and dependency in risk man-agement: properties and pitfalls. *Review of Economics and Statistics*, 86, 2000.

Faravelli, L. Response-surface approach for reliability analysis. *J. Eng. Mech.*, 115:2763–2781, 1989.

Forrester, A. I.J. and A. SObester and A. J. Keane Engineering design via surrogate modelling. John Wiley & Sons, 2006.

Gaspar, B. and A. P. Teixeira and C. Guedes Soares Assessment of the efficiency of Kriging surrogate models for structural reliability analysis. *Prob. Eng. Mech*, 37:24–34, 2014.

Huang, X. and J. Chen and H. Zhu Assessing small failure probabilities by AK–SS: An active learning method combining Kriging and Subset Simulation. *Structural Safety*, 59:86–95, 2016.

Kaymaz, I. Application of kriging method to structural reliability problem. *Struct. Saf.*, 27(2):133–151, 2005.

Li, G. and C. Rosenthal and H. Rabitz High-dimensional model representation. *J. Physical Chem. A*, 106:7765–7777, 2001.

Linfoot, E. An informational measure of correlation. *Information and Control*, 1:85–89, 1957.

LV, Z. and Z. Lu and P. Wang A new learning function for Kriging and its applications to solve reliability problems in engineering. *Computers and Mathematics with Applications*, 70:1182–1197, 2015.

Madsen, H. O., S. Krenk and N. C. Lind. Methods of Structural Safety. Prentice-Hall, 1986.

Melchers, R. E. Importance Sampling in structural systems. *Structural Safety*, 6(1):3–10, 1989.

Melchers, R. E. Structural Reliability, analysis and prediction. New York: John Wiley & Sons, 1999.

Rajashekhar, M. R. and B. R. Ellingwood A new look at the response surface approach for reliability analysis. *Structural Safety*, 12:205–220, 1993.

Rasmussen, C. E. and C. K.I. Williams Gaussian Processes for Machine Learning. MIT Press, 2006.

Rubinstein, R. and D. Krose Simulation and the Monte Carlo method. Wiley Series in Probability and Statistics: Wiley, 2008.

Sudret, B. and A. Der Kiureghian Comparison of finite element reliability methods. *Prob. Eng. Mech.*, 17:337–348, 2002.

Zhang, X. and L. Wang and J. D. Sørensen REIF: A novel active-learning function toward adaptive Kriging surrogate models for structural reliability analysis. *Reliability Engineering and System Safety*, 185:440–454, 2019.

Zheng, Y. and P. K. Das Improved response surface method and its application to stiffened plate reliability analysis. *Eng. Struct.*, 22:544–551, 2000.