HAL open science

Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu, Dominique Lesselier, Bruno B Sudret, Joe Wiart

To cite this version:

Zicheng Liu, Dominique Lesselier, Bruno B Sudret, Joe Wiart. Surrogate modeling based on resampled polynomial chaos expansions. Reliability Engineering and System Safety, 2020, 202, pp.107008. 10.1016/j.ress.2020.107008 . hal-01889651

HAL Id: hal-01889651

https://centralesupelec.hal.science/hal-01889651v1

Submitted on 22 Aug 2022

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

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

Distributed under a Creative Commons Attribution - NonCommercial 4.0 International License

Version of Record: https://www.sciencedirect.com/science/article/pii/S0951832020305093
Manuscript_cdf362b607f195d5b1169af0f6d5c97e

Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu[a,b,*], Dominique Lesselier[b], Bruno Sudret[c], Joe Wiart[a]

[a] *Chaire C2M, LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France.*
[b] *Laboratoire des Signaux et Systèmes, UMR8506 (CNRS-CentraleSupelec-Université Paris-Sud), Université Paris-Saclay, Gif-sur-Yvette cedex 91192, France*
[c] *ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification, Stefano-Franscini-Platz 5, Zürich 8093, Switzerland*

Abstract

In surrogate modeling, polynomial chaos expansion (PCE) is popularly utilized to represent the random model responses, which are computationally expensive and usually obtained by deterministic numerical modeling approaches including finite-element and finite-difference time-domain methods. Recently, efforts have been made on improving the prediction performance of the PCE-based model and building efficiency by only selecting the influential basis polynomials (e.g., via the approach of least angle regression). This paper proposes an approach, named as resampled PCE (rPCE), to further optimize the selection by making use of the knowledge that the true model is fixed despite the statistical uncertainty inherent to sampling in the training. By simulating data variation via resampling (k-fold division utilized here) and collecting the selected polynomials with respect to all resamples, polynomials are ranked mainly according to the selection frequency. The resampling scheme (the value of k here) matters much and various configurations are considered and compared. The proposed resampled PCE is implemented with two popular selection techniques, namely least angle regression and orthogonal matching pursuit, and a combination thereof. The performance of the proposed algorithm is demonstrated on two analytical examples, a benchmark problem in structural mechanics, as well as a realistic case study in computational dosimetry.

Keywords:
Surrogate modeling, Sparse polynomial chaos expansion, Resampled polynomial chaos expansion, Data resampling, Sensitivity analysis, Double cross validation

1. Introduction

Mathematical modeling is common practice nowadays for better understanding real-world phenomena. However, a closed-form solution of the governing equations is unavailable in general and numerical modeling schemes, such as finite-difference time-domain (FDTD) [1] and finite element method (FEM) [2], are commonly employed. The computational method can be considered as a black-box code that takes a vector of parameters as input and yields a vector of quantities of interest that can be further used to assess the system under consideration. However, the real-world system may not be accurately modeled, one critical factor being the uncertainty of input parameters [3], which can be taken into account by setting a probabilistic model of these parameters.

Describing inputs by random variables which follow specific probabilistic density functions (PDFs) [4], the propagation of such random inputs through the system yields random outputs and the investigation of such uncertainty propagation is one of the major problems in uncertainty quantification (UQ) [5]. Monte Carlo simulations (MCS) can be applied/used to run the UQ analysis, however, it becomes intractable when the computational cost of a single simulation is high (which corresponds with the cases described here). Surrogate

*Corresponding author at Chaire C2M, LTCI, Télécom ParisTech, 46 rue Barrault, 75013 Paris. E-mail address: zicheng.liu@telecom-paristech.fr;

© 2020 published by Elsevier. This manuscript is made available under the CC BY NC user license
https://creativecommons.org/licenses/by-nc/4.0/

model (a.k.a. metamodel) is popularly utilized as a remedy to emulate the system response. Among various approaches, such as Gaussian process (Kriging method) [6], neural networks [7], etc., surrogate modeling based on polynomial chaos expansion (PCE) [8, 9, 10, 11, 12] is of interest here due to its advantages in both interpretation and versatility.

Representing the finite-variance random output on a Hilbert space spanned by multivariate basis polynomials orthogonal to the joint PDF of input variables, the numerical modeling of the system response is replaced by the computation of a PCE, while the expansion coefficients can be obtained by two different methodologies. For the so-called intrusive methods, taking the spectral finite element method [13] as an example, the classical FEM is combined with the Karhunen-Loève expansion of input random fields and the coefficients are obtained by a Galerkin scheme which results in a system of deterministic equations [14]. In contrast, without modifying the underlying code, hence as non-intrusive methods, coefficients can be obtained based on an experimental design (ED) by two popularly utilized approaches. While minimizing the mean square error of data discrepancy leads to the solution of regression method [15], projection method [16, 17] exploits the orthogonality of basis functions, the expansion coefficient being the solution of multidimensional integrations which can computed by quadrature methods.

A PCE, as an infinite series, should be truncated for computational purpose. How to perform this truncation optimally is the major issue, which is addressed in this paper. In the literature, a maximum value is commonly set to the total degree of multivariate polynomials [18]. However, the number of basis polynomials, as well as the required ED size, dramatically increases with the number of input variables, which is known as the curse-of-dimensionality. Thus, the so-called sparse PCE [19, 20, 21, 22] has been developed by only including the most influential polynomials in the truncation. Measuring this influence by correlation, the classical greedy algorithms, orthogonal matching pursuit (OMP) [23] and least angle regression (LARS) [24], have been utilized to rank the polynomials.

This contribution is aimed at stabilizing the constructed sparse PCE model with respect to small changes in the training data. Bagging (a.k.a. bootstrap aggregating) [25] is a popular approach, especially for decision tree methods, to stabilize the modeling approach by training multiple regression models based on bootstrap resamples [26] and taking the final prediction as the mean of all predictions. In the study of variable selection, rather than treating the resamples independently in the construction of regression models, the so-called *inclusion frequency* [27] (or *inclusion fraction* [28]) is computed as the criterion for the importance of a variable. With the knowledge that resamples are perturbed versions of the same original data, the truly important variables should be included in the built model for most bootstrap resamples since all models should reflect the same underlying data structure. The utilization of inclusion frequency improves the replication stability of selected variables [27, 29].

In this paper, the idea of inclusion frequency is applied to the construction of a sparse PCE model. Based on LARS or OMP, multiple PCE models are constructed based on resamples and involved basis polynomials are ranked according to the inclusion frequency. The replication stability of selected polynomials in the final model is expected to be enhanced. Since the PCE model is highly determined by the basis, the stability of the built model would be increased as well. Such construction method of a sparse PCE model is named as *resampled PCE* (rPCE).

Improvements and adjustments are made in rPCE based on the application procedure of inclusion frequency on variable selection. First, recent work in [30] shows subsampling [31] is superior to bootstrapping in the ability of distinguishing important and redundant variables and in the favor of sparse models. Remark that subsampling consists of randomly drawing part of samples without replacement while bootstrapping approach generates observations of the same size as the original data but with replacement. Here, an efficient subsampling technique, k-fold division, is applied, where the original data is divided into k parts and a resampling data set is composed of any $k-1$ parts. This procedure ensures the original data is fully explored with k resamples. In variable selection, variables are roughly labeled "important" or "redundant" by comparing the associated inclusion frequency with a cut-off value, the choice of which is still an open problem [32]. In rPCE, while the basis polynomials are ranked by inclusion frequency, the number of included polynomials in the final model is decided by cross validation. Moreover, for polynomials with the same inclusion frequency, the associated cross-validation errors are taken as an extra criterion for further ranking. Such ranking approach provides the possibility to combine different basis pursuit methods. Efforts

trying to merge the selection results of LARS and OMP are made.

This paper itself is organized as follows. A general framework of the PCE-based surrogate modeling is introduced in Section 2. Section 3 gives the concept of the full and sparse PCE truncation, where the building processes based on LARS and OMP are briefly described, respectively. The methodology of rPCE is illustrated in Section 4. Resampling data through the random division into k parts, based on the generated candidate polynomials by LARS and/or OMP, the importance of polynomials is evaluated through the inclusion frequency. The value of k matters and the determination strategy is discussed in Section 5, where the strategy to select the source of candidate polynomials (LARS, OMP, or their combination) is also presented. The improved performances in prediction and sensitivity analysis by rPCE are shown via application to two classical analytical functions, one finite-element model and one finite-difference-time-domain model in Section 6. Conclusions and perspectives follow in Section 7.

2. Surrogate model based on polynomial chaos expansion

2.1. Probabilistic modeling

Consider a physical model represented by a deterministic function $\boldsymbol{y} = \mathcal{M}(\boldsymbol{x})$, where $\boldsymbol{x} \in \mathbb{R}^M$ and $\boldsymbol{y} \in \mathbb{R}^Q$, M, Q being the number of input and output quantities, respectively. The uncertainty of inputs and the propagation to responses lead to the description of $\boldsymbol{x}$ and $\boldsymbol{y}$ as random vectors, $\boldsymbol{X}$ and $\boldsymbol{Y}$. Here, since each component of $\boldsymbol{Y}$ can be separately analyzed in statistical learning, only cases with scalar response, i.e., $Q = 1$, are considered for simplicity.

Describing the random vector $\boldsymbol{X}$ by the joint probability density function (PDF) $p_{\boldsymbol{X}}$ and assuming that Y has a finite variance, the latter belongs to a Hilbert space $L^2(\mathbb{R}^M, \mathcal{B}_M, \mathbb{P}_{\boldsymbol{X}})$, $\mathcal{B}_M$ being the Borel σ-algebra of the event space $\mathbb{R}^M$ and $\mathbb{P}_{\boldsymbol{X}}$ being the probability measure of $\boldsymbol{X}$. The Hilbert space is equipped with the following inner product

$$\langle f, g \rangle = E\left[f(\boldsymbol{X})g(\boldsymbol{X})\right] = \int_{\mathbb{X}} f(\boldsymbol{x})g(\boldsymbol{x})p_{\boldsymbol{X}}(\boldsymbol{x})d\boldsymbol{x}, \tag{1}$$

and can be represented by a complete set of orthogonal basis functions.

2.2. Polynomial chaos expansion

Polynomial chaos expansion is a spectral representation of Y taking polynomials as basis functions,

$$Y = \sum_{\boldsymbol{\alpha} \in \mathbb{N}^M} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}), \tag{2}$$

where $\boldsymbol{\alpha}$ is a vector of non-negative integers indicating the order of multivariate polynomials $\psi_{\boldsymbol{\alpha}}$ and $\beta_{\boldsymbol{\alpha}}$ is the corresponding expansion coefficient.

The construction of $\psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$ is briefly recalled now [8, 11]. Assuming that the input random variables are independent, the multivariate polynomials is a tensor product of univariate polynomials π_{α_i}, i.e.,

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \pi^{(1)}_{\alpha_1}(X_1) \times \ldots \times \pi^{(M)}_{\alpha_M}(X_M), \tag{3}$$

where $\pi^{(i)}_{\alpha_i}$'s are univariate orthonormal polynomials with respect to the PDF of the i-th parameter X_i, with degree α_i (e.g., Hermite polynomials for Gaussian distributions). This methodology is referred to as generalized PCE (gPCE) [11, 12]. For PDFs not included in gPCE, a nonlinear mapping of input variables to the known ones can be made with the technique of isoprobabilistic transformation [33, 34] or specific orthogonal polynomials are computed numerically via the Stieltjes procedure [35].

The PCE coefficients $\beta_{\boldsymbol{\alpha}}$ are obtained in a non-intrusive way by the regression approach. A data set $\{\boldsymbol{x}^{(n)}, n = 1, \ldots, N\}$ sampled from the input PDF $p_{\boldsymbol{X}}$ and the corresponding response $\{y^{(n)} = \mathcal{M}(\boldsymbol{x}^{(n)})\}$ compose altogether the ED. With notations of column vector $\boldsymbol{y} = [y^{(n)}]$, $\boldsymbol{\beta} = [\beta_{\boldsymbol{\alpha}}]$ and matrix $\boldsymbol{\psi} = [\psi_{\boldsymbol{\alpha}}(\boldsymbol{x}^{(n)})]$, the PCE coefficients can be obtained from

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{\psi}\boldsymbol{\beta}||_2^2, \tag{4}$$

which yields the ordinary least square (OLS) [36] solution as the normal equation

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\psi}^T \boldsymbol{\psi}\right)^{-1} \boldsymbol{\psi}^T \boldsymbol{y}, \tag{5}$$

the superscript "T" denoting the transpose operation. Remark that overfitting problems may be suffered for the OLS solution but can be avoided by implementing regularization [22, 37] techniques, as provided by the LARS algorithm described later on.

Remark that, although only cases with independent inputs are considered in the above analysis, it is possible to describe the mutual dependence by a copula [38] and use Rosenblatt transformation [33] to cast the problem as a function of auxiliary independent variables.

2.3. Estimation of prediction performance

The model assessment is often performed by Monte Carlo simulations with a large test dataset, which is independent from the experimental design. Denote $\widehat{\mathcal{M}}$ as the surrogate model, the input vector and response of the n-th test data as $\boldsymbol{x}_{\text{test}}^{(n)}$ and $y_{\text{test}}^{(n)}$, respectively. The performance of the constructed model is assessed by computing the mean square error of data discrepancy

$$\epsilon_{\text{test}} = \frac{1}{N_{\text{test}}} \sum_{n=1}^{N_{\text{test}}} \left(\mathcal{M}(\boldsymbol{x}_{\text{test}}^{(n)}) - \widehat{\mathcal{M}}(\boldsymbol{x}_{\text{test}}^{(n)})\right)^2. \tag{6}$$

For an easier interpretation of ϵ_{test}, the associated coefficient of determination R^2_{test} is computed by

$$R^2_{\text{test}} = 1 - \frac{\epsilon_{\text{test}}}{\text{Var}(\boldsymbol{y}_{\text{test}})}, \tag{7}$$

where $\text{Var}(\boldsymbol{y}_{\text{test}}) = \sum_{n=1}^{N_{\text{test}}} (y_{\text{test}}^{(n)} - \bar{y}_{\text{test}})^2/(N_{\text{test}} - 1)$ and $\bar{y}_{\text{test}} = \sum_{n=1}^{N_{\text{test}}} y_{\text{test}}^{(n)}/N_{\text{test}}$. Therefore, the closer R^2_{test} is to one, the more accurate is the prediction by $\widehat{\mathcal{M}}$.

However, in scenarios with high computational cost for a single simulation, it is usually intractable to have a large test dataset. Then, the same data as for training are often reused for model assessment. However, the underestimation of the generalization error is well-known in the case of overfitting [18]. Cross-validation was thus proposed and is commonly advocated [39, 40]. Here, leave-one-out cross-validation (LOOCV) is applied and the corresponding cross-validation error reads:

$$\epsilon_{\text{LOO}} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{M}(\boldsymbol{x}^{(n)}) - \widehat{\mathcal{M}}^{-(n)}(\boldsymbol{x}^{(n)})\right)^2, \tag{8}$$

where $\widehat{\mathcal{M}}^{-(n)}$ denotes the surrogate model trained by leaving the n-th data out. Remark that ϵ_{LOO} is also known as predicted residual of squares (PRESS) or jacknife error [41] and it can be computed fast in single training process [18] by

$$\epsilon_{\text{LOO}} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\mathcal{M}(\boldsymbol{x}^{(n)}) - \widehat{\mathcal{M}}(\boldsymbol{x}^{(n)})}{1 - h_n}\right)^2, \tag{9}$$

where h_n is the n-th diagonal element of the matrix $\boldsymbol{\psi}\left(\boldsymbol{\psi}^T\boldsymbol{\psi}\right)^{-1}\boldsymbol{\psi}^T$.

3. Surrogate modeling based on full PCE and sparse PCE

The accurate PCE of the true model is an infinite series and needs a truncation for the sake of computation. From Eq. (2), one sees that truncating a PCE is actually selecting a subset of $\mathbb{N}^M$ for $\boldsymbol{\alpha}$ such that the system response can be represented by the associated polynomials at a sufficient accuracy. Assuming the selected $\boldsymbol{\alpha}$ vectors compose the set $\mathbb{A}$, the truncated PCE can be written as

$$\widehat{\mathcal{M}}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{10}$$

Setting a maximum value to the total degree of polynomials leads to the so-called full PCE model, which suffers from the curse-of-dimensionality [42], meaning that the cardinality of $\mathbb{A}$ sharply increases with the number of input parameters, as explained below. While the problem of curse-of-dimensionality can be moderated by the algorithm of Smolyak sparse quadrature [43], recently least angle regression (LARS) [20, 44] and orthogonal matching pursuit (OMP) [23, 44] have been used to downsize the truncation and achieve the so-called sparse PCE model.

3.1. Full PCE model

$\mathbb{A}$ is commonly selected by setting a maximum to the total degree of multivariate polynomials, i.e., $\mathbb{A}_{\text{full}} = \{\boldsymbol{\alpha} \in \mathbb{N}^M, \sum_{i=1}^{M} \alpha_i \leq p\}$, p a positive integer. The PCE-based surrogate model with this setup is named in the sequel as the full PCE model. However, the cardinality of $\mathbb{A}_{\text{full}}$, denoted by P_{full}, equals $\binom{p+M}{p}$ and polynomially increases with the value of p and M. Moreover, to ensure the well-conditioning of the information matrix $\boldsymbol{\psi}$ in Eq. (5), the ED size N should be larger than P_{full}. As a result, the resulting curse-of-dimensionality prevents the application of the full PCE model in scenarios with large p and M.

3.2. Sparse PCE model

The problem of curse-of-dimensionality during the construction of full PCE models is addressed by constructing the so-called sparse PCE models, where $\mathbb{A}$ is downsized through the use of greedy algorithms, so that only the most influential polynomials are included in the truncated PCE.

For $p = 1, \ldots, p_{\max}$,

1. $\mathbb{A}_{\text{full}} = \{\boldsymbol{\alpha} \in \mathbb{N}^M, \sum_{i=1}^{M} \alpha_i \leq p\}$ and set active set $\mathbb{A}_0^a = \varnothing$;
2. Rank basis polynomials in $\{\psi_{\boldsymbol{\alpha}}, \boldsymbol{\alpha} \in \mathbb{A}_{\text{full}}^p\}$ by LARS, OMP, or rPCE. The $\boldsymbol{\alpha}$ corresponding with the first $J_{\max}$ most influential basis polynomials compose the set $\{\boldsymbol{\alpha}_j, j = 1, \ldots, J_{\max}\}$.
3. For $j = 1, \ldots, J_{\max}$,
   Update $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$. Based on $\boldsymbol{\psi}_{\mathbb{A}_j^a}$, compute $\boldsymbol{\beta}_j$ as the OLS solution and associated ϵ_{LOO}^j.
   End
4. $J = \arg\min_j \{\epsilon_{\text{LOO}}^j\}$ and $\epsilon_{\text{LOO}}^{p,\min} = \epsilon_{\text{LOO}}^J$. When $p \geq 3$, if $\epsilon_{\text{LOO}}^{p,\min} > \epsilon_{\text{LOO}}^{p-1,\min} > \epsilon_{\text{LOO}}^{p-2,\min}$, stop the model-construction process and output the PCE model corresponding with $\boldsymbol{\psi}_{\mathbb{A}_J^a}$.

End

Table 1: Procedures of constructing a sparse PCE model, $J_{\max} = \min\{N-1, \text{card}(\mathbb{A}_{\text{full}})\}$.

Table 1 presents the procedures to construct a sparse PCE model. Based on candidate $\boldsymbol{\alpha}$ from the full PCE model, i.e., $\boldsymbol{\alpha} \in \mathbb{A}_{\text{full}}$, the associated basis polynomials are ranked (e.g., by correlation with response data $\boldsymbol{y}$ for OMP in Table A.1) and the first $J_{\max}$ most influential ones are selected by OMP (refer to algorithm in Table A.1) or LARS (refer to Table A.2), $J_{\max}$ being the maximum number of included polynomials in the final constructed PCE model and set as $\min\{N-1, \text{card}(\mathbb{A}_{\text{full}})\}$ (otherwise the least-square problem becomes ill-posed).

Assessing the model performance by leave-one-out cross-validation, the optimal number of selected polynomials, J, corresponds with the PCE model with the minimal ϵ_{LOO}, the computation of which follows (9), where only the surrogate model constructed with the whole set of data is required.

The optimal value for the total degree of polynomials follows an early-stopping criterion. Setting the maximum value for p, a progressive increase stops when the minimal ϵ_{LOO} increase with two consequent p.

4. Surrogate modeling based on resampled PCE

During replications with resampled training data, different PCE truncations are obtained by LARS or OMP and the inclusion frequency of involved polynomials can be computed. Resampled PCE (rPCE) is

proposed to refine standard PCE truncation schemes by making use of the inclusion frequency. Cross-validation error associated with each polynomial is an additional factor to further rank polynomials with the same inclusion frequency. Efforts to combine selection results by LARS and OMP to further improve the performance of rPCE are also presented.

4.1. Resampled PCE based on LARS or OMP

Inclusion frequency is defined as the percentage of replications [27] in which a given basis polynomial is selected by LARS or OMP. The variation of training data is simulated by the subsampling technique, k-fold division, considering its efficiency in exploiting the information of original data, i.e., the training process makes use of all data in k replications.

Dividing the whole set of data into k subsets, all with approximately same size. Of k subsets, the l-th subset is left out and the remaining $k-1$ subsets are used for the PCE construction. Varying l from 1 to k, one has k PCE models built by LARS/OMP and the associated active sets are denoted by $\mathbb{A}^a_{P,(l)}$, $l = 1, \ldots, k$. The subscript "P" and superscript "a" are ignored in $\mathbb{A}^a_{P,(l)}$ to be $\mathbb{A}_{(l)}$ in the followings.

To search for the most frequent $\boldsymbol{\alpha}$ indices within the k different sets $\mathbb{A}_{(l)}$, $l = 1, \ldots, k$, one can merge the latter into a multiset $\mathbb{A}^{\text{Mul}} = \{\mathbb{A}_{(1)}, \ldots, \mathbb{A}_{(k)}\}$, the superscript "Mul" denoting a multiset (rather than set), which allows for multiple instances for each $\boldsymbol{\alpha}$. Then the selection frequency of $\boldsymbol{\alpha}$ in the k building processes is equal to the number of its duplicates in $\mathbb{A}^{\text{Mul}}$. Denote $\mathbb{A}$ as the set (thus no duplicate elements) composed of elements in $\mathbb{A}^{\text{Mul}}$. The selection frequency corresponding with each of element in $\mathbb{A}$ is an integer in the interval $[1, k]$ and saved in the vector $\boldsymbol{s}_f$. The inclusion frequency is computed as the normalized frequency, i.e., $\boldsymbol{s}_f/k$.

For applications wherein the idea of inclusion frequency has been applied, the final model keeps components (e.g., influential variables for the variable-selection problem [27]) for which the inclusion frequency exceeds the cutpoint ν. The value of ν impacts much on the stability and complexity of the final model but is usually arbitrarily taken [30, 32], and no conclusive method seems available for an optimal choice of ν [27, 29]. To avoid this problem, based on the ranked polynomials, the total number of included polynomials (rather than the cutpoint ν) in the final model is chosen by cross validation following the procedures in Table 1.

However, during the running of rPCE, different multi-indices $\boldsymbol{\alpha}$ might have the same frequency, which introduces some uncertainty in the ranking of polynomials. To avoid this uncertainty, one more factor, namely the effect of each basis polynomial on ϵ_{LOO}, is considered.

From the LARS/OMP procedures, one can see that the correlated polynomials are sequentially added into the active set, thus the increment of ϵ_{LOO} by adding $\boldsymbol{\alpha}_j$ into $\mathbb{A}^a_{j-1}$ equals $\Delta\epsilon^j_{\text{LOO}} = \epsilon^j_{\text{LOO}} - \epsilon^{j-1}_{\text{LOO}}$ for $j \geq 1$, where ϵ^0_{LOO} is set as 0. Thus, each $\boldsymbol{\alpha}$ in $\mathbb{A}^{\text{Mul}}$ corresponds with a $\Delta\epsilon_{\text{LOO}}$.

Add the superscript "(l)" to the notation standing for the quantity obtained by leaving the l-th subset out from model construction. Then, the so-called *error score* $\mathbf{s}_e$ can be computed as the mean of all terms $\Delta\epsilon^{(l),j}_{\text{LOO}}$ mapping to the same element of $\mathbb{A}$, i.e.,

$$s^i_e = \frac{1}{s^i_f \Delta\epsilon^{\max}_{\text{LOO}}} \sum_{\{(l),j|\boldsymbol{\alpha}^{(l),j}=\boldsymbol{\alpha}^i\}} \Delta\epsilon^{(l),j}_{\text{LOO}},\ i = 1, \ldots, \text{card}\{\mathbb{A}\}. \tag{11}$$

where the superscript "i" stands for the i-th element of a vector or set. The normalization by $\Delta\epsilon^{\max}_{\text{LOO}}$, the maximum element of $|\Delta\epsilon^{(l),j}_{\text{LOO}}|$, is to confine the value of s^i_e between -1 and 1 such that the ranking of polynomials by the total score

$$\mathbf{s} = \mathbf{s}_f + \mathbf{s}_e, \tag{12}$$

is mainly affected by $\mathbf{s}_f$ in rPCE. Remark that $\mathbf{s}_f$ is used instead of inclusion frequency (the normalized $\mathbf{s}_f$) and is subsequently named *frequency score*.

4.2. Resampled PCE combining LARS and OMP

The way to rank polynomials in rPCE allows the possibility to combine the results by LARS and OMP. Following the procedures in Section 4.1, $\mathbb{A}^{\text{Mul}}$ and $\mathbb{E}^{\text{Mul}}$ (multiset of $\Delta\epsilon^{(l),j}_{\text{LOO}}$) can be obtained by LARS

and OMP separately, denoted by $\mathbb{A}^{\text{Mul,LARS}}$, $\mathbb{E}^{\text{Mul,LARS}}$ and $\mathbb{A}^{\text{Mul,OMP}}$, $\mathbb{E}^{\text{Mul,OMP}}$, respectively. Then, merging results by LARS and OMP into a single multiset, $\mathbb{A}^{\text{Mul}} = \{\mathbb{A}^{\text{Mul,LARS}}, \mathbb{A}^{\text{Mul,OMP}}\}$ and $\mathbb{E}^{\text{Mul}} = \{\mathbb{E}^{\text{Mul,LARS}}, \mathbb{E}^{\text{Mul,OMP}}\}$, from which $\mathbb{A}$ and the associated total score $\boldsymbol{s}$ can be computed. Then, the basis polynomials associated with $\mathbb{A}$ are ranked according to $\boldsymbol{s}$ and the construction of a sparse PCE model follows procedures in Table 1.

5. Parameter settings

5.1. Resampling scheme

The k-fold division is used to simulate the data variation in rPCE and the value of k matters on the performance. A tradeoff lies behind the determination of k. With a small k (e.g., $k = 2$), a large portion (half) of data is apart from the building process. As a result, some information of the true system might be lost or not accurately learned by the surrogate model and the selected polynomials may not be truly influential. On the other side, a large k, (e.g., $k = N$) cannot sufficiently simulate the data statistical variation and the selected polynomials in the construction of the k different PCEs might have a high correlation. This way, the polynomials selected by rPCE would be almost the same as those with LARS or OMP and the prior knowledge, from which rPCE is to benefit, cannot be well exploited.

The proposed strategy is to merge $\mathbb{A}^{\text{Mul}}$ obtained for different values of k. Considering that the validation error on the data left out is used to estimate the prediction performance in Section 5.2 and values of $3, 5, 10, 20, N$ (leave-one-out), are usually recommended [39, 45, 46] for k-fold cross validation, rPCE will run based on the multiset $\mathbb{A}^{\text{Mul}} = [\mathbb{A}_3^{\text{Mul}}, \mathbb{A}_5^{\text{Mul}}, \mathbb{A}_{10}^{\text{Mul}}, \mathbb{A}_{20}^{\text{Mul}}, \mathbb{A}_N^{\text{Mul}}]$, where the subscript of $\mathbb{A}_q^{\text{Mul}}$ corresponds with the value of k. Data variation is fully simulated via $k = 3, 5$ and the bias error is small considering that in average about $0.86N$ resamples (without replacement) are used to generate candidate polynomials. It seems not easy to optimize the setting of k, especially considering that the optimal value may differ w.r.t. scenarios. However, the proposed setting is revealed robust in the various application examples.

With respect to a set of k values, i.e., $k = \{3, 5, 10, 20, N\}$, the total score can be computed based on $\boldsymbol{s}_{f,k}$ and $\boldsymbol{s}_{e,k}$, the subscript "k" indicating the quantity for a specific value of k. Denote $\mathbb{A}$ as the copy of $\mathbb{A}^{\text{Mul}}$ but without element duplication. For each $\boldsymbol{\alpha}$ in $\mathbb{A}$, its selection frequency can be computed by

$$f^i = \sum_{k=\{3,5,10,20,N\}} s_{f,k}^i,\ i = 1, \ldots, \text{card}(\mathbb{A}), \tag{13}$$

where the superscript "i" stands for the i-th element of a vector and $s_{f,k}^i$ equals zero if the i-th $\boldsymbol{\alpha}$ of $\mathbb{A}$ is not in $\mathbb{A}_k^{\text{Mul}}$. Since $s_{f,k}^i$ is upper bounded by k, the polynomials selected with small values of k (e.g., elements in $\mathbb{A}_3^{\text{Mul}}$) will have small values of f^i and be less likely to have high ranks in rPCE.

To solve this problem, instead of (13), the frequency score is computed as a summation of weighted $s_{f,k}^i$:

$$s_f^i = \sum_{k=\{3,5,10,20,N\}} s_{f,k}^i \frac{\text{lcm}(3, 20, N)}{k},\ i = 1, \ldots, \text{card}(\mathbb{A}), \tag{14}$$

where $\text{lcm}(3, 20, N)$ computes the least common multiple of $3, 20, N$ (same for $3, 5, 10, 20, N$). The weights give rise to the same maximum value of the summands in (14). Consequently, the candidate polynomials w.r.t. different values of k are equally considered in rPCE.

Finally, the set of k values, i.e., $\{3, 5, 10, 20, N\}$, needs an adjustment for a small N. For instance, k can only be $3, 5, 10, N$ when $N = 15$.

The computation of error score follows as:

$$s_e^i = \frac{1}{f^i} \sum_{k=\{3,5,10,20,N\}} s_{e,k}^i,\ i = 1, \ldots, \text{card}\{\mathbb{A}\}, \tag{15}$$

where $s_{e,k}^i$ equals zero if the i-th $\boldsymbol{\alpha}$ of $\mathbb{A}$ is not in $\mathbb{A}_k^{\text{Mul}}$.

5.2. Source of candidate polynomials

Section 4 presents the rPCE based on candidate polynomials generated by three sources, LARS, OMP or their combination, and one needs to decide which source is the optimal option. The polynomials commonly and frequently selected by two different approaches are believed influential and more likely to be included in rPCE. However, if one approach has a much worse performance than the other, the combination scheme would not be recommended, since the candidate polynomials generated by the worse approach might deteriorate the performance of rPCE. Therefore, if LARS is much better than OMP, only candidate polynomials by LARS participate into the ranking in rPCE, and vice versa. Otherwise, the combination scheme is used.

The criterion of "much better" should be properly set. Assuming a large set of validation data is available, as illustrated in Section 2.3, R^2_{test} can be computed as the unbiased estimation of the prediction performance. Here, the comparison of two building approaches is conducted with the analysis of the distribution of R^2_{test}. Varying the training data, a sequence of surrogate models is built and the associated R^2_{test} values are computed. Representing $\mathbb{R}^2_{\text{test,LARS}}$ and $\mathbb{R}^2_{\text{test,OMP}}$ as the sets of R^2_{test} values obtained by LARS and OMP respectively, the first and third quartile of these two sets are computed and denoted by Q_1^{LARS}, Q_1^{OMP}, Q_3^{LARS}, Q_3^{OMP}. Then, if $Q_1^{\text{LARS}} > Q_3^{\text{OMP}}$, one considers that LARS is much better than OMP, and vice versa. Otherwise, LARS and OMP are considered with similar performances and the combination scheme would be adopted.

However, again a large set of validation data is usually not available due to the high computational costs. Here, R^2_{test} is approximated through the validation on the data left out in the k-fold division. With different values of k and l, the validations generate a set of determination coefficient $R^2_{k,(l)}$ as the approximations to R^2_{test}, $l = 1, \ldots, k$, $k \in \{3, 5, 10, 20, N\}$. Denoting $\mathbb{R}^2_{\text{LARS}}$ and $\mathbb{R}^2_{\text{OMP}}$ as the sets of $R^2_{k,(l)}$ values obtained by LARS and OMP, the distribution of sets $\mathbb{R}^2_{\text{test}}$ is then simulated by $\mathbb{R}^2_{\text{LARS}}$ and $\mathbb{R}^2_{\text{OMP}}$. Remark that two layers of cross validations now have been operated in rPCE. The outer cross validation is just illustrated to simulate the distribution of $\mathbb{R}^2_{\text{LARS}}$ and $\mathbb{R}^2_{\text{OMP}}$. The inner one is embedded in the running of LARS and OMP to compute ϵ_{LOO} in Table A.1 and A.2. The two-layer cross validation here is indeed an realization of the known *double-cross-validation* (DCV) [47] or *cross model validation* (CMV) [45, 48]. The related literature shows the unbiased estimation of R^2_{test} by the determination coefficient from the outer cross-validation errors, i.e., $R^2_{k,(l)}$.

The procedures to rank basis polynomials by rPCE are summarized in Fig. 1. Then, the construction of sparse PCE models follows the steps in Table 1.

Benefiting from the obtained PCE model, the global sensitivity analysis, which measures the impacts of input variables to the response, can be conducted via the computation of Sobol' indices [49, 50] for independent variables or Kucherenko indices [51] for dependent cases by Monte-Carlo simulations. Note that in the case of independent inputs, Sobol' indices are readily available from PCE coefficients, as shown in [52].

6. Application examples

The knowledge that the influential polynomials are to be frequently selected during replications is first checked on a specially designed function, the true basis polynomials of which are known. Then, to present the performance of surrogate modeling based on rPCE and the comparisons to LARS and OMP, two benchmark functions (with dimension $M = 3$ and $M = 8$, respectively), a finite-element model (with $M = 10$) and a finite-difference-time-domain model (with $M = 4$) are analyzed. The PCE models based on LARS and OMP are obtained with the Matlab package UQLab (www.uqlab.com) [53, 54], where the maximum degree of multivariate polynomials p is set as 20. Using resampling, UQLab provides the candidate polynomials to rPCE. Remark that if no specific configurations are given in the following examples, resampled PCE is performed with the suggested configurations in Section 5, i.e., optimized source (LARS, OMP, or both) of candidate polynomials and candidate polynomials from $k = \{3, 5, 10, 20, N\}$.

Latin-Hypercube sampling [55] is used to sample the input random variables. Since cases with a small ED are concerned in this paper, the size of ED N is chosen between 10 and 50 here. As mentioned in Section 2.2, dependent variables can be analyzed after the transformation into the corresponding independent ones

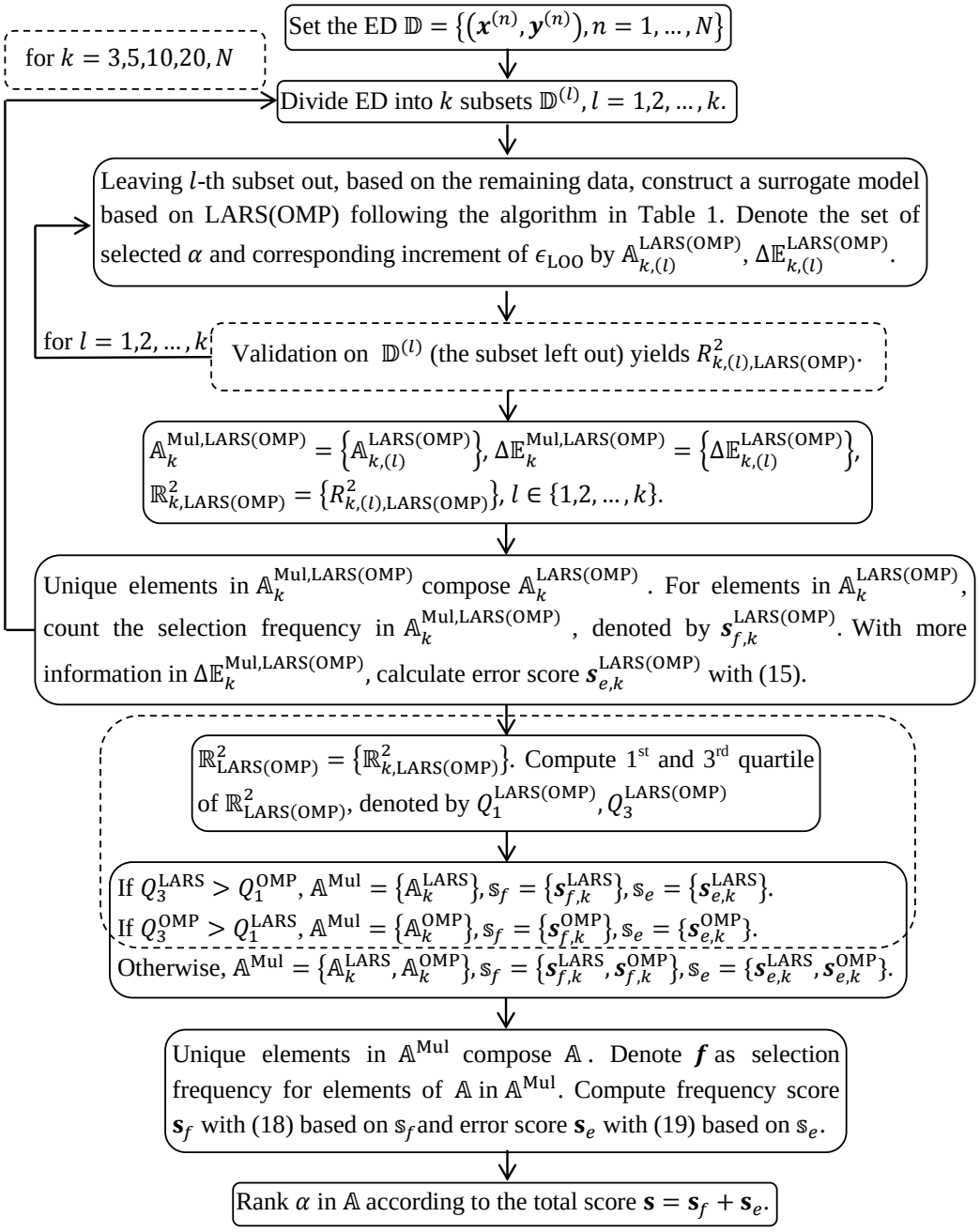


Fig. 1: Flow chart for ranking basis polynomials based on resampled PCE, where steps enclosed by dashed lines are with the suggested configurations in Section 5.

through the generalized Nataf transformation, so only examples with independent variables are presented in this section and the global sensitivity is analyzed with the computation of Sobol' indices.

6.1. Summation of multivariate polynomials

To show that the influential polynomials associated with the true model are frequently selected, the surrogate modeling of the following expression,

$$Y = 1 + X_1 + X_1X_2 + X_1X_2^2 + X_1X_2^3, \tag{16}$$

which is a summation of five multivariate polynomials (including the constant term), is conducted. X_1 and X_2 are independent variables that follow the Gaussian distributions $\mathcal{N}(0,1)$ and $\mathcal{N}(6,1)$, respectively. OMP is used to build a sparse PCE model with 12 data points for training and 10^4 data for independent testing. A total of 100 PCE constructions are made to check the selection frequency of polynomials.

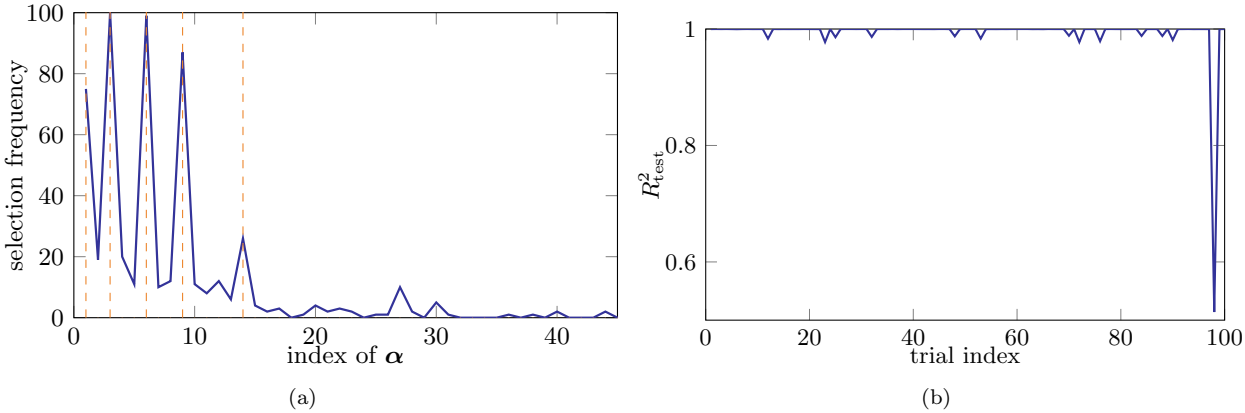


Fig. 2: Example 1: Summation of multivariate polynomials - (a) the selection frequency of $\boldsymbol{\alpha}$ by OMP and (b) the associated R^2_{test} in all replications

Due to the Gaussian distribution of input variables, Hermite polynomials are used to compose the basis, where the bivariate polynomials are indexed by $\boldsymbol{\alpha} = (\alpha_1, \alpha_2)$. The constant term corresponds with $\boldsymbol{\alpha} = (0,0)$, while the other four terms in Eq. (16) are with $(1,0), (1,1), (1,2), (1,3)$, respectively. Labeling $\boldsymbol{\alpha}$ by integers, the selection frequency during the 100 PCE constructions is plotted in Fig. 2(a), where the dashed lines indicate the five true $\boldsymbol{\alpha}$ indices. Remark that, the selection frequency is smaller than 2 when the labels are larger than 45 and only the results with labels ≤ 45 are displayed for a better visualization. As observed, although the true indices of $\boldsymbol{\alpha}$ are not always selected, they are the most frequent ones during replications. Making use of this knowledge and selecting the most frequent $\boldsymbol{\alpha}$ (also the associated polynomial) may improve the performance of the obtained PCE model and avoid the outliers (for example the 98-th replication with $R^2_{\text{test}} = 0.51$ in Fig. 2(b), where X_2, X_1, X_1^3 are selected as the basis).

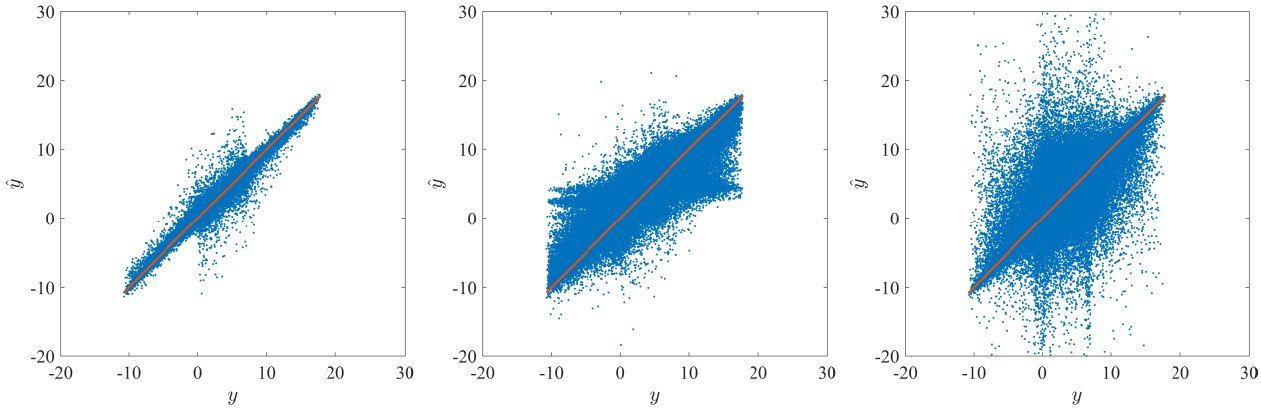
6.2. Ishigami function

The Ishigami function, which is defined by

$$Y = \sin X_1 + a \sin^2 X_2 + bX_3^4 \sin X_1, \tag{17}$$

is widely used for benchmarking in uncertainty and sensitivity analysis. The parameters are set to $a = 7$, $b = 0.1$ and the input random variables X_i, $i = 1,2,3$, are independent and uniformly distributed over $[-\pi, \pi]$. Legendre polynomials are thus used as the basis according to the principle of the generalized PCE.

First, 50 data points are used for building the surrogate model and 10^4 points for estimating the prediction performance. The analysis is repeated 100 times in order to investigate the statistical uncertainty of different modeling approaches. The prediction of all validation data (10^6 data over 100 replications) by the surrogate models built based on LARS, OMP and rPCE is shown in Fig. 3, where y stands for the true value, $\hat{y}$ for the predicted one, and the solid line indicates the case when $\hat{y}$ exactly equals y. As observed, although rPCE



(a) rPCE ($R^2_{\text{test}} = 0.9971$) (b) LARS ($R^2_{\text{test}} = 0.8724$) (c) OMP ($R^2_{\text{test}} = 0.8790$)

Fig. 3: Ishigami function - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 50 data points (100 replications).

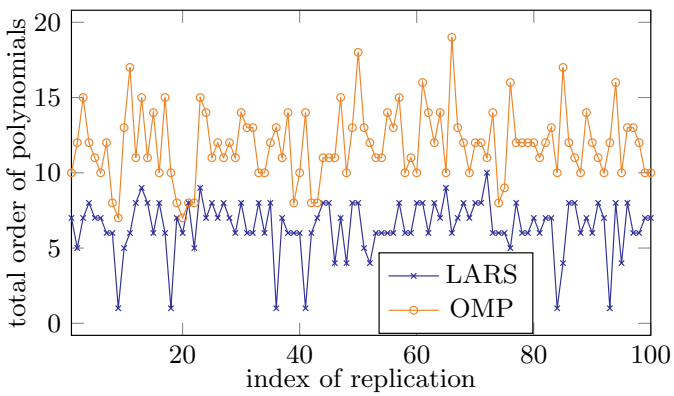


Fig. 4: Ishigami function - optimal total order of polynomials selected by LARS and OMP in 100 replications.

and OMP provide unbiased estimations of the Ishigami function, OMP suffers from more outliers and a higher variance. LARS tends to have larger predictions (relative to the true values) when $y < 0$ and smaller predictions when $y > 8$. Meanwhile, the prediction variance of LARS is not as small as rPCE.

The reason for different performances of surrogate modeling based on LARS and OMP may be seen from Fig. 4. The PCE models are with higher orders when constructed based on OMP than based on LARS. The larger value of the total order p leads to a bigger polynomial basis and thus a more flexible surrogate model, which tends to have less biased but high-variance predictions.

As mentioned in Section 5, statistical uncertainty is emulated via the k-fold division in rPCE and the value of k matters. The suggested configuration of rPCE is combining the polynomial-selection results with $k = \{3, 5, 10, 20, N\}$. To show the effects of k, R^2_{test} is computed at each replication and 100 values of R^2_{test} yield the box plots of Fig. 5, where $k = 1$ indicates the surrogate modeling with the whole set of training data but without the refinement by rPCE and "all k" denotes the rPCE results by combining results with different values of k. As observed, when $k = 1$, although the interquartile range (IQR), i.e., the span between the first quartile to the third quartile, of LARS is larger than that of OMP, more outliers appear with OMP and the minimum R^2_{test} is even smaller than -1.5. With rPCE, except the case of $k = 3$, improvements can be observed from the reduced outliers and/or prediction variance. The combination of LARS and OMP, denoted by "LARS+OMP" (see Section 4.2), seems to have advantages over the rPCE based on LARS or OMP and the advantages are more obvious with cases $k = 3$ and 5.

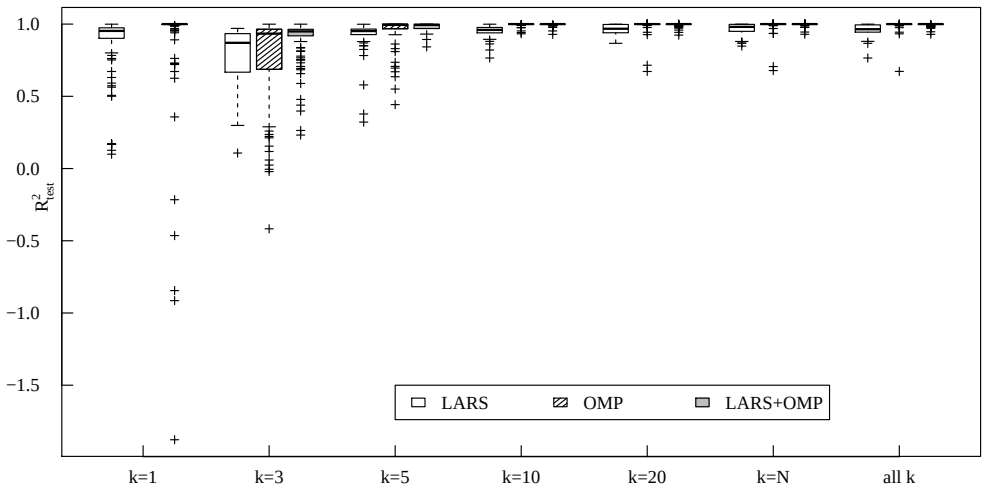


Fig. 5: Ishigami function - box plots of R^2_{test} using different values of k in k-fold division with 50 data points (100 replications).

Table 2: Ishigami function - mean of R^2_{test} over 100 replications with 50 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.8723	0.8788	
$k = 3$	0.7890	0.7734	0.8935
$k = 5$	0.9281	0.9566	0.9817
$k = 10$	0.9542	0.9972	0.9974
$k = 20$	0.9630	0.9919	0.9969
$k = N$	0.9686	0.9918	0.9978
all k	0.9619	0.9947	0.9971

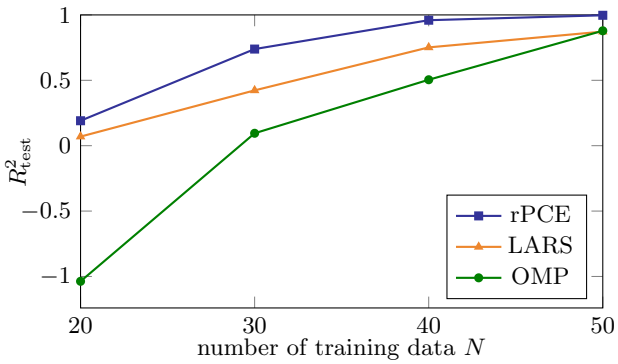


Fig. 6: Ishigami function - mean of R^2_{test} versus different values of N (100 replications).

As quantitative comparisons, Table 2 gives the mean of R^2_{test} over 100 replications. Generally, OMP is better than LARS. However, the advantage of OMP is not large and, as a result, the combination of LARS and OMP in rPCE generates better surrogate models. Remark that the means in Table 2 are obtained by fixing the value of k and the source of candidate polynomials (LARS, OMP, or LARS+OMP) during all replications. Selecting the "all k" option and optimizing the polynomial source at each replication with the suggested configuration in Section 5, the obtained mean of R^2_{test} equals 0.9972, only 6×10^{-4} smaller than the highest value when $k = N$ with LARS+OMP.

Simulations with $N = 20, 30, 40$ are also operated with the same configurations and the means of R^2_{test} are plotted as the line graph in Fig. 6, which shows the better performance of rPCE compared to LARS and OMP in the cases with small EDs.

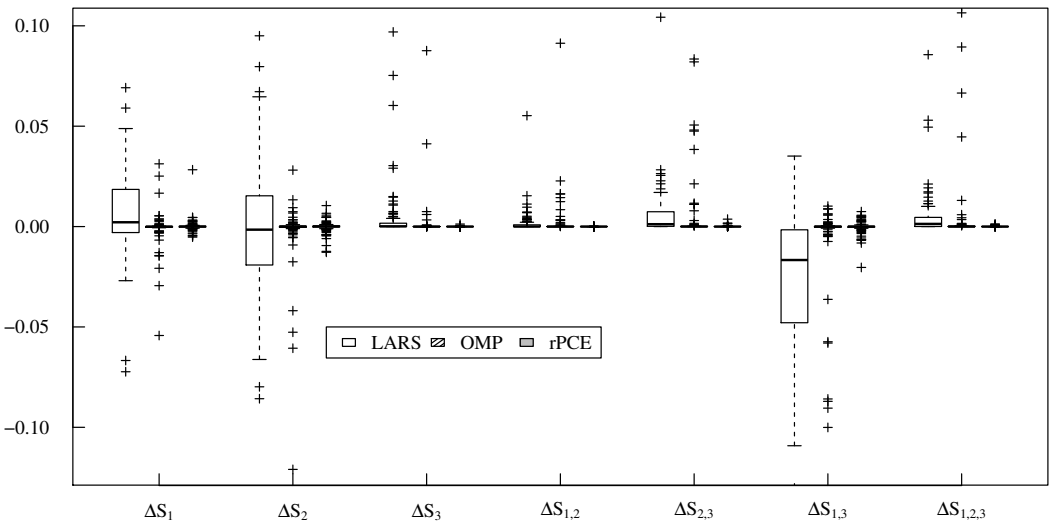


Fig. 7: Ishigami function - the estimation error of Sobol' indices with 50 data points (100 replications).

	Reference	rPCE	LARS	OMP
S_1	0.3139	0.3141	0.3553	0.3017
S_2	0.4424	0.4422	0.4152	0.4239
S_3	0.0000	0.0000	0.0114	0.0028
$S_{1,2}$	0.0000	0.0000	0.0017	0.0052
$S_{2,3}$	0.0000	0.0001	0.0096	0.0042
$S_{1,3}$	0.2437	0.2435	0.2019	0.2363
$S_{1,2,3}$	0.0000	0.0001	0.0049	0.0258

Table 3: Ishigami function - mean of Sobol' indices 50 data points (100 replications).

The Sobol' sensitivity indices can be analytically computed according to

$$\begin{aligned} D &= \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{18} + \frac{1}{2}, \\ D_1 &= \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50} + \frac{1}{2}, \\ D_2 &= \frac{a^2}{8},\ D_{1,3} = \frac{8b^2\pi^8}{225}, \\ D_3 &= D_{1,2} = D_{2,3} = D_{1,2,3} = 0. \end{aligned} \tag{18}$$

Taking the analytical solution as the reference, the estimation error of the Sobol' indices by the PCE-based surrogate model is computed by

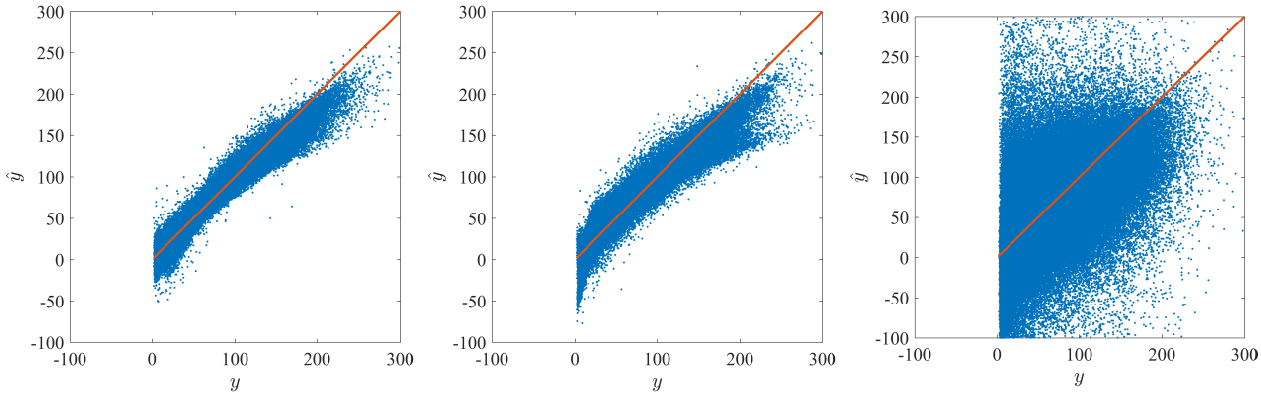
$$\Delta S_i = S_i^{\mathrm{PCE}} - S_i^{\mathrm{ref}}, \tag{19}$$

where the superscripts of S indicate the generation approach. With $N = 50$ and 100 replications, the box plots of all ΔS_i are shown in Fig. 7, where only values between -0.12 and 0.1 are presented for a better view and several outliers are absent. The variance of ΔS_i is relatively large with LARS when the Sobol' indices are non zero, i.e., ΔS_1, ΔS_2, $\Delta S_{1,3}$, and the outliers are efficiently avoided by rPCE. The mean of S_i is given by Table 3, from which the superiority of rPCE in the sensitivity analysis of the Ishigami function is obviously observed. The accuracy of rPCE for estimating Sobol' indices is in the order of 10^{-4} when using 50 data points in the experimental design.

6.3. Borehole function

Name	Distribution	Bounds	Description
r_w (m)	$\mathcal{N}(0.10, 0.0161812)$	$[0.05, 0.15]$	radius of borehole
r (m)	Lognormal(7.71, 1.0056)	$[100, 50000]$	radius of influence
T_u (m^2/yr)	Uniform	$[63070, 115600]$	transmissivity of upper aquifer
H_u (m)	Uniform	$[990, 1110]$	potentiometric head of upper aquifer
T_l (m^2/yr)	Uniform	$[63.1, 116]$	transmissivity of lower aquifer
H_l (m)	Uniform	$[700, 820]$	potentiometric head of lower aquifer
L (m)	Uniform	$[1120, 1680]$	length of borehole
K_w (m/yr)	Uniform	$[1500, 15000]$	hydraulic conductivity of borehole

Table 4: Borehole function - description and distribution of input variables [56].



(a) rPCE ($R^2_{\text{test}} = 0.9723$) (b) LARS ($R^2_{\text{test}} = 0.9517$) (c) OMP ($R^2_{\text{test}} = 0.1472$)

Fig. 8: Borehole function - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 40 data points (100 replications).

The Borehole function with expression

$$Y = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\,(1 + T_u/T_l) + 2LT_u/r_w^2 K_w} \tag{20}$$

models the water flow through a borehole and is a benchmark for emulation and prediction tests. This function has 8 independent variables, the description and distribution of which are presented in Table 4, where the range of k_w is set as $[1\,500, 15\,000]$, rather than the usual $[9\,855, 12\,045]$, to make this function more nonlinear and non-additive. For the construction of multivariate polynomials, Hermite polynomials are used for r_w and r (after an isoprobabilistic transformation into a standard normal variable) whereas Legendre polynomials are used for the other variables.

Making use of 40 data points in the model training and 10^4 points for validation at each replication, the prediction of validation data obtained from 100 replications is shown in Fig. 8. As seen, the PCE models constructed by the three methods are unbiased approximations of the Borehole function when $y < 150$. The underestimation when $y > 150$ is due to the small portion (1.63 percents for all replications) of data in this range. In prediction variance, rPCE is much better than OMP and slightly superior to LARS. The latter may be explained by observing the box plots in Fig. 9.

The poor performance with OMP may be explained from the procedures for constructing sparse PCE models in Table 1. One sees that leave-one-out cross validation error ϵ_{LOO} is used for tuning two hyperparameters, the total order p and the number of included polynomials J. p is decided by an early-stopping

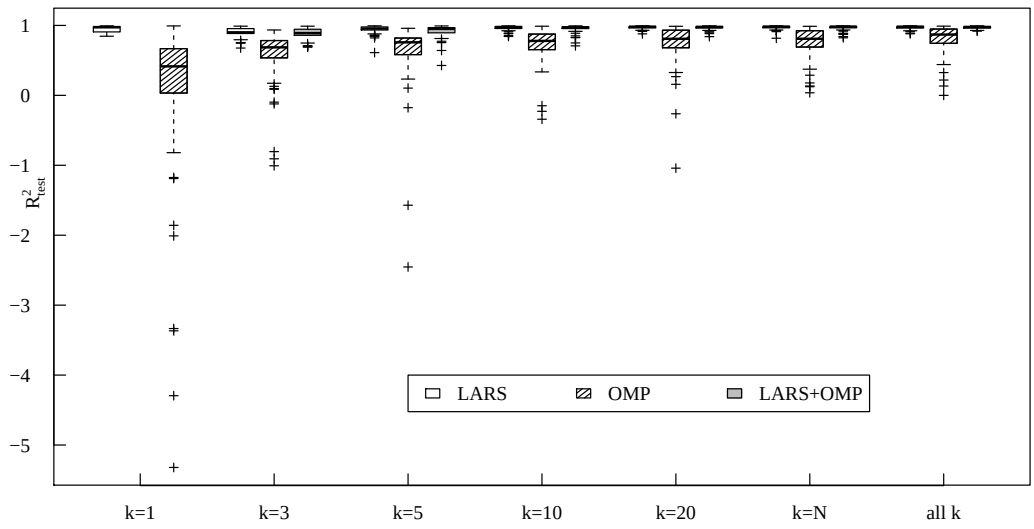


Fig. 9: Borehole function - box plots of R^2_{test} using different values of k with 40 data points (100 replications).

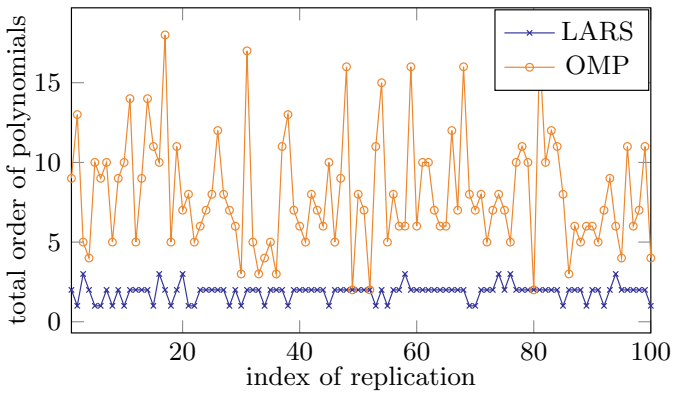


Fig. 10: Borehole function - optimal total order of polynomials selected by LARS and OMP in 100 replications.

strategy based on the minimum value of ϵ_{LOO}, i.e., $\epsilon^{\text{min}}_{\text{LOO}}$, which is obtained during the tuning of J. Since the estimation of leave-one-out cross validation is with a high variance [47] (compared with e.g., 10-fold cross-validation) and small values of ϵ_{LOO} are preferred, the resulted $\epsilon^{\text{min}}_{\text{LOO}}$ tends to be optimistic for the model assessment. However, the optimistic $\epsilon^{\text{min}}_{\text{LOO}}$ is then used to tune the value of p. The biased model-assessment by $\epsilon^{\text{min}}_{\text{LOO}}$ may lead to an improper value for p. With OMP, obviously the optimal value for p is overvalued as shown by Fig. 10 and thus the prediction is with a high variance as already seen in Fig. 8 and 9.

Building the PCE model with the whole set of data, i.e., $k = 1$, the third quartile of R^2_{test} with LARS is obviously larger than the first quartile with OMP. As explained in Section 5.2, the candidate polynomials will be generated by LARS in rPCE, rather than OMP and LARS+OMP, and the results in Fig. 9 provide good arguments for this strategy. As seen, the performance of OMP is remarkably improved after the refinement by rPCE. However, no matter the value of k, OMP is still the worst polynomial selection scheme for rPCE and LARS seems to be the best option, except that LARS+OMP is slightly better than LARS when taking the "all k" option.

Similar phenomena maybe more clearly observed from Table 5. The mean of R^2_{test} with LARS is significantly larger than the one with OMP and consequently the rPCE based on LARS is preferred. Applying this strategy and automatically selecting the candidate-polynomial source at each replication, the obtained mean of $R^2_{\text{test}} = 0.9724$. Fig. 11 provides more results when $N \in \{20, 30, 40, 50\}$. Since OMP has been

Table 5: Borehole function - mean of R^2_{test} with 40 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.9517	0.1467	
$k = 3$	0.9072	0.5852	0.8859
$k = 5$	0.9451	0.6434	0.9239
$k = 10$	0.9673	0.7293	0.9587
$k = 20$	0.9736	0.7506	0.9704
$k = N$	0.9743	0.7633	0.9697
all k	0.9719	0.8112	0.9723

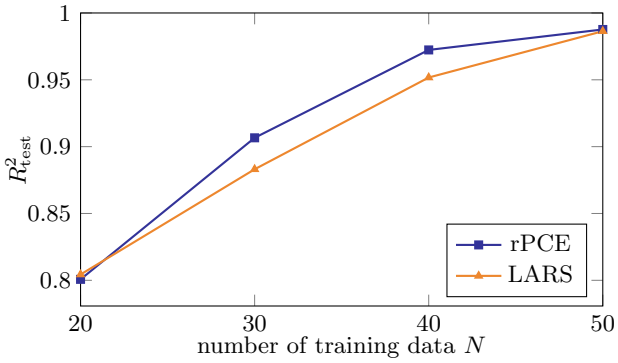


Fig. 11: Borehole function - mean of R^2_{test} versus different values of N (100 replications).

shown much worse than rPCE and LARS, its associated line graph is not displayed for a clear view of the comparison between LARS and rPCE. The improvements are reached with rPCE in general except for the case of $N = 20$.

	Reference	rPCE	LARS	OMP
r_w	0.3127	0.3072	0.2962	0.4127
r	0.0000	0.0010	0.0023	0.1967
T_u	0.0000	0.0010	0.0015	0.1635
H_u	0.0487	0.0418	0.0420	0.1995
T_l	0.0000	0.0011	0.0018	0.1802
H_l	0.0487	0.0431	0.0427	0.1751
L	0.0472	0.0423	0.0427	0.2026
K_w	0.6369	0.6376	0.6322	0.6259
$\sum$	1.0942	1.0751	1.0614	2.1562

Table 6: Borehole function - mean of the total Sobol' indices with 40 data points (100 replications).

Global sensitivity analysis is then considered and the total Sobol' indices are computed from the various PC expansions. The reference values are obtained by the Monte Carlo method with 10^7 data and presented in Table 6. The importance of variables r, T_u and T_l can be neglected and the response uncertainty mainly comes from the variation of r_w and K_w. The same conclusions can be drawn from the estimation results by rPCE and LARS. The summation of reference values is close to 1, which indicates weak variable interactions. However, the estimation by OMP leads to the opposite conclusion. The stochastic property of the estimation deviation ΔS^T is revealed by Fig. 12. The estimation variance by OMP is large, especially when the true value of S^T is small, and rPCE outperforms LARS in terms of the estimation variance and the control of outliers.

6.4. Maximum deflection of a truss structure

In Fig. 13, six vertical loads denoted by $P_1 \sim P_6$ are put on a truss structure composed of 23 bars, the cross-sectional area and Young's modulus of which are respectively denoted by A and E, the subscripts "h" and "o" standing for the horizontal and oblique bars. The response quantity of interest, the mid-span deflection V, is computed with the finite-element method.

To analyze the uncertainty of the response, the input parameters are modeled by ten independent random variables following the distributions in Table 7. Transforming the input variables into standard normal ones

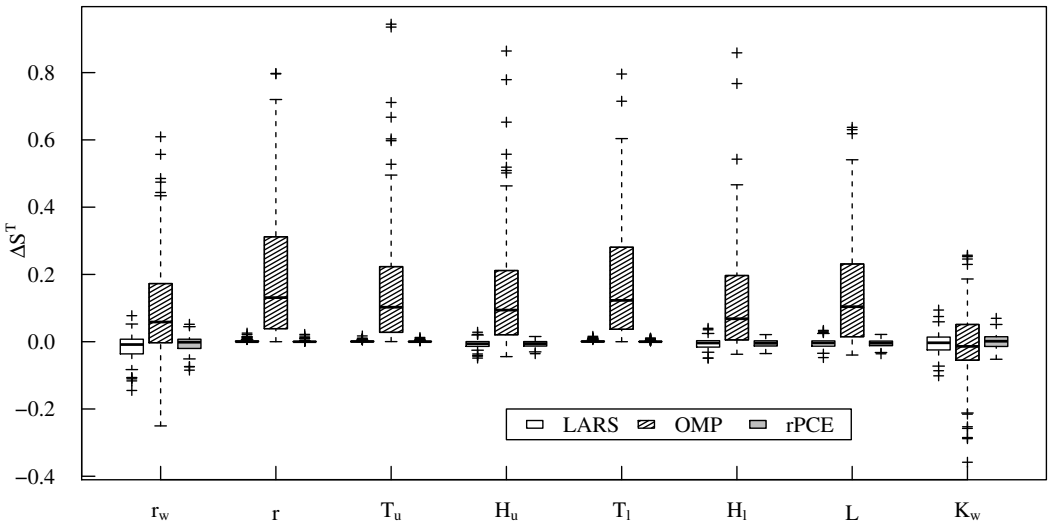


Fig. 12: Borehole function - the estimation error of total Sobol' indices with 40 data points (100 replications).

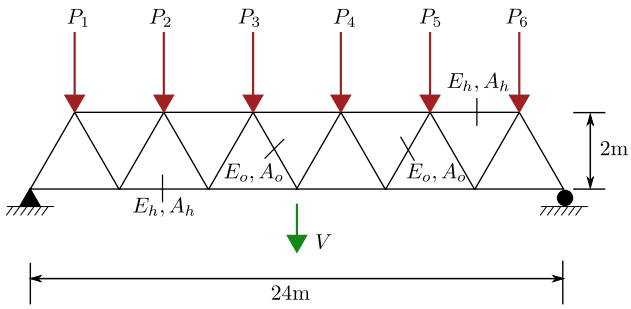


Fig. 13: Sketch of a truss structure made of 23 bars [20].

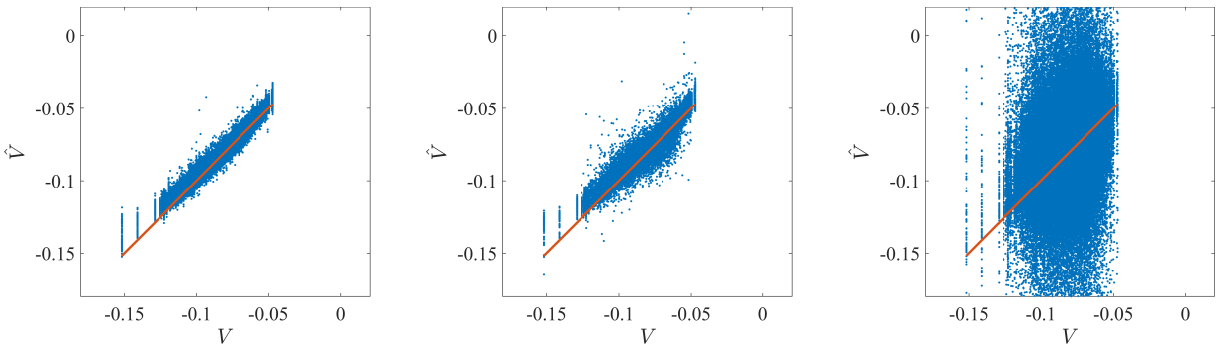
Variable	Distribution	Mean	Std	Description
E_h, E_o (Pa)	Lognormal	2.1×10^{11}	2.1×10^{10}	Young's moduli
A_h (m^2)	Lognormal	2.0×10^{-3}	2.0×10^{-4}	cross-section area of horizontal bars
A_o (m^2)	Lognormal	1.0×10^{-3}	1.0×10^{-4}	cross-section area of oblique bars
$P_1 \sim P_6$ (N)	Gumbel	5.0×10^4	7.5×10^3	vertical loads

Table 7: Truss deflection - description and distribution of input variables [20].

with the isoprobabilistic transformation, LARS, OMP and rPCE surrogate models are built with basis composed of Hermite polynomials.

With $N = 50$ and 10^4 data for validation at each replication, Fig. 14 shows the prediction results by the surrogate models over 100 replications and the solid line indicates the true values of V. OMP definitely fails in this scenario. Although the predictions are unbiased, the variance is high due to the too much flexibility of the PCE model built by OMP. In contrast, LARS and rPCE achieve a much better trade-off between the variance and bias. Moreover, rPCE is slightly superior to LARS in variance and the number of outliers. The poor prediction performance when $V < -0.11$ is a consequence of a small portion (0.78 percent for all replications) of data in this range.

Based on the validation data, R^2_{test} is computed at each replication and the distribution of R^2_{test} over 100 replications is given in Fig. 15. The results with $k = 1$ indicate the running of LARS and OMP with the whole set of data, thus no refinement of the basis by rPCE and "all k" means that rPCE is run based on



(a) rPCE ($R^2_{\text{test}} = 0.9770$) (b) LARS ($R^2_{\text{test}} = 0.9631$) (c) OMP ($R^2_{\text{test}} = -6.2257$)

Fig. 14: Truss deflection - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 50 data points (100 replications).

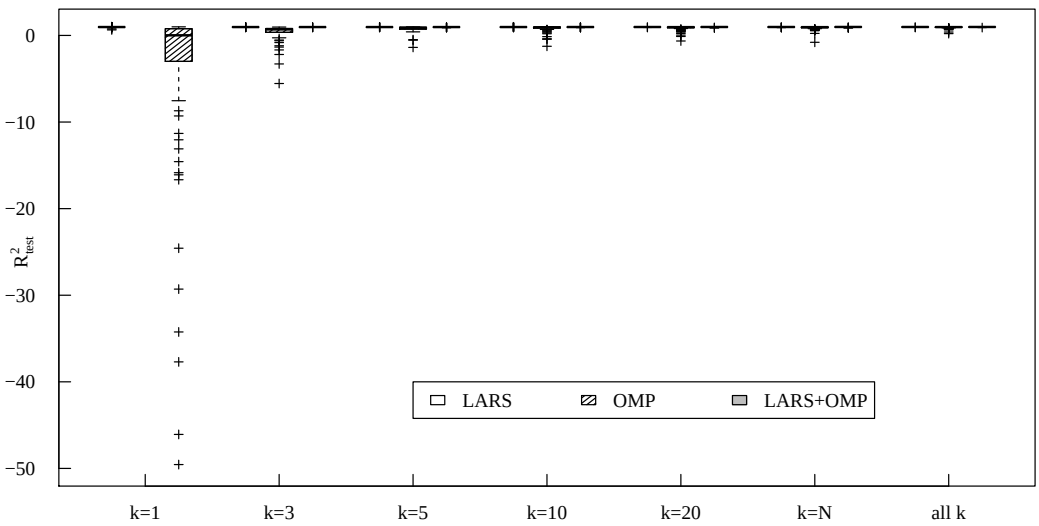


Fig. 15: Truss deflection - box plots of R^2_{test} using different values of k with 50 data points (100 replications).

Table 8: Truss deflection - mean of R^2_{test} with 50 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.9631	-6.2248	
$k = 3$	0.9651	0.3873	0.9641
$k = 5$	0.9658	0.7915	0.9660
$k = 10$	0.9692	0.8273	0.9693
$k = 20$	0.9726	0.8721	0.9735
$k = N$	0.9735	0.8974	0.9741
all k	0.9744	0.9315	0.9762

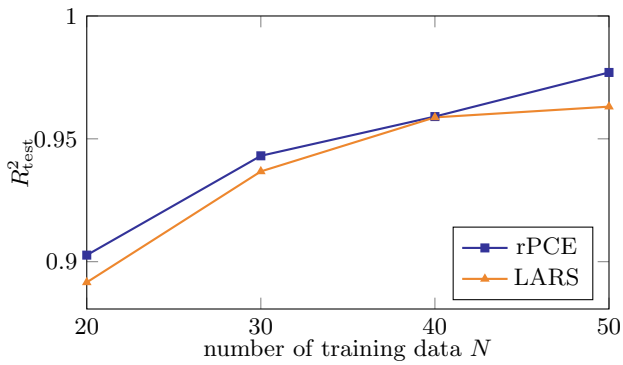


Fig. 16: Truss deflection - mean of R^2_{test} versus different values of N (100 replications).

the combination of candidate polynomials generated with $k = [3, 5, 10, 20, N]$. Although the performance of OMP is much enhanced with the application of rPCE, LARS is still better than OMP, whatever the value of k. The rPCE model combining LARS and OMP seems to have the same performance with the rPCE model based on LARS itself. Table 8 presents the associated mean of R^2_{test}. As seen, the highest mean appears with LARS+OMP when all k values are considered, but, with the same configurations, the difference between LARS and LARS+OMP is only 0.0018. Optimizing the selection of candidate polynomials at each replication, as displayed in Fig. 16, the mean value reaches 0.9770 for the "all k" option. The slight superiority of rPCE to LARS is also seen with $N = 20, 30, 40$.

	Reference	rPCE	LARS	OMP
E_h	0.367	0.3713	0.3748	0.4295
E_o	0.010	0.0121	0.0135	0.2290
A_h	0.388	0.3695	0.3715	0.4037
A_o	0.014	0.0127	0.0135	0.2291
P_1	0.004	0.0046	0.0057	0.2105
P_2	0.031	0.0359	0.0365	0.2251
P_3	0.075	0.0750	0.0759	0.2808
P_4	0.079	0.0756	0.0751	0.2557
P_5	0.035	0.0355	0.0361	0.2271
P_6	0.005	0.0048	0.0061	0.1891
$\sum$	1.008	0.9969	1.0086	2.6795

Table 9: Truss deflection - mean of the total Sobol' indices with 50 data points (100 replications).

Global sensitivity analysis is conducted by computing the total Sobol' indices based on the PCE coefficients. The reference values listed in Table 9 are obtained with 5.5×10^6 Monte Carlo simulations [20]. Since the characteristics of the horizontal bars impact more the displacement at midspan than the oblique ones, the total Sobol' indices of E_h and A_h are much larger than those of E_o and A_o. Moreover, due to the same type of probabilistic distribution and the fact that the products $E_h A_h$ (resp. $E_o A_o$) are the physically meaningful quantities in the analysis, E_h and A_h (resp. E_o and A_o) have similar importance to the response. Considering the variables of P_i, $i = 1, \ldots, 6$, P_i and P_{7-i} play the same role due to the geometric symmetry of the structure and greater sensitivities are observed for loads closer to the midspan. The above conclusions are clearly supported by the estimations of rPCE and LARS. In contrast, the largely biased estimation by OMP might give a wrong understanding of the physical phenomena. For instance, one may falsely conclude that the actually negligible interactions among inputs have great effects on the midspan deflection, since the sum of the total Sobol' indices obtained by OMP is much larger than 1.

The distribution of the prediction error of total Sobol' indices ΔS^T is given in Fig. 17. In addition to the largely biased and scattered OMP, rPCE and LARS has similar ΔS^T distribution with relatively small variances.

6.5. Estimation of specific absorption rate

The population is surrounded by a increasing number of wireless local area networks (WLAN) and the electromagnetic exposure of human body by WLAN access points needs to be estimated to make sure the exposure level is under the limit [57]. Here, an indoor down-link scenario is considered, as sketched in Fig. 18. A high-resolution model of a 8-year girl (1.36 m high), named as "Eartha", from the Virtual Classroom [58], is standing inside a $4 \times 3 \times 2$ m^3 room, which is equipped with a WLAN source operating at 2.4 GHz. The field emitted by the source is measured using the StarLab near-field-measurement system, which is based on spherical wave expansion [59], by Microwave Vision Group (MVG®). With an in-house finite-difference-time-domain (FDTD) code, the whole-body specific absorption rate (SAR) [60], which is the system response here, is computed as the ratio of the total power absorbed in the body to the mass of the human model and with the unit mW/kg.

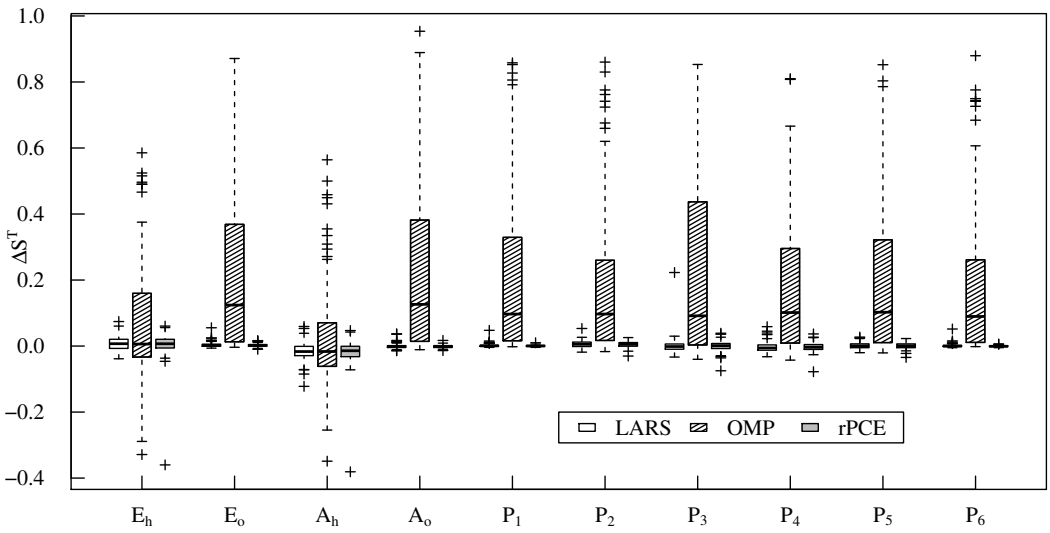


Fig. 17: Truss deflection - the estimation error of total Sobol' indices with 50 data points (100 replications).

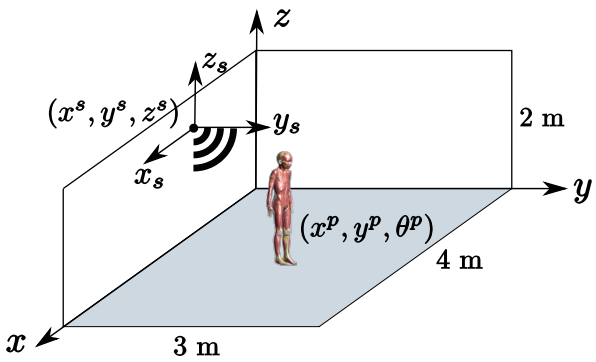
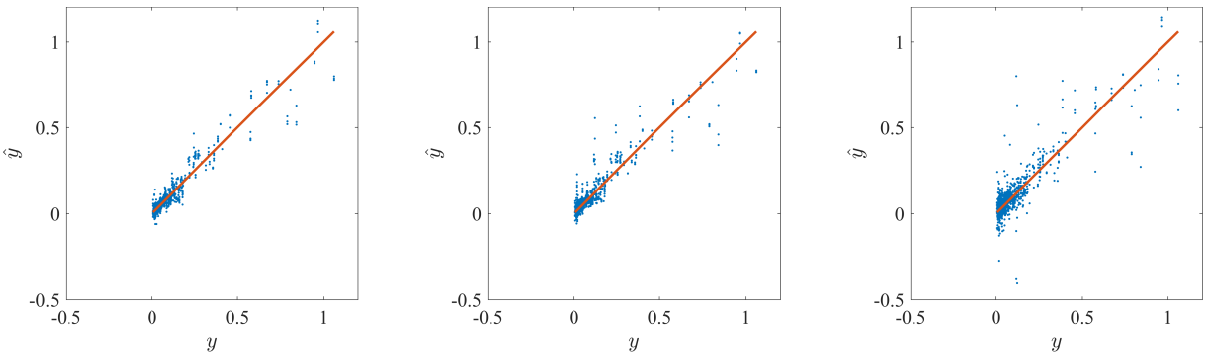


Fig. 18: Sketch of the human-exposure estimation in an indoor down-link scenario.

The parameters considered are the position of the emitting source and the human model, whose coordinates are denoted by (x^s, y^s, z^s) and (x^p, y^p, z^p), respectively. z^p is set as 0, since we consider that the human model is standing on the ground. The human orientation θ^p, which is defined as the angle between the direction faced by the human model and x-axis, may matter and is taken into account.The reflection by the walls, ceiling and ground is neglected in the simulation and the WLAN source is attached to the walls. Thus, six parameters are involved. x^s, y^s, z^s, x^p, y^p are assumed to be uniformly distributed over $[0.3, 3.7]$, $[0.3, 2.7]$, $[0.25, 2]$, $[0.05, 3.95]$, $[0.05, 2.95]$ in meters and θ^p over $[0, 360)$ in degrees, where the lower bound value 0.3 m is the minimum distance between the human model and the wall, 0.25 m is the minimum height of the source and 0.05 m is the minimum distance of the WLAN source to the wall.

The number of input variables can be reduced via a coordinate transformation. Without the reflection by the walls, the system response is actually driven by the relative position between the source and the human model. The relative position is represented in the (x, y) plane. In the local coordinate system of the source, as shown in Fig. 18, position and orientation of the human model are denoted by polar coordinates (r_s^p, ϕ_s^p) and θ_s^p. Thus, four parameters r_s^p, ϕ_s^p, θ_s^p, and z^s are used in the following uncertainty analysis.

Sampling 350 points from the input space with the Latin-Hypercube sampling method, the prediction performance of the obtained surrogate models is estimated with the leave-many-out approach, where 10 data are randomly chosen from the experimental design for validation and an approximation of R^2_{test} is yielded by repeating this process 100 times. Consequently, with the remaining 340 data, surrogate models



(a) rPCE ($R^2_{\text{test}} = 0.9102$) (b) LARS ($R^2_{\text{test}} = 0.8688$) (c) OMP ($R^2_{\text{test}} = 0.7269$)

Fig. 19: SAR estimation - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 340 data points (100 replications).

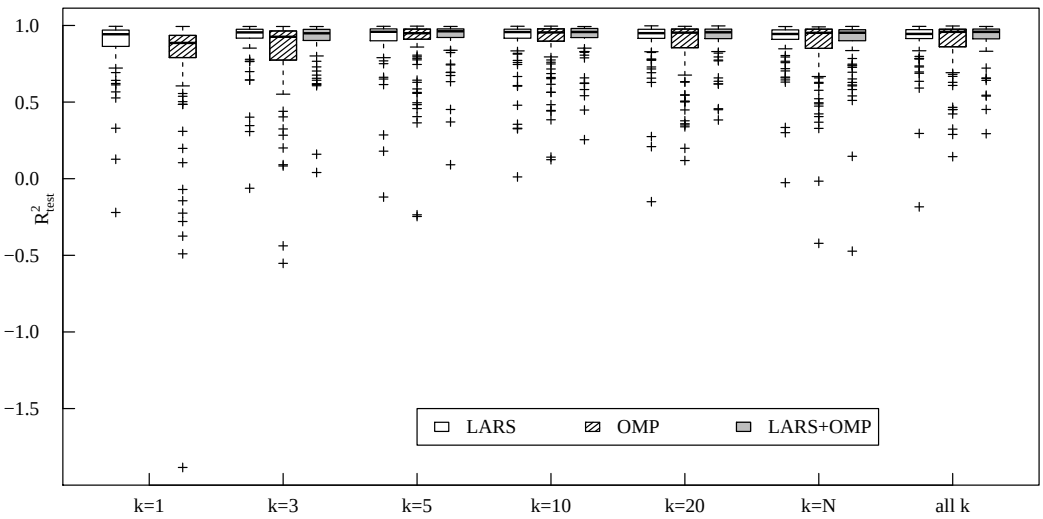


Fig. 20: SAR estimation - box plots of R^2_{test} with different values of k (100 replications).

are obtained with LARS, OMP and rPCE. Then, a validation set of size 10^3 is computed and the results are shown in Fig. 19. As seen, the whole-body SAR is smaller than 0.2 for most of cases (90 percents for all replications) in this scenario. However, the three approaches can provide unbiased estimations when the SAR value is larger than 0.2, in addition to the the superiority of rPCE to LARS and OMP in variance and suppression of outliers. The associated box plots of R^2_{test} is given in Fig. 20. The refinement by rPCE reduces the variance of modeling by LARS and OMP with different values of k, except for the case with OMP and $k = 3$. The combination of LARS and OMP seems to be the best option for rPCE and actually is selected by the suggested scheme in Section 4.2 during all replications (although three options are available at each replication), since LARS has the same-level performances with OMP. Table 10 shows the mean of R^2_{test}.

The total Sobol' indices are computed based on the PCE coefficients and the mean values are presented in Table 11. As seen, the whole-body human exposure is mainly impacted by the relative distance r^p_s and the height of the source z^s has a smaller influence. The small value w.r.t. the relative angle between the human model and the source, ϕ^p_s, might be explained by looking at the contours of electric-field intensity in Fig. 21, where the WLAN source locates at the center of a wall and field values are sampled in the (x_s, y_s) plane with $z_s = 0$. As observed, the dependency of wave strength on radiation directions is weak. The

Table 10: SAR estimation - mean of R^2_{test} with 340 data points (100 replications).

	LARS	OMP	LARS+OMP
$k = 1$	0.8799	0.7500	
$k = 3$	0.9085	0.8186	0.9046
$k = 5$	0.9067	0.8771	0.9182
$k = 10$	0.8995	0.8854	0.9171
$k = 20$	0.9033	0.8628	0.9157
$k = N$	0.8995	0.8521	0.8893
all k	0.9068	0.8794	0.9178

Table 11: SAR estimation - mean of the total Sobol' indices with 340 data points (100 replications).

	rPCE	LARS	OMP
r_s^p	0.9809	0.9714	0.9761
ϕ_s^p	0.0128	0.0357	0.0984
z^s	0.2175	0.1954	0.2925
θ_s^p	0.0098	0.0316	0.0743
$\sum$	1.2210	1.2341	1.4412

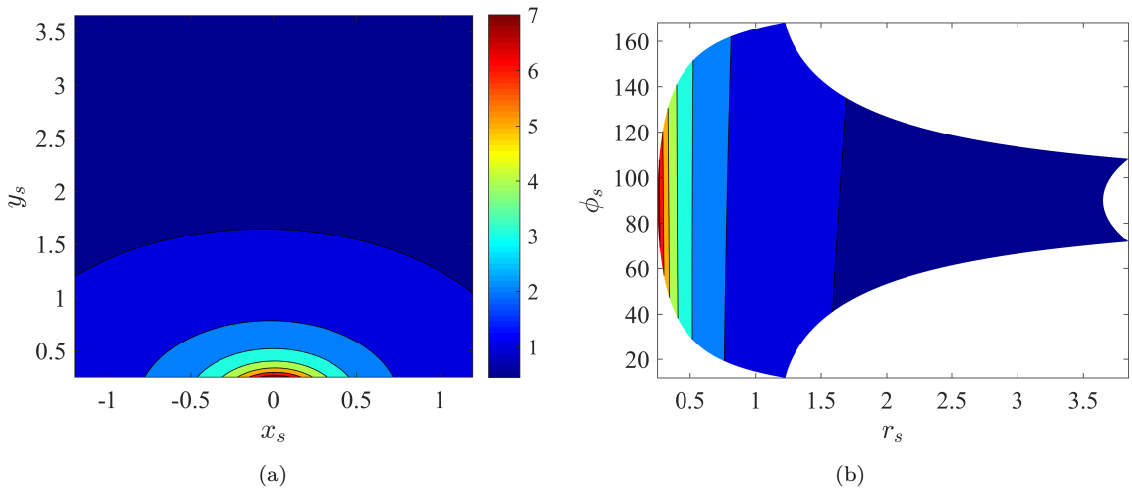


Fig. 21: SAR estimation - contour of electric-field intensity (a) in the (x, y) plane and (b) its representation in the polar coordinate system, $z_s = 0$.

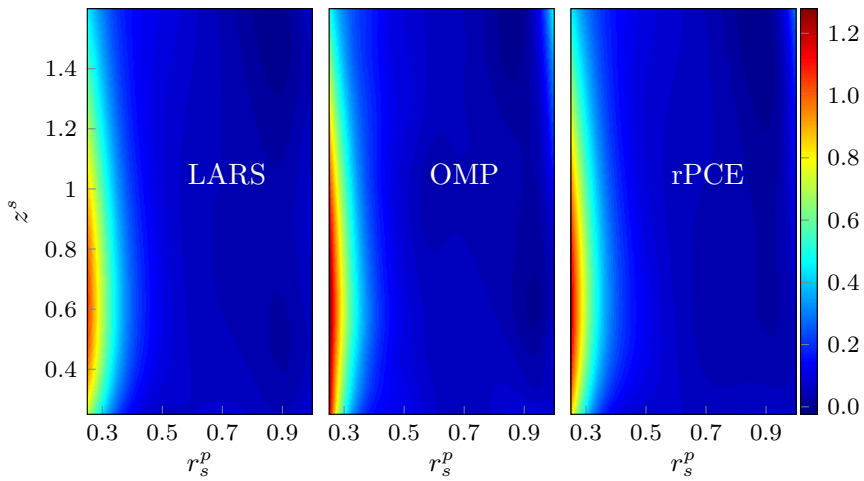


Fig. 22: SAR estimation - the prediction of whole-body SAR with 340 data points (100 replications).

human orientation θ_s^p affects the distribution of SAR in the human body. However, as the mean value of this distribution, the whole-body SAR is not much affected by θ_s^p. The sum of the total Sobol' indices in

Table 11 is larger than 1 and the excess values indicate that z^s impacts the response mainly through its interaction with r_s^p. Such an interaction can be viewed from the map of predicted SAR in Fig. 22, where ϕ_s^p, θ_s^p are fixed to zero and r_s^p, z^s are uniformly sampled over $[0.25, 1]$, $[0.25, 2]$, respectively. The amplitude of each pixel in the map is a mean of 100 predictions by the built PCE models during all replications. The three approaches provide similar results.

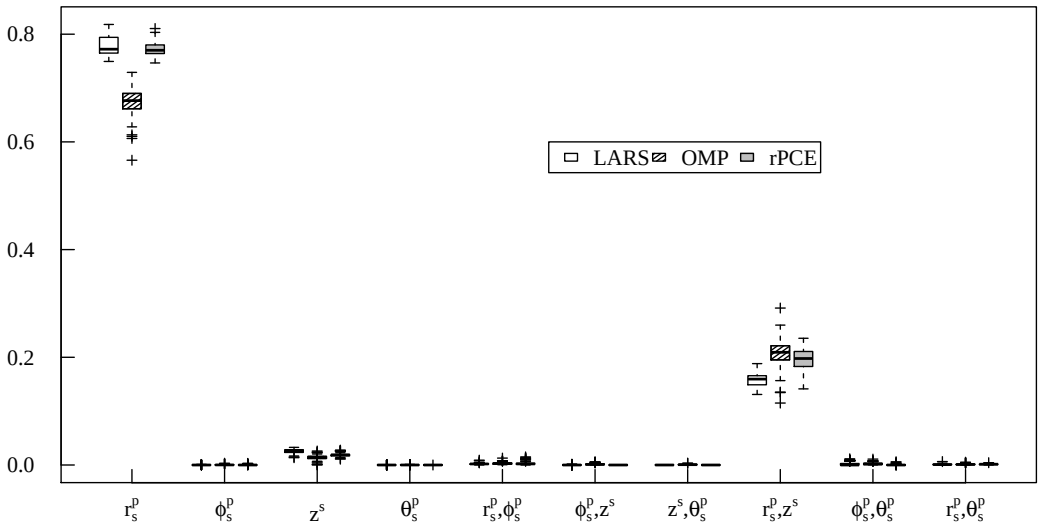


Fig. 23: SAR estimation - the estimation of first-order and second-order Sobol' indices with 340 data points (100 replications).

Considering the height of the human model is 1.36 m, tissues mainly locate at the heights between 0.4 m $\leq z^s \leq$ 1.0 m. One observes that the whole-body SAR is rather small when the source is farther from this influential region of the human model. r_s^p and z^s model the distance between the source and this influential region together and their interactions happen. The distribution of the estimated first-order and second-order Sobol' indices is proposed in Fig. 23, which presents that r_s^p and its interaction with z^s contribute the most to the uncertainty of the response.

6.6. Example with varied input dimension

To investigate the effects of the dimension of input on the modeling performance, the following test function [53] is used,

$$y = 3 + \frac{1}{M}\sum_{k=1}^{M} k(x_k^3 - 5x_k) + \ln\left(\frac{1}{3M}\sum_{k=1}^{M} k(x_k^2 + x_k^4)\right) + x_1x_2^2 - x_3x_5 + x_2x_4 + x_{M-4} + x_{M-4}x_M^2, \quad (21)$$

where M denotes the number of variables, which are independent and uniformly distributed in the range of $[1, 2]$. To increase the non-linearity, the range of x_{20} (when $M \geq 20$) is changed as $[1, 3]$.

The value of M changes from 11 to 41 with a step 5 and the size of experimental design N is fixed as 200 independently of M. For a statistical assessment of the modeling performance (accuracy and efficiency), 50 replications are performed and 10^3 data are used for the independent test at each replication. Remark that, due to randomness of the LHS method, different training and test datasets are used in replications.

From the methodology of rPCE, one knows that the computational cost is proportional to the number of resampled datasets. In Section 5.1, the suggested (not obliged) setting of k is a set of values, i.e., $k = \{3, 5, 10, 20, N\}$. As a result, the corresponding computational cost would be high when the ED size N is large. Here, a lighter setting of k, $k = \{3, 5, 10, 20\}$, is applied and improved modeling performances are still observed as presented by the following results. For the configuration of UQLab, which the running of LARS and OMP is based on, the maximum value of total degree p is set as 5.

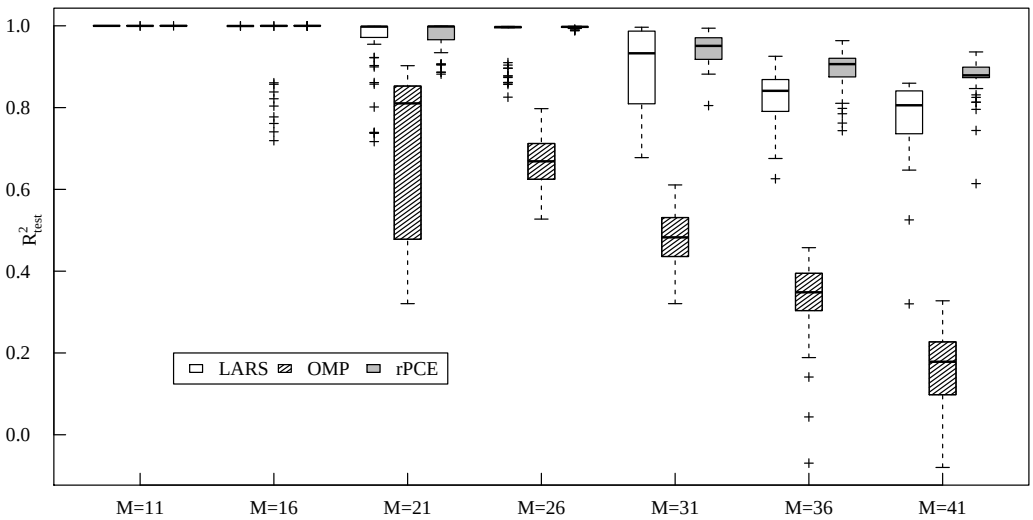


Fig. 24: Example with varied dimension - box plots of R^2_{test} with different values of M (50 replications).

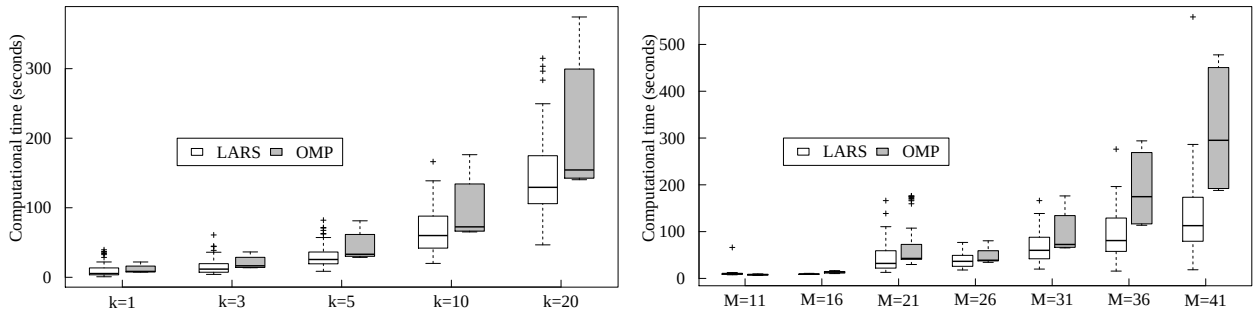
The effects of M on the modeling accuracy can be observed from the distribution of R^2_{test} in Fig. 24. When M equals 11 and 16, accurate models are constructed with the three approaches, although outliers appear with OMP and $M = 16$. As the dimension of input increases, the modeling accuracy becomes poorer in both variance and bias. While LARS performs much better than OMP when $M \geq 21$, substantial advantages of rPCE (with suggested configurations) are observed. When $M \in \{31, 36, 41\}$, the mean value of R^2_{test} with rPCE is larger than the value with LARS and OMP, and the variance of rPCE is also superior to the other two approaches.

		M=11	M=16	M=21	M=26	M=31	M=36	M=41
$k = 1$	LARS	0.9998	0.9995	0.9573	0.9679	0.8985	0.8260	0.7761
	OMP	0.9998	0.9634	0.6940	0.6679	0.4832	0.3308	0.1536
$k = 3$	LARS	0.9997	0.9996	0.9422	0.9249	0.8646	0.8322	0.8125
	OMP	0.9998	0.8072	0.7810	0.7737	0.6514	0.5358	0.3870
	L+O	0.9998	0.9996	0.8929	0.8771	0.7810	0.7262	0.6805
$k = 5$	LARS	0.9998	0.9995	0.9600	0.9726	0.8899	0.8574	0.8351
	OMP	0.9999	0.9552	0.8171	0.7915	0.6935	0.5894	0.4826
	L+O	0.9999	0.9996	0.9511	0.9651	0.8630	0.8110	0.7681
$k = 10$	LARS	0.9999	0.9995	0.9714	0.9945	0.9316	0.8724	0.8445
	OMP	0.9999	0.9963	0.8395	0.8194	0.7252	0.6239	0.5340
	L+O	0.9999	0.9998	0.9668	0.9937	0.9210	0.8557	0.8193
$k = 20$	LARS	0.9999	0.9995	0.9824	0.9971	0.9523	0.8947	0.8714
	OMP	0.9999	0.9999	0.8392	0.8195	0.7197	0.6149	0.5165
	L+O	0.9999	0.9999	0.9784	0.9971	0.9404	0.8692	0.8391
all k	LARS	0.9999	0.9996	0.9765	0.9965	0.9437	0.8904	0.8725
	OMP	0.9999	0.9987	0.8423	0.8248	0.7316	0.6191	0.5011
	L+O	0.9999	0.9998	0.9738	0.9961	0.9371	0.8790	0.8604

Table 12: Example with varied dimension - mean of R^2_{test} with varied values of k and M (50 replications), "L+O" denoting the combination of LARS and OMP.

From the mean value of R^2_{test} (w.r.t. 50 replications) in Table 12, one finds the effects of varied M on rPCE with different configurations. Remark that the cases with $k = 1$ correspond with the modeling results

based on LARS or OMP without the refinement by rPCE and the other cases are results of rPCE with different configurations. "all k" here means the combination of selection results with $k = \{3, 5, 10, 20\}$. As seen, when $M \geq 21$ and k is configured as 10, 20, or "all k", significant improvements are observed. When only OMP is applied, the mean value of R^2_{test} increases from 0.6940 to 0.8423 when $M = 21$. With LARS, the mean value increases from 0.7761 to 0.8725 when $M = 41$. Remark that these two peak values are reached with the configuration of "all k". Concerning on the source of candidate polynomials, since LARS performs much better than OMP, rPCE makes use of the selection results by LARS more often than OMP or their combination. Consequently, rPCE with the suggested setting (denoted by "L+O" in Table 12), is slightly (not significantly) inferior to the setting only based on LARS in modeling accuracy.



(a) $M = 31$ (b) $k = 10$

Fig. 25: Example with varied dimension - computational time versus varied values of k and M (50 replications).

With a laptop (dual cores, clock speeds 2.6 GHz, memory 16 GB), the computational time of rPCE is shown in Fig. 25. Effects of k and M are studied by fixing $M = 31$ and $k = 10$, respectively. Remark that this figure gives the total time cost for each configuration, i.e., for a specific value of k, the total time for constructing k PCE models is given. As observed, since about two thirds of the dataset is used for model construction, the computation time with $k = 3$ is only slightly longer than (rather than three times as) with $k = 1$. As the value of k increases, since the size of training datasets and the number of model constructions gets larger, the computational time increases fast. Moreover, in general the computational cost with OMP is higher than that with LARS. Similar phenomena are observed when the value of M increases. When $M = 41$, the maximum time cost on running both LARS and OMP is below 18 minutes which is usually much shorter than the time cost on getting new samples (e.g., ≈ 3 hours are required for getting a new sample in the example of Section 6.5).

7. Conclusions

A new polynomial selection approach, called resampled PCE, has been investigated herein to refine the ranking of importance of candidate polynomials in the context of sparse polynomial chaos expansions. Based on the selected polynomials by LARS and OMP, with the simulation of data variation by resampling, both the selection frequency and the increment on cross-validation error associated with each basis polynomial are arguments in the computation of a total score used in the ranking process. With the PCE model based on rPCE, sensitivity analysis is conveniently performed via the analytical computation of the Sobol' indices based on the expansion coefficients.

Two factors impact the performance of rPCE. First, the data resampling is conducted by dividing the whole set of data into k similar-sized subsets. The value of k needs to be optimized and set as a combination of good candidates $\{3, 5, 10, 20, N\}$. Second, the candidate polynomials can be generated by LARS, OMP or both. If LARS performs much better than OMP, the resulting selection of polynomials is based on LARS, and vice versa. Otherwise, both the polynomials selected by LARS and OMP would all be treated as candidates in rPCE.

The performance of rPCE, LARS and OMP is tested on two analytical functions, the maximum deflection of a truss structure and the estimation of the whole-body SAR (specific absorption rate). In terms of prediction and sensitivity analysis, OMP-based PCE modeling seems the worst among these three methods, especially when the size of ED is small. In contrast, the LARS-based approach generally generates a better model and the refinements by rPCE are obvious in terms of prediction variance and the number of outliers. In any case, rPCE performs as least as well as LARS for global sensitivity analysis.

Although the size of ED is fixed here, the samples can be automatically enriched to reach a certain accuracy in a specific estimation (e.g., moments) [61, 20, 62, 63]. Moreover, since the building processes with multiple resamples are independent in rPCE, the technique of parallel computations can be applied to ensure the building efficiency of rPCE at the same level with LARS or OMP.

In resampled PCE, a high computational cost may be suffered, especially with the suggested setting of k, i.e., $k = \{3, 5, 10, 20, N\}$ and a large N. As shown by the flow chart in Fig. 1, the possible high computational cost is due to the loop about k and l. However, one should realize that the loop about k and l are not necessarily performed in sequence and can proceed in parallel. All resampled datasets with different values of k can be first easily generated with LHS method. Since the surrogate modeling with respect to different sets of resamples is separable, techniques of parallel computation (e.g., computation with GPU and distributed computation) can be applied. Moreover, the suggested setting is not compulsory in the running of resampled PCE. Actually, from the results in Section 6, we can see the improved modeling performance has been observed with $k = 10$, 20, or $k = \{3, 5, 10, 20\}$, even better performances can be obtained if with the suggested setting. Thus, if the additional computational cost is considered high (especially relative to the cost of obtaining new data), 10 or 20 would be suggested for the setting of k.

In forthcoming investigations, more complex scenarios (e.g., electromagnetic dosimetry for human models in the telecommunications network [64, 65, 66]) are to be analyzed, where a high-order PCE model is often required and the classical approaches easily sink into the overfitting problem. Resampled PCE has the potential to avoid this problem. The refined selection of polynomials reduces the possibility of including redundant or irrelevant basis polynomials in the expansion, thus would have better chances to reach a model with a proper complexity. Here, rPCE combines two forward basis pursuit approaches and the improvements may be slight due to similar selected polynomials. The combination of different kinds of approaches (e.g. forward selection, backward elimination [67, 19] and sparsity-based approach [22]) is open to investigation.

Appendix A. Ranking basis polynomials based on LARS or OMP

1. Initialization: residual $\boldsymbol{R}_0 = \boldsymbol{y}$, active set $\mathbb{A}_0^a = \varnothing$, candidate set $\mathbb{A}_0^c = \mathbb{A}_{full}$.
2. For $j = 1, \ldots, P_{max} = \min\{N-1, \mathrm{card}(\mathbb{A}_{full})\}$,
   1) Find the basis most correlated with $\boldsymbol{R}_{j-1}$, $\boldsymbol{\alpha}_j = \arg\max_{\boldsymbol{\alpha} \in \mathbb{A}_{j-1}^c} \left|\boldsymbol{R}_{j-1}^T \boldsymbol{\psi}_{\boldsymbol{\alpha}}\right|$.
   2) Update $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$ and $\mathbb{A}_j^c = \mathbb{A}_{j-1}^c \setminus \boldsymbol{\alpha}_j$.
   3) With $\boldsymbol{\psi}_{\mathbb{A}_j^a}$, compute $\boldsymbol{\beta}_j$ as the OLS solution.
   4) Update residual $\boldsymbol{R}_j = \boldsymbol{y} - \boldsymbol{\psi}_{\mathbb{A}_j^a}^T \boldsymbol{\beta}_j$.

   End

Table A.1: Ranking basis polynomials based on orthogonal matching pursuit (OMP).

The PCE model based on orthogonal matching pursuit (OMP) is iteratively built and the iterative procedure is summarized in Table A.1. At each iteration, the influence of each polynomial term $\psi_{\boldsymbol{\alpha}}$ is measured by its correlation with the data residual $\boldsymbol{R}$ (the initial value being $\boldsymbol{y}$). The $\boldsymbol{\alpha}$ corresponding with the most correlated basis term $\psi_{\boldsymbol{\alpha}}$ becomes a member of the active set $\mathbb{A}^a$. Then, computing the basis function $\boldsymbol{\psi}_{\mathbb{A}^a}$ supported by the active set, the associated coefficients are obtained by minimizing the least-square error and $\boldsymbol{R}$ is updated as the new residual. The most influential polynomials are sequentially selected by repeating the procedure above.

1. Initialization: residual $\boldsymbol{R}_0 = \boldsymbol{y}$, active set $\mathbb{A}_0^a = \varnothing$, candidate set $\mathbb{A}_0^c = \mathbb{A}_{full}$;
2. For $j = 1, \ldots, P_{max} = \min\{N-1, \mathrm{card}(\mathbb{A}_{full})\}$,
   If j equals 1, define $\boldsymbol{u}_1 = \boldsymbol{\psi}_{\boldsymbol{\alpha}_1}$, $\boldsymbol{\alpha}_1 = \arg\max_{\boldsymbol{\alpha}\in\mathbb{A}_0^c} \left|\boldsymbol{R}_0^T \boldsymbol{\psi}_{\boldsymbol{\alpha}}\right|$, and update $\mathbb{A}_1^a = \{\boldsymbol{\alpha}_1\}$, $\mathbb{A}_1^c = \mathbb{A}_0^c \setminus \boldsymbol{\alpha}_1$.
   Otherwise,
   1) update $\boldsymbol{R}_{j-1} = \boldsymbol{R}_{j-2} + \gamma_{j-1}\boldsymbol{u}_{j-1}$, γ_{j-1} the smallest step length when $\boldsymbol{R}_{j-1}$ has the same correlation with a basis polynomial (denoted by $\boldsymbol{\psi}_{\boldsymbol{\alpha}_j}$, $\boldsymbol{\alpha}_j \in \mathbb{A}_{j-1}^c$) as those with all polynomials in $\boldsymbol{\psi}_{\mathbb{A}_{j-1}^a}$.
   2) update $\mathbb{A}_j^a = \mathbb{A}_{j-1}^a \cup \boldsymbol{\alpha}_j$ and $\mathbb{A}_j^c = \mathbb{A}_{j-1}^c \setminus \boldsymbol{\alpha}_j$.
   3) compute the equiangular vector of all polynomials in $\boldsymbol{\psi}_{\mathbb{A}_j^a}$ as $\boldsymbol{u}_j$.
   End

Table A.2: Ranking basis polynomials based on least angle regression (LARS).

Least angle regression (LARS) is a less greedy version of traditional forward selection methods. It is known that different flavors of LARS yield efficient solutions of LASSO [37] (which constrains both the data discrepancy by ordinary least square and the sparsity of regression coefficients by ℓ_1-norm) and forward stagewise linear regression [68] (another promising model-selection method), respectively.

The iterative algorithm of sparse PCE modeling based on LARS (originally proposed in [20]) is given in Table A.2, where details on how to compute step length γ_{j-1} and equiangular vector $\boldsymbol{u}_j$ can be found in [24]. As seen from this short summary, the building process is similar with the one based on OMP, except that from the second iteration since the residual $\boldsymbol{R}$ evolves along the equiangular directions of basis functions other than along basis functions themselves.

Acknowledgments

The first author has been supported in part by the Emergence programme of the Science and Technologies of Information and Communication (STIC) department, University Paris-Saclay.

References

[1] A. Taflove, S. C. Hagness, Computational Electrodynamics: the Finite-Difference Time-Domain Method, Artech House, 2005.
[2] K.-J. Bathe, E. L. Wilson, Numerical Methods in Finite Element Analysis, Prentice-Hall, 1976.
[3] R. R. Barton, Tutorial: Input uncertainty in output analysis, in: Proc. Winter Simulation Conference, WSC2012, Berlin, Germany, 2012.
[4] A. Kolmogorov, Foundations of the Theory of Probability: Second English Edition, Dover Publications, 1956.
[5] R. L. Iman, J. C. Helton, An investigation of uncertainty and sensitivity analysis techniques for computer models, Risk Anal. 8 (1) (1988) 71–90.
[6] J. P. Kleijnen, Kriging metamodeling in simulation: A review, Eur. J. Oper. Res. 192 (3) (2009) 707–716.
[7] D. J. MacKay, Bayesian methods for adaptive models, Ph.D. thesis, California Institute of Technology, CA, USA (1992).
[8] B. Sudret, Uncertainty propagation and sensitivity analysis in mechanical models–contributions to structural reliability and stochastic spectral methods, Habilitation à diriger des recherches, Université Blaise Pascal, Clermont-Ferrand, France.
[9] K. Sepahvand, S. Marburg, H.-J. Hardtke, Uncertainty quantification in stochastic systems using polynomial chaos expansion, Int J. Appl. Mech. 2 (02) (2010) 305–353.
[10] P. Kersaudy, S. Mostarshedi, B. Sudret, O. Picon, J. Wiart, Stochastic analysis of scattered field by building facades using polynomial chaos, IEEE Trans. Antennas Propag. 62 (12) (2014) 6382–6393.
[11] C. Soize, R. Ghanem, Physical systems with random uncertainties: chaos representations with arbitrary probability measure, SIAM J. Sci. Comput. 26 (2) (2004) 395–410.
[12] D. Xiu, G. E. Karniadakis, The Wiener–Askey polynomial chaos for stochastic differential equations, SIAM J. Sci. Comput. 24 (2) (2002) 619–644.
[13] R. G. Ghanem, P. D. Spanos, Stochastic Finite Elements: A Spectral Approach, Dover Publications, 2003.
[14] B. Sudret, M. Berveiller, M. Lemaire, A stochastic finite element method in linear mechanics, CR Mécanique 332 (7) (2004) 531–537.
[15] M. Berveiller, B. Sudret, M. Lemaire, Stochastic finite element: a non intrusive approach by regression, Eur. J. Comput. Mech. 15 (1-3) (2006) 81–92.

[16] O. P. Le Maître, M. T. Reagan, H. N. Najm, R. G. Ghanem, O. M. Knio, A stochastic projection method for fluid flow: II. Random process, J. Comput. Phys. 181 (1) (2002) 9–44.
[17] L. Gilli, D. Lathouwers, J. Kloosterman, T. Van der Hagen, A. Koning, D. Rochman, Uncertainty quantification for criticality problems using non-intrusive and adaptive polynomial chaos techniques, Ann. Nucl. Energy 56 (2013) 71–80.
[18] G. Blatman, Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis, Ph.D. thesis, Université Blaise Pascal, Clermont-Ferrand, France (2009).
[19] G. Blatman, B. Sudret, An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis, Probab. Eng. Mech. 25 (2) (2010) 183–197.
[20] G. Blatman, B. Sudret, Adaptive sparse polynomial chaos expansion based on least angle regression, J. Comput. Phys. 230 (6) (2011) 2345–2367.
[21] A. Doostan, H. Owhadi, A non-adapted sparse approximation of PDEs with stochastic inputs, J. Comput. Phys. 230 (8) (2011) 3015–3034.
[22] J. D. Jakeman, M. S. Eldred, K. Sargsyan, Enhancing ℓ_1-minimization estimates of polynomial chaos expansions using basis selection, J. Comput. Phys. 289 (2015) 18–34.
[23] J. A. Tropp, A. C. Gilbert, Signal recovery from random measurements via orthogonal matching pursuit, IEEE Trans. Inf. Theory 53 (12) (2007) 4655–4666.
[24] B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression, Ann. Stat. 32 (2) (2004) 407–499.
[25] L. Breiman, Bagging predictors, Mach. Learn. 24 (2) (1996) 123–140.
[26] B. Efron, R. J. Tibshirani, An Introduction to The Bootstrap, CRC Press, 1994.
[27] W. Sauerbrei, M. Schumacher, A bootstrap resampling procedure for model building: application to the Cox regression model, Stat. Med. 11 (16) (1992) 2093–2109.
[28] P. Royston, W. Sauerbrei, et al., Bootstrap assessment of the stability of multivariable models, Stata J. 9 (4) (2009) 547.
[29] E. Anna, Variable selection for the Cox proportional hazards model: A simulation study comparing the stepwise, lasso and bootstrap approach, Master's thesis, Umeå University (2017).
[30] R. De Bin, S. Janitza, W. Sauerbrei, A.-L. Boulesteix, Subsampling versus bootstrapping in resampling-based model selection for multivariable regression, Biometrics 72 (1) (2016) 272–280.
[31] S. Geisser, The predictive sample reuse method with applications, J. Am. Stat. Assoc. 70 (350) (1975) 320–328.
[32] M. Walschaerts, E. Leconte, P. Besse, Stable variable selection for right censored data: comparison of methods, arXiv preprint arXiv:1203.4928.
[33] R. Lebrun, A. Dutfoy, A generalization of the Nataf transformation to distributions with elliptical copula, Probab. Eng. Mech. 24 (2) (2009) 172–178.
[34] M. Lemaire, Structural Reliability, John Wiley & Sons, 2013.
[35] W. Gautschi, Orthogonal Polynomials: Computation and Approximation, Oxford University Press on Demand, 2004.
[36] C. R. Rao, C. R. Rao, M. Statistiker, C. R. Rao, C. R. Rao, Linear Statistical Inference and Its Applications, Wiley, 1973.
[37] R. Tibshirani, Regression shrinkage and selection via the lasso, J. Royal Stat. Soc. Series B (1996) 267–288.
[38] R. B. Nelsen, An Introduction to Copulas, Springer Science & Business Media, 2007.
[39] R. Kohavi, et al., A study of cross-validation and bootstrap for accuracy estimation and model selection, in: Proc. 14th International Joint Conference on Artificial Intelligence, IJCAI1995, Vol. 14, Montreal, Canada, 1995, pp. 1137–1145.
[40] K. Konakli, B. Sudret, Polynomial meta-models with canonical low-rank approximations: numerical insights and comparison to sparse polynomial chaos expansions, J. Comput. Phys. 321 (2016) 1144–1169.
[41] B. Efron, The Jackknife, The Bootstrap, and Other Resampling Plans, SIAM, 1982.
[42] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Springer Series in Statistics, 2001.
[43] S. A. Smolyak, Quadrature and interpolation formulas for tensor products of certain classes of functions, Dokl. Akad. Nauk SSSR 148 (5) (1963) 1042–1045.
[44] S. Marelli, B. Sudret, UQLab user manual–polynomial chaos expansions, Report UQLab-V1.1-104, Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich.
[45] E. Anderssen, K. Dyrstad, F. Westad, H. Martens, Reducing over-optimism in variable selection by cross-model validation, Chemom. Intell. Lab. Syst. 84 (1-2) (2006) 69–74.
[46] T. Fushiki, Estimation of prediction error by using k-fold cross-validation, Statistics and Computing 21 (2) (2011) 137–146.
[47] D. Baumann, K. Baumann, Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation, J. Cheminf. 6 (1) (2014) 47.
[48] L. Gidskehaug, E. Anderssen, B. K. Alsberg, Cross model validation and optimisation of bilinear regression models, Chemom. Intell. Lab. Syst. 93 (1) (2008) 1–10.
[49] I. M. Sobol, Sensitivity estimates for nonlinear mathematical models, Mathem. Mod. Comput. Exp. 1 (4) (1993) 407–414.
[50] T. Homma, A. Saltelli, Importance measures in global sensitivity analysis of nonlinear models, Reliab. Eng. Syst. Safe. 52 (1) (1996) 1–17.
[51] S. Kucherenko, S. Tarantola, P. Annoni, Estimation of global sensitivity indices for models with dependent variables, Comput. Phys. Commun. 183 (4) (2012) 937–946.
[52] B. Sudret, Global sensitivity analysis using polynomial chaos expansions, Reliab. Eng. Syst. Safe. 93 (7) (2008) 964–979.
[53] S. Marelli, B. Sudret, UQLab: A framework for uncertainty quantification in Matlab, in: Proc. 2nd International Conference on Vulnerability, Risk Analysis and Management, ICVRAM2014, Liverpool, United Kingdom, 2014, pp. 2554–2563.
[54] Chair of Risk, Safety and Uncertainty Quantification of ETH Zurich, [online] Available at: https://www.uqlab.com [Accessed Aug. 18, 2019].
[55] M. D. McKay, R. J. Beckman, W. J. Conover, Comparison of three methods for selecting values of input variables in the analysis of output from a computer code, Technometrics 21 (2) (1979) 239–245.

[56] S. Xiong, P. Z. Qian, C. J. Wu, Sequential design and analysis of high-accuracy and low-accuracy computer codes, Technometrics 55 (1) (2013) 37–46.
[57] E. Van Deventer, E. Van Rongen, R. Saunders, WHO research agenda for radiofrequency fields, Bioelectromagnetics 32 (5) (2011) 417–421.
[58] M.-C. Gosselin, E. Neufeld, H. Moser, E. Huber, S. Farcito, L. Gerber, M. Jedensjoe, I. Hilber, F. Di Gennaro, B. Lloyd, et al., Development of a new generation of high-resolution anatomical models for medical device evaluation: the virtual population 3.0, Phys. Med. Biol. 59 (18) (2014) 5287.
[59] J. E. Hansen, Spherical Near-Field Antenna Measurements, Peter Peregrinus Ltd, 1988.
[60] I. Liorni, M. Parazzini, N. Varsier, A. Hadjem, P. Ravazzani, J. Wiart, Exposure assessment of one-year-old child to 3G tablet in uplink mode and to 3G femtocell in downlink mode using polynomial chaos decomposition, Phys. Med. Biol. 61 (8) (2016) 3237.
[61] V. Picheny, D. Ginsbourger, O. Roustant, R. T. Haftka, N.-H. Kim, Adaptive designs of experiments for accurate approximation of a target region, J. Mech. Des. 132 (7) (2010) 071008.
[62] S. Dubreuil, M. Berveiller, F. Petitjean, M. Salaün, Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion, Reliab. Eng. Sys. Safety 121 (2014) 263–275.
[63] N. Fajraoui, S. Marelli, B. Sudret, Sequential design of experiment for sparse polynomial chaos expansions, SIAM/ASA J. Unc. Quant. 5 (1) (2017) 1061–1085.
[64] I. Liorni, M. Parazzini, S. Fiocchi, P. Ravazzani, Study of the influence of the orientation of a 50-Hz magnetic field on fetal exposure using polynomial chaos decomposition, Int. J. Environ. Res. Public. Health 12 (6) (2015) 5934–5953.
[65] P. Kersaudy, B. Sudret, N. Varsier, O. Picon, J. Wiart, A new surrogate modeling technique combining Kriging and polynomial chaos expansions–application to uncertainty analysis in computational dosimetry, J. Comput. Phys. 286 (2015) 103–117.
[66] Y. Huang, J. Wiart, Simplified assessment method for population RF exposure induced by a 4G network, IEEE J. Electromagn. RF Microw. Med. Biol. 1 (1) (2017) 34–40.
[67] N. Mantel, Why stepdown procedures in variable selection, Technometrics 12 (3) (1970) 621–625.
[68] S. Weisberg, Applied Linear Regression, John Wiley & Sons, 2005.