ETH zürich

SURROGATE MODELING BASED ON RESAMPLED POLYNOMIAL CHAOS EXPANSIONS

Z. Liu, D. Lesselier, B. Sudret, J. Wiart

CHAIR OF RISK, SAFETY AND UNCERTAINTY QUANTIFICATION
STEFANO-FRANSCINI-PLATZ 5
CH-8093 ZÜRICH

Risk, Safety & Uncertainty Quantification

Data Sheet

Journal:	
Report Ref.:	RSUQ-2018-006
Arxiv Ref.:	http://arxiv.org/abs/1810.09116 - [stat.CO] [stat.ME]
DOI:	10.13140/RG.2.2.14747.31522
Date submitted:	22 October 2018
Date accepted:	-

Surrogate modeling based on resampled polynomial chaos expansions

Zicheng Liu[1,2], Dominique Lesselier[2], Bruno Sudret[3], and Joe Wiart[1]

[1]Chaire C2M, LTCI, Télécom ParisTech, Université Paris-Saclay, Paris 75013, France.

[2]Laboratoire des Signaux et Systèmes, UMR8506 (CNRS-CentraleSupelec-Université Paris-Sud), Université Paris-Saclay, Gif-sur-Yvette cedex 91192, France

[3]ETH Zürich, Chair of Risk, Safety and Uncertainty Quantification, Stefano-Franscini-Platz 5, Zürich 8093, Switzerland

October 22, 2018

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) Taflove and Hagness (2005) and finite element method (FEM) Bathe and Wilson (1976), 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 Barton (2012), 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) Kolmogorov (1956), the propagation of such random inputs through the system yields random outputs and the investigation of such uncertainty propagation is known as uncertainty quantification (UQ) Iman and Helton (1988). 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 to the cases focused onto here). Surrogate 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) Kleijnen (2009), neural networks MacKay (1992), etc., surrogate modeling based on polynomial chaos expansion (PCE) Sudret (2007); Sepahvand et al. (2010); Kersaudy et al. (2014) 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 Ghanem and Spanos (2003) 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 Sudret et al. (2004). 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 Berveiller et al. (2006), projection method Le Maître et al. (2002); Gilli et al. (2013) 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 the computational purpose. How to make 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 Blatman (2009). 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 Blatman and Sudret (2010, 2011); Doostan and Owhadi (2011); Jakeman et al. (2015) 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) Tropp and Gilbert (2007) and least angle regression (LARS) Efron et al. (2004), have been utilized to rank the polynomials.

This contribution is aimed at reducing the variance of the sparse PCE model while keeping or improving the level of bias, through refining the truncation of a PCE. Based on the available sparse PCE modeling methods, the refined modeling approach takes advantage of the knowledge that the true model is fixed despite the variation of training data and the associated influential polynomials should be frequently selected during replications. Simulating the variation of ED via the resampling technique (e.g., bootstrap Efron and Tibshirani (1994), leave-many-out Geisser (1975)), a PCE truncation can be generated for each resampled data set by LARS or OMP. Merging involved polynomials of all PCE truncations into one set, the influence of each candidate base is mainly decided by the frequency of appearance. Such a building process of the PCE-based surrogate model is named as resampled PCE (rPCE). This name refers to the fact that the variation of ED through resampling is fully exploited in the analysis.

Let us emphasize that the technique to stabilize the PCE modeling method in rPCE is different from bagging (a.k.a. bootstrap aggregating) Breiman (1996), which is a popular approach, especially for decision tree methods, to reduce the modeling variance by training multiple regression models based on resampling samples and taking the final prediction as the mean of all predictions. Since different sets of resampling samples are based on the same experimental design, the optimal polynomials in the construction of different PCEs might have a correlation, which is however not considered by bagging. In contrast, such correlation is considered as an additional knowledge in rPCE to refine the selection of expansion bases. Once the basis polynomials have been selected, the expansion coefficients are computed with the whole set of original data and the prediction is only made by the refined PCE model.

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 relative importance of polynomials is evaluated through the selection 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.

How to build $\psi_{\boldsymbol{\alpha}}(\boldsymbol{X})$ is recalled now Soize and Ghanem (2004); Sudret (2007). Assuming that the input random variables are independent, the joint PDF is a multiplication of the marginal ones,

$$p_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{M} p_{X_i}(x_i), \tag{3}$$

where p_{X_i} is the marginal PDF of random variable X_i. Such a property gives the composition of orthogonal multivariate polynomials as a tensor product of univariate polynomials π_{α_i}, i.e.,

$$\psi_{\boldsymbol{\alpha}}(\boldsymbol{X}) = \pi_{\alpha_1}(X_1) \times \ldots \times \pi_{\alpha_M}(X_M), \tag{4}$$

while π_{α_i} should satisfy the following orthogonality

$$\langle \pi_{\alpha_j}, \pi_{\alpha_k} \rangle = E\left[\pi_{\alpha_j}(X_i)\pi_{\alpha_k}(X_i)\right] = \delta_{j,k}, \tag{5}$$

where $\delta_{j,k}$ equals 1 when $j = k$ and 0 otherwise. The following derivation

$$\langle \psi_{\boldsymbol{\alpha}}, \psi_{\boldsymbol{\gamma}} \rangle = E\left[\psi_{\boldsymbol{\alpha}}(\boldsymbol{X})\psi_{\boldsymbol{\gamma}}(\boldsymbol{X})\right] = \delta_{\boldsymbol{\alpha},\boldsymbol{\gamma}} \tag{6}$$

can be made, which indicates the orthogonality of the constructed multivariate polynomials $\psi_{\boldsymbol{\alpha}}$.

Earlier on, the normal input variables have been considered with the corresponding Hermite polynomial basis Wiener (1938). Then, the extension to other types of random variables has been proposed as the generalized PCE (gPCE) Xiu and Karniadakis (2002); Soize and Ghanem (2004), where a specific family of polynomials is matched to some common PDFs (uniform, exponential, etc.). 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 Lebrun and Dutfoy (2009); Lemaire (2013) or specific orthogonal polynomials are computed numerically via the Stieltjes procedure (Gautschi, 2004).

The PCE coefficients $\beta_{\boldsymbol{\alpha}}$ is 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

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} ||\boldsymbol{y} - \boldsymbol{\psi}\boldsymbol{\beta}||_2^2, \tag{7}$$

which yields the ordinary least square (OLS) Rao et al. (1973) solution as

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{\psi}^T\boldsymbol{\psi}\right)^{-1}\boldsymbol{\psi}^T\boldsymbol{y}, \tag{8}$$

the superscript "T" denoting the transpose operation.

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 Nelsen (2007) and use Rosenblatt transformation Lebrun and Dutfoy (2009) to cast the problem as a function of auxiliary independent variables.

2.3 Estimation of prediction performance

Once the surrogate model is obtained, the prediction performance can be estimated via the generalization error,

$$\epsilon = E\left[\left(\mathcal{M}(\boldsymbol{X}) - \widehat{\mathcal{M}}(\boldsymbol{X})\right)^2\right], \tag{9}$$

where $\widehat{\mathcal{M}}$ denotes the surrogate model. The estimation of ϵ is often obtained by Monte Carlo simulations when a large set of validation data, which are independent from the experimental

design, are available. Denoting the input vector and response of the n-th validation data as $\boldsymbol{x}_{\text{val}}^{(n)}$ and $y_{\text{val}}^{(n)}$, respectively, the mean square error of data discrepancy

$$\epsilon_{\text{val}} = \frac{1}{N_{\text{val}}} \sum_{n=1}^{N_{\text{val}}} \left(\mathcal{M}(\boldsymbol{x}_{\text{val}}^{(n)}) - \widehat{\mathcal{M}}(\boldsymbol{x}_{\text{val}}^{(n)}) \right)^2, \tag{10}$$

converges to ϵ if the data size N_{val} tends to infinity. For an easier interpretation of ϵ_{val}, the associated coefficient of determination R_{val}^2 is computed by

$$R_{\text{val}}^2 = 1 - \frac{\epsilon_{\text{val}}}{\text{Var}(\boldsymbol{y}_{\text{val}})}, \tag{11}$$

where $\text{Var}(\boldsymbol{y}_{\text{val}}) = \sum_{n=1}^{N_{\text{val}}} (y_{\text{val}}^{(n)} - \bar{y}_{\text{val}})^2/(N_{\text{val}} - 1)$ and $\bar{y}_{\text{val}} = \sum_{n=1}^{N_{\text{val}}} y_{\text{val}}^{(n)}/N_{\text{val}}$. Therefore, the closer R_{val}^2 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 often intractable to have a large validation set. Then, the same data as for training are often reused for validation. However, the underestimation of the generalization error is well-known in the case of overfitting Blatman (2009). Cross validation was thus proposed and is commonly advocated Kohavi et al. (1995); Konakli and Sudret (2016).

Randomly dividing the data of the ED into a training set and a validation set, cross validation means the prediction of the validation set by the surrogate model built from the training set. Here, rather than random division, k-fold cross validation is utilized so that the information in the whole set of data is fully considered in both training and validation. For this purpose, one divides the ED into k approximately equal-sized subsets. Leaving the l-th subset out for validation, a surrogate model is trained with the remaining data. Varying l from 1 to k, the cross-validation error is computed based on all validation results. Specifically, letting $k = N$, one obtains the so-called leave-one-out cross validation, whose error reads:

$$\epsilon_{LOO} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathcal{M}(\boldsymbol{x}^{(n)}) - \widehat{\mathcal{M}}^{-(n)}(\boldsymbol{x}^{(n)}) \right)^2, \tag{12}$$

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 Efron (1982) and it can be computed fast in single training process Blatman (2009) by

$$\epsilon_{LOO} = \frac{1}{N} \sum_{n=1}^{N} \left(\frac{\mathcal{M}(\boldsymbol{x}^{(n)}) - \widehat{\mathcal{M}}(\boldsymbol{x}^{(n)})}{1 - h_n} \right)^2, \tag{13}$$

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

Table 1: Sparse PCE model based on orthogonal matching pursuit.

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 $j = 1, \ldots, P_{max} = \min\{N-1, \text{card}(\mathbb{A}_{full})\}$,
   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
3. Based on $\boldsymbol{\psi}_{\mathbb{A}_j^a}$ and $\boldsymbol{\beta}_j$, compute ϵ_{LOO}^j, $j = 1, \ldots, P_{max}$.
4. $P = \arg\min_j \epsilon_{LOO}^j$ and the PCE model corresponding to $\boldsymbol{\psi}_{\mathbb{A}_P^a}$ is selected.

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

$$\widehat{\mathcal{M}}(\boldsymbol{X}) = \sum_{\boldsymbol{\alpha} \in \mathbb{A}} \beta_{\boldsymbol{\alpha}} \psi_{\boldsymbol{\alpha}}(\boldsymbol{X}). \tag{14}$$

The usual way to decide about $\mathbb{A}$ leads to the so-called full PCE model, which suffers from the curse-of-dimensionality Friedman et al. (2001), meaning that the cardinality of $\mathbb{A}$ sharply increases with the number of input parameters, as explained below. Recently, least angle regression (LARS) Blatman and Sudret (2011); Marelli and Sudret (2018) and orthogonal matching pursuit (OMP) Tropp and Gilbert (2007); Marelli and Sudret (2018) 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., $\mathbb{A}_{full} = \{\boldsymbol{\alpha} \in \mathbb{N}^M, \sum_{i=1}^M \alpha_i \leq p\}$, 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 $\mathbb{A}_{full}$, 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. (8), 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. This problem is addressed by downsizing $\mathbb{A}$ through the use of greedy algorithms, so that only the most influential polynomials are included in the truncated PCE.

Table 2: Sparse PCE model based on least angle regression.

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 $j = 1, \ldots, P_{max} = \min\{N-1, \text{card}(\mathbb{A}_{full})\}$,
   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
3. Based on $\boldsymbol{\psi}_{\mathbb{A}_j^a}$, compute $\boldsymbol{\beta}_j$ as the OLS solution and associated ϵ_{LOO}^j, $j = 1, \ldots, P_{max}$.
4. $P = \arg\min_j \epsilon_{LOO}^j$ and the PCE model corresponding to $\boldsymbol{\psi}_{\mathbb{A}_P^a}$ is selected.

3.2 Sparse PCE model based on orthogonal matching pursuit

The PCE model based on orthogonal matching pursuit (OMP) is iteratively built and the iterative procedure is summarized in Table 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 to 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.

The number of selected polynomials, P, needs an optimization, considering that the flexibility of the surrogate model increases with P and too flexible a model might lead to the overfitting problem. Leave-one-out cross-validation is utilized in Table 1 to assess the quality of the obtained surrogate models. Setting the maximum number of P as $\min\{N-1, \text{card}(\mathbb{A}_{full})\}$ (otherwise the least-square problem becomes ill-posed), the optimal P corresponds to the PCE model with the minimal ϵ_{LOO}.

3.3 Sparse PCE model based on least angle regression

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 Tibshirani (1996) (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 Weisberg (2005) (another promising model-selection method), respectively.

The iterative algorithm of sparse PCE modeling based on LARS (originally proposed in Blatman and Sudret (2011)) is given in Table 2, where details on how to compute step length γ_{j-1} and equiangular vector $\boldsymbol{u}_j$ can be found in Efron et al. (2004). 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.

4 Surrogate modeling based on resampled PCE

Resampled PCE (rPCE) is proposed to refine standard PCE truncation schemes by taking advantage of the heuristic observation that the influential polynomials associated with the true model are frequently selected by LARS and OMP during replications with resampled training data. Simulating the data resampling via k-fold division, the rank of polynomials is mainly decided by the selection frequency. Efforts to combine results by LARS and OMP to further improve the performance of rPCE are now presented.

4.1 Resampled PCE based on LARS or OMP

A fixed set of data is only available in practice for surrogate modeling and the statistical variation of training data is simulated by 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 have 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 single set $\mathbb{A}_m = \{\mathbb{A}_1, \ldots, \mathbb{A}_k\}$. Then the selection frequency of $\boldsymbol{\alpha}$ in the k building processes is equal to the number of its duplicates in $\mathbb{A}_m$. Then we can save the selection frequency as the vector $\boldsymbol{s}_f$, the elements of which are integers in the interval $[1, k]$, and denote $\mathbb{A}$ as the copy of $\mathbb{A}_m$ but without duplications. The elements in $\mathbb{A}$ are sorted according to the descending $\boldsymbol{s}_f$, named in the sequel the *frequency score*, and compose the new set $\mathbb{A}^*$. Then, following the same procedures as the step 3 and 4 in Table 2 for LARS or Table 1 for OMP, except for replacing the active set by $\mathbb{A}^*$, the refined truncation of PCE is obtained, while the optimal number of basis polynomials is still

determined by cross-validation errors.

However, during the running of rPCE, different muti-indicies $\boldsymbol{\alpha}$ 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 $\Delta\epsilon^j_{LOO} = \epsilon^j_{LOO} - \epsilon^{j-1}_{LOO}$ for $j \geq 1$, where ϵ^0_{LOO} is set as 0. Denoting $\Delta\epsilon^{l,j}_{LOO}$ as the $\Delta\epsilon^j_{LOO}$ at the l-th PCE construction and $\mathbb{E}_m$ the set composed of all $\Delta\epsilon^{l,j}_{LOO}$, the elements in $\mathbb{A}_m$ have a unique mapping to those in $\mathbb{E}_m$. Based on this relation, the so-called *error score* $\mathbf{s}_e$ can be computed as the mean of all terms $\Delta\epsilon^{l,j}_{LOO}$ mapping to the same element of $\mathbb{A}$, i.e.,

$$s^i_e = \frac{1}{s^i_f \Delta\epsilon^{\max}_{LOO}} \sum_{\{l,j|\boldsymbol{\alpha}^{l,j}=\boldsymbol{\alpha}^i\}} \Delta\epsilon^{l,j}_{LOO},\ i = 1, \ldots, \mathrm{card}\{\mathbb{A}\} \tag{15}$$

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

$$\mathbf{s} = \mathbf{s}_f + \mathbf{s}_e, \tag{16}$$

is mainly affected by $\mathbf{s}_f$ in rPCE.

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, $\mathbb{A}_m$ and $\mathbb{E}_m$ can be obtained by LARS and OMP separately, denoted by $\mathbb{A}^{LARS}_m$, $\mathbb{E}^{LARS}_m$ and $\mathbb{A}^{OMP}_m$, $\mathbb{E}^{OMP}_m$, respectively. Then, merging results by LARS and OMP into a single set, $\mathbb{A}^{LARS+OMP}_m = \{\mathbb{A}^{LARS}_m, \mathbb{A}^{OMP}_m\}$ and $\mathbb{E}^{LARS+OMP}_m = \{\mathbb{E}^{LARS}_m, \mathbb{E}^{OMP}_m\}$, from which $\mathbb{A}^{LARS+OMP}$ and the associated total score $\boldsymbol{s}^{LARS+OMP}$ can be computed. Then, the basis polynomials of the surrogate model are selected based on $\mathbb{A}^{LARS+OMP}$, $\boldsymbol{s}^{LARS+OMP}$ and the coefficients are the OLS solution.

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 $\mathbb{A}_m$ obtained for different values of k suggested in literature, including $k = 3, 5, 10, 20, N$. Denoting $\mathbb{A}_m^q$ the set $\mathbb{A}_m$ when $k = q$, rPCE will run based on the set $\mathbb{A}_m^C = [\mathbb{A}_m^3, \mathbb{A}_m^5, \mathbb{A}_m^{10}, \mathbb{A}_m^{20}, \mathbb{A}_m^N]$. This strategy considers both the data variation and the bias of generated candidates polynomials and is revealed robust in the various application examples.

Denote $\mathbb{A}^C$ as the copy of $\mathbb{A}_m^C$ but without element duplication and $\boldsymbol{s}_f^k$ as the frequency score w.r.t. $\mathbb{A}_m^k$. For each $\boldsymbol{\alpha}$ in $\mathbb{A}^C$, the frequency score can be computed by

$$s_f^i = \sum_{k=\{3,5,10,20,N\}} s_f^{k,i},\ i = 1, \ldots, \mathrm{card}(\mathbb{A}^C), \tag{17}$$

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 $\mathbb{A}^C$ is not in $\mathbb{A}_m^k$. However, since $s_f^{k,i}$ is upper bounded by k, the polynomials selected with small values of k (e.g., elements in $\mathbb{A}_m^3$) will have small frequency scores and be less likely to have high ranks in rPCE.

To solve this problem, instead of (17), the frequency score of elements in $\mathbb{A}_m^C$ is computed as a summation of weighted $s_f^{k,i}$:

$$s_f^i = \sum_{k=\{3,5,10,20,N\}} s_f^{k,i} \frac{\mathrm{lcm}(3, 20, N)}{k},\ i = 1, \ldots, \mathrm{card}(\mathbb{A}^C), \tag{18}$$

where $\mathrm{lcm}(3, 20, N)$ 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 (18). 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 $N = 15$.

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_{val} 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_{val}. Varying the training data, a sequence of surrogate models is built and the associated R^2_{val} values are computed. Representing $\mathbb{R}^2_{LARS}$ and $\mathbb{R}^2_{OMP}$ as the sets of R^2_{val} 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 $Q_3^{LARS} > Q_1^{OMP}$, 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_{val} 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 $\hat{R}^2_{l,k}$ as the approximations to R^2_{val}, $l = 1, \ldots, k$, $k \in \{3, 5, 10, 20, N\}$. Denoting $\hat{\mathbb{R}}^2_{LARS}$ and $\hat{\mathbb{R}}^2_{OMP}$ as the sets of $\hat{R}^2_{l,k}$ values obtained by LARS and OMP, the distribution of sets $\mathbb{R}^2_{LARS}$ and $\mathbb{R}^2_{OMP}$ is then simulated by $\hat{\mathbb{R}}^2_{LARS}$ and $\hat{\mathbb{R}}^2_{OMP}$, respectively.

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 $\mathbb{R}^2_{LARS}$ and $\mathbb{R}^2_{OMP}$. The inner one is embedded in the running of LARS and OMP to compute ϵ_{LOO} in Table 1 and 2. The two-layer cross validation here is indeed an realization of the known *double-cross-validation* (DCV) Baumann and Baumann (2014) or *cross model validation* (CMV) Anderssen et al. (2006); Gidskehaug et al. (2008). The related literature shows the unbiased estimation of R^2_{val} by the determination coefficient from the outer cross-validation errors, i.e., $\hat{R}^2_{l,k}$.

The procedures to build a PCE-based surrogate model by rPCE are summarized in Fig. 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 Sobol (1993); Homma and Saltelli (1996) for independent variables or Kucherenko indices Kucherenko et al. (2012) 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 Sudret (2008).

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-

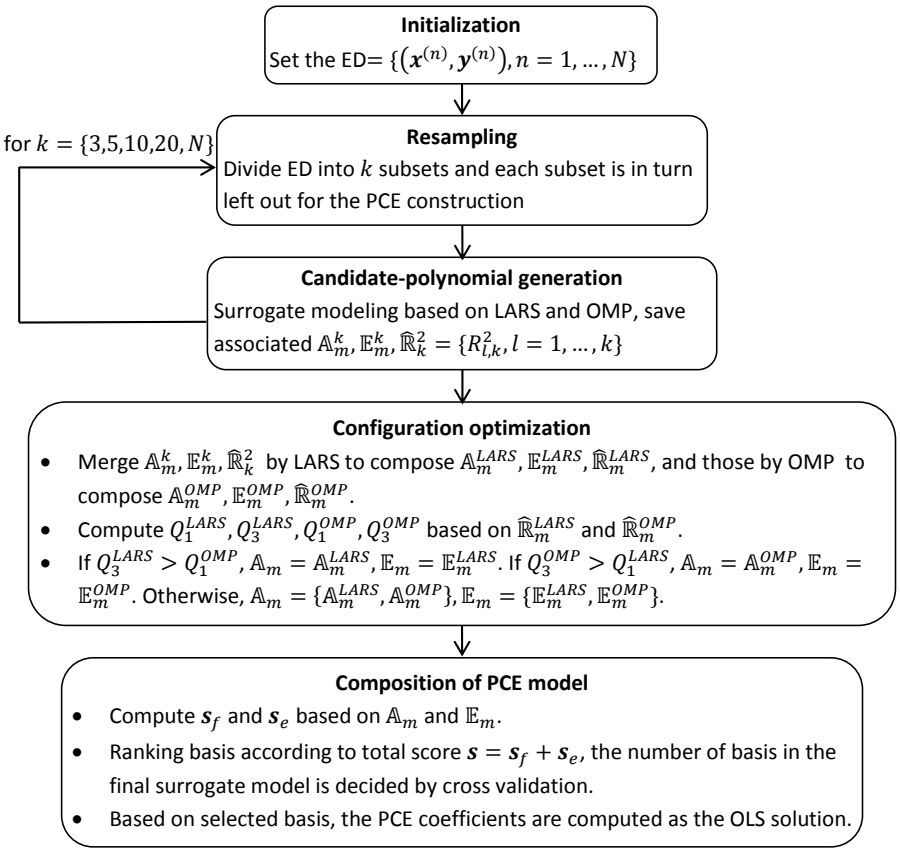


Figure 1: Computational flowchart of building a PCE model based on resampled PCE.

domain model (with $M = 4$) are analyzed. The PCE models based on LARS and OMP are obtained with the Matlab package UQLab Marelli and Sudret (2014). Using resampling, UQLab provides the candidate polynomials to rPCE.

Latin-Hypercube sampling McKay et al. (1979) 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. The maximum degree of multivariate polynomials p decides the flexibility of the obtained PCE model based on LARS and OMP, through the setting of $\mathbb{A}_{full}$ in Table 1 and 2. To optimize the value of p, the modeling process starts with $p = 1$ and early stops when ϵ_{LOO}^P increases for two consecutive degrees. The maximum degree p is equal to 20.

As mentioned in Section 2.2, dependent variables can be analyzed after the transformation into the corresponding independent ones 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.

Sobol' indices can be simply computed based on the PCE coefficients, following the ANOVA (analysis of variance) decomposition of the PCE expansion Eq. (14) as (see Sudret

(2008))

$$\begin{aligned}\widehat{\mathcal{M}}(\boldsymbol{x}) =&\beta_{\mathbf{0}} + \sum_{i=1}^{M} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_i) + \sum_{1\le i<j\le M} \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i,j\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_i,x_j) + \ldots + \\ &\sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{1,\ldots,M\}}} \beta_{\boldsymbol{\alpha}}\psi_{\boldsymbol{\alpha}}(x_1,\ldots,x_M),\end{aligned} \tag{19}$$

where $\mathbb{A}_{\{i\}}$ is defined as a subset of $\mathbb{A}$, within which only the i-th component α_i has a nonzero order:

$$\mathbb{A}_{\{i\}} = \{\boldsymbol{\alpha}\in\mathbb{A}, \alpha_i \neq 0, \alpha_{j\neq i} = 0\}. \tag{20}$$

$\mathbb{A}_{\{i,j\}}, \ldots, \mathbb{A}_{\{1,\ldots,M\}}$ have similar definitions:

$$\mathbb{A}_{\{i_1,\ldots,i_s\}} = \{\boldsymbol{\alpha}\in\mathbb{A}, \alpha_k \neq 0 \,\text{if}\, k\in\{i_1,\ldots,i_s\}; \alpha_k = 0 \,\text{otherwise}\}\,. \tag{21}$$

The orthogonality of basis polynomials gives the estimation of the total and partial variances,

$$\hat{D} = \sum_{\boldsymbol{\alpha}\in\mathbb{A}} \beta_{\boldsymbol{\alpha}}^2 - \beta_{\mathbf{0}}^2,\ \hat{D}_{i_1,\ldots,i_s} = \sum_{\boldsymbol{\alpha}\in\mathbb{A}_{\{i_1,\ldots,i_s\}}} \beta_{\boldsymbol{\alpha}}^2 - \beta_{\mathbf{0}}^2, \tag{22}$$

and the ratio between them yields the Sobol' indices

$$S_{i_1,\ldots,i_s} = \frac{D_{i_1,\ldots,i_s}}{D}. \tag{23}$$

While S_i estimates the influence of the i-th variable taken alone, the so-called total Sobol' indices Homma and Saltelli (1996) assess the total influence of an input variable and are computed as the summation of all Sobol' indices involving this variable, i.e.,

$$S_i^T = \sum_{\mathbb{I}_i} S_{i_1,\ldots,i_s},\ \mathbb{I}_i = \{\{i_1,\ldots,i_s\}\supset\{i\}\}\,\text{and}\,s\in\{1,\ldots,M\}. \tag{24}$$

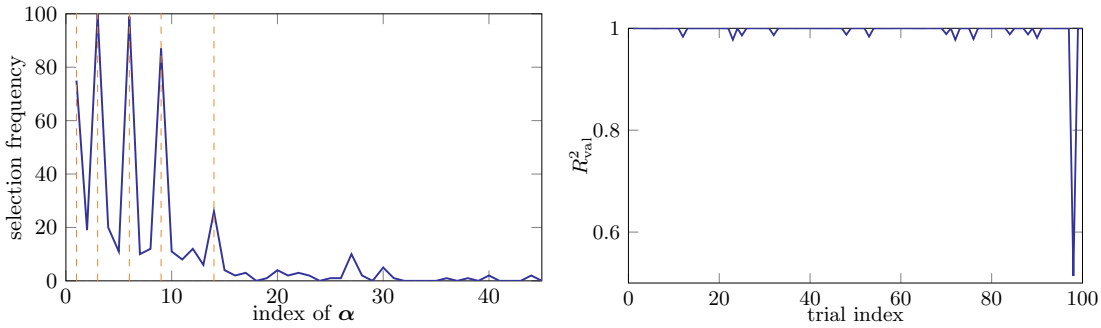
6.1 Summation of multivariate polynomials

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

$$Y = 1 + X_1 + X_1X_2 + X_1X_2^2 + X_1X_2^3, \tag{25}$$

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 validation. A total of 100 PCE constructions are made to test the selection frequency of polynomials.

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 to $\boldsymbol{\alpha} = (0,0)$, while the other four terms in Eq. (25) are with $(1,0), (1,1), (1,2), (1,3)$, respectively. Labeling $\boldsymbol{\alpha}$ by integers, the selection frequency during



(a) (b)

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

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 $R^2_{\text{val}} = 0.51$ in Fig. 2(b), where X_2, X_1, X_1^3 are selected as the basis).

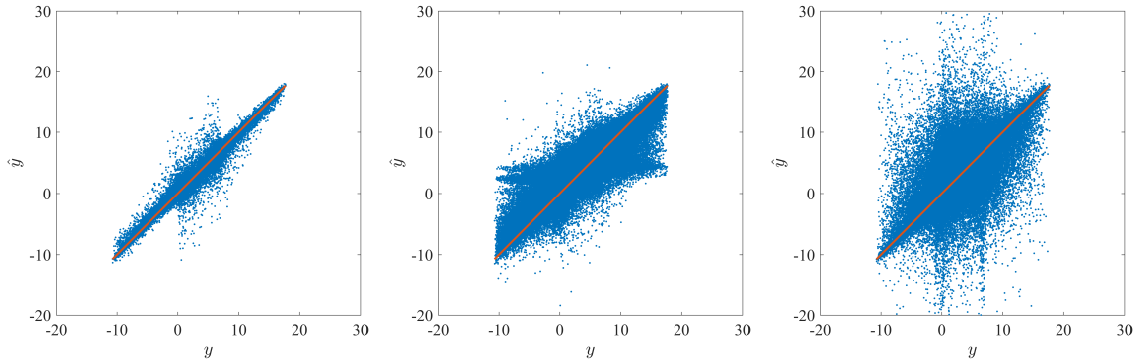
6.2 Ishigami function

The Ishigami function, which is defined by

$$Y = \sin X_1 + a \sin^2 X_2 + b X_3^4 \sin X_1, \tag{26}$$

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 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.



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

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

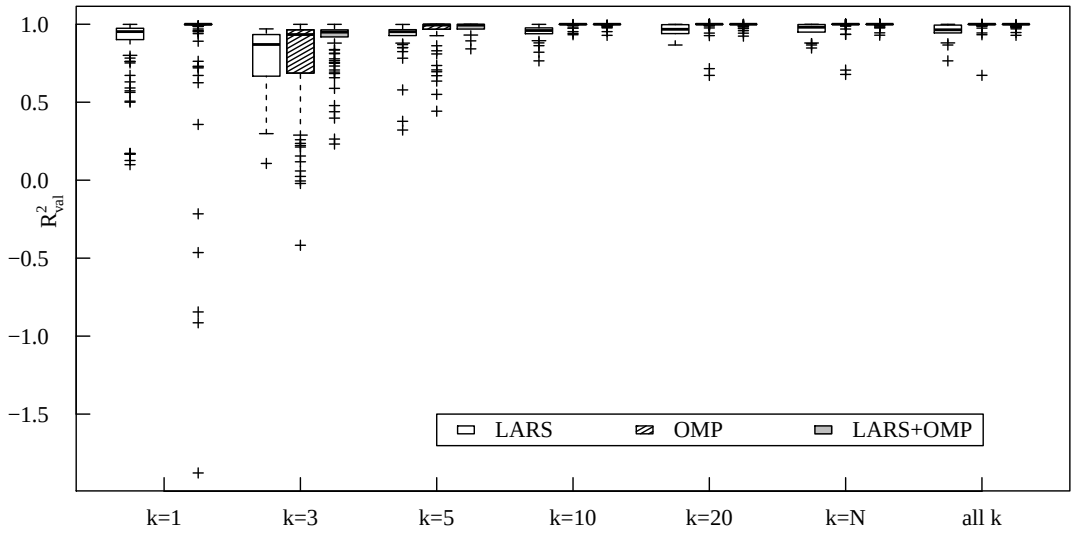


Figure 4: Ishigami function - box plots of R^2_{val} using different values of k in k-fold division with 50 data points (100 replications).

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_{val} is computed at each replication and 100 values of R^2_{val} yield the box plots of Fig. 4, 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_{val} 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.

Table 3: Ishigami function - mean of R^2_{val} 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

As quantitative comparisons, Table 3 gives the mean of R^2_{val} 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 3 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_{val} 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_{val} are plotted as the line graph in Fig. 5, which shows the better performance of rPCE compared to LARS and OMP in the cases with small EDs.

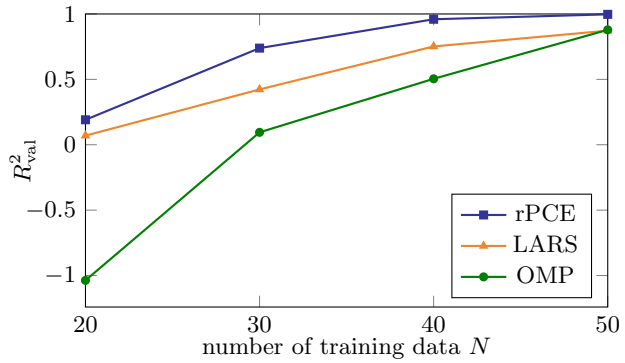


Figure 5: Ishigami function - mean of R^2_{val} versus different values of N (100 replications).

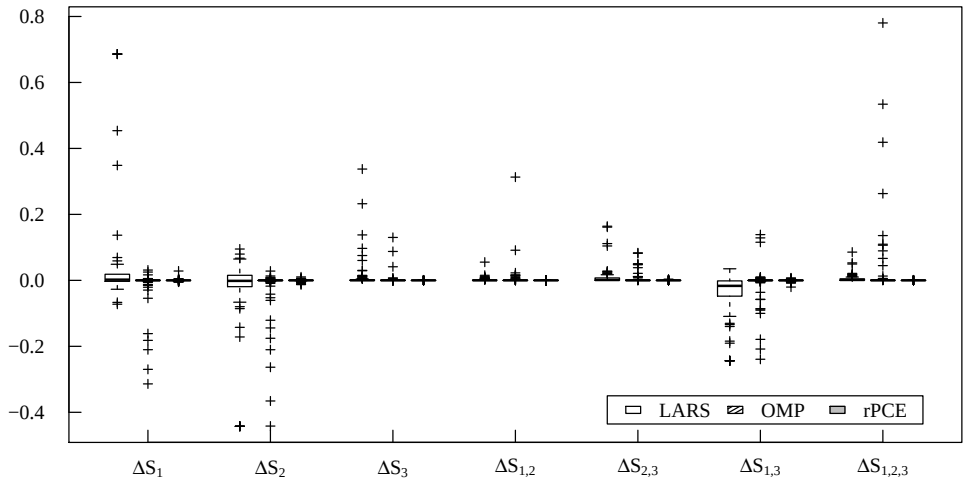


Figure 6: Ishigami function - the estimation error of Sobol' indices with 50 data points (100 replications).

Table 4: Ishigami function - mean of Sobol' indices 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

The Sobol' sensitivity indices can be analytically computed according to

$$D = \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{18} + \frac{1}{2}, D_1 = \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50} + \frac{1}{2}, D_2 = \frac{a^2}{8}, D_{1,3} = \frac{8b^2\pi^8}{225}, \\ D_3 = D_{1,2} = D_{2,3} = D_{1,2,3} = 0. \tag{27}$$

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

$$\Delta S_i = S_i^{\mathrm{PCE}} - S_i^{\mathrm{ref}}, \tag{28}$$

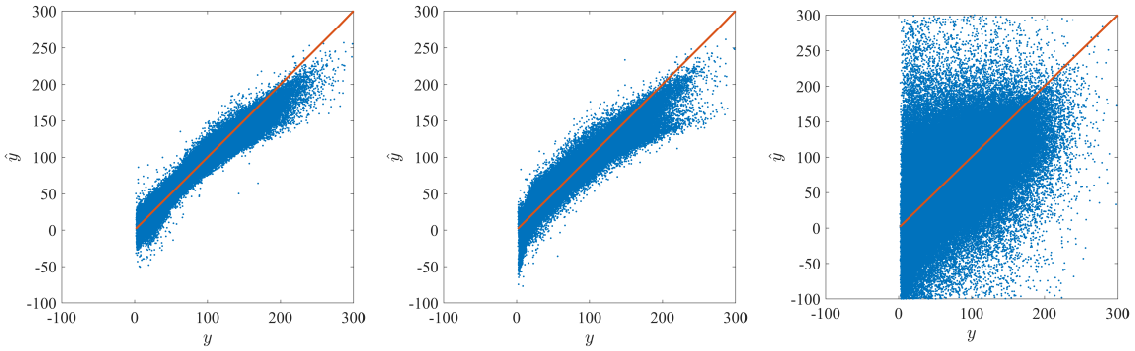
where the superscripts of S indicate the generation approach. With $N = 50$ and 100 replications, the box plots of all ΔS_i are shown in Fig. 6. The variance of ΔS_i is relatively large with LARS when the Sobol' indices are non zero, i.e., ΔS_1, ΔS_2, $\Delta S_{1,3}$, and the outliers are efficiently avoided by rPCE. The mean of S_i is given by Table 4, from which the superiority of rPCE in the sensitivity analysis of the Ishigami function is obviously observed. The accu-

racy 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

Table 5: Borehole function - description and distribution of input variables Xiong et al. (2013).

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



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

Figure 7: Borehole function - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 40 data points (100 replications).

The Borehole funtion with expression

$$Y = \frac{2\pi T_u (H_u - H_l)}{\ln(r/r_w)\,(1 + T_u/T_l) + 2LT_u/r_w^2 K_w} \tag{29}$$

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 5, 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 composition 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.

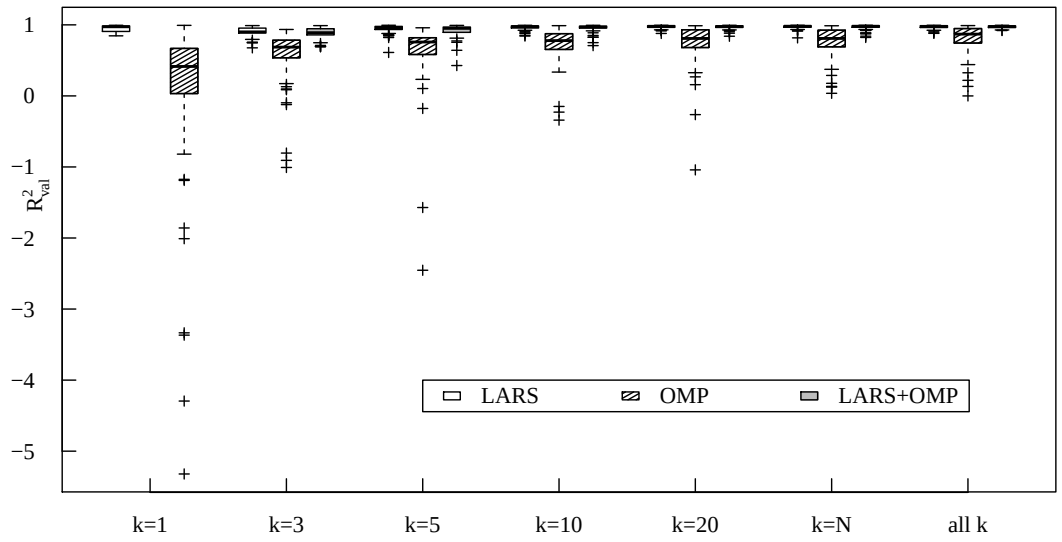


Figure 8: Borehole function - box plots of R^2_{val} using different values of k with 40 data points (100 replications).

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. 7. 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. 8.

Building the PCE model with the whole set of data, i.e., $k = 1$, the third quartile of R^2_{val} 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. 8 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 6. The mean of R^2_{val} 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 $R^2_{\text{val}} = 0.9724$. Fig. 9 provides more results when $N \in \{20, 30, 40, 50\}$. Since OMP has been shown much worse than rPCE and LARS, its associated line graph is not displayed for a clear view of the comparison

Table 6: Borehole function - mean of R^2_{val} 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

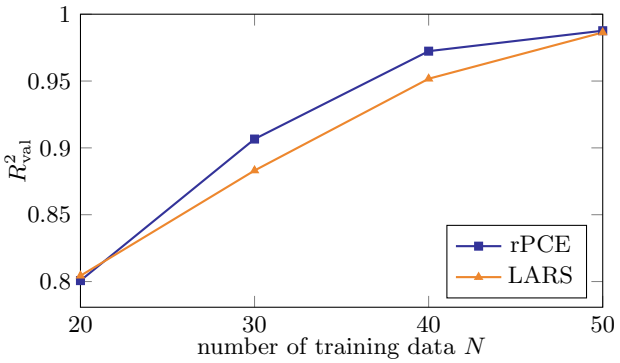


Figure 9: Borehole function - mean of R^2_{val} versus different values of N (100 replications).

between LARS and rPCE. The improvements are reached with rPCE in general except for the case of $N = 20$.

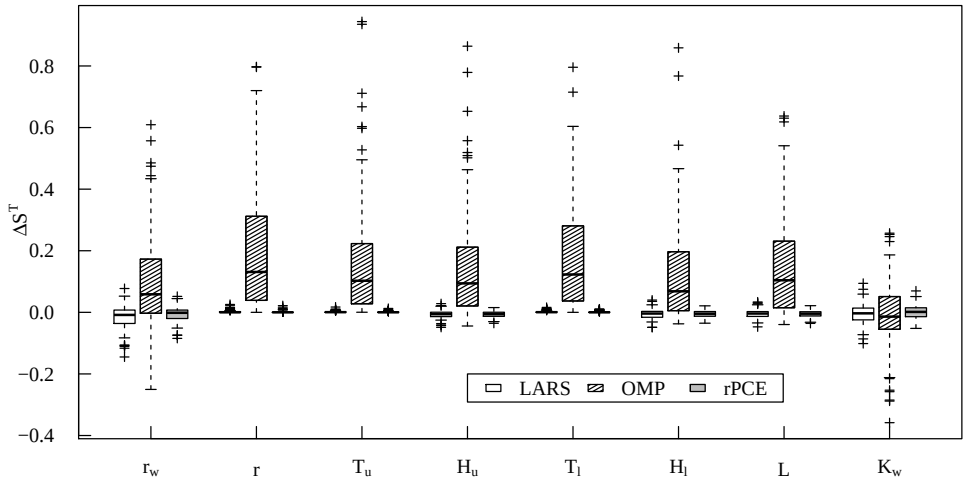


Figure 10: Borehole function - the estimation error of total Sobol' indices with 40 data points (100 replications).

Global sensitivity analysis is then considered and the total Sobol' indices are computed

Table 7: Borehole function - mean of the total Sobol' indices with 40 data points (100 replications).

	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

from the various PC expansions. The reference values are obtained by the Monte Carlo method with 10^7 data and presented in Table 7. 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. 10. 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

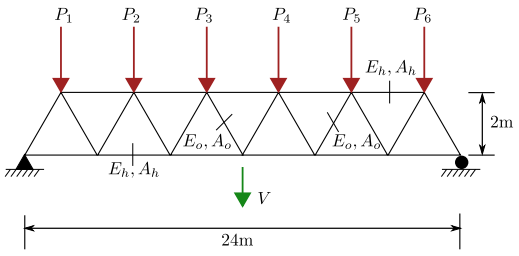


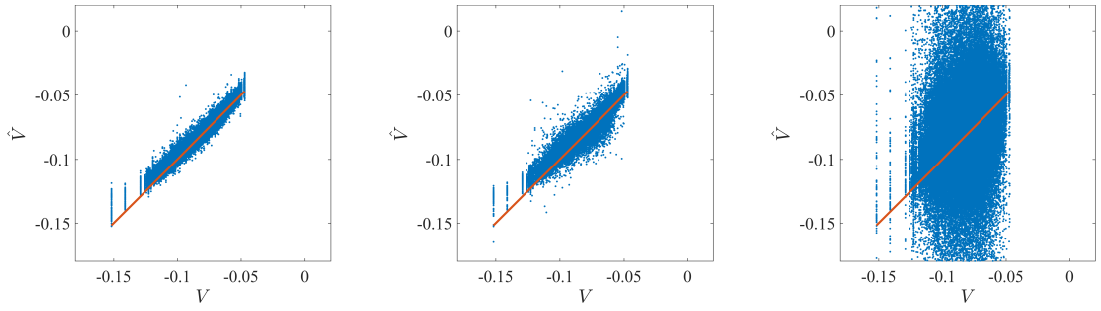
Figure 11: Sketch of a truss structure made of 23 bars Blatman and Sudret (2011).

In Fig. 11, 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.

Table 8: Truss deflection - description and distribution of input variables Blatman and Sudret (2011).

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

To analyze the uncertainty of the response, the input parameters are modeled by ten independent random variables following the distributions in Table 8. Transforming the input variables into standard normal ones with the isoprobabilistic transformation, LARS, OMP and rPCE surrogate models are built with basis composed of Hermite polynomials.



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

Figure 12: Truss deflection - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 50 data points (100 replications).

With $N = 50$ and 10^4 data for validation at each replication, Fig. 12 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_{val} is computed at each replication and the distribution of R^2_{val} over 100 replications is given in Fig. 13. 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 the combination of candidate polynomials generated with $k = [3, 5, 10, 20, N]$. Although the performance of OMP is much enhanced

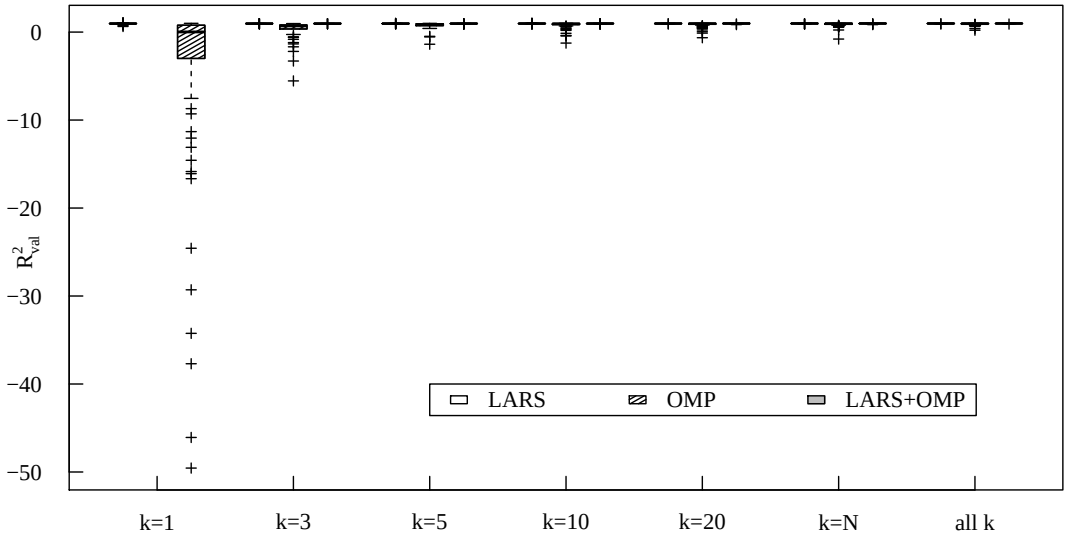


Figure 13: Truss deflection - box plots of R^2_{val} using different values of k with 50 data points (100 replications).

	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

Table 9: Truss deflection - mean of R^2_{val} with 50 data points (100 replications).

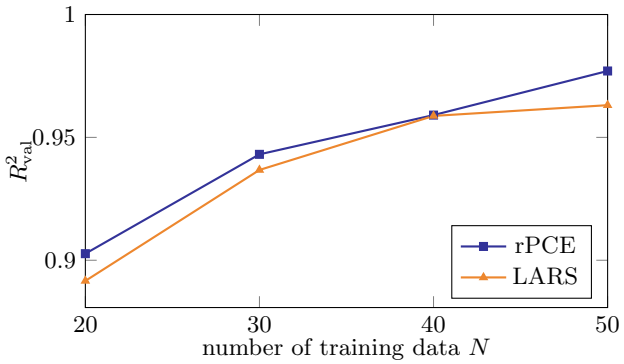


Figure 14: Truss deflection - mean of R^2_{val} versus different values of N (100 replications).

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 9 presents the associated mean of R^2_{val}. 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. 14, the mean value reaches 0.9770 for the "all k" option. The slight superiority of rPCE to LARS is also seen with $N = 20, 30, 40$.

Table 10: Truss deflection - mean of the total Sobol' indices with 50 data points (100 replications).

	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

Global sensitivity analysis is conducted by computing the total Sobol' indices based on the PCE coefficients. The reference values listed in Table 10 are obtained with 5.5×10^6 Monte Carlo simulations Blatman and Sudret (2011). 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. 15.

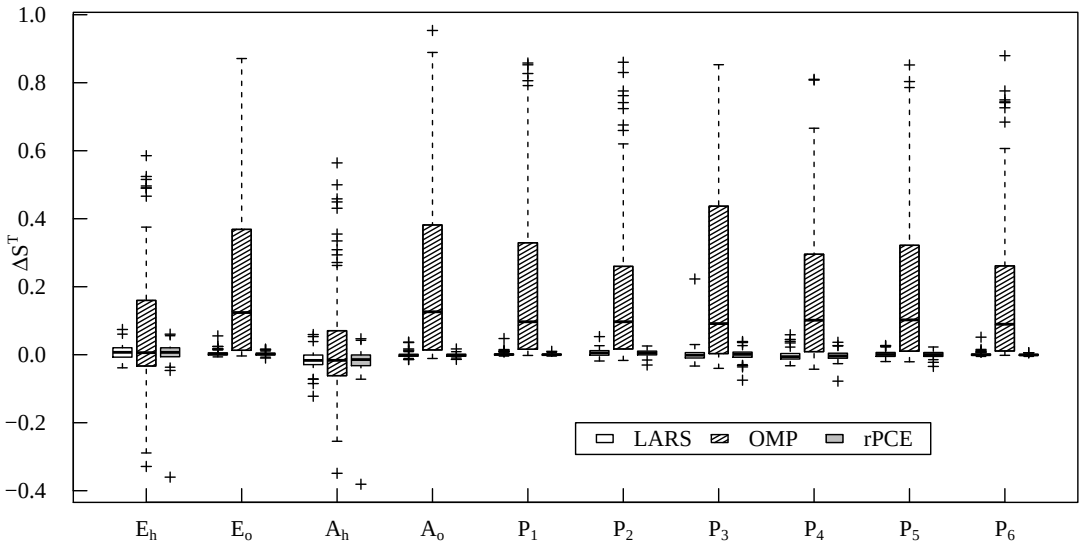


Figure 15: Truss deflection - the estimation error of total Sobol' indices with 50 data points (100 replications).

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

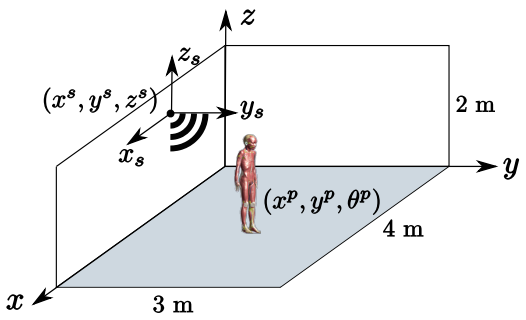


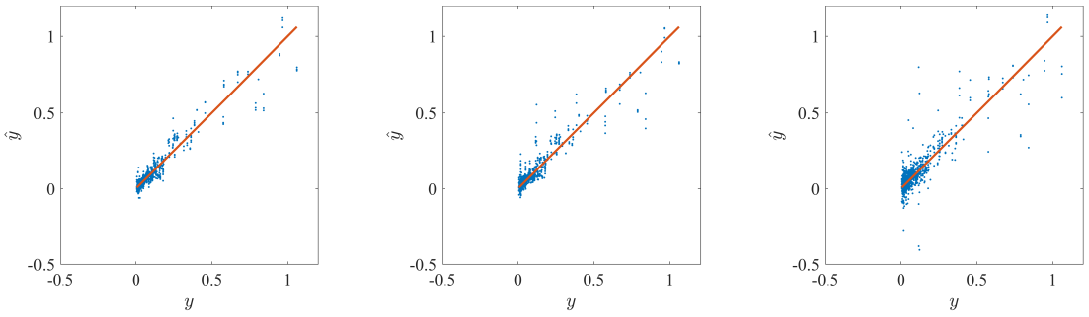
Figure 16: Sketch of the human-exposure estimation in an indoor down-link scenario.

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 Van Deventer et al. (2011). Here, an indoor down-link scenario is considered, as sketched in Fig. 16. A high-resolution model of a 8-year girl (1.36 m high), named as "Eartha", from the Virtual Classroom Gosselin et al. (2014), 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 Hansen (1988), by Microwave Vision Group (MVG®). With an in-house finite-difference-time-domain (FDTD) code, the whole-body specific absorption rate (SAR) Liorni et al. (2016), 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.

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. Leaving z^s as an independent input, the relative position is represented in the (x, y) plane. In the local coordinate system of the source, as shown in Fig. 16, 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.



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

Figure 17: SAR estimation - prediction of validation data by (a) rPCE, (b) LARS and (c) OMP with 340 data points (100 replications).

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_{val} is yielded by repeating this process 100 times. Consequently, with the remaining 340 data, surrogate models are obtained with LARS, OMP and rPCE. Then, a validation set of size 10^3 is computed and the results are shown in Fig. 17. As seen, the whole-body SAR is smaller than 0.2 for most of cases (90 percents

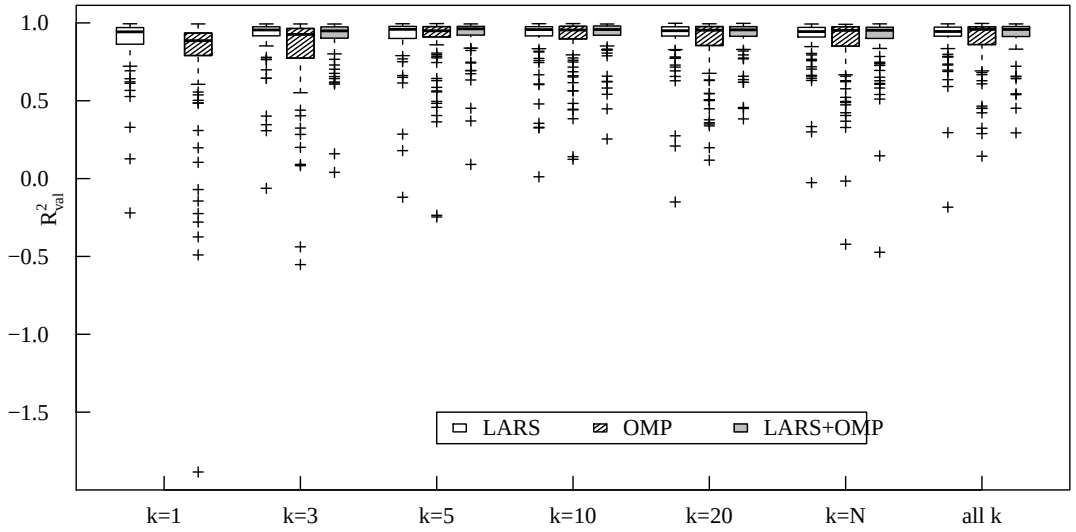


Figure 18: SAR estimation - box plots of R^2_{val} with different values of k (100 replications).

	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

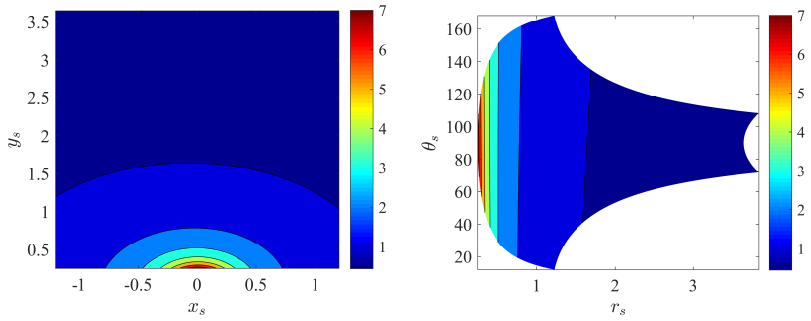
Table 11: SAR estimation - mean of R^2_{val} with 340 data points (100 replications).

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_{val} is given in Fig. 18. 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 11 shows the mean of R^2_{val}.

The total Sobol' indices are computed based on the PCE coefficients and the mean values are presented in Table 12. As seen, the whole-body human exposure is mainly impacted by the relative distance r_s^p 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_s^p, might be

	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

Table 12: SAR estimation - mean of the total Sobol' indices with 340 data points (100 replications).



(a) (b)

Figure 19: SAR estimation - contour of electric-field intensity (a) in the (x, y) plane and (b) its representation in the polar coordinate system, $z_s = 0$.

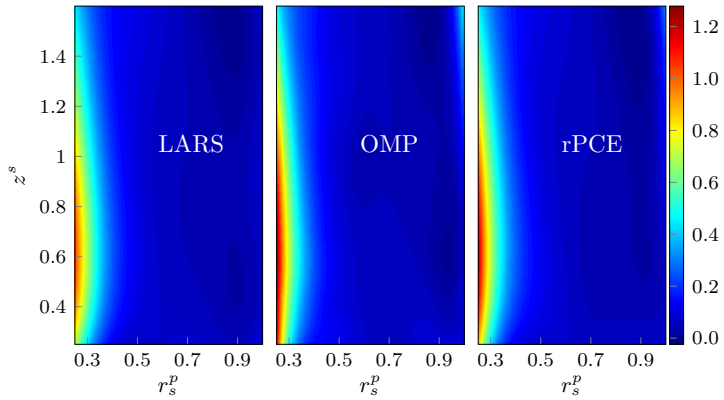


Figure 20: SAR estimation - the prediction of whole-body SAR with 340 data points (100 replications).

explained by looking at the contours of electric-field intensity in Fig. 19, 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 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 12 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. 20, 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.

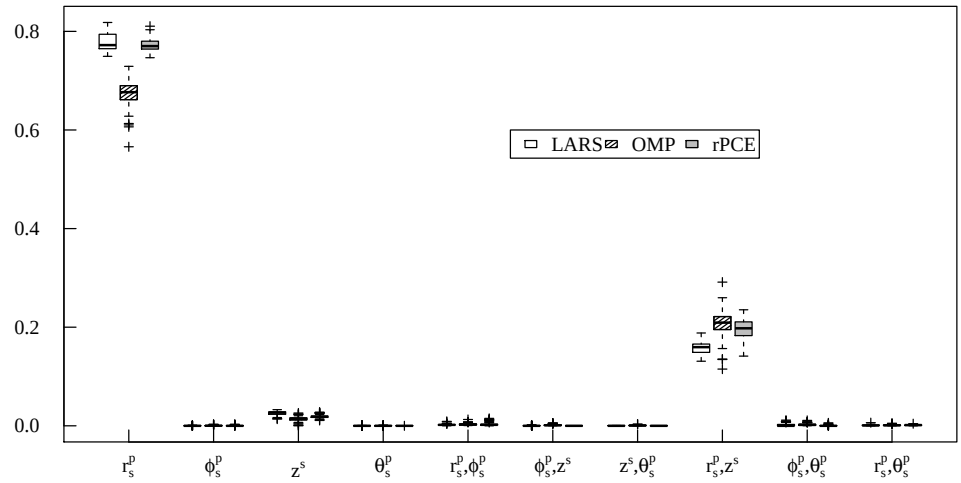


Figure 21: SAR estimation - the estimation of first-order and second-order Sobol' indices with 340 data points (100 replications).

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. 21, which presents that r_s^p and its interaction with z^s contribute the most to the uncertainty of the response.

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) Picheny et al. (2010); Blatman and Sudret (2011); Dubreuil et al. (2014); Fajraoui et al. (2017). 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 forthcoming investigations, more complex scenarios (e.g., electromagnetic dosimetry for human models in the telecommunications network Liorni et al. (2015); Kersaudy et al. (2015); Huang and Wiart (2017)) 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.

Acknowledgments

Support is from the Emergence programme of the Science and Technologies of Information and Communication (STIC) department, University Paris-Saclay.

References

Anderssen, E., K. Dyrstad, F. Westad, and H. Martens (2006). Reducing over-optimism in variable selection by cross-model validation. *Chemom. Intell. Lab. Syst. 84*(1-2), 69–74.

Barton, R. R. (2012). Tutorial: Input uncertainty in output analysis. In *Proc. Winter Simulation Conference, WSC2012*. Berlin, Germany.

Bathe, K.-J. and E. L. Wilson (1976). *Numerical Methods in Finite Element Analysis*. Prentice-Hall.

Baumann, D. and K. Baumann (2014). Reliable estimation of prediction errors for QSAR models under model uncertainty using double cross-validation. *J. Cheminf. 6*(1), 47.

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

Blatman, G. (2009). *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. Ph. D. thesis, Université Blaise Pascal, Clermont-Ferrand, France.

Blatman, G. and B. Sudret (2010). An adaptive algorithm to build up sparse polynomial chaos expansions for stochastic finite element analysis. *Probab. Eng. Mech. 25*(2), 183–197.

Blatman, G. and B. Sudret (2011). Adaptive sparse polynomial chaos expansion based on least angle regression. *J. Comput. Phys. 230*(6), 2345–2367.

Breiman, L. (1996). Bagging predictors. *Mach. Learn. 24*(2), 123–140.

Doostan, A. and H. Owhadi (2011). A non-adapted sparse approximation of PDEs with stochastic inputs. *J. Comput. Phys. 230*(8), 3015–3034.

Dubreuil, S., M. Berveiller, F. Petitjean, and M. Salaün (2014). Construction of bootstrap confidence intervals on sensitivity indices computed by polynomial chaos expansion. *Reliab. Eng. Sys. Safety 121*, 263–275.

Efron, B. (1982). *The Jackknife, The Bootstrap, and Other Resampling Plans*. SIAM.

Efron, B., T. Hastie, I. Johnstone, and R. Tibshirani (2004). Least angle regression. *Ann. Stat. 32*(2), 407–499.

Efron, B. and R. J. Tibshirani (1994). *An Introduction to The Bootstrap*. CRC Press.

Fajraoui, N., S. Marelli, and B. Sudret (2017). Sequential design of experiment for sparse polynomial chaos expansions. *SIAM/ASA J. Unc. Quant. 5*(1), 1061–1085.

Friedman, J., T. Hastie, and R. Tibshirani (2001). *The Elements of Statistical Learning*. Springer Series in Statistics.

Gautschi, W. (2004). *Orthogonal Polynomials: Computation and Approximation*. Oxford University Press on Demand.

Geisser, S. (1975). The predictive sample reuse method with applications. *J. Am. Stat. Assoc. 70*(350), 320–328.

Ghanem, R. G. and P. D. Spanos (2003). *Stochastic Finite Elements: A Spectral Approach*. Dover Publications.

Gidskehaug, L., E. Anderssen, and B. K. Alsberg (2008). Cross model validation and optimisation of bilinear regression models. *Chemom. Intell. Lab. Syst. 93*(1), 1–10.

Gilli, L., D. Lathouwers, J. Kloosterman, T. Van der Hagen, A. Koning, and D. Rochman (2013). Uncertainty quantification for criticality problems using non-intrusive and adaptive polynomial chaos techniques. *Ann. Nucl. Energy 56*, 71–80.

Gosselin, M.-C., E. Neufeld, H. Moser, E. Huber, S. Farcito, L. Gerber, M. Jedensjoe, I. Hilber, F. Di Gennaro, B. Lloyd, et al. (2014). Development of a new generation of high-resolution anatomical models for medical device evaluation: the virtual population 3.0. *Phys. Med. Biol. 59*(18), 5287.

Hansen, J. E. (1988). *Spherical Near-Field Antenna Measurements*. Peter Peregrinus Ltd.

Homma, T. and A. Saltelli (1996). Importance measures in global sensitivity analysis of nonlinear models. *Reliab. Eng. Syst. Safe. 52*(1), 1–17.

Huang, Y. and J. Wiart (2017). Simplified assessment method for population RF exposure induced by a 4G network. *IEEE J. Electromagn. RF Microw. Med. Biol. 1*(1), 34–40.

Iman, R. L. and J. C. Helton (1988). An investigation of uncertainty and sensitivity analysis techniques for computer models. *Risk Anal. 8*(1), 71–90.

Jakeman, J. D., M. S. Eldred, and K. Sargsyan (2015). Enhancing ℓ_1-minimization estimates of polynomial chaos expansions using basis selection. *J. Comput. Phys. 289*, 18–34.

Kersaudy, P., S. Mostarshedi, B. Sudret, O. Picon, and J. Wiart (2014). Stochastic analysis of scattered field by building facades using polynomial chaos. *IEEE Trans. Antennas Propag. 62*(12), 6382–6393.

Kersaudy, P., B. Sudret, N. Varsier, O. Picon, and J. Wiart (2015). A new surrogate modeling technique combining Kriging and polynomial chaos expansions–application to uncertainty analysis in computational dosimetry. *J. Comput. Phys. 286*, 103–117.

Kleijnen, J. P. (2009). Kriging metamodeling in simulation: A review. *Eur. J. Oper. Res. 192*(3), 707–716.

Kohavi, R. et al. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proc. 14th International Joint Conference on Artificial Intelligence, IJCAI1995*, Volume 14, pp. 1137–1145. Montreal, Canada.

Kolmogorov, A. (1956). *Foundations of the Theory of Probability: Second English Edition*. Dover Publications.

Konakli, K. and B. Sudret (2016). Polynomial meta-models with canonical low-rank approximations: numerical insights and comparison to sparse polynomial chaos expansions. *J. Comput. Phys. 321*, 1144–1169.

Kucherenko, S., S. Tarantola, and P. Annoni (2012). Estimation of global sensitivity indices for models with dependent variables. *Comput. Phys. Commun. 183*(4), 937–946.

Le Maître, O. P., M. T. Reagan, H. N. Najm, R. G. Ghanem, and O. M. Knio (2002). A stochastic projection method for fluid flow: II. Random process. *J. Comput. Phys. 181*(1), 9–44.

Lebrun, R. and A. Dutfoy (2009). A generalization of the Nataf transformation to distributions with elliptical copula. *Probab. Eng. Mech. 24*(2), 172–178.

Lemaire, M. (2013). *Structural Reliability*. John Wiley & Sons.

Liorni, I., M. Parazzini, S. Fiocchi, and P. Ravazzani (2015). 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), 5934–5953.

Liorni, I., M. Parazzini, N. Varsier, A. Hadjem, P. Ravazzani, and J. Wiart (2016). 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), 3237.

MacKay, D. J. (1992). *Bayesian methods for adaptive models*. Ph. D. thesis, California Institute of Technology, CA, USA.

Marelli, S. and B. Sudret (2014). UQLab: A framework for uncertainty quantification in Matlab. In *Proc. 2nd International Conference on Vulnerability, Risk Analysis and Management, ICVRAM2014*, pp. 2554–2563. Liverpool, United Kingdom.

Marelli, S. and B. Sudret (2018). UQLab user manual–polynomial chaos expansions, Report UQLab-V1.1-104. *Chair of Risk, Safety & Uncertainty Quantification, ETH Zürich*.

McKay, M. D., R. J. Beckman, and W. J. Conover (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics 21*(2), 239–245.

Nelsen, R. B. (2007). *An Introduction to Copulas*. Springer Science & Business Media.

Picheny, V., D. Ginsbourger, O. Roustant, R. T. Haftka, and N.-H. Kim (2010). Adaptive designs of experiments for accurate approximation of a target region. *J. Mech. Des. 132*(7), 071008.

Rao, C. R., C. R. Rao, M. Statistiker, C. R. Rao, and C. R. Rao (1973). *Linear Statistical Inference and Its Applications*. Wiley.

Sepahvand, K., S. Marburg, and H.-J. Hardtke (2010). Uncertainty quantification in stochastic systems using polynomial chaos expansion. *Int J. Appl. Mech. 2*(02), 305–353.

Sobol, I. M. (1993). Sensitivity estimates for nonlinear mathematical models. *Mathem. Mod. Comput. Exp. 1*(4), 407–414.

Soize, C. and R. Ghanem (2004). Physical systems with random uncertainties: chaos representations with arbitrary probability measure. *SIAM J. Sci. Comput. 26*(2), 395–410.

Sudret, B. (2007). 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*.

Sudret, B. (2008). Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Syst. Safe. 93*(7), 964–979.

Sudret, B., M. Berveiller, and M. Lemaire (2004). A stochastic finite element method in linear mechanics. *CR Mécanique 332*(7), 531–537.

Taflove, A. and S. C. Hagness (2005). *Computational Electrodynamics: the Finite-Difference Time-Domain Method*. Artech House.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Royal Stat. Soc. Series B*, 267–288.

Tropp, J. A. and A. C. Gilbert (2007). Signal recovery from random measurements via orthogonal matching pursuit. *IEEE Trans. Inf. Theory 53*(12), 4655–4666.

Van Deventer, E., E. Van Rongen, and R. Saunders (2011). WHO research agenda for radiofrequency fields. *Bioelectromagnetics 32*(5), 417–421.

Weisberg, S. (2005). *Applied Linear Regression*. John Wiley & Sons.

Wiener, N. (1938). The homogeneous chaos. *Am. J. Math. 60*(4), 897–936.

Xiong, S., P. Z. Qian, and C. J. Wu (2013). Sequential design and analysis of high-accuracy and low-accuracy computer codes. *Technometrics 55*(1), 37–46.

Xiu, D. and G. E. Karniadakis (2002). The Wiener–Askey polynomial chaos for stochastic differential equations. *SIAM J. Sci. Comput. 24*(2), 619–644.