Implementation of Supervised PCA for Global Sensitivity Analysis of Models with Correlated Inputs

M. A. M. J. Sharbaf
Shiraz University
Mohammad Javad Abedini (abedini@shirazu.ac.ir)
Shiraz University https://orcid.org/0000-0002-0756-3872

Research Article

Keywords: global sensitivity analysis, correlated inputs, Supervised PCA, RKHS, variance-based SA

Posted Date: July 29th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-679733/v1

Implementation of Supervised PCA for global sensitivity analysis of models with correlated inputs

Mohammad Ali Mohammad Jafar Sharbaf [1] - Mohammad Javad Abedini [2]

Abstract

Global Sensitivity Analysis (GSA) plays a significant role in quantifying the tangible impact of model inputs on the uncertainty of response variable. As GSA results are strongly affected by correlated inputs, several studies considered this issue, but most of them are quite time-consuming, labor-intensive, and difficult to implement. Accordingly, this paper puts forward a novel strategy based on the Supervised Principal Component analysis (Supervised PCA), benefiting from the Reproducing Kernel Hilbert Space (RKHS). Indeed, by conducting one kind of variance-based sensitivity analysis (SA), a renowned method exclusively customized for models with orthogonal inputs, on Supervised PCA (SPCA) regression, the impact of correlation structure of input variables is effectively taken into account. The ability of the suggested technique is evaluated with five test cases as well as two hydrologic and hydraulic models, and the results are compared and contrasted with those obtained from the correlation ratio method taken as a robust benchmark solution. It is found that the proposed method satisfactorily identifies the sensitivity ordering of model inputs. Furthermore, it is proved in this study that the performance of the proposed approach is also supported by the total contribution index in the derived covariance decomposition equation. Moreover, the proposed method compared to the correlation ratio method, is found to be time efficient and easy to implement. Overall, the proposed scheme is appropriate for high dimensional, relatively nonlinear or expensive models with correlated inputs whose coefficient of determination is larger than 0.5.

Keywords: global sensitivity analysis, correlated inputs, Supervised PCA, RKHS, variance-based SA

1 Introduction

The identification, quantification, and propagation of uncertainty are considered to be three indispensable tasks in model building and construction. As the process of modeling and simulation of water resources and environmental projects become more advanced and computationally expensive particularly for high

[1] PhD Candidate, Department of Civil and Environmental Engineering, School of Engineering, Shiraz University, Shiraz, Iran. Email: ma.mj.sharbaf@shirazu.ac.ir

[2] Professor, Department of Civil and Environmental Engineering, School of Engineering, Shiraz University, Shiraz, Iran. Email: abedini@shirazu.ac.ir (Corresponding author)

dimensional problems, the process of quantification and propagation of uncertainty in both parameter estimation and model structure identification become quite time consuming and labor-intensive. In a sense, in a typical modeling and simulation with a highly nonlinear model structure, a small error in input variables might result in a large uncertainty (or error) in model output (Helton et al. 2005; Iman et al. 2002). Thus, it is precious to investigate how the uncertainty in the output of a model can be proportioned to uncertainty in model inputs. Global Sensitivity Analysis is devised to address these issues in model building and construction.

Indeed, GSA is utilized in many applications in water resources engineering and environmental studies (Ciriello et al. 2013; Zheng and Wang 2015; Razavi and Gupta 2016) such as model calibration, uncertainty reduction of model output, model validation, identification of redundant parameters and decision-making processes that try to explore which model inputs have the most influential impact on output variability (Crosetto and Tarantola 2001). Due to the numerous applications including the one cited above, GSA can be effectively utilized to conduct preliminary exploratory data analysis to build the corresponding model. As an example, in a typical rainfall-runoff process, being able to implement GSA in order to build a parsimonious model structure would not only help to eliminate the redundant parameters but also help to reduce the uncertainty in model output which somehow lead to saving in time and capital. This task is considered as an inevitable part of any efforts conducted to convert rainfall into runoff in an efficient way.

There is a wide range of GSA methods in the market based on different philosophies and theoretical definitions of sensitivity measures, such as regression technique, Morris, variance-based, regional SA, and density-based methods. The details of these approaches can be found elsewhere (Saltelli et al. 2004; Borgonovo 2007). Moreover, in this field, the breakthrough discovery has been made by Razavi and Gupta (2016), who proposed the methodology called Variogram Analysis of Response Surface (VARS) method. However, the cited approaches suffer from a fundamental problem with regard to correlation structure of input variables. Nowadays, non-orthogonal model input variables can be considered to be a rule rather than an exception in GSA because it is apparent in the shadow of belittling and ignorance of model factor dependence; GSA outcome will be fallacious. As a result, on no account is it satisfactory to accept the result of GSA without considering the dependent nature of input variables. Accordingly, several scientists had deliberately attempted to present or develop GSA methods in light of dependent factors. As an example, a few studies applied the parametric and/or non-parametric methods to decompose the variance of the response variable in the presence of dependency among model inputs (Saltelli et al. 2001 and 2004; Xu and Gertner 2008; Da Veiga et al. 2009; Chastaing et al. 2012; Li et al. 2010; Xu 2013; Zhang et al. 2015; Mara et al. 2015), whereas some other investigators extended the analytical formulations for non-orthogonal input variables and the associated numerical approaches to calculate the variance-based sensitivity measure

and/or more advanced Sobol's sensitivity indices (Mara and Tarantola 2012; Kucherenko et al. 2012; Kucherenko et al. 2017; Wang et al. 2018). As the computational budgets associated with most of the above approaches are quite expensive due to high dimensionality and/or more complex model structure, development of more innovative, time-efficient approaches in light of non-orthogonal inputs is strongly recommended for practical purposes (Ge and Menendez 2017; Zhou et al. 2019; Lamboni and Kucherenko 2021). One should also acknowledge the fact that how intricate it is for engineers and other investigators, with rudimentary information on GSA, to use the above methods because of their complicated procedures and concepts for practical applications on a routine basis.

In light of the above elaboration, this paper intended to couple the variance-based SA, a class of methods customized for independent inputs, with Supervised PCA, based on Hilbert-Schmidt Independence Criterion (HSIC), to address issues concerning CPU time and methods' complexity. Indeed, implementation of this kind of variance-based SA on a regression model derived from the Supervised PCA would lead to an analytical method in terms of components of dominant eigenvector to estimate the first-order sensitivity measure with no additional computational budget for a system whose inputs probability distribution functions have several correlated input variables.

The rest of the paper is organized as follows: first, an overview of the GSA based on the variance-based approach is presented in the next section. Then, the paper gives a detail account of Supervised PCA, followed by the proposed method along with the proof of its validity in the next section. The subsequent section is devoted to assessing the performance of the proposed method via five test function as well as two more realistic models in hydrologic and hydraulic domains, for which the corresponding results of all test cases are compared and contrasted with the correlation ratio method based on McKay's scheme. Finally, the summary and the conclusions which can be drawn from the paper are summarized in the last section.

2 Theoretical Background

The methodology presented in this work is intended to couple Supervised PCA with the variance-based SA. It is important to make this coupling process quite clear. In essence, the variance-based (e.g., Sobol and/or FAST) methodology is limited to transfer functions for which the input variables are independent. On the other hand, Supervised PCA considers the impact of correlation among input variables and dependent variable using regression-based analysis. Could it be possible to effectively benefit from the advantage of Supervised PCA and combine it with the particular variance-based SA to come up with a better tool for prioritizing the input variables when the input variables are correlated? In light of this question, in what follows, at first, a concise account of the basic concept of variance-based SA is provided, followed by

touching on the theoretical background of Supervised PCA in some detail to see how this coupling exercise can be implemented.

2.1 Variance-based sensitivity analysis

For a computer model with p number of input variables given by $Y = f(\boldsymbol{X})$, $\boldsymbol{X} = (X_1, X_2 \dots, X_p)$, if the input variables are random, then the output becomes a random variable as well. For such a model, decomposition of the total unconditional variance of output variable can be written as:

$$V(Y) = V[E(Y|X_i)] + E[V(Y|X_i)] \tag{1}$$

V conveys the notion of variance, and E represents the expectation operator. Also, $V[E(Y|X_i)]$ and $E[V(Y|X_i)]$ are called the main and residual effect, respectively (Saltelli et al. 2008). Thus, in order to achieve a reduction in the variance of model output, one can compute the first-order sensitivity index by defining the ratio of the main variance (i.e., V_i) as a variance of conditional expectation to variance of model output as:

$$S_i = V_i/V(Y) = V[E(Y|X_i)]/V(Y) \tag{2}$$

And consequently, the sensitivity analysis of higher-order terms in light of the interaction of model factors such as X_i and X_j on model output uncertainty can be written as:

$$S_{ij} = V_{ij}/V(Y) = \left(V\left[E\left(Y\middle|X_i, X_j\right)\right] - V_i - V_j\right)/V(Y) \tag{3}$$

Saltelli et al. (2004) can be consulted for further detail regarding computation of higher order interactions. Under the independent assumption, the variance-based sensitivity indices can be estimated numerically via the two most popular techniques, e.g., Sobol (1993) and FAST (Cukier et al. 1973).

If a mathematical model under consideration is assumed to be approximately linear due to the relatively linear impact of model factors on output, the given model, i.e., $Y = f(X_1, X_2 \dots, X_p)$ can be simplified by eliminating the nonlinear or interaction terms as follows:

$$Y = a_0 + \sum_{i=1}^{p} a_i X_i \tag{4}$$

Later on, we will be exposed to how the original nonlinear function can be converted to Eq. (4). Hence, in light of this simplification, one has to acknowledge the fact that we consider only the main effect of each model input in the corresponding regression model. When the components of this linear model inputs are mutually independent, it can be proved that the main effect in reference to variance decomposition can be stated as follows (Saltelli et al. 2008):

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)} = {\theta_i}^2 \quad (5)$$

In which $\theta_i = a_i\, \sigma_{X_i}/\sigma_Y$. The $\theta's$ are standardized regression coefficients (SRCs). Indeed, in the case of orthogonal model factors, the SRC is considered as a method of sensitivity analysis for the model inputs, and in accordance with the identity of Eq. (5), the squared SRCs describe the contribution of each factor to total variance (Saltelli et al. 2008).

In order to evaluate the performance of the SRCs, it is necessary to appreciate the connection between SRCs and the measure of goodness so called the model coefficient of determination, R^2 as:

$$R^2 = \sum_{i=1}^{p} (\hat{Y}_i - \overline{Y}) \Big/ \sum_{i=1}^{p} (Y_i - \overline{Y}) \quad (6)$$

Where Y_i is the output of the original model i.e., $Y = f(X_1, X_2 \dots, X_p)$, and $\hat{Y}_i$ stands for an estimate of Y_i using the regression model i.e., Eq. (4). This coefficient is an attempt to recognize how well the model under study can be approximated by regression in the form of Eq. (4) (Weisberg 2005). When the model is linear, using SRC as a GSA method can exactly quantify the amount of response surface variation explained by each model input. Needless to say, a moderate nonlinear model should honor the coefficient of determination to be greater than 0.7 to be applicable in this context (Cariboni et al. 2007; Saltelli et al. 2004).

2.2 The Supervised PCA

The supervised PCA, initially pioneered by Barshan et al. (2011) laid the foundations for this approach in the field of supervised methods. This is because before the cited contribution, a considerable number of dimensionality reduction techniques based on supervised methods could only take into account similarities and differences for classification purposes. This property of the cited supervised methods is in contrast with Barshan et al.'s approach, which also examines the quantitative value of the target variables. As a result, it is applicable to both classification and regression problems. Their ideas can be considered as a paradigm shift in researchers' way of carrying out prediction based on regression approaches. Their studies makes reference to the work of Gretton et al. (2005), who proposed an independent criterion in Reproducing Kernel Hilbert Spaces. This criterion measures the dependency between two random variables according to the Hilbert-Schmidt Independence Criterion (HSIC). In reference to HSIC, the necessary and sufficient conditions for the two random variables to be independent, can be achieved if the value of this statistic (i.e., HSIC) is zero.

There is a theoretical relationship for HSIC but it is impractical for actual settings. Consequently, the empirical estimate of HSIC proposed by Gretton et al. (2005) as a practical criterion, can be implemented to check the dependence or independence of two random variables with finite number of observations. Thus, this criterion, i.e., empirical estimate of HSIC for a series of n observations such as $\mathcal{Z} \coloneqq \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$, is as follows (Gretton et al. 2005):

$$HSIC(\mathcal{Z}, \mathcal{F}, \mathcal{G}) \coloneqq (n-1)^{-2} tr(KHLH) \tag{7}$$

Where $\mathcal{F}$ is a Reproducing Kernel Hilbert Space 3 in which for each point $x \in X$, there is an element $\emptyset(x) \in \mathcal{F}$, such that $\langle \emptyset(x_i), \emptyset(x_j) \rangle_{\mathcal{F}} = k(x_i, x_j) = K_{ij}$ and K *is* considered to be a positive definite kernel for $n \geq 1$ and $b_1, b_2, \dots, b_n \in \mathbb{R}$, i.e.,

$$\sum_{i=1}^{n} \sum_{j=1}^{n} b_i \, b_j k(x_i, x_j) > 0 \tag{8}$$

Likewise, $\mathcal{G}$ is a RKHS in which for each point $y \in Y$, there is an element $\omega(y) \in \mathcal{G}$, such that $\langle \omega(y_i), \omega(y_j) \rangle_{\mathcal{G}} = l(y_i, y_j)$ and L is considered as a positive definite kernel. It is necessary to recall that $\mathcal{F}$ and $\mathcal{G}$ have to be separable. Indeed, they must have complete orthonormal systems. In a nutshell, $K, L, H \in R^{n \times n}$ and $K_{ij} \coloneqq k(x_i, x_j)$, $L_{ij} \coloneqq l(y_i, y_j)$. In addition, $H_{ij} \coloneqq I - \frac{1}{n} ee^T$ (centering matrix and e is a unit vector). Finally tr stands for the trace of a matrix.

In Supervised PCA, the subspace spanned by new features are examined such that the principal components of input variables with maximum dependency on response surface are reserved through the empirical criterion, i.e., HSIC. Indeed, for a model with p standardized inputs, $\widetilde{\boldsymbol{X}} = (\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p)$ and ℓ standardized outputs, $\widetilde{\boldsymbol{Y}} = (\tilde{Y}_1, \tilde{Y}_2, \dots, \tilde{Y}_\ell)$, to maximize the dependency among the projected data (i.e., $\boldsymbol{Z} = \widetilde{\boldsymbol{X}} U$) [4] and the response surface (i.e., $\widetilde{\boldsymbol{Y}}$), it requires to maximize the $tr(KHLH)$. This could be justified by resorting to empirical HSIC cited above (Barshan et al. 2011). In order to maximize $tr(KHLH)$, $\boldsymbol{Z}\boldsymbol{Z}^T$ is replaced with the kernel matrix K to get:

$$tr(KHLH) = tr(\boldsymbol{Z}\boldsymbol{Z}^T HLH) = tr\left(\widetilde{\boldsymbol{X}} U U^T \widetilde{\boldsymbol{X}}^T HLH\right) \tag{9}$$

After replacing the dimension of each matrix in Eq. (9), this equation can be rewritten as follows:

$$tr(KHLH) = tr\left(\widetilde{\boldsymbol{X}}_{n*p} U_{p*m} {U^T}_{m*p} {\widetilde{\boldsymbol{X}}^T}_{p*n} H_{n*n} L_{n*n} H_{n*n}\right) \tag{10}$$

[3] The necessary background of this concept is provided by Gretton et al (2005)

[4] Matrix U is a modal matrix consisting of m eigenvectors that maps the data sets to a new space in which features are uncorrelated.

$$= tr([\widetilde{\boldsymbol{X}}U]_{n*m}\,[U^T\widetilde{\boldsymbol{X}}^T HLH]_{m*n})$$

Based on the unique property of trace in matrix algebra[5]:

$$tr(KHLH) \;=\; tr(\,[U^T\widetilde{\boldsymbol{X}}^T HLH]_{m*n}[\widetilde{\boldsymbol{X}}U]_{n*m}) \tag{11}$$

Consequently, we have:

$$tr(KHLH) \;=\; tr(\,U^T\widetilde{\boldsymbol{X}}^T HLH\widetilde{\boldsymbol{X}}U) \tag{12}$$

L is a kernel matrix of $\widetilde{\boldsymbol{Y}}$ (e.g., $\widetilde{\boldsymbol{Y}}\widetilde{\boldsymbol{Y}}^{\boldsymbol{T}}$), and $H_{ij} = I - \frac{1}{n}ee^T$.

Ultimately, the optimization for HSIC objective function is accompanied by the following constraint (Barshan et al. 2011):

$$\begin{aligned} &\underset{U}{argmax} && tr(U^T\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}HLH\widetilde{\boldsymbol{X}}U) \\ &subject\ to && U^TU = I \end{aligned} \tag{13}$$

The optimal solution for this optimization problem is considered to be the eigenvectors of the real and symmetric matrix $\mathbb{Q} = \widetilde{\boldsymbol{X}}^{\boldsymbol{T}}HLH\widetilde{\boldsymbol{X}}$. Taking the components of the eigenvectors as the decision variables, the optimal solution will be $U = [U_1, U_2, \dots, U_m]$ associated with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$, which are selected among p eigenvalues. Here m is the dimension of eigenspace (Barshan et al. 2011). The beauty of the above mathematical argument can be justified as follows: If kernel L is equal to the identity matrix I, the matrix $\widetilde{\boldsymbol{X}}^{\boldsymbol{T}}HLH\widetilde{\boldsymbol{X}}$ is equivalent with the covariance of matrix $\tilde{X}$. i.e., the PCA method (Barshan et al. 2011). Since $H^T = H$:

$$\mathbb{Q} = \widetilde{\boldsymbol{X}}^{\boldsymbol{T}}HIH\widetilde{\boldsymbol{X}} = (H\widetilde{\boldsymbol{X}})^T(H\widetilde{\boldsymbol{X}}) \tag{14}$$

$$= \left(\left(I - \frac{1}{n}ee^T\right)\widetilde{\boldsymbol{X}}\right)^T\left(I - \frac{1}{n}ee^T\right)\widetilde{\boldsymbol{X}} = \left(\widetilde{\boldsymbol{X}} - \mu\right)^T(\widetilde{\boldsymbol{X}} - \mu) \;= COV(\widetilde{\boldsymbol{X}})$$

3 Proposed method

As mentioned in the preceding section, the novelty in this study is based on implementing variance-based SA (applicable under the assumption of independent inputs) on the regression model extracted from Supervised PCA for models with correlated inputs. For this reason, consider the regression model cited in Eq. (4), as mentioned, this linear regression model is obtained in reference to the approximate linear impact of inputs on the response variable i.e., $Y = f(X_1, X_2, \dots, X_p)$. In order to use this regression model, first, the

[5] $tr(\,\tilde{\mathcal{A}}_{\,n*m}\,\widetilde{\mathcal{B}}_{m*n}\,) = tr(\,\widetilde{\mathcal{B}}_{m*n}\,\tilde{\mathcal{A}}_{\,n*m}\,)$

samples of model inputs are required. For this purpose, Latin Hypercube Sampling (LHS) can be used to generate various realizations of the input variables for both correlated and/or uncorrelated model inputs (Iman and Conover 1982). This method is also recognized as an inverse Nataf transformation (Nataf 1962). Then, the original model is run using these samples, and a one-dimensional response surface (Y) will be generated. As soon as the multi-inputs-single output realizations are generated, one can cast the regression model stipulated in Eq. (4). When the independent assumption is violated, i.e., the model inputs are correlated, using ordinary least square gives a poor estimate of model parameters of this equation, i.e., a_i. Thus, performing SA on this linear model will produce unjustified solution. Although the Principal Component Regression (PCR) method can overcome this problem, coupling it with variance-based approach (e.g., Sobol and/or FAST) has a drawback that will be highlighted in the following study. As a result, we have decided to use the Supervised PCA using linear kernel L on standardized response variable $\tilde{Y}$ (i.e., $\tilde{Y}\tilde{Y}^T$) to estimate the simple regression model coefficients. For this purpose, $\tilde{Y}$ is expressed as a linear combination of new features called Z_j as stipulated in Eq. (15).

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_m Z_m + \beta_0 \tag{15}$$

In which Z_j $(j = 1, 2, \dots, m \quad m < p)$ are a set of new features taking into account the impact of dependent variable, $\tilde{Y}$. In addition, m is the number of supervised principal components. The procedure of deriving Eq. (15) is explained in appendix A. In order to express Y as a linear combination of X_i, a linear transformation was utilized to convert Z_j to X_i as follows:

$$[Z_1, Z_2, \dots, Z_m] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1m} \\ u_{21} & u_{22} & \dots & u_{2m} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_{p1} & u_{p2} & \dots & u_{pm} \end{bmatrix} \tag{16}$$

It is proved in appendix B, after replacing matrix L with $\tilde{Y}\tilde{Y}^T$ in matrix $\mathbb{Q}$, eigenvalue analysis results in only one nonzero eigenvalue (λ_1). In that appendix, a unique eigenvalue and the corresponding eigenvector for matrix $\mathbb{Q}$ is rationalized. Thus, its associated eigenvector, i.e., $U_1 = [u_{11}, u_{21}, \dots, u_{p1}]^T$ is selected for projecting explanatory variables, i.e., $\boldsymbol{X}$. As the corresponding new feature (Z_1) has a maximum linear dependency on $\tilde{Y}$, this means other new features can be eliminated from the model due to zero correlation with output which is proved in appendix B. In the meanwhile, in linear regression, when model variables are uncorrelated with zero means, the regression coefficients of Eq. (15) can be estimated as follows (Bedford 1998; Xu and Gertner 2008):

$$\hat{\beta}_j = \frac{COV(\tilde{Y}, Z_j)}{V(Z_i)}, \quad j = 1,2, \dots, m \tag{17}$$

As we proved in appendix B, $COV(\tilde{Y}, Z_j) = 0$, for $j \neq 1$. Thus, $\hat{\beta}_j = 0$ for $j \neq 1$. Consequently, the regression-based SPCA model will be simplified as follows:

$$\tilde{Y} = \beta_1 Z_1 + \beta_0 \tag{18}$$

After replacing Z_1 and β_1 with their equivalences [i.e., Eq. (16) and Eq. (17)] into Eq. (18), the regression-based SPCA model can be expressed in terms of standardized variables as:

$$\tilde{Y} = \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)}[u_{11}\tilde{X}_1 + u_{21}\tilde{X}_2 + \cdots + u_{p1}\tilde{X}_p] + \beta_0 \tag{19}$$

Where u_{i1} are the entries of the U_1 eigenvector corresponding to the maximum eigenvalue, i.e., λ_1. As a result, Eq. (19) can be expressed in terms of the original variables Y and X_i as:

$$\frac{Y - E(Y)}{\sigma_Y} = \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)}\left[u_{11}\frac{X_1 - E(X_1)}{\sigma_{X_1}} + u_{21}\frac{X_2 - E(X_2)}{\sigma_{X_2}} + \cdots + u_{p1}\frac{X_p - E(X_p)}{\sigma_{X_p}}\right] + \beta_0 \tag{20}$$

The above equation can be written as:

$$Y = \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)} * \sigma_Y * \sum_{i=1}^{p} u_{i1}\frac{X_i}{\sigma_{X_i}} + a_0 \tag{21}$$

$$a_0 = \beta_0\sigma_Y + E(Y) - \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)}\sigma_Y \sum_{i=1}^{p} u_{i1}\frac{E(X_i)}{\sigma_{X_i}}, \qquad i = 1,2, \dots, p$$

Hence, Eq. (21) is considered as a regression-based SPCA model for $Y = f(\boldsymbol{X})$. It is quite important to note that the seemingly observed Y [i.e., $Y = f(\boldsymbol{X})$] was shown to have coefficient of determination greater than 0.7 with the computed Y based on Eq. (21) to get meaningful results (Saltelli et al. 2004). This means that a large part of the variation of output variable can be effectively described by the regression model. In order to add another dimension to what they found, we examined the reliability of the proposed method for the case when R^2 is smaller than 0.7 with positive feedback.

At this stage, the variance-based SA can be effectively utilized to investigate the impact of X_i on Y. Before this purpose, it might help to summarize the main point of the above mathematical manipulations for practical purposes. If we are given a functional thereby the predictor variables are correlated. We have two choices to investigate the impact of various predictor variables on output. According to the first option, one can implement the variance-based SA (e.g., Sobol and/or FAST) directly on original model and report the most important variables for which the results are not justified. However, one can also implement regression-based SPCA to come up with Eq. (21) and then impose the variance-based SA (applicable under

the assumption of the independent inputs) on this regression [6] model to get a meaningful and justifiable results. We argue that prioritization in reference to the second option would give rise to better and acceptable results.

Following the second option, estimation of the sensitivity measures based on Eq. (5) using regression-based SPCA model can be written as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = \frac{\left[\frac{COV(\tilde{Y}, Z_1)}{V(Z_1)} * \sigma_Y\right]^2 * \left(\frac{u_{i1}}{\sigma_{X_i}}\right)^2 * V(X_i)}{[\frac{COV(\tilde{Y}, Z_1)}{V(Z_1)} * \sigma_Y]^2 * [\left(\frac{u_{11}}{\sigma_{X_1}}\right)^2 * V(X_1) + \cdots + \left(\frac{u_{p1}}{\sigma_{X_p}}\right)^2 * V(X_p)]} \quad (22)$$

Where $V(Y)_{unc}$ stands for estimation of $V(Y)$ without considering the correlation among inputs. Since $U^T U = I$, the constraint of ${U_1}^T U_1 = 1$ is satisfied. Thus:

$$\left[{u_{11}}^2 + {u_{21}}^2 + {u_{31}}^2 + \cdots + {u_{p1}}^2\right] = U_1 . U_1 = {U^T}_1 U_1 = 1 \quad (23)$$

After proper simplification, we have:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = {u_{i1}}^2, \qquad i = 1,2,\ldots,p \quad (24)$$

This equation demonstrates the simplicity of the proposed method in light of having a linear or nonlinear function with correlated input variables. The simplicity claim concerns with the fact that sensitivity measures, i.e., S_i equals to the square of i[th] component of the eigenvector corresponding to the dominant eigenvalue (i.e., first new feature).

At this stage, we might want to raise a serious question. While variance-based SA (e.g., Sobol and/or FAST) cannot be implemented on correlated data emerging from a linear or nonlinear model, how and why such tool can be safely implemented on the regression-based SPCA model for evaluating the importance of correlated predictor variables. The following mathematical elaboration is intended to address this question. To this end, at first, we showed in appendix B, there is a relationship between the first eigenvalue of matrix $\mathbb{Q}$ assuming $L = \tilde{Y}\tilde{Y}^T$, its eigenvector components associated with the first eigenvalue , u_{i1}, and $COV(Y, X_i)$ as follows:

[6] "*The regression algorithm returns a regression meta-model, whereby the output Y is described in terms of a linear combination of the input factors"* (Saltelli et al. 2004). Indeed, in a meta-modeling approach, instead of an original system, the effective model maps output and inputs. Then, the importance of model inputs was examined using the constructed surrogate model. On the contrary, in a classical approach, the original system and the corresponding model output values are directly considered to estimate the sensitivity indices (Li et al. 2010).

$$u_{i1} = \frac{COV(\tilde{Y},\ \tilde{X}_i)}{(\lambda_1)^{.5}} = \frac{1}{(\lambda_1)^{.5} * \sigma_Y * \sigma_{X_i}}\ COV(Y,\ X_i)\,, \qquad i = 1,2,\dots,p \tag{25}$$

Eq. (25) is substituted into Eq. (24). As a result, the first-order sensitivity measure can also be computed as follows:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = \frac{1}{\lambda_1 * V(Y) * V(X_i)}\ COV(Y,\ X_i)^2, \qquad i = 1,2,\dots.,p \tag{26}$$

Eq. (26) can be restated as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = \frac{COV(Y,\ X_i)}{V(Y)}\left[\frac{COV(Y,\ X_i)}{\lambda_1 * V(X_i)}\right],\ \ i = 1,2,\dots.,p$$
$$b_i = \left[\frac{COV(Y,\ X_i)}{\lambda_1 * V(X_i)}\right] \tag{27}$$

In what follows, we managed to prove that the coefficient b_i is approximately equivalent to the corresponding coefficient obtained from the regression-based SPCA model, i.e., a_i. In light of this, one may want to start with the coefficients of the regression-based SPCA model in reference to Eq. (21):

$$a_i = \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)} * \sigma_Y * \left[\frac{u_{i1}}{\sigma_{X_i}}\right], \qquad i = 1,2,\dots.,p \tag{28}$$

The coefficient a_i, can be simplified via the following mathematical manipulations. After substituting the standardized variables into Z_1, $COV(\tilde{Y}, Z_1)$ can be written as:

$$\begin{aligned} COV(\tilde{Y}, Z_1) &= COV(\tilde{Y}, u_{11}\tilde{X}_1 + u_{21}\tilde{X}_2 + \cdots + u_{p1}\tilde{X}_p) \\ &= COV(\tilde{Y}, u_{11}\tilde{X}_1) + COV(\tilde{Y}, u_{21}\tilde{X}_2) + \cdots + COV(\tilde{Y}, u_{p1}\tilde{X}_p) \\ &= u_{11}COV(\tilde{Y}, \tilde{X}_1) + u_{21}COV(\tilde{Y}, \tilde{X}_2) + \cdots + u_{p1}COV(\tilde{Y}, \tilde{X}_p) \end{aligned} \tag{29}$$

Following Eq. (25), the $COV(\tilde{Y}, Z_1)$ becomes:

$$COV(\tilde{Y}, Z_1) = \frac{[COV(\tilde{Y},\ \tilde{X}_1)]^2}{(\lambda_1)^{.5}} + \frac{[COV(\tilde{Y},\ \tilde{X}_2)]^2}{(\lambda_1)^{.5}} + \cdots + \frac{[COV(\tilde{Y},\ \tilde{X}_p)]^2}{(\lambda_1)^{.5}} \tag{30}$$

As shown in appendix B, the dominant eigenvalue, λ_1 can be shown to be:

$$\lambda_1 = \sum_{i=1}^{p} [COV(\tilde{Y}, \tilde{X}_i)]^2 \tag{31}$$

After combining Eq. (30) and (31), $COV(\tilde{Y}, Z_1)$ can be found to be:

$$COV(\tilde{Y}, Z_1) = (\lambda_1)^{.5} \tag{32}$$

Now we can switch to computation of $V(Z_1)$ in order to further simplify the coefficients of the regression-based SPCA model. For this purpose, considering Eq. (18) once again and taking variance of both sides results in:

$$\tilde{Y} = \beta_1 Z_1 + \beta_0$$

$$V(\tilde{Y}) = V(\beta_1 Z_1 + \beta_0)$$

$$V(\tilde{Y}) = {\beta_1}^2 V(Z_1)$$

$$V(\tilde{Y}) = \left(\frac{COV(\tilde{Y}, Z_1)}{V(Z_1)}\right)^2 * V(Z_1) \tag{33}$$

Since $\tilde{Y}$ is a standardized variable, $V(\tilde{Y}) = 1$. Therefore, by substituting Eq. (32) into Eq. (33) we have:

$$1 = \frac{\lambda_1}{V(Z_1)}$$

$$V(Z_1) = \lambda_1 \tag{34}$$

Eq. (34) is very applicable to linear models. However, as one departs from linearity, then $V(Z_1)$ departs from the dominant eigenvalue, λ_1. As a result, $V(Z_1) \approx \lambda_1$.

After replacing u_{i1} and $COV(\tilde{Y}, Z_1)$ with their equivalences, i.e., Eq. (25) and Eq. (32), respectively into Eq. (28), the coefficients a_i are simplified to the following relation:

$$a_i = \frac{COV(\tilde{Y}, Z_1)}{V(Z_1)} * \sigma_Y * \left[\frac{u_{i1}}{\sigma_{X_1}}\right] = \frac{(\lambda_1)^{.5}}{V(Z_1)} * \sigma_Y * \left[\frac{COV(Y,\ X_i)}{(\lambda_1)^{.5} * \sigma_Y * \sigma_{X_i} * \sigma_{X_i}}\right]$$

$$a_i = \frac{1}{V(Z_1)} * \left[\frac{COV(Y,\ X_i)}{V(X_i)}\right], \qquad i = 1,2, \ldots., p \tag{35}$$

Based on Eq. (34) and subsequent comment, the coefficients a_i can be further simplified to:

$$a_i \approx \frac{COV(Y,\ X_i)}{\lambda_1 * V(X_i)}, \qquad i = 1,2, \ldots., p \tag{36}$$

Therefore, in reference to the above equation, after implementing the variance-based SA on the regression-based SPCA model, the first-order sensitivity measure can be written as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = {u_{i1}}^2 = \frac{COV(Y,\ X_i)}{V(Y)}\left[\frac{COV(Y,\ X_i)}{\lambda_1 * V(X_i)}\right] \approx \frac{COV(Y,\ X_i)}{V(Y)} * a_i \tag{37}$$

Finally:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = {u_{i1}}^2 \approx \frac{COV(Y,\ a_i X_i)}{V(Y)},\ \ i = 1,2,\dots.,p \tag{38}$$

In the meanwhile, in appendix C, we managed to relate the first-order sensitivity measure [i.e., Eq. (38)] to two sources of variability, i.e., the variability due to each variable and the variability due to co-variation among the variable under consideration and other predictor variables as stipulated below:

$$S_i^{tc} = \frac{COV(\ Y, a_i X_i\ )}{V(Y)} = \frac{V(a_i X_i)}{V(Y)} + \frac{COV\left(a_i X_i\ , \sum_{j=1, j\neq i}^{p} a_j X_j\right)}{V(Y)},\quad i = 1,2,\dots.,p \tag{39}$$

It is worth noting that the derived covariance decomposition is in line with covariance decomposition, first proposed by Li et al. (2010). In short, back to the question raised in the theoretical background section, the above mathematical manipulations imply that while variance-based SA (under the assumption of the independent factors) cannot be used directly on original model with correlated input, it is possible to effectively benefit from the advantage of Supervised PCA and combine it with this type of variance-based SA to come up with a better tool for prioritizing the input variables when the input variables are correlated.

After proposing the above approach, an important question can be raised by the reader: is it possible to conduct the variance-based SA (e.g., Sobol and/or FAST) on Principal Component Regression (a special form of the Supervised PCA) to differentiate between the important and irrelevant variables in models with correlated inputs. It depends on the nature of the problem at hand. Correlation of new predictor variables with the output and/or its lack could be the cause of success or failure. The process to address this issue is very similar to what we did while coupling variance-based SA with regression-based SPCA model. In order to keep the integrity of material in place, interested readers might want to check appendix D for further detail regarding the coefficients in Eq. (40). In summary, the PCR model can be expressed as a linear combination of original input variables as follows:

$$Y = \frac{c_1\sigma_Y}{\sigma_{X_1}}X_1 + \frac{c_2\sigma_Y}{\sigma_{X_2}}X_2 + \cdots + \frac{c_p\sigma_Y}{\sigma_{X_p}}X_p + E(Y) + \beta_0\sigma_Y - \sum_{i=1}^{p} c_i\frac{E(X_i)}{\sigma_{X_i}} \tag{40}$$

$$c_i = \frac{COV\left(\tilde{Y}, {Z'}_1\right)}{\lambda_1}{u'}_{i1} + \frac{COV\left(\tilde{Y}, {Z'}_2\right)}{\lambda_2}{u'}_{i2} + \cdots + \frac{COV\left(\tilde{Y}, {Z'}_m\right)}{\lambda_m}{u'}_{im},\qquad i = 1,2,\dots,p$$

Thus, after coupling the variance-based SA (under assumption of the independent inputs) with PCR, the first-order sensitivity measure can be written as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = \frac{\left(\frac{c_i\sigma_Y}{\sigma_{X_i}}\right)^2 * V(X_i)}{(\frac{c_1\sigma_Y}{\sigma_{X_1}})^2 * V(X_1) + \left(\frac{c_2\sigma_Y}{\sigma_{X_2}}\right)^2 * V(X_2) + \cdots + \left(\frac{c_p\sigma_Y}{\sigma_{X_p}}\right)^2 * V(X_p)} \quad (41)$$

After some manipulations, the above equation is simplified as:

$$S_i = \frac{V[E(Y|X_i)]}{V(Y)_{unc}} = \frac{c_i^2}{[{c_1}^2 + {c_2}^2 + \cdots . + {c_p}^2]}, \qquad i = 1,2, \dots, p \quad (42)$$

In subsequent sections, a few numerical test cases are devised to either confirm or refute the validity of the above coupling exercise.

4 Applications and results

In this section, the proposed method based on coupling variance-based SA (applicable under the assumption of independent inputs) with either regression-based SPCA or PCR is applied to five test cases as well as two simple hydrologic and hydraulic models to evaluate the effectiveness of the corresponding modified model. From the theoretical background section, it became quite clear that implementation of either PCR and/or regression-based SPCA calls for generation of realizations for various input variables. As the governing regression model (e.g., regression-based SPCA) is linear, the number of realizations to build this regression can be kept quite small, e.g., 500 (Song et al. 2015). In the meanwhile, a few test cases were run with different sample sizes to examine the impact of various number of realizations on results of the proposed scheme.

It is worth recalling that after conducting eigenvalue analysis on $\mathbb{Q}$ extracted from Supervised PCA method, the aforementioned matrix has only one dominant eigenvalue and associated eigenvector. However, when it comes to PCR, it has more than one dominant eigenvalue. For this reason, the impact of different number of components on model performance was investigated. It is worth noting that if all components are included in the regression model, the PCR coefficients are equivalent to ordinary least square (Jolliffe 1986).

Even though we had a chance to analytically touch on validity of the proposed scheme, we also used the correlation ratio method as a benchmark solution implemented on the original model to further evaluate the hypothesis incorporated into various test cases. This benchmark, which is based on McKay's approach (McKay 1997; Saltelli et al. 2001), is always recognized as a valid approach due to its nonparametric nature (Xu and Gertner 2008). For this reason, this benchmark solution is suitable for nonlinear models even in the presence of strong nonlinear effects. However, as the method uses replicated LHS, it requires a large sample size to acquire an acceptable precision. In this work, like Xu and Gertner (2008), we suggest 100

replications with each replication having a sample size of 500 (a total of 50000 model runs). Finally, the negative impact of ignoring the correlation between inputs is examined by performing Sobol numerical variance-based SA (which assumed the model inputs are independent) on all test functions.

4.1 Test case 1: A linear model with constant coefficients

This test case, which had also been used by Li et al. (2010), is a simple and additive model with five inputs:

$$Y = X_1 + X_2 + X_3 + X_4 + X_5 \tag{43}$$

In this case, the marginal distribution of each input variable is normal with a mean of 0.5 and a unit standard deviation. i.e. $X_i \sim N(0.5,1)$. The Spearman Rank Correlation Coefficient (SRCC), better to say Spearman Rank Correlation matrix, of input variables is defined as follows:

$$\begin{bmatrix} 1 & 0.6 & 0.2 & 0 & 0 \\ 0.6 & 1 & 0.2 & 0 & 0 \\ 0.2 & 0.2 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0.2 \\ 0 & 0 & 0 & 0.2 & 1 \end{bmatrix}$$

Intuitively, in the case of independent inputs, it is straightforward to deduce each variable having the same impact on model output uncertainty due to their equal SRCs. No doubt, taking into account the correlation structure of input variables, each variable has different impact. Table 1 summarizes the result of sensitivity analysis using four different approaches including Sobol on the original model, the proposed approach, three flavors of variance-based PCR, and the benchmark solution implemented on the original model as well as the approximated equivalence of the proposed method (S^{tc}) obtained from covariance decomposition implemented on the regression-based SPCA model.

Table 1 Summary of SA--Test case 1

			Coupling variance-based SA with PCR				
	Variance -based (Sobol)	Proposed method $R^2 = 0.9836$	(2-comp) $R^2 = 0.9977$ $\frac{\sum \lambda_i}{5} = 0.614$	(3-comp) $R^2 = 0.9995$ $\frac{\sum \lambda_i}{5} = .776$	(4-comp) $R^2 = 1$ $\frac{\sum \lambda_i}{5} = .924$	Benchmark	S^{tc}
X_1	0.2035	0.2663	0.2200	0.2009	0.1999	0.4641	0.2889
X_2	0.2011	0.2705	0.2250	0.2057	0.2047	0.4820	0.2933
X_3	0.2021	0.1938	0.1507	0.1905	0.2021	0.3128	0.1940
X_4	0.2041	0.1323	0.2061	0.1826	0.2041	0.2381	0.1093
X_5	0.2039	0.1371	0.2130	0.2202	0.2038	0.2406	0.1146

In reference to the numerical values of S_i for various approaches and different input variables in Table 1, after clustering the input variables into three clusters, the solution associated with the benchmark solution has X_1, X_2 in cluster (1), X_3 in cluster (2), and X_4, X_5 in cluster three. It is crystal clear that the sorting in the proposed approach and its equivalence (S^{tc}) is quite consistent with the benchmark solution. However, neither the Sobol, implemented on the original model, nor three flavors of the variance-based PCR managed to reproduce the benchmark solution.

Needless to say, the change in numerical values of S_i from one approach to another is highly coined with the approach itself and one should not expect similar values emerging from each approach. Indeed, for sensitivity analysis purposes, detecting the importance of input variables and their rankings in each approach will become quite important.

Fig. 1 investigates the impact of number of realizations on sensitivity measures based on the proposed method. As the figure clearly shows, the number of realizations has minimal effect on S_i (i.e., the aforementioned clustering) and one can safely choose a minimum number of realizations (e.g., 500) for sensitivity analysis.

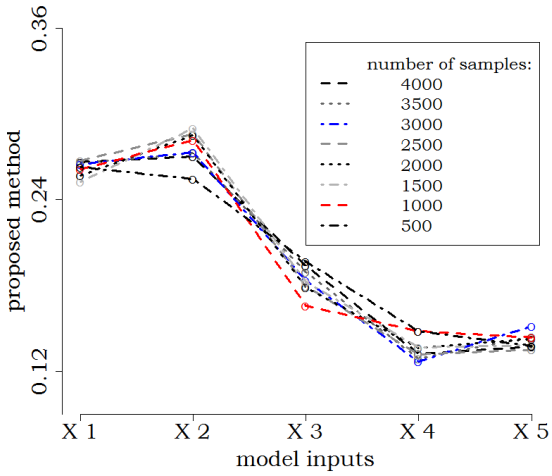

Fig. 1 Sensitivity measure (S_i) versus input variables for various number of realizations--Test case 1

4.2 Test case 2: A linear model with variable coefficients

Like the preceding test case, the second test case is also adopted from Li et al. (2010) as follows:

$$Y = 5X_1 + 4X_2 + 3X_3 + 2X_4 + X_5 \tag{44}$$

Except the equation, all other conditions of this case are very similar to the first test case. The results of the estimated sensitivity measures based on the aforementioned methods cited in Test case 1 are summarized in Table 2**.**

Table 2 Summary of SA of--Test case 2

	Variance -based (Sobol)	Proposed method $R^2 = 0.997$	Coupling variance-based SA with PCR (2-comp) $R^2 = 0.9934$ $\frac{\sum \lambda_i}{5} = 0.614$	Coupling variance-based SA with PCR (3-comp) $R^2 = 0.9953$ $\frac{\sum \lambda_i}{5} = .776$	Coupling variance-based SA with PCR (4-comp) $R^2 = 0.9979$ $\frac{\sum \lambda_i}{5} = .9243$	Benchmark	S^{tc}
X_1	0.4607	0.3897	0.3673	0.3837	0.3680	0.7128	0.3831
X_2	0.2927	0.3588	0.3683	0.3871	0.3763	0.6629	0.3662
X_3	0.1653	0.1792	0.1929	0.1443	0.1646	0.3018	0.1827
X_4	0.0741	0.0457	0.0397	0.0480	0.0722	0.0894	0.0399
X_5	0.0188	0.0268	0.0435	0.0368	0.0188	0.0509	0.0280

According to the content of this table, the rankings of model variables based on the proposed method are very similar to the correlation ratio scheme. Furthermore, the importance of inputs in reference to the proposed method are the same as those obtained from S^{tc}. Likewise, the performance of coupling PCR and variance-based in recognizing the sensitivity ordering of input variables agrees with the benchmark. In this test case, it seems the correlation structure has a minimal impact on prioritizing the input variables. However, the model structure seems to change the destiny of variables as regard to their importance.

4.3 Test case 3: A simple nonlinear model

In order to evaluate the performance of the proposed method when the transfer function is nonlinear with correlated inputs, the simplest nonlinear model with three-variable inputs, first proposed by Xu and Gertner (2008), is considered in this study:

$$Y = \frac{X_1 X_2}{X_3} \tag{45}$$

The marginal distribution of each input is uniform i.e., $X_i \sim U(1,10)$, and the SRCC of the model inputs are:

$$\begin{bmatrix} 1 & 0.4 & 0.2 \\ 0.4 & 1 & 0.4 \\ 0.2 & 0.4 & 1 \end{bmatrix}$$

Table 3 shows the results of SA regarding the aforementioned methods cited earlier.

Table 3 Summary of SA--Test case 3

	Variance -based (Sobol)	Proposed method $R^2 = 0.6096$	Coupling variance-based SA with PCR (1-comp) $R^2 = 0.1131$ $\frac{\sum \lambda_i}{3} = .579$	Coupling variance-based SA with PCR (2-comp) $R^2 = 0.5932$ $\frac{\sum \lambda_i}{3} = .837$	Benchmark	S^{tc}
X_1	0.1706	0.5369	0.2611	0.7541	0.3183	0.6228
X_2	0.1789	0.3309	0.3032	0.0265	0.2057	0.3975
X_3	0.4176	0.1322	0.4351	0.2186	0.1236	0.0203

As the table shows, the most important input variable according to the Sobol scheme is X_3. However, both the proposed scheme as well as the benchmark solution found the most important variable to be X_1. Even more, the result is quite consistent with S^{tc}. By contrast, conducting the variance-based approach on PCR using either one and/or two components, cannot capture the degree of importance of input variables.

4.4 Test case 4: A nonlinear model--A typical version of the Portfolio model

This nonlinear test function, which is a typical version of the portfolio model, has four variables with the following equation:

$$Y = X_1X_3 + X_2X_4 \tag{46}$$

This relatively strong test case with respect to the degree of nonlinearity, as relates to interactions among the input variables, can be examined in the following three scenarios.

4.4.1 The first scenario

This scenario was utilized by Kucherenko et al. (2012). The marginal distributions of variables are normal, i.e., $X_i \sim N(\mu, \sigma)$ with $\mu \equiv (0, 0, 250, 400)$, $\sigma \equiv (4, 2, 200, 300)$ and Spearman rank correlation among factors are: $\rho_{12}^s = 0.8$, $\rho_{34}^s = 0.8$ and the SRCC for the remaining variables are assumed to be zero. The importance of each input based on the aforementioned approaches is quantified and documented in Table 4.

Table 4 Summary of SA of the first scenario--Test case 4

			Coupling variance-based SA with PCR			
	Variance-based (Sobol)	Proposed method $R^2 = .591$	(2-comp) $R^2 = .593$ $\frac{\sum \lambda_i}{4} = .905$	(3-comp) $R^2 = .594$ $\frac{\sum \lambda_i}{4} = .955$	Benchmark	S^{tc}
X_1	0.3802	0.5140	0.5024	0.5564	0.5791	0.5045
X_2	0.2399	0.4822	0.4975	0.4412	0.5550	0.4880
X_3	0.0074	0.0012	0.0001	0.0009	0.0105	0.0029
X_4	0.0049	0.0026	2.6871e-07	0.0015	0.0181	0.0047

As Table 4 clearly demonstrates, the suggested implemented technique along with the results summarized in the last column (S^{tc}) and yardstick solution are in good agreement. This could be attributed to the fact that all schemes take care the impact of correlation structure among input variables effectively. What's more, this ranking can be achieved using PCR based on two components. In contrast, performing variance-based SA on PCR regression using three components only ranks the inputs correctly and cannot differentiate between the importance of X_1 and X_2. Similar to the test case 2, adding the correlation structure cannot

change the rank of the inputs, although with respect to the correlation among inputs, X_2 becomes slightly less important than X_1.

4.4.2. The second scenario

With the exception of the correlation structure, all conditions of this test case are the same as the preceding scenario. The correlation structure is: $\rho_{14}^{S} = 0.8$, $\rho_{34}^{S} = 0.3$ and the remaining pairs are assumed to be independent. Fig. 2 demonstrates the scatter plot matrix for this test case. The results of SA are summarized in Table 5.

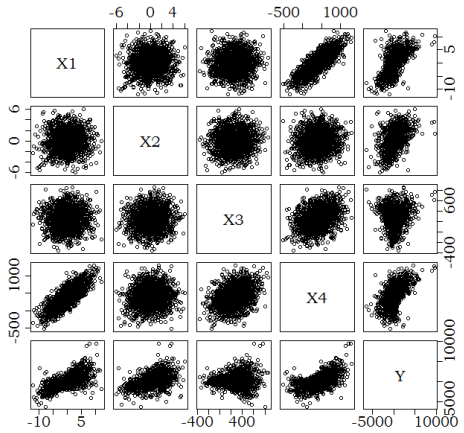

Fig. 2 Scatterplot matrix for the second scenario of test case 4

Table 5 Summary of SA of the second scenario--Test case 4

	Variance-based (Sobol)	Proposed method $R^2 = 0.512$	Coupling variance-based SA with PCR (2-comp) $R^2 = 0.371$ $\frac{\sum \lambda_i}{4} = .730$	Coupling variance-based SA with PCR (3-comp) $R^2 = 0.554$ $\frac{\sum \lambda_i}{4} = .965$	Benchmark	S^{tc}
X_1	0.3802	0.4296	0.2539	0.2832	0.3755	0.4375
X_2	0.2399	0.2622	0.1298	0.4857	0.2548	0.1813
X_3	0.0074	0.0014	0.2113	0.0343	0.0104	0.0064
X_4	0.0050	0.3068	0.4051	0.1967	0.3200	0.3748

In reference to the content of Table 5, as far as the coupling of PCR with variance-based SA is concerned, this coupling cannot successfully delineate the most important variables and their proper rankings. However, the proposed method managed to capture the benchmark solution and the results are quite compatible with S^{tc} summarized in the last column of the table, albeit with the small Multiple R-squared: 0.5116. The low value of the coefficient of determination stems from the nonlinear relationship between inputs and output as well as the interaction among variables. The scatterplot depicted in Fig. 2 clearly demonstrate this moderate trend in data which somehow leads to the low value of R^2.

4.4.3. The third scenario

This scenario is proposed by Ge and Menendez (2017) for which they assumed the PDF associated with each random variable has different form. In other words, in this case, the distributions of X_1, X_2, X_3 and X_4 are respectively, normal $N(0,1)$, gamma $\Gamma(2,1)$, uniform $\mathcal{U}(0,1)$, and lognormal $log\ N(0,1)$. The rank correlation coefficients are the same as the preceding scenario, i.e., $\rho_{14}^{s} = 0.8$, $\rho_{34}^{s} = 0.3$ and $\rho_{ij}^{s} = 0\ for\ i \neq 1,3\ and\ j \neq 4$. The first-order sensitivity measures based on the adopted approaches can be seen in Table 6.

Table 6 Summary of SA of the third scenario--Test case 4

			Coupling variance-based SA with PCR			
	Variance-based (Sobol)	Proposed method $R^2 = 0.685$	(2-comp) $R^2 = 0.613$ $\frac{\sum \lambda_i}{4} = .691$	(3-comp) $R^2 = 0.664$ $\frac{\sum \lambda_i}{4} = .925$	Benchmark	S^{tc}
X_1	0.0087	0.2447	0.2148	0.2933	0.3181	0.3014
X_2	0.1695	0.1632	0.2979	0.3591	0.1890	0.1211
X_3	0.0014	0.0352	0.1666	0.0280	0.0562	0.0496
X_4	0.5057	0.5569	0.3207	0.3197	0.5835	0.5280

In light of different governing PDF for each predictor variable, as Table 6 demonstrates, the proposed method managed to mimic the variation of sensitivity measures documented by the benchmark solution as well as S^{tc}, while conducting variance-based on PCR has failed to reproduce the proper ranking of model inputs.

4.5 Test case 5: The high-dimensional model similar to Sobol G function

As an additional more comprehensive example, this test case is designed to evaluate the ability of the proposed scheme in the presence of high dimensional problem. This model has the following equation:

$$F = \prod_{i=1}^{p} f_i(X_i) \quad , \quad f_i(X_i) = \frac{2X_i + d_i}{1 + d_i} \tag{47}$$

The variables incorporated into this test case have uniform standard marginal distribution and the parameter d_i is always nonnegative. Eq. (47), by its nature, could accommodate interaction of all order. The Sobol G function, a strongly nonlinear and non-monotonic model, can be generated by replacing X_i with $|2T_i - 1|$. The test case considered in this study has a lower degree of nonlinearity and is not as complex as the Sobol G function. In the Sobol G function, by increasing d_i, the importance of the corresponding variable will be reduced (Saltelli et al. 2008). This mathematical feature is very applicable to the proposed test function as

well. In this test case, we assume the F-function comprises of 12 variables and their associated coefficients are: [$d_1, d_2, \ldots, d_{12}$] = [0.01, 0.3, 0.6, 18, 25, 32, 39, 57, 77, 83, 90, 99]. Furthermore, it is assumed that the SRCCs of (X_1, X_{12}), (X_2, X_{11}) and (X_3, X_{10}) are 0.8, 0.75 and 0.7, respectively. By considering these correlation structure and other pertinent information, the estimated first-order sensitivity measures are summarised in Table 7 and Table 8 for all schemes considered in this study.

Table 7 Summary of SA--Test case 5

	Variance-based (Sobol)	Proposed method	Benchmark	S^{tc}
		$R^2 = 0.7439$		
X_1	0.4069	0.2598	0.4194	0.2363
X_2	0.2411	0.1734	0.2731	0.1559
X_3	0.1623	0.1197	0.1986	0.1079
$X_4, \ldots, X_9$	0.00	0.00	0.0	0.00
X_{10}	5.15E-05	0.0793	0.1227	0.0932
X_{11}	4.15E-05	0.1346	0.2043	0.1463
X_{12}	4.63E-05	0.1950	0.3295	0.2062

Table 8 Summary of SA--Test case 5

	Coupling variance-based SA with PCR						
	(5-comp)	(6-comp)	(7-comp)	(8-comp)	(9-comp)	(10-comp)	(11-comp)
	$R^2 = 0.7367$	$R^2 = 0.7368$	$R^2 = 0.7379$	$R^2 = 0.7381$	$R^2 = 0.7381$	$R^2 = 0.7657$	$R^2 = 0.786$
	$\frac{\sum \lambda_i}{12} = .636$	$\frac{\sum \lambda_i}{12} = .717$	$\frac{\sum \lambda_i}{12} = .795$	$\frac{\sum \lambda_i}{12} = .871$	$\frac{\sum \lambda_i}{12} = .946$	$\frac{\sum \lambda_i}{12} = .971$	$\frac{\sum \lambda_i}{12} = .988$
X_1	0.2407	0.2405	0.2397	0.2397	0.2397	0.1984	0.1830
X_2	0.1482	0.1481	0.1477	0.1477	0.1477	0.1348	0.4007
X_3	0.0996	0.1006	0.1003	0.1003	0.1003	0.3623	0.2735
$X_4, \ldots, X_9$	0	0	0	0	0	0	0
X_{10}	0.1073	0.1073	0.1080	0.1080	0.1080	0.0010	4.68E-05
X_{11}	0.1575	0.1574	0.1556	0.1556	0.1556	0.1106	2.11E-05
X_{12}	0.2523	0.2521	0.2530	0.2530	0.2530	0.1944	0.1436

Once again, Table 7 shows that both the proposed scheme as well as the S^{tc} approach are capable of reproducing the ranking stipulated by the benchmark solution. If one ignores the correlation structure inherent in input variables, the conventional scheme, Sobol, fails to highlight the importance of the last three variables. However, the correlation structure inherent in input variables triggers both the proposed scheme as well as the benchmark solution to give appropriate ranking to input variables. Upon coupling the variance-based SA with PCR (Table 8), the first-order sensitivity measures of the last three variables become more distinct but the coupling exercise did not manage to regenerate the ranking associated with those variables compared to benchmark solution.

4.6 Practical test cases

In this part, two practical test cases are considered. One of them is based on a nonlinear hydrologic model, and another considers a hydraulic model to simulate flood inundation. The details of these models are discussed below:

4.6.1 Test case 6- A simple hydrologic model

This test case represents a hydrologic model that relates instantaneous peak discharge to watershed characteristics in India by a power model as given below:

$$Q_{10} = 5.23A^{0.6965}\, S^{0.6565} {L_c}^{0.223} {P_i}^{0.7009} {H_r}^{0.1497} d^{-0.1053} C^{1.1976} \quad (48)$$

The parameters of this model are calibrated against the characteristics of 58 watersheds (McCuen and Snyder 1986). These characteristics are considered as model input variables and their probability density functions are described in Table 9.

Table 9 Description of output-inputs--Test case 6 (McCuen and Snyder 1986)

variable	meaning	Distribution
A	Drainage area (mi^2)	Beta (0.93, 1.28, 14.6, 203)
S	Channel longitudinal slope (ft/ft)	Gamma (1.66, 0.2421, 1.05)
L_c	Channel length (mi), distance along a stream from the watershed divide to the mouth of the watershed	Beta (1.84, 5.31, 5.64, 70.1)
P_i	Precipitation index (in), mean annual precipitation minus the sum of average annual evapotranspiration and mean annual snowfall (water equivalent)	Beta (2.21, 2.63, 3.69, 18.5)
H_r	Watershed relief (ft), the difference in elevation between the highest point on watershed divide and the stream at the point of discharge	Gamma(1.6, 0.0089, 39.89)
D	Drainage density (mi/mi^2), the total stream length divided by watershed area	Triangular (1.29, 2.5, 14.35)
C	Runoff coefficient, rainfall volume divided by the total volume of runoff	Triangular (0.19, 0.7, 1.06)
Q_{10}	Instantaneous peak discharge (ft^3/s) for a 10-year return period	----

The correlation structure of the model input variables based on SRCC is:

$$\begin{bmatrix} & \boldsymbol{A} & \boldsymbol{S} & \boldsymbol{L_c} & \boldsymbol{P_i} & \boldsymbol{H_r} & \boldsymbol{D} & \boldsymbol{C} \\ \boldsymbol{A} & 1 & -.44 & 0.75 & -0.01 & 0.09 & -.01 & .05 \\ \boldsymbol{S} & -.44 & 1 & -0.22 & 0.61 & 0.63 & 0.57 & 0.52 \\ \boldsymbol{L_c} & 0.75 & -0.22 & 1 & 0.15 & 0.40 & 0.10 & 0.15 \\ \boldsymbol{P_i} & -0.01 & 0.61 & 0.15 & 1 & 0.81 & 0.73 & 0.79 \\ \boldsymbol{H_r} & 0.09 & 0.63 & 0.40 & 0.81 & 1 & 0.76 & 0.62 \\ \boldsymbol{D} & -0.01 & 0.57 & 0.10 & 0.73 & 0.76 & 1 & 0.55 \\ \boldsymbol{C} & 0.05 & 0.52 & 0.15 & 0.79 & 0.62 & 0.55 & 1 \end{bmatrix}$$

After implementing the adopted approaches, including Sobol on the original model, the proposed approach, five flavors of variance-based PCR, and the benchmark solution implemented on the original model as well as the approximated equivalence of the proposed method (S^{tc}), on the respective governing models, the first-order sensitivity measures are summarized in Table 10 and Table 11.

Table 10 Summary of SA--Test case 6

	Variance-based (Sobol)	Proposed method R^2=0.738	Benchmark	S^{tc}
A	0.2697	0.0484	0.1940	0.0313
S	0.3010	0.1046	0.3008	0.1023
L_c	0.0164	0.0814	0.2947	0.0707
P_i	0.0677	0.1936	0.6399	0.2054
H_r	0.0139	0.2453	0.6923	0.2360
D	0.0050	0.1432	0.4385	0.1640
C	0.1607	0.1835	0.6392	0.1902

Table 11 Summary of SA--Test case 6

	Coupling variance-based SA with PCR				
	(2-comp) R^2 =0.7374 $\frac{\sum \lambda_i}{7} = 0.807$	(3-comp) R^2 = 0.7376 $\frac{\sum \lambda_i}{7}$ =0.88	(4-comp) R^2 =0.7714 $\frac{\sum \lambda_i}{7} = 0.93$	(5-comp) R^2 =0.7816 $\frac{\sum \lambda_i}{7}$ =0.96	(6-comp) R^2 =0.7847 $\frac{\sum \lambda_i}{7}$ =.986
A	0.0803	0.0767	0.0558	0.2032	0.2540
S	0.0435	0.0511	0.2415	0.3959	0.3696
L_c	0.1405	0.1509	0.2345	0.1056	0.0599
P_i	0.1874	0.1746	0.0881	0.0394	0.1297
H_r	0.2191	0.2384	0.1814	0.0872	0.1039
D	0.1675	0.1751	0.0040	0.0004	0.0042
C	0.1722	0.1435	0.2063	0.1790	0.0887

As can be seen in Table 10, the most influential parameters are H_r, P_i, and C based on the proposed method. Nevertheless, one can hardly differentiate among variables such as P_i and C when it comes to the sensitivity measures. In addition, parameters such as (D), (S, L_c), and (A) are divided into the third, fourth, and fifth groups in terms of their influences, respectively. This classification, borne out by the proposed approach is

very consistent with the correlation ratio method and S^{tc}. It seems in coupling variance-based SA with PCR (Table 11), the number of principal components considered is quite influential in delineating the important parameters. In our case, beyond the two principal components, the PCR approach cannot manage to reproduce the sensitivity ordering inherent in model input variables. It is quite interesting to acknowledge the fact that upon ignoring the correlation structure of input variables, Sobol approach, implemented on the original model, found the drainage area to be the most influential model input. Indeed, according to the Sobol approach, important variables are found to be unimportant (i.e., type II error) while less influential parameters are considered to be quite important (i.e., type I error).

Fig. 3 considers the impact of number of realizations on sensitivity measures based on the proposed scheme. As the figure clearly demonstrates, the number of realizations has minimal effect on S_i and one can safely choose a minimum number of realizations (e.g., 500) for sensitivity analysis.

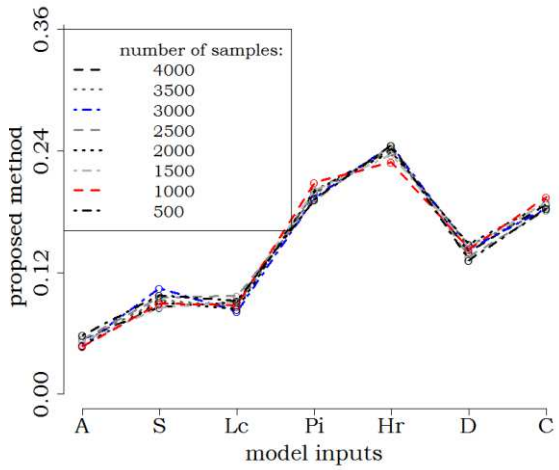

Fig. 3 Sensitivity measure (S_i) versus input variables for various number of realizations--Test case 6

4.6.2 Test case 7- A hydraulic model: flood inundation in a diversion channel

This practical example is a simplified version of a diversion channel subjected to uniform flow under flood inundation scenario. In order to protect the service road from flood inundation, a dyke is built in between the diversion channel and the service road. Indeed, this simple application, which is used as an instructive test in Iooss and Lemaître (2015), and Chastaing et al. (2012), try to simulate the height of water in the diversion channel with respect to the height of dyke intended to protect the service road, agricultural and industrial sites adjacent to the diversion channel bank from inundation. This model that comprises the characteristics of a typical channel reach has the following equation (See Fig. 4):

$$S = Z_v + h - H_d - C_b \tag{49}$$

Upon assuming uniform flow in a wide rectangular channel, h can be derived as (de Rocquigny 2006):

$$h = \left(\frac{Q}{BK_s\sqrt{\left(\frac{Z_m - Z_v}{L}\right)}}\right)^{.6} \tag{50}$$

Where S is the maximum overflow (in meter) being a function of eight inputs. Symbols used in Eq. (49) and (50) were defined in Table 12. Computation of maximum overflow calls for monitoring these eight variables. As these field variables are highly corrupted by noise and some of them also exhibit both spatial and temporal variability (e.g., Z_m and Z_v), the dependent variable S has to be considered as a random variable of its own. In Table 12, after defining each variable, the probability density function associated with each variable is also documented.

Table 12 Description of input variables--Test case 7 (Limbourg and De Rocquigny 2010; Chastaing et al. 2012; Iooss and Lemaître 2015)

random variables	meaning	unit	marginal distribution
h	Maximum annual water height in a diversion channel	m	-
Q	Maximum annual flow rate	m^3/s	Gumbel G(1013,558) truncated to [500, 3000]
K_s	Strickler coefficient	$m^{\frac{1}{3}}/s$	Normal N(30, 8) truncated to [15, ∞]
Z_v	Downstream bed channel	m	Normal N(50.19, 0.38)
Z_m	Upstream bed channel	m	Normal N(55.03, 0.45)
H_d	Dyke height	m	Uniform U(7, 9)
C_b	Pedestrian level	m	Triangular T(55, 55.5, 56)
L	Length of channel	m	Triangular T(4990, 5000, 5010)
B	Channel width	m	Triangular T(295, 300, 305)

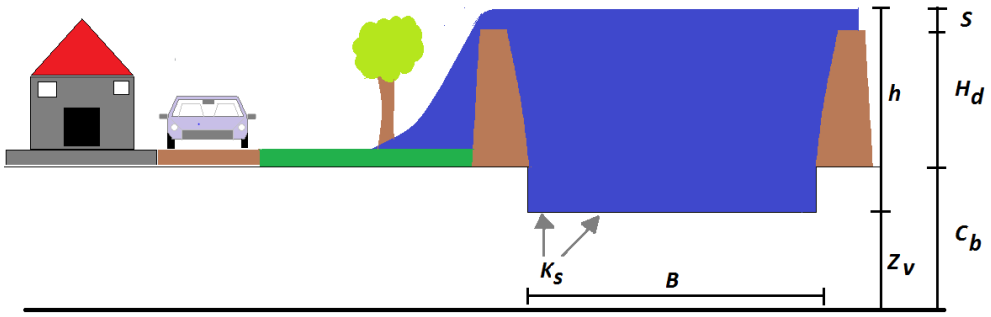

Fig. 4 flood inundation in diversion channel

As a general rule, the flow rate is highly correlated with channel roughness. As one increases the channel roughness, the flow discharge will decrease. Needless to say, Strickler coefficient is inversely proportional to channel roughness. As a result, a correlation coefficient of 0.45 is assumed for subsequent computations. Furthermore, Z_v and Z_m can be considered to be correlated with correlation coefficient 0.3. From

educational point of view, other parameters are considered to be uncorrelated. After implementing the adopted approaches on the governing equations, the first-order sensitivity measures are summarized in Table 13.

Table 13 Summary of SA--Test case 7

			Coupling variance-based SA with PCR				
	Variance -based (Sobol)	Proposed method R^2=0.9048	(5-comp) R^2=0.3482 $\frac{\sum \lambda_i}{8} = 0.7364$	(6-comp) R^2=0. 5634 $\frac{\sum \lambda_i}{8} = 0.8509$	(7-comp) R^2=0. 693 $\frac{\sum \lambda_i}{8} = 0.9339$	Benchmark	S^{tc}
Q	0.3648	0.2579	0.0673	0.0374	0.0360	0.2335	0.2314
K_s	0.1298	0.0118	0.0823	0.0450	0.0217	0.0110	0.0120
Z_v	0.1653	0.2089	0.1001	0.0693	0.3399	0.1630	0.2146
Z_m	0.0054	0.00003	0.0984	0.0610	0.0225	0.0017	0.0005
H_d	0.2903	0.4560	0.2693	0.6851	0.5175	0.3764	0.4897
C_b	0.0362	0.0596	0.0448	0.1002	0.0599	0.0485	0.0652
L	0.0000	0.0007	0.3087	0.0006	0.0002	0.0013	0.0011
B	0.0001	0.0050	0.0293	0.0011	0.0008	0.0007	0.0092

As Table 13 illustrates, the results of the proposed method agree quite well with the assessment reported in benchmark and S^{tc} solutions. According to the table content, H_d is found to have an influential impact on overflow depth followed by Q and Z_v. Indeed, in practice setting, that would be the case as far as protection of adjacent agricultural as well as industrial sites is concerned. In light of this, the height of dyke along with the maximum annual flow rate has to be chosen with proper care in practice. It is quite puzzling if one implements the variance-based SA on PCR, the methodology can trigger the height of dyke and leave the maximum annual flow rate intact. In addition, this coupling compared to the proposed method, cannot effectively demonstrate the other variables importance. Once again, the above procedural scheme was also implemented for various number of realizations. As Fig. 5 shows it seems the number of realizations has minimal impact on ranking process.

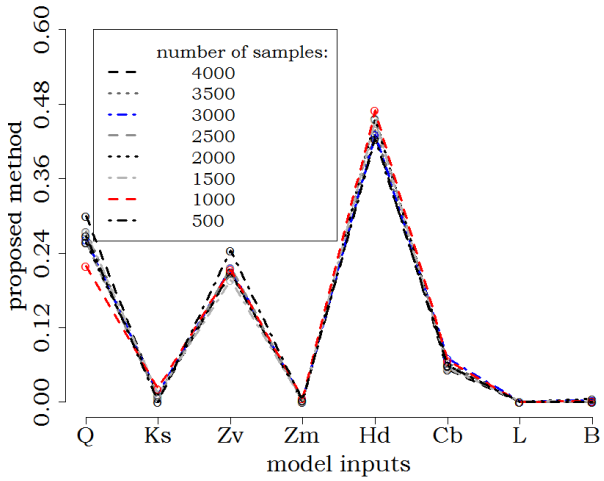

Fig. 5 Sensitivity measure (S_i) versus input variables for various number of realizations--Test case 7

5 Summary and Discussion

Nowadays, the GSA is considered to be a potent approach to organize the input variables in terms of their degree of importance as they affect the dependent variable. Over the last two decades, one of the most important challenging task in GSA is to consider the impact of correlated input variables on the variability of model output. More recent scientific endeavors and the associated literature tried to do their best to develop approaches to address this issue. In light of this, when it comes to quantitative approaches (e.g., correlation ratio method), the methodology is usually accompanied by a great deal of computational cost and is quite time consuming, particularly in the case of having high dimensionality. For this reason and due to the complexity of the developed approaches, the majority of research activities assume the decision variables to be independent for prioritization purposes. In this paper, an innovative methodology is developed thereby the conventional variance-based approach (under the assumption of orthogonal input variables) is coupled with a regression-based SPCA model to evaluate and assess the impact of correlated input variables.

In order to evaluate the effectiveness of the proposed scheme, altogether seven test cases are considered. All test cases have correlation among input variables in common. However, the model structure (i.e., linear versus nonlinear and/or low versus high dimensional) could differ from one test case to another. When it comes to the governing probability density function, a variety of scenarios are assumed. After implementing the proposed scheme on various test cases, the scheme managed to capture the ranking structure in input variables with even less computational cost and being quite consistent with the results obtained from the benchmark solution. In addition, the proposed scheme can mimic the variation among the first-order sensitivity measures very similar to the benchmark solution. It is worth mentioning, in

reference to the test case 3, first and second scenario of test case 4, the proposed method is recommended when the coefficient of determination is larger than 0.5. Research is underway to delineate a threshold for the coefficient of determination thereby one can safely make judgement on the effectiveness of the proposed approach.

Apart from the aforementioned advantages, it is necessary to touch on one disadvantage of the proposed method. As a drawback, the proposed approach cannot replicate the acceptable results in light of more complex model. This implies that this approach cannot reliably differentiate and rank the input variables when the coefficient of determination is very small. In such scenarios, it is recommended to use the benchmark solution or conventional approaches cited in the literature (e.g., Ge and Menendez 2017).

In the meanwhile, after coupling the variance-based approach (under the assumption of orthogonal input variables) with PCR, it was found that implementation of this scheme on most of the test functions, gave rise to solutions which are incompatible with the benchmark solution, the correlation ratio method. On the other hand, based on the second test function and the first scenario of the fourth test function, it seems that if results of Sobol and benchmark solutions are approximately the same, then this coupling may be somewhat successful.

We would like to conclude our paper by just noting that the developed methodology is remarkably simple to implement, time efficient particularly for high dimensional problems due to its analytical nature and it is approximately equivalent to total contribution index in the covariance decomposition equation, i.e., S^{tc} which could be assessed by noting how S^{tc} is related to Variance-Covariance structure. This equivalency can be justified by the fact that the regression-based SPCA can take into account the impact of output variable on the new features. In conclusion, the proposed scheme can be considered to be an enlightening approach for sensitivity analysis modelers. In future, it is recommended to modify and implement the proposed methodology on more complex transfer function and monitor the CPU time in comparison with the more time consuming approaches in the literature such as correlation ratio method.

Acknowledgments: We would like to thank Prof. G. Strang from MIT for his enlightening remark on symmetric matrix and uniqueness of dominant eigenvalue of matrix $\mathbb{Q}$ documented in appendix B. Furthermore, extensive implementation of SIMLAB 2.2 in this research is greatly acknowledged.

Funding: This manuscript is prepared without receiving support from any organization.

Disclosure statement: No potential conflict of interest was reported by the authors.

References

Barshan E, Ghodsi A, Azimifar Z, Zolghadri Jahromi M (2011) Supervised principal component analysis: visualization, classification and regression on subspaces and submanifolds. Pattern Recognit 44(7):1357–1371. https://doi.org/10.1016/j.patcog.2010.12.015

Bedford T (1998) Sensitivity indices for (tree)-dependent variables. In: Proceedings of the second international symposium on sensitivity analysis of model output, Venice(Italy), pp 17-20.

Borgonovo E (2007) A new uncertainty importance measure. Reliab Eng Syst Saf 92(6):771-784. https://doi.org/10.1016/j.ress.2006.04.015

Cariboni J, Gatelli D, Liska R, Saltelli A (2007) The role of sensitivity analysis in ecological modelling. Ecol Modell 203(1):167–182. https://doi.org/10.1016/j.ecolmodel.2005.10.045

Chastaing G, Gamboa F, Prieur C (2012) Generalized Hoeffding-Sobol decomposition for dependent variables-application to sensitivity analysis. Electron J Statist 6:2420–2448. https://doi.org/10.1214/12-EJS749

Ciriello V, Di Federico V, Riva M et al (2013) Polynomial chaos expansion for global sensitivity analysis applied to a model of radionuclide migration in a randomly heterogeneous aquifer. Stoch Environ Res Risk Assess 27(4): 945–954. https://doi.org/10.1007/s00477-012-0616-7

Crosetto M, Tarantola S (2001) Uncertainty and sensitivity analysis: tools for GIS-based model implementation. Int J Geogr Inf Sci 15(5):415–437. https://doi.org/10.1080/13658810110053125

Cukier R, Fortuin C, Shuler K, Petschek A, Schaibly J (1973) Study of the sensitivity of coupled reaction systems to uncertainties in rate coefficients. I Theory. J Chem Phys 59(8):3873–3878. https://doi.org/10.1063/1.1680571

Da Veiga S, Wahl F, Gamboa F (2009) Local polynomial estimation for sensitivity analysis on models with correlated inputs. Technometrics 51(4):452–463. https://doi.org/10.1198/TECH.2009.08124

De Rocquigny E (2006) La maˆıtrise des incertitues dans un context industriel-1ere partie: une approche m´ethodologique globale bas´ee sue des exemples. Journal de la Soci´et´e Fran¸caise de Statistique 147(3):33-71.

Ge Q, Menendez M (2017) Extending Morris method for qualitative global sensitivity analysis of models with dependent inputs. Reliab Eng Syst Saf 162:28–39. https://doi.org/10.1016/j.ress.2017.01.010

Gretton A, Bousquet O, Smola A J, Scholkopf B (2005) Measuring statistical dependence with hilbert-schmidt norms. In: Proceedings Algorithmic Learning Theory (ALT), springer- Berlin, pp 63–77,.

Helton JC, Davis FJ, Johnson JD (2005) A comparison of uncertainty and sensitivity analysis results obtained with random and Latin hypercube sampling. Reliab Eng Syst Saf 89(3):305–330. https://doi.org/10.1016/j.ress.2004.09.006

Iman RL, Conover WJ (1982) A distribution-free approach to inducing rank correlation among input variables. Commun Stat Simul Comput 11(3):311–334. https://doi.org/10.1080/03610918208812265

Iman RL, Johnson ME, Schroeder TA (2002) Assessing hurricane effects. Part 1. Sensitivity analysis. Reliab Eng Syst Saf 78(2):131–145. https://doi.org/10.1016/S0951-8320(02)00133-3

Iooss B, Lemaître P (2015) A review on global sensitivity analysis methods. In: Meloni C, Dellino G (Eds) Uncertainty Management in Simulation optimization of Complex Systems: Algorithms and Applications. Springer, Boston MA, pp 101-122. https://doi.org/10.1007/978-1-4899-7547-8_5

Jolliffe IT (1986) Principal Component Analysis. Springer-Verlag, New York.

Kucherenko S, Klymenko OV, Shah N (2017) Sobol'indices for problems defined in non-rectangular domains. Reliab Eng Syst Saf 167:218–231. https://doi.org/10.1016/j.ress.2017.06.001

Kucherenko S, Tarantola S, Annoni P (2012) Estimation of global sensitivity indices for models with dependent variables. Comput Phys Commun 183(4):937–946. https://doi.org/10.1016/j.cpc.2011.12.020

Lamboni M, Kucherenko S (2021) Multivariate sensitivity analysis and derivative-based global sensitivity measures with dependent variables. Reliab Eng Syst Saf 212:107519. https://doi.org/10.1016/j.ress.2021.107519

Li G, Rabitz H, Yelvington PE, Oluwole OO, Bacon F, Kolb CE, Schoendorf J (2010) Global sensitivity analysis for systems with independent and/or correlated inputs. J Phys Chem A 114(19):6022–6032. https://doi.org/10.1021/jp9096919

Limbourg P, De Rocquigny E (2010) Uncertainty analysis using evidence theory – confronting level-1 and level-2 approaches with data availability and computational constraints. Reliab Eng Syst Saf 95(5): 550-564. https://doi.org/10.1016/j.ress.2010.01.005

Mara TA, Tarantola S (2012) Variance-based sensitivity indices for models with dependent inputs. Reliab Eng Syst Saf 107:115–121. https://doi.org/10.1016/j.ress.2011.08.008

Mara TA, Tarantola S, Annoni P (2015) Non-parametric methods for global sensitivity analysis of model output with dependent inputs. Environ Model Softw 72:173–183. https://doi.org/10.1016/j.envsoft.2015.07.010

McCuen RH, Snyder WM (1986) Hydrologic modeling: statistical methods and applications. Prentice-Hall, Englewood, NJ.

McKay MD (1997) Nonparametric variance-based methods of assessing uncertainty importance. Reliab Eng Syst Saf 57(3):267–279. https://doi.org/10.1016/S0951-8320(97)00039-2

Nataf A (1962) De´termination des distributions dont les marges sont donne´es. C R Acad Sci 225:42–43.

Razavi S, Gupta HV (2016) A new framework for comprehensive, robust, and efficient global sensitivity analysis: 1. Theory. Water Resour Res 52(1):423–439. https://doi.org/10.1002/2015WR017558

Razavi S, Gupta HV (2016) A new framework for comprehensive, robust, and efficient global sensitivity analysis: 2. Application. Water Resour Res 52(1): 440– 455 https://doi.org/10.1002/2015WR017559

Saltelli A, Ratto M, Tarantola S (2001) Model-free importance indicators for dependent input. In: Proceedings of SAMO 2001, third international symposium on sensitivity analysis of model output, Madrid.

Saltelli A, Ratto M, Andres T, Campolongo F, Cariboni J, Gatelli D, Saisana M, Tarantola S (2008) Global sensitivity analysis: the primer. Wiley.

Saltelli A, Tarantola S, Campolongo F, Ratto M (2004) Sensitivity analysis in practice: a guide to assessing scientific models.Wiley.

Sobol' IM (1993) Sensitivity analysis for non-linear mathematical models, Math Modeling Comput Exp 1(4): 407–414; Translated from Russian: Sobol' IM (1990) Sensitivity estimates for nonlinear mathematical models, Matematicheskoe Modelirovanie 2:112–118 (in Russian).

Song X, Zhang J, Zhan C, Xuan Y, Ye M, Xu C (2015) Global sensitivity analysis in hydrological modeling: Review of concepts, methods, theoretical framework, and applications. J Hydrol 523: 739-757. https://doi.org/10.1016/j.jhydrol.2015.02.013

Wang P, Lu Z, Zhang K, Xiao S and Yue Z (2018) Copula-based decomposition approach for the derivative-based sensitivity of variance contributions with dependent variables. Reliab Eng Syst Saf 169:437-450. https://doi.org/10.1016/j.ress.2017.09.012

Weisberg S (2005) Applied linear regression. Wiley, New York.

Xu C (2013) Decoupling correlated and uncorrelated parametric uncertainty contributions for nonlinear models. Appl Math Model 37(24): 9950–9969. https://doi.org/10.1016/j.apm.2013.05.036

Xu C, Gertner GZ (2008) Uncertainty and sensitivity analysis for models with correlated parameters. Reliab Eng Syst Saf 93(10):1563–1573. https://doi.org/10.1016/j.ress.2007.06.003

Zhang K, Lu Z, Cheng L, Xu F (2015) A new framework of variance based global sensitivity analysis for models with correlated inputs. Struct Saf 55:1–9. https://doi.org/10.1016/j.strusafe.2014.12.005

Zheng C, Wang Q (2015) Spatiotemporal pattern of the global sensitivity of the reference evapotranspiration to climatic variables in recent five decades over China. Stoch Environ Res Risk Assess 29(8): 1937–1947. https://doi.org/10.1007/s00477-015-1120-7

Zhou Y, Lu Z, Xiao S and Yun W (2019) Distance correlation-based method for global sensitivity analysis of models with dependent inputs. Struct Multidiscip Optim 60(3):1189-1207. https://doi.org/10.1007/s00158-019-02257-z

Appendix A

The linear regression model, which can be used in place of the original model, i.e., $Y = f(X_1, X_2, \dots X_p)$ can be written as follows:

$$Y = \sum_{i=1}^{p} a_i X_i + a_0 \tag{A1}$$

In Supervised PCA approach, it is recommended to standardize the original variables (Y, X_i). Subsequently, the functional relationship between the standardized variables $(\tilde{Y}, \tilde{X}_i)$ is:

$$\tilde{Y} = \sum_{i=1}^{p} \theta_i \tilde{X}_i + \theta_0$$

$$\tilde{Y} = \left[\tilde{X}_1, \dots, \tilde{X}_p\right] \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} + \theta_0 \tag{A2}$$

For the sake of briefness, we consider:

$$\widetilde{\boldsymbol{X}} = \left[\tilde{X}_1, \dots, \tilde{X}_p\right] \; and \; \boldsymbol{\theta} = \begin{bmatrix} \theta_1 \\ \vdots \\ \theta_p \end{bmatrix} \tag{A3}$$

Thus,

$$\tilde{Y} = \widetilde{\boldsymbol{X}}\boldsymbol{\theta} + \theta_0 \tag{A4}$$

$\widetilde{\boldsymbol{X}}$ is a standardized random vector that comprises of p standardized random variables, i.e., $\tilde{X}_i$ and $\boldsymbol{\theta}$ is a vector with p scalar values, i.e., θ_i. Indeed, the dimension of $\widetilde{\boldsymbol{X}}$ and $\boldsymbol{\theta}$ are $1 * p$ and $p * 1$, respectively. This means:

$$\tilde{Y} = \widetilde{\boldsymbol{X}}_{1*p} \boldsymbol{\theta}_{p*1} + \theta_0 \tag{A5}$$

Since the identity Matrix can be multiplied by a matrix without any changing in results, we have:

$$\tilde{Y} = \widetilde{\boldsymbol{X}}_{1*p} \, I_{p*p} \, \boldsymbol{\theta}_{p*1} + \theta_0 \tag{A6}$$

Since ${U^T}_{p*p} U_{p*p} = U_{p*p} {U^T}_{p*p} = I$ (Jolliffe 1986),

$$\tilde{Y} = \widetilde{\boldsymbol{X}}_{1*p} \, U_{p*p} \; {U^T}_{p*p} \, \boldsymbol{\theta}_{p*1} + \theta_0 \tag{A7}$$

In the equation above, $\widetilde{\boldsymbol{X}}_{1*p} \, U_{p*p}$ are the new features. In other words, $\boldsymbol{Z}_{1*p} = \widetilde{\boldsymbol{X}}_{1*p} \, U_{p*p}$ or

$$[Z_1, Z_2, \dots, Z_p] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1p} \\ u_{21} & u_{22} & \dots & u_{2p} \\ . & . & . & . \\ . & . & . & . \\ u_{p1} & u_{p2} & \dots & u_{pp} \end{bmatrix} \quad \text{(A8)}$$

$\boldsymbol{Z}$ is a random vector consisting of p random variables of new feature, i.e., Z_i, and ${U^T}_{p*p}\, \boldsymbol{\theta}_{p*1}$ is a new vector, i.e., $\boldsymbol{\beta}$ that includes p values, i.e.,:

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} \quad \text{(A9)}$$

As a result:

$$\tilde{Y} = \boldsymbol{Z\beta} + \theta_0 \quad \text{(A10)}$$

Thus:

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_p Z_p + \beta_0 \quad \text{(A11)}$$

Where $\beta_0 = \theta_0$

Since the new features are uncorrelated, we can use variable reduction and define the model above based on the reduced model:

$$\tilde{Y} = \beta_1 Z_1 + \beta_2 Z_2 + \cdots + \beta_m Z_m + \beta_0 \quad \text{(A12)}$$

Appendix B

After replacing L with $\tilde{Y}\tilde{Y}^T$ in the matrix $\mathbb{Q} = \widetilde{\boldsymbol{X}}^T HLH\widetilde{\boldsymbol{X}}$**,** we have:

$$\mathbb{Q} = \widetilde{\boldsymbol{X}}^T H\, \tilde{Y}\tilde{Y}^T H\widetilde{\boldsymbol{X}} \quad \text{(B1)}$$

Since $\widetilde{\boldsymbol{X}}$ and $\tilde{Y}$ are standardized, and matrix H is symmetric, i.e., $H = H^T$, we have:

$$\widetilde{\boldsymbol{X}}^T H = (H\widetilde{\boldsymbol{X}})^T = \left(\left(I - \frac{1}{n} e_{n*1} {e^T}_{1*n}\right)\widetilde{\boldsymbol{X}}\right)^T = \left(\widetilde{\boldsymbol{X}} - \mu_{\widetilde{X}}\right)^T = \widetilde{\boldsymbol{X}}^T \quad \text{(B2)}$$

And

$$\tilde{Y}^T H = (H\tilde{Y})^T = \left(\left(I - \frac{1}{n} e_{n*1} {e^T}_{1*n}\right)\tilde{Y}\right)^T = \left(\tilde{Y} - \mu_{\tilde{Y}}\right)^T = \tilde{Y}^T \quad \text{(B3)}$$

As a result, the matrix $\mathbb{Q}$ is simplified as follows:

$$\mathbb{Q} = \widetilde{\boldsymbol{X}}^T\, \tilde{Y}\, \tilde{Y}^T \widetilde{\boldsymbol{X}} \quad \text{(B4)}$$

Indeed, the terms $\widetilde{\boldsymbol{X}}^T\ \tilde{Y}$ and $\tilde{Y}^T\widetilde{\boldsymbol{X}}$ are two column and row vectors, respectively for which their arrays are $COV(\tilde{Y}, \tilde{X}_i)$:

$$\widetilde{\boldsymbol{X}}^T{}_{p*n}\ \tilde{Y}_{n*1} = \begin{bmatrix} COV(\tilde{Y},\ \tilde{X}_1) \\ COV(\tilde{Y},\ \tilde{X}_2) \\ \\ COV(\tilde{Y},\ \tilde{X}_p) \end{bmatrix}, \quad \tilde{Y}^T{}_{1*n}\ \widetilde{\boldsymbol{X}}_{n*p} = [COV\left(\tilde{Y}, \tilde{X}_1\right), COV\left(\tilde{Y}, \tilde{X}_2\right), \dots, COV\left(\tilde{Y}, \tilde{X}_p\right)] \tag{B5}$$

In reference to Eq. (B5), Eq. (B4) implies that the matrix $\mathbb{Q}$ can be considered as multiplication of a vector i.e., $\widetilde{\boldsymbol{X}}^T{}_{p*n}\ \tilde{Y}_{n*1}$ (with p entries) by its transpose.

As a general rule, multiplication of a vector (with dimension of $p * 1$) by its transpose would always result in a symmetric matrix with rank [7] of one. In the following, the eigenvalues and eigenvectors of this matrix are further discussed (Strang 2021). For this reason, consider the vector W as follows:

$$W = \begin{bmatrix} w_1 \\ w_2 \\ \cdot \\ \cdot \\ w_p \end{bmatrix} \tag{B6}$$

And a Matrix:

$$\tilde{A} = WW^T \tag{B7}$$

For this matrix, $\|W\|^2$ can be claimed to be an eigenvalue of matrix $\tilde{A}$ corresponding to eigenvector W. This can be proved as follows:

$$\tilde{A}U = \lambda U \tag{B8}$$

After replacing U with W and $\tilde{A}$ with WW^T, we have:

$$\tilde{A}W = (WW^T)W = W(W^TW) = W\|W\|^2 = \|W\|^2W \tag{B9}$$

In summary, based on the above equation:

$$\tilde{A}W = \|W\|^2W \tag{B10}$$

As for the remaining eigenvalues and corresponding eigenvectors, such as:

$$M = \{M_2, M_3 \dots, M_p\} \tag{B11}$$

As these eigenvectors are perpendicular to W, one can show that:

[7] The definition of matrix rank can be found in almost all linear algebra book

$$W.M_i = 0 \;\rightarrow\; W^T M_i = 0, \;\; i = 2,3,\dots,p \tag{B12}$$

Thus,

$$\tilde{A}M_i = \; WW^T M_i = W(W^T M_i) = W * 0 = 0 = 0 * M_i \tag{B13}$$

This implies that the remaining eigenvalues are zero.

Based on the above theory, since the matrix $\mathbb{Q}$ is the multiplication of a vector i.e., $\tilde{\boldsymbol{X}}^T\,\tilde{Y}$ (with p entries) by its transpose, its dominant eigenvalue can be written as:

$$\begin{aligned}\lambda_1 = \left\|\tilde{\boldsymbol{X}}^T\,\tilde{Y}\right\|^2 = \left[COV\left(\tilde{Y},\;\tilde{X}_1\right)\right]^2 + \left[COV\left(\tilde{Y},\;\tilde{X}_2\right)\right]^2 + \left[COV\left(\tilde{Y},\;\tilde{X}_3\right)\right]^2 + \cdots \\ + \left[COV\left(\tilde{Y},\;\tilde{X}_p\right)\right]^2\end{aligned} \tag{B14}$$

Thus:

$$\lambda_1 = \left\|\tilde{\boldsymbol{X}}^T\,\tilde{Y}\right\|^2 = \sum_{i=1}^{p}[COV(\tilde{Y},\tilde{X}_i)]^2 \tag{B15}$$

,and other eigenvalues are zero. In addition, the eigenvector of this matrix corresponding to the dominant eigenvalue, λ_1 is $\tilde{\boldsymbol{X}}^T\,\tilde{Y}$ with its components as follows:

$$u'_{i1} = \; COV\left(\tilde{Y},\;\tilde{X}_i\right) \tag{B16}$$

As in subsequent application, we will be using the standard eigenvector, i.e. ${U_1}^T U_1 = 1$, each component of the dominant eigenvector will be divided by the magnitude of the eigenvector to make it standard as:

$$u_{i1} = \frac{COV\left(\tilde{Y},\;\tilde{X}_i\right)}{\left\|\,COV\left(\tilde{Y},\;\tilde{X}_i\right)\right\|}\,, \qquad i = 1,2,\dots,p \tag{B17}$$

Thus:

$$u_{i1} = \frac{COV\left(\tilde{Y},\;\tilde{X}_i\right)}{\left(\sum_{i=1}^{p}[COV(\tilde{Y},\tilde{X}_i)]^2\right)^{.5}} \tag{B18}$$

After using Eq. (B15), the eigenvector u_{i1} can be simplified to:

$$u_{i1} = \frac{COV\left(\tilde{Y},\;\tilde{X}_i\right)}{(\lambda_1)^{.5}}, \qquad i = 1,2,\dots,p \tag{B19}$$

The $\tilde{Y}$ and $\tilde{X}_i$ are standardized values of the original variables, thus the u_{i1} can be expressed in terms of the original variables (Y, X_i) as:

$$u_{i1} = \frac{1}{(\lambda_1)^{.5} * \sigma_Y * \sigma_{X_i}} \, COV(Y, \ X_i), \ \ i = 1,2, \dots, p \tag{B20}$$

In addition, in the following, we prove that if L is replaced by $\tilde{Y}\tilde{Y}^T$ in matrix $\mathbb{Q} = \tilde{\boldsymbol{X}}^T HLH\tilde{\boldsymbol{X}}$, the correlation between standardized output and all new features with the exception of the dominant new feature (i.e., $Z_j \ \ j \neq 1$) is zero. For this purpose, we substitute the standardized variables into Z_j, thus, $COV(\tilde{Y}, Z_j)$ ($j = 2, 3, \dots, m$) can be expanded as:

$$\begin{aligned} COV(\tilde{Y}, Z_{\mathrm{j}}) &= COV(\tilde{Y}, u_{1j}\tilde{\mathrm{X}}_1 + u_{2j}\tilde{\mathrm{X}}_2 + \cdots + u_{pj}\tilde{\mathrm{X}}_{\mathrm{p}}) \\ &= COV(\tilde{Y}, u_{1j}\tilde{X}_1) + COV(\tilde{Y}, u_{2j}\tilde{X}_2) + \cdots + COV(\tilde{Y}, u_{pj}\tilde{X}_p) \\ &= u_{1j}COV(\tilde{Y}, \tilde{X}_1) + u_{2j}COV(\tilde{Y}, \tilde{X}_2) + \cdots + u_{pj}COV(\tilde{Y}, \tilde{X}_p) \\ &= U_j \, . COV(\tilde{Y}, \tilde{X}_{\mathrm{i}}) = \langle \, U_j \, , COV(\tilde{Y}, \tilde{X}_{\mathrm{i}}) \, \rangle \, , i = 1,2, \dots, p \text{ and } j = 2,3 \dots, m \end{aligned} \tag{B21}$$

Where $\langle \, . , . \, \rangle$ denotes the inner product of two vectors.

Based on Eq. (B19), Eq. (B21) can be written as:

$$\begin{aligned} COV(\tilde{Y}, Z_{\mathrm{j}}) &= \langle \, U_j \, , COV(\tilde{Y}, \tilde{X}_{\mathrm{i}}) \rangle = \langle \, U_j \, , {\lambda_1}^{0.5} \frac{COV(\tilde{Y}, \tilde{X}_{\mathrm{i}})}{{\lambda_1}^{0.5}} \rangle \ , \qquad \mathrm{j} \neq 1 \\ &= {\lambda_1}^{0.5} * \langle \, U_j \, , \frac{COV(\tilde{Y}, \tilde{X}_{\mathrm{i}})}{{\lambda_1}^{0.5}} \rangle = {\lambda_1}^{0.5} * \langle \, U_j \, , U_1 \rangle \, , \qquad \mathrm{j} \neq 1 \end{aligned} \tag{B22}$$

Since the eigenvector U_1 is perpendicular to other eigenvectors ($U_j, j \neq 1$), $\langle \, U_j \, , U_1 \rangle = 0$. Thus, $COV(\tilde{Y}, Z_{\mathrm{j}}) = 0$ for $\mathrm{j} \neq 1$. This implies only the first new feature has maximum linear dependency on output.

Appendix C

For a linear model such as:

$$Y = a_0 + \sum_{i=1}^{p} a_i X_i \tag{C1}$$

The variance of model output with correlated inputs can be computed as:

$$\begin{aligned} V(Y) &= \sum_{i=1}^{p}\sum_{j=1}^{p} a_i a_j COV(X_i, X_j) = \sum_{i=1}^{p}\sum_{j=1}^{p} COV(a_i X_i, a_j X_j) \\ &= V(a_1 X_1) + COV(a_1 X_1 \, , a_2 X_2) + \cdots + COV(a_1 X_1 \, , a_p \, X_p) \end{aligned} \tag{C2}$$

$$+ COV(a_2X_2\,, a_1X_1) + V(a_2X_2) + \cdots + COV\left(a_2X_2\,, a_pX_p\right)$$
$$+ \cdots$$
$$+ COV\left(a_pX_p\,, a_1X_1\right) + COV\left(a_pX_p\,, a_2X_2\right) + \cdots + V\left(a_pX_p\right)$$

Based on the following equation:

$$COV\left(X, \sum_{i=1}^{M} X_i\right) = \sum_{i=1}^{M} COV(X, X_i) \tag{C3}$$

Where X and X_i are random variables. Thus, the $V(Y)$ can be rewritten as follow:

$$\begin{aligned} V(Y) = {} & V(a_1X_1) + COV\left(a_1X_1\,, a_2X_2 + a_3X_3 + \cdots + a_pX_p\right) \\ & + V(a_2X_2) + COV\left(a_2X_2\,, a_1X_1 + a_3X_3 \ldots + a_pX_p\right) \\ & + \cdots \\ & + V\left(a_pX_p\right) + COV\left(a_pX_p\,, a_1X_1 + a_2X_2 + \cdots + a_{p-1}X_{p-1}\right) \end{aligned} \tag{C4}$$

Finally, the $V(Y)$ is decomposed by the following equation:

$$V(Y) = \sum_{i=1}^{p}\left[V(a_iX_i) + COV\left(a_iX_i\,, \sum_{j=1\,, j\neq i}^{p} a_jX_j\right)\right] \tag{C5}$$

One possible interpretation of the above equation is that the impact of each variable on the uncertainty of model output is related to variance of each factor and its correlation structure with other variables. This equation can be rewritten as follows:

$$\begin{aligned} V(Y) &= \sum_{i=1}^{p}\left[COV(a_iX_i\,, a_iX_i) + COV\left(a_iX_i\,, \sum_{j=1, j\neq i}^{p} a_jX_j\right)\right] \\ &= \sum_{i=1}^{p}\left[COV\left(a_iX_i\,, \sum_{j=1, j\neq i}^{p} a_jX_j + a_iX_i\right)\right] \\ &= \sum_{i=1}^{p}\left[COV\left(a_iX_i\,, \sum_{j=1}^{p} a_jX_j\right)\right] \\ &= \sum_{i=1}^{p}[COV(a_iX_i\,, Y - a_0)] \\ V(Y) &= \sum_{i=1}^{p}[COV(Y, a_iX_i\,)] \end{aligned} \tag{C6}$$

This equation indicates that the variance of linear model output is partitioned by all $COV(Y, a_i X_i)$. Finally, we can express the variance of model output via the following equation in reference to (C5) and (C6):

$$V(Y) = \sum_{i=1}^{p} [COV(Y, a_i X_i)] = \sum_{i=1}^{p} \left[V(a_i X_i) + COV\left(a_i X_i , \sum_{j \neq i}^{p} a_j X_j \right) \right] \tag{C7}$$

In reference to the above equation, we have:

$$\sum_{i=1}^{p} [COV(Y, a_i X_i)] = \sum_{i=1}^{p} \left[V(a_i X_i) + COV\left(a_i X_i , \sum_{j \neq i}^{p} a_j X_j \right) \right] \tag{C8}$$

The main conclusion which can be drawn from the above equation is:

$$\frac{COV(Y, a_i X_i)}{V(Y)} = \frac{V(a_i X_i)}{V(Y)} + \frac{COV\left(a_i X_i , \sum_{j=1, j \neq i}^{p} a_j X_j \right)}{V(Y)}, \qquad i = 1,2, \dots, p \tag{C9}$$

It is worth mentioning, $COV(Y, a_i X_i)$ is the total contribution of random variable X_i composed of variance structure i.e., $V(a_i X_i)$ and covariance structure i.e., $COV\left(a_i X_i , \sum_{j=1, j \neq i}^{p} a_j X_j \right)$.

Appendix D

In this appendix, a rationale is offered to prove Eq. (40) in the text. After regressing the $\tilde{Y}$ on the new uncorrelated features based on PCR, we have:

$$\tilde{Y} = \beta_1 Z'_1 + \beta_2 Z'_2 + \cdots . + \beta_m Z'_m + \beta_0 \tag{D1}$$

Where Z'_j is a new feature on the basis of PCR. As mentioned in the main text, the coefficient of linear regression when the inputs are uncorrelated with zero means can be estimated by Eq.(17). Thus:

$$\hat{\beta}_i = \frac{COV(\tilde{Y}, Z'_j)}{V(Z'_i)}, \quad j = 1,2, \dots, m \tag{E2}$$

Since in PCA, $V(Z'_j) = \lambda_j$, the PCR is:

$$\tilde{Y} = \frac{COV(\tilde{Y}, Z'_1)}{\lambda_1} Z'_1 + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2} Z'_2 + \cdots . + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m} Z'_m + \beta_0 \tag{D3}$$

Where:

$$[Z'_1, Z'_2, \dots, Z'_m] = [\tilde{X}_1, \tilde{X}_2, \dots, \tilde{X}_p] \begin{bmatrix} u'_{11} & u'_{12} & \dots & u'_{1m} \\ u'_{21} & u'_{22} & \dots & u'_{2m} \\ . & . & . & . \\ . & . & . & . \\ u'_{p1} & u'_{p2} & \dots & u'_{pm} \end{bmatrix} \tag{D4}$$

Then, it can be easily deduced that the expression of the PCR in terms of the standardized variables .i.e. $\tilde{X}_i$ is:

$$\tilde{Y} = \frac{COV(\tilde{Y}, Z'_1)}{\lambda_1}(u'_{11}\tilde{X}_1 + u'_{21}\tilde{X}_2 + \cdots + u'_{p1}\tilde{X}_p) + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2}(u'_{12}\tilde{X}_1 + u'_{22}\tilde{X}_2 + \cdots + u'_{p2}\tilde{X}_p) + \cdots. + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m}(u'_{1m}\tilde{X}_1 + u'_{2m}\tilde{X}_2 + \cdots + u'_{pm}\tilde{X}_p) + \beta_0 \quad \text{(D5)}$$

Therefore:

$$\tilde{Y} = (\frac{COV(\tilde{Y}, Z'_1)}{\lambda_1}u'_{11} + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2}u'_{12} + \cdots + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m}u'_{1m})\tilde{X}_1 + \left(\frac{COV(\tilde{Y}, Z'_1)}{\lambda_1}u'_{21} + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2}u'_{22} + \cdots + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m}u'_{2m}\right)\tilde{X}_2 + \cdots + \left(\frac{COV(\tilde{Y}, Z'_1)}{\lambda_1}u'_{p1} + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2}u'_{p2} + \cdots + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m}u'_{pm}\right)\tilde{X}_p + \beta_0 \quad \text{(D6)}$$

The PCR can be expressed as regards to the original variables (Y, X_i), after some manipulations:

$$Y = \frac{c_1\sigma_Y}{\sigma_{X_1}}X_1 + \frac{c_2\sigma_Y}{\sigma_{X_2}}X_2 + \cdots + \frac{c_p\sigma_Y}{\sigma_{X_p}}X_p + E(Y) + \beta_0\sigma_Y - \sum_{i=1}^{p} c_i\frac{E(X_i)}{\sigma_{X_i}} \quad \text{(D7)}$$

$$c_i = \frac{COV(\tilde{Y}, Z'_1)}{\lambda_1}u'_{i1} + \frac{COV(\tilde{Y}, Z'_2)}{\lambda_2}u'_{i2} + \cdots + \frac{COV(\tilde{Y}, Z'_m)}{\lambda_m}u'_{im}, \qquad i = 1,2, \ldots, p$$