Adaptive surrogate modeling by ANOVA and sparse polynomial dimensional decomposition for global sensitivity analysis in fluid simulation

Kunkun Tang\textsuperscript{a,b,}\textsuperscript{*}, Pietro M. Congedo\textsuperscript{b}, Rémi Abgrall\textsuperscript{c}

\textsuperscript{a} The Center for Exascale Simulation of Plasma-Coupled Combustion (XPACC), University of Illinois at Urbana–Champaign, 1308 W Main St, Urbana, IL 61801, USA

\textsuperscript{b} Inria Bordeaux – Sud-Ouest, Team Cardamom, 200 avenue de la Vieille Tour, 33405 Talence, France

\textsuperscript{c} Institut für Mathematik, Universität Zürich, Winterthurerstrasse 190, CH-8057 Zürich, Switzerland

\textbf{A R T I C L E   I N F O}

Article history:
Received 7 June 2015
Received in revised form 14 January 2016
Accepted 11 March 2016
Available online 16 March 2016

Keywords:
Uncertainty quantification
Non-intrusive approach
Global sensitivity analysis
Analysis of Variance (ANOVA)
Polynomial dimensional decomposition (PDD)
Regression approach
Adaptive sparse polynomial surrogate model
Atmospheric reentry

\textbf{A B S T R A C T}

The Polynomial Dimensional Decomposition (PDD) is employed in this work for the global sensitivity analysis and uncertainty quantification (UQ) of stochastic systems subject to a moderate to large number of input random variables. Due to the intimate connection between the PDD and the Analysis of Variance (ANOVA) approaches, PDD is able to provide a simpler and more direct evaluation of the Sobol’ sensitivity indices, when compared to the Polynomial Chaos expansion (PC). Unfortunately, the number of PDD terms grows exponentially with respect to the size of the input random vector, which makes the computational cost of standard methods unaffordable for real engineering applications. In order to address the problem of the curse of dimensionality, this work proposes essentially variance-based adaptive strategies aiming to build a cheap meta-model (i.e. surrogate model) by employing the sparse PDD approach with its coefficients computed by regression. Three levels of adaptivity are carried out in this paper: 1) the \textit{truncated dimensionality} for ANOVA component functions, 2) the \textit{active dimension} technique especially for second- and higher-order parameter interactions, and 3) the \textit{stepwise regression} approach designed to retain only the most influential polynomials in the PDD expansion. During this adaptive procedure featuring stepwise regressions, the surrogate model representation keeps containing few terms, so that the cost to resolve repeatedly the linear systems of the least-squares regression problem is negligible. The size of the finally obtained sparse PDD representation is much smaller than the one of the full expansion, since only significant terms are eventually retained. Consequently, a much smaller number of calls to the deterministic model is required to compute the final PDD coefficients.

© 2016 Elsevier Inc. All rights reserved.

1. Introduction

Compared to local sensitivity analysis methods, Global Sensitivity Analysis (GSA) has the advantage of taking into account the overall influence of input parameters and their interactions onto the output quantity of interest, by considering the entire input space rather than a specific nominal point (see for example [1]). However, the main difficulty encountered

\textsuperscript{*} Corresponding author at: The Center for Exascale Simulation of Plasma-Coupled Combustion (XPACC), University of Illinois at Urbana–Champaign, 1308 W Main St, Urbana, IL 61801, USA. Tel.: +1 (217) 418 4934.

E-mail address: ktg@illinois.edu (K. Tang).

http://dx.doi.org/10.1016/j.jcp.2016.03.026

0021-9991/© 2016 Elsevier Inc. All rights reserved.
when employing the global methods is the required high cost of numerical computations, since the Monte Carlo simulation (MC) or a quasi-Monte Carlo method (QMC) is usually applied to estimate the sensitivity indices. For complex problems in real life, relying upon MC or QMC can be very expensive from a computational point of view.

Iman & Hora [2] compared three GSA uncertainty importance measures that are based on the expected variance reduction assuming that some input is known with certainty, the same importance measure based on logarithmic scale of the model output, and a measure based on shifts in the quantiles of the output distribution. Later, GSA was performed using methods based on the decomposition of the output variance (see for instance [3–6]), i.e. Analysis of Variance (ANOVA), which is nowadays one of the most commonly used GSA techniques in the literature. Indeed, ANOVA relies on a functional decomposition that incorporates component functions involving a single uncertain parameter, or a group of parameters, and the computation of the sensitivity measures of each component function usually requires MC or QMC simulations. In [3,4,6], they used this formulation to define global sensitivity indices, displaying the relative variance contributions of different ANOVA terms. Note that the same sensitivity measures have been developed independently in [7] in operations research. In [8,9], they introduced two High-Dimensional Model Representation (HDMR) techniques to capture input–output relationships of physical systems with many input variables. These techniques are based on ANOVA-type decompositions. Since it usually requires a large number of function evaluations to perform this analysis, several techniques have been developed to compute the different so-called sensitivity indices at low cost [10]. In [11–13], the generalized Polynomial Chaos expansions (gPC) are used to build surrogate models for computing the Sobol’ indices analytically as a post-processing of the PC coefficients. In [14], they combine the multi-element polynomial chaos with the ANOVA functional decomposition to enhance the convergence rate of polynomial chaos in high dimensions and in problems with low stochastic regularity. The use of adaptive ANOVA decomposition is investigated in [15] as an effective dimension–reduction technique in modeling incompressible and compressible flows with a high dimensionality of random space. In Sudret [13], the sparse Polynomial Chaos (PC) expansions are introduced in order to compute the global sensitivity indices. An adaptive algorithm allows the construction of a PC-based meta-model that only contains the significant terms, with the PC coefficients being computed by the least-squares regression. Other approaches are developed if the assumption of independence of the input parameters is not valid. New indices have been proposed to address the dependence [16,17], but these attempts are limited to a linear correlation. In [18], they introduce a global sensitivity indicator which looks at the influence of input uncertainty on the entire output distribution without reference to a specific moment of the output (moment independence) and which can also be defined in the presence of correlations among the parameters. In [19], a gPC methodology to address the GSA for this kind of problems is introduced. A moment-independent sensitivity index that suits problems with dependent parameters is reviewed. Recently, in [20], a numerical procedure is set up for moment-independent sensitivity methods. An exhaustive review of recent advances in sensitivity analysis methods is given in [21], where several types of both local and global approaches have been analyzed.

Recently, the anchored ANOVA method (i.e. cut-HDMR) has been widely used in the literature (see for instance [22–27, 15,28]). In particular, the method proposed in [28] is based on the covariance decomposition of the output variance to obtain accurate results which are less sensitive to the choice of the reference point, and meanwhile preserves the main advantage of the anchored method, i.e. a much smaller number of deterministic simulations is needed compared to the standard ANOVA method combined with MC or QMC. A disadvantage of this anchor based global method is that a surrogate model cannot be built in a straightforward way.

The objective of this work consists in building an efficient UQ and GSA method featuring a surrogate model representation that is affordable for complex numerical simulation problems. An accurate surrogate representation is useful for accelerating the model evaluations when using, for instance, the MC type techniques. We emphasize again that in [13], the generalized Polynomial Chaos (gPC) is combined with the ANOVA approach to perform the global analysis and to build a surrogate sparse representation. We recall a traditional gPC decomposition of a $N$-dimensional function $f(\xi_1, \xi_2, \cdots, \xi_N)$ of degree not exceeding $p$ can be written as follows

$$f(\xi_1, \xi_2, \cdots, \xi_N) = a_0\xi_0 + \sum_{i_1=1}^{N} a_{1i_1}\xi_1(\xi_{i_1}) + \sum_{i_2=1}^{N} \sum_{i_1=1}^{N} a_{1i_2i_1}\xi_2(\xi_{i_1}, \xi_{i_2})$$

$$+ \cdots + \sum_{i_p=1}^{N} \sum_{i_{p-1}=1}^{N} \sum_{i_{p-2}=1}^{N} \cdots \sum_{i_1=1}^{N} a_{1i_2\cdots i_p}\xi_p(\xi_{i_1}, \xi_{i_2}, \cdots, \xi_{i_p}),$$

(1)

where $\xi_0(\xi_{i_1}, \xi_{i_2}, \cdots, \xi_{i_p})$ denote the orthogonal polynomials of order $n$ from the Askey scheme (see [29,30]) in terms of the multivariate random variables $(\xi_1, \xi_2, \cdots, \xi_n)$. On the one hand, note the gPC formulation (1) is organized with respect to an increasing degree of multivariate polynomials, and not to an increasing order of parameter interactions. For instance, the group of polynomials of order $n$ can contain polynomials with random variables subject to a dimensionality equal to or less than $n$. For this reason in order to compute the Sobol’ indices, one needs additionally to reorder the PC terms according to the random variables they depend on. This task can be annoying in our practice for complex multivariate problems. On the other hand, the gPC expansions are known to succumb to the curse of dimensionality for high-dimensional problems. To address these issues in the present work, we employ the Polynomial Dimensional Decomposition (PDD) (introduced and developed by Rahman and coworkers in e.g. [31–35]) instead of PC and combine it with the ANOVA decomposition due to
their direct link with each other. Indeed, the PDD expansion of the model output relies on the ANOVA functional decomposition in a very direct way: a $T$-dimensional ANOVA component function is represented by a summation of $T$-dimensional multivariate polynomials from lowest to highest degree using a specific polynomial basis subject to the input parameter distributions. Thus, the PDD prioritizes the low order parameter interactions, which matches perfectly the principle of the ANOVA decomposition where we suppose that the low order component functions are dominant for most engineering cases. Consequently, even with a slightly truncated PDD expansion, one can take advantage of employing relatively high-order uni- or multi-variate polynomials. Similar to the gPC approach, an inevitable task is to determine the polynomial coefficients of PDD. The least-squares regression approach is an efficient tool for this purpose, by minimizing the error of the surrogate model representation in the mean square sense (see e.g. [36,11,13]). Compared to the projection approach (see e.g. [37–39]) where each polynomial coefficient is obtained by computing a multi-dimensional integral, the regression approach is more flexible (in choosing sampling points) for problems involving a moderate number of uncertain parameters. It is known that the number of ANOVA component functions increases exponentially with respect to the uncertain parameter dimensionality, and meanwhile the imposed polynomial order for the PDD expansion involves a polynomial increase of the number of PDD terms for each component function. This phenomenon causes one main limitation of the regression approach even for a truncated low-order ANOVA expansion, namely the high number of deterministic model evaluations for problems characterized by a moderate to large number of uncertainties; indeed, for the regression problem to be well posed, the number of deterministic model evaluations is necessary to be larger than the total polynomial expansion size [13,40]. In this respect, this paper proposes to combine the active dimension strategy in the framework of the adaptive ANOVA method [15,28] and the stepwise regression technique [13,40] to obtain an efficient sparse surrogate model representation. This method has similarities with the adaptive sparse PC approach presented in [13,40], but differs by the use of the PDD expansion, the truncation order/active dimension technique used in the ANOVA type methods, and the selection criteria allowing to retain the most important polynomial terms. Once the sparse surrogate model is built, the Sobol’ sensitivity indices can be easily obtained by manipulating the polynomial coefficients without any reordering, which is required for the PC expansion.

The paper is organized as follows. The classical ANOVA functional decomposition and the related variance-based global sensitivity indices are summarized in Section 2. The Polynomial Dimensional Decomposition (PDD) expansion, the determination of its coefficients, and its link with the sensitivity analysis are presented in Section 3. The proposed adaptive sparse meta-modeling approach is then presented in Section 4 by using a combination of the active dimension technique of the ANOVA expansion and the stepwise regression approach. As far as the strategy of selecting the most significant polynomial terms is concerned, the variance-based selection criterion is detailed (Criterion 1). Three academic benchmark problems as well as an atmospheric reentry spacecraft problem are addressed in Section 5. Conclusions follow. Appendix A is devoted to the presentation of the estimator of accuracy used in this work. We also present Criterion 2 that allows to retain the most important terms by comparing the surrogate representation error.

2. ANOVA decomposition and global sensitivity analysis

Let us introduce some notation before introducing ANOVA expansion. The upper-case letters, $X = (X_1, \cdots, X_N)$ and $Y$, denote a list of input random variables (random vector) and a scalar random output, respectively; the lower-case letters $x$ and $y$ represent the realizations. We assume the $N$ input random variables $X$ are independent, and admit a product joint probability density function (pdf) such that

$$p_X(x) = \prod_{i=1}^{N} p_{X_i}(x_i),$$

(2)

where $p_{X_i}(x_i)$ is the marginal pdf of $X_i$.

The expectation and variance of an integrable function of random vector $X$, denoted by $E(g(X))$ and $\text{Var}(g(X))$ respectively, are given by

$$E(g(X)) = \int_{\mathbb{R}^N} g(x)p_X(x) \, dx,$$

$$\text{Var}(g(X)) = \int_{\mathbb{R}^N} (g(x) - E(g(X)))^2 p_X(x) \, dx,$$

where $dx = \prod_{i=1}^{N} dx_i$. The model output expectation $E(g(X)|x_i)$ and variance $\text{Var}(g(X)|x_i)$, conditioned on $x_i$, employed in this paper are defined in the standard way with respect to the conditional pdf.

Let us suppose that the response of a given system of interest can be represented by a $N$-dimensional function $y = f(x)$

$$y = f(x) = f(x_1, x_2, \cdots, x_N).$$

(3)
We consider our model output (3) in its functional expansion form as follows

\[ y = f_0 + \sum_{1 \leq i \leq N} f_i(x_i) + \sum_{1 \leq i < j \leq N} f_{ij}(x_i, x_j) + \cdots + f_{1, 2, \ldots, N}(x_1, x_2, \ldots, x_N), \]  

or in compact form

\[ y = f_0 + \sum_{I=1}^{N} \sum_{i_1 < \cdots < i_T} f_{i_1, \ldots, i_T}(x_{i_1}, \ldots, x_{i_T}). \]  

With independent inputs, the ANOVA (Analysis Of Variance) representation \([3–6]\) is a unique functional decomposition defined in such a way that any component function in (5) satisfies

\[ \int \prod_{i=1}^{T} f_{i_1, \ldots, i_T}(x_{i_1}, \ldots, x_{i_T}) \, dx_I = 0, \quad \forall x_I \in [x_{i_1}, \ldots, x_{i_T}]. \]  

It follows from (6) all the terms in ANOVA (5) are orthogonal. Thus, we also know the expectation of any component function (excluding \(f_0\) of course) vanishes: \(E(f_{i_1, \ldots, i_T}) = 0\). It is well known the terms in the ANOVA can be expressed as unconditonal and conditional expectation of \(f(x)\)

\[
E(Y) = f_0, \\
E(Y|x_i) = f_0 + f_i(x_i), \\
E(Y|x_i, x_j) = f_0 + f_i(x_i) + f_j(x_j) + f_{ij}(x_i, x_j),
\]

and so on.

The ANOVA expansion decomposes the original high-dimensional space into a union of several low-dimensional spaces that are orthogonal between each other. From computational viewpoint, working on low-dimensional spaces can be easier and potentially avoids the overhead computational cost of considering a single high-dimensional space.

2.1. Variance-based global sensitivity measures

Squaring the ANOVA representation (5) and integrating over \(\mathbb{R}^N\), the total variance \(V(Y)\) of output is found to be the summation of all partial variances

\[ V(Y) = \sum_{I=1}^{N} \sum_{i_1 < \cdots < i_T} \text{Var}(f_{i_1, \ldots, i_T}). \]

(8)

The global sensitivity indices (SI) \([3–6]\) are defined by ratios

\[ S_{i_1, \ldots, i_T} = \frac{\text{Var}(f_{i_1, \ldots, i_T})}{V(Y)}. \]

In particular, we note the first-order sensitivity indices are defined by

\[ S_i = \frac{\text{Var}_X(E(Y|X_i))}{V(Y)}. \]

From [8], all the sensitivity indices are non-negative and their sum equals unity.

Furthermore, the total effect of the variable \(X_i\) (i.e. total sensitivity index (TSI)) is estimated by

\[ S_i^T = 1 - \frac{\text{Var}_X(E(Y|X_i))}{V(Y)} = \frac{E_X(\text{Var}_X(Y|X_i))}{V(Y)} \]

which is indeed the sum of all sensitivity indices containing \(X_i\). Here \(X_{-i} = (X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_N)\). Note we have

\[ 0 \leq S_i^T \leq 1, \quad \sum_{i=1}^{N} S_i^T \geq 1. \]

The sensitivity indices presented above can be computed by Monte Carlo (MC) type sampling methods (see \([6, 41–43]\) and the references therein). The convergence of the MC type methods are known to be slow. In order to obtain a faster convergence, we employ a polynomial spectral expansion method to compute the global sensitivity indices as a post-processing of the expansion coefficients. For this purpose, let us introduce the Polynomial Dimensional Decomposition in next section.
3. Polynomial Dimensional Decomposition (PDD)

The previous section deals with the functional decomposition of a model response aiming to compute the moments and sensitivity indices. However, this approach does not provide any strategy to build a meta-model. For time-consuming numerical simulations, the meta-modeling is of importance to approximate and thus to accelerate the model evaluations. To seek an efficient way for this purpose, the polynomials can be used to represent the component functions in the ANOVA expansion. Roughly speaking, two techniques are widely used in the literature: polynomial chaos (PC) [29], and Polynomial Dimensional Decomposition (PDD) [31]. We prefer to employ the PDD representation in this work to take advantage of the close dimensional structure between the PDD and ANOVA. Note however the PC can also be used in a similar way, and no additional difficulty appears.

3.1. PDD representation

A set of orthogonal univariate polynomials is an infinite sequence of polynomials, \{\psi^j(x); j = 0, 1, \cdots\}, where \psi^j(x) has degree \( j \) and any two polynomials in the set are orthogonal to each other, i.e.

\[
\int_{\mathbb{R}} \psi^j(x) \psi^k(x) p_X(x) \, dx = \gamma_{j, X} \delta_{jk}.
\]

where \( \delta_{jk} = 1 \) if \( j = k \), otherwise \( \delta_{jk} = 0 \). The normalization constant \( \gamma_{j, X} \) can be determined as follows

\[
\gamma_{j, X} = \int_{\mathbb{R}} (\psi^j(x))^2 p_X(x) \, dx.
\]

The classical continuous orthogonal polynomials widely used in UQ context consist of for instance Hermite polynomials, Laguerre polynomials, Jacobi polynomials, and Legendre polynomials.

It is well known in the literature of generalized Polynomial Chaos (gPC) [29] that common distributions can be associated to specific families of polynomials [29]. For instance, a Gaussian distribution can be associated to Hermite polynomials, and a uniform distribution to Legendre polynomials.

For notational convenience in this section, we write a general \( T \)-dimensional component function \( (1 \leq T \leq N) \) of the ANOVA decomposition as

\[
f_{\mathbf{r}}(\mathbf{x}_T) = f_{1, j_2, \cdots, j_T}(x_{i_1}, x_{i_2}, \cdots, x_{i_T}).
\]

Due to the assumption of independence between member variables of the random vector \( \mathbf{X} \), it can be proven that

\[
\psi^{j_1}_{i_1} = \prod_{k=1}^{T} \psi^{j_k}(x_{i_k}) \quad \text{(14)}
\]

is a multivariate polynomial basis in the \( T \)-dimensional subspace. \( j_k \geq 1 \) is the degree of polynomial for the variable \( x_{i_k} \), and \( j_T = \{j_1, j_2, \cdots, j_T\} \).

Keeping in mind its zero mean property, the component function (13) can be expanded using an infinite number of orthogonal basis functions (as done in [31]) by a tensor product for multi-indices \( \mathbf{j}_T \):

\[
f_{\mathbf{r}}(\mathbf{x}_T) = \sum_{j_1, \cdots, j_T} C^{j_1,\cdots,j_T}_{\mathbf{i}_T} \psi^{j_1}_{i_1} \psi^{j_2}_{i_2} \cdots \psi^{j_T}_{i_T} \quad \text{(15)}
\]

Here \( C^{j_1,\cdots,j_T}_{\mathbf{i}_T} \) is a coefficient. As will be discussed later, these coefficients will be determined by the regression approach in this work.

Due to the fact that any two multivariate bases from different subspaces (same or different dimensionality) are always orthogonal, i.e.

\[
\mathbb{E} \left[ \psi^{j_1}_{i_1} \psi^{j_T}_{i_T} \right] = 0, \forall i_1 \neq i_T,
\]

the orthogonality of ANOVA component functions is guaranteed.

In practice, the expansion with an infinite number of terms in (15) must be truncated. As was done in [31], for the sake of simplicity we truncate (15) by \( m \) terms for each dimension, i.e.

\[
f_{\mathbf{r}}(\mathbf{x}_T) = \sum_{j_1, \cdots, j_T \leq m} C^{j_1,\cdots,j_T}_{\mathbf{i}_T} \psi^{j_1}_{i_1} \psi^{j_2}_{i_2} \cdots \psi^{j_T}_{i_T} \quad \text{(16)}
\]
In particular, the first-order, second-order, and third-order component functions can be expressed respectively by

\[ f_1(x_i) = \sum_{j=1}^{m} C_i^j \psi^j(x_i), \]

\[ f_{k_1}(x_{k_2}) = \sum_{j_2=1}^{m} \sum_{j_1=1}^{m} C_{k_2}^{j_1} \psi^{j_1}(x_{k_1}) \psi^{j_2}(x_{k_2}), \]

\[ f_{k_1}(x_{k_3}) = \sum_{j_3=1}^{m} \sum_{j_2=1}^{m} \sum_{j_1=1}^{m} C_{k_3}^{j_1} \psi^{j_1}(x_{k_1}) \psi^{j_2}(x_{k_2}) \psi^{j_3}(x_{k_3}). \]  

(17)

In conclusion, the polynomial dimensional decomposition of order \( m \) of the model output \( f(x) \) can be written as

\[ f(x) \approx f_m(x) = f_0 + \sum_{i=1}^{N} \sum_{j=1}^{m} C_i^j \psi^j(x_i) \]

\[ + \sum_{i_1<i_2}^{N} \sum_{j_2=1}^{m} \sum_{j_1=1}^{m} C_{i_2}^{j_1} \psi^{j_1}(x_{i_1}) \psi^{j_2}(x_{i_2}) \]

\[ + \sum_{i_1<i_2<i_3}^{N} \sum_{j_3=1}^{m} \sum_{j_2=1}^{m} \sum_{j_1=1}^{m} C_{i_3}^{j_1} \psi^{j_1}(x_{i_1}) \psi^{j_2}(x_{i_2}) \psi^{j_3}(x_{i_3}) \]

\[ + \cdots \]

\[ + \sum_{i_1<\cdots<i_N}^{N} \sum_{j_N=1}^{m} \cdots \sum_{j_1=1}^{m} C_{i_N}^{j_1} \prod_{k=1}^{N} \psi^{j_k}(x_{i_k}). \]  

(18)

Hence, the total size \( P \) of the \( m \)-th order full PDD expansion of an \( N \)-dimensional function is found to be the following

\[ P = 1 + Nm + \binom{N}{2} m^2 + \cdots + \binom{N}{m} m^m = (1 + m)^N. \]  

(19)

Note in the particular case of \( m = 1 \), the PDD representation has the same expansion size as ANOVA.

The following section is devoted to the discussion of the expansion coefficient computation.

3.2. PDD coefficient computation by regression

The computation of the coefficients involved in (18) can be obtained by projection [37,44]. Indeed, we have

\[ C_{l_r}^{j_r} = \frac{E[f(x) \psi_{l_r}^{j_r}]}{E[(\psi_{l_r}^{j_r})^2]}. \]  

(20)

As it is well-known, the formulation (20) can be expensive to evaluate in case of a large number of input variables, since high-dimensional integrations are required to be computed. Typically, one can employ random sampling type methods (e.g. Monte Carlo (MC), Latin Hypercube) which are costly in general, or Gauss quadrature methods which are prohibitive in the high-dimensional case even when using a sparse grid.

As done in [40,13], we use instead the regression approach in this work to determine the expansion coefficients, that can also afford an attractive flexibility in choosing the sampling points.

The regression approach in this work can be regarded as a response surface aiming to provide an optimized PDD expansion to the considered model problem.

Let us write the finite PDD expansion (18) in a vector form as follows

\[ f_m(x) = \mathbf{C}_\alpha \Phi_\alpha(x) \]  

(21)

where

\[ \mathbf{C}_\alpha = (C_{\alpha_0}, \cdots, C_{\alpha_{P-1}})^T \]

is a vector containing all the coefficients, and

\[ \Phi_\alpha(x) = (\Phi_{\alpha_0}, \cdots, \Phi_{\alpha_{P-1}})^T \]
gathering all the multivariate basis polynomials including the unity basis \( \Phi_{\alpha_0} = 1 \). For the sake of notation clarity, we recall \( C_{\alpha}^T \in \{ C_{\alpha} \} \) and \( \Psi_{\alpha}^T \in \{ \Phi_{\alpha} \} \). Here \( P = | C_{\alpha} | = | \Phi_{\alpha} | \) is the total size of PDD expansion, given by (19).

When using a regression method to determine the expansion coefficients, it is necessary to choose a set of realizations of input random vector (i.e. an experimental design \([40,13]\)), for instance by a Sobol’ quasi-random sequence \([45]\), denoted by

\[
\mathbf{x} = (\mathbf{x}^1, \mathbf{x}^2, \cdots, \mathbf{x}^Q)^T,
\]

the size of which has to be larger than the PDD expansion size. As indicated in \([46]\), in order for the regression problem to be stable,\(^1\) the experimental design size \(Q\), in practice, is usually taken as

\[
Q = kP \text{ with } 2 \leq k \leq 3.
\] (22)

We denote the corresponding model outputs by

\[
\mathbf{y} = (y^1, y^2, \cdots, y^Q)^T.
\]

The idea is to determine the coefficients \( C_{\alpha} \) by minimizing the projection error in \( L_2 \) norm. That is,

\[
C_{\alpha} = \arg \min_{C_{\alpha} \in \mathbb{R}^Q} \sum_{i=1}^{Q} \left( y^i - \tilde{C}_{\alpha}^T \Phi_{\alpha} (\mathbf{x}^i) \right)^2.
\] (23)

The solution of the least-squares regression (LSR) problem (23) can be easily found by using a variational approach

\[
\frac{\partial}{\partial \tilde{C}_{\alpha}^j} \left[ \sum_{i=1}^{Q} \left( y^i - \tilde{C}_{\alpha}^T \Phi_{\alpha} (\mathbf{x}^i) \right)^2 \right] = 0, \text{ for } j = 0, \cdots, P - 1.
\]

and can be obtained by solving the following linear system

\[
\langle \mathcal{M}^T \mathcal{M} \rangle C_{\alpha} = \mathcal{M}^T \mathbf{y}.
\] (24)

\( \mathcal{M}_{Q \times P} \) represents the following matrix involving basis polynomials evaluated at the realizations in the experimental design:

\[
\mathcal{M}_{ij} = \Phi_{\alpha_j} (\mathbf{x}^i), \quad i = 1, \cdots, Q, \quad j = 0, \cdots, P - 1
\] (25)

It is well known the real matrix \( \langle \mathcal{M}^T \mathcal{M} \rangle_{P \times P} \) is symmetric and positive-definite. If this matrix is well-conditioned, the linear system (24) can be in general resolved efficiently by Cholesky factorization (Normal Methods Equation). Despite of the efficiency, Cholesky factorization is known to be sensitive to rounding errors; so when the LSR matrix is ill-conditioned, more sampling points will need to be added. To this purpose, a nested QMC sampling technique is applied in this work to generate the multi-dimensional sampling points of experimental design, e.g. the Sobol’ quasi-random sequence.

We emphasize the design size \(Q\) is directly linked to the global computational cost of uncertainty quantification for numerical simulations. As a consequence from (22) for the choice of experimental design size, the main objective of this work is to minimize the PDD expansion size \(P\), by including only the most influential polynomials terms.

### 3.3. PDD based global sensitivity indices

Once the coefficients \( C_{\alpha} \) are determined by the regression approach for the PDD expansion (21), the second-order moment and the global sensitivity indices can be obtained in a straightforward way by post-processing.

Indeed, keeping in mind

\[
E(f_m(\mathbf{X})) = C_{\alpha_0},
\]

the approximated variance of the model output is then

\[
\text{Var}(f_m(\mathbf{X})) = \sum_{j=1}^{P-1} C_{\alpha_j}^2 \gamma_{\alpha_j},
\]

with the multivariate normalization constant determined by

\[
\gamma_{\alpha_j} = E \left[ \Phi_{\alpha_j}^2 (\mathbf{X}) \right].
\]

\(^1\) As will be explained later in the same section, in order for the Cholesky factorization to be numerically stable, the matrix of a least-squares problem needs to be well-conditioned.
If one employs normalized basis polynomials, i.e. \( y_{ij} = 1 \), the output variance formulation can be further simplified to

\[
\text{Var}(f_m(\mathbf{X})) = \sum_{j=1}^{p-1} c_{aj}^2.
\]

It is straightforward to write the variance-based global sensitivity indices by using the PDD expansion. Indeed,

\[
S_{i_1, \ldots, i_T} = \frac{\text{Var}(f_{i_1, \ldots, i_T})}{\text{Var}(f_m(\mathbf{X}))} = \frac{1}{\text{Var}(f_m(\mathbf{X}))} \sum_{\alpha_j \leq (i_1, \ldots, i_T)} c_{aj}^2 y_{aj}
\]

The total sensitivity index \( S_i^T \) can be obtained simply by adding all the measures \( S_{i_1, \ldots, i_T} \) whose index involves the variable \( X_i \).

4. Variance-based dimension reduction for the model representation

For practical problems, in particular for the ones with a large number of stochastic parameters, the size of the PDD representation given in Section 3 must be reduced to make the uncertainty analysis feasible.

[40,13] proposed a gPC approach coupled with a stepwise regression method to build sparse polynomial representation of the model output. In order to efficiently deal with a moderate to large dimensional parameter space, in addition to the stepwise regression technique proposed in [40,13], the PDD approach used in this work is further coupled with an adaptive ANOVA decomposition, which allows to model a given number of low-dimensional spaces instead of a single high-dimensional one.

The purpose of this section is to present our overall adaptive technique, which can be viewed as a variant of the one used in [40,13].

More in detail, apart from applying a truncated dimensionality widely used for the ANOVA component functions, we consider additionally two levels of dimension reduction in this section, both of which relies on the relative importance of expansion functions in terms of their variance contribution compared to the total variance. We emphasize here that this variance-based criterion is different from [40,13] where the adaptive strategy is based on the modeling error of the polynomial representation. We think a variance-based criterion can be more reliable if one cares more about the statistical properties (particularly second-order moment) of the model output.

The rest of this section is devoted to the description of the three levels of adaptivity proposed in this paper, namely the truncation dimension \( \nu \) (i.e. the largest input interaction order), the active dimension \( DT \) of ANOVA, and the stepwise regression to retrieve the most significant polynomials using a variance-based selection criterion.

4.1. Adaptive ANOVA – retaining active dimensions

First of all, we use some of the adaptive strategies presented in [15] in order to retain only the most important dimensions for second- and higher-order interaction terms in ANOVA.

**Truncation dimension \( \nu \).** The low order interactions of input variables often have the main impact upon the output [6]. Thus, the full ANOVA expansion (5) can be approximated by

\[
f(\mathbf{x}) = f_0 + \sum_{T=1}^{\nu} \sum_{i_1 < \cdots < i_T} f_{i_1, i_2, \ldots, i_T}(x_{i_1}, x_{i_2}, \cdots, x_{i_T}), \text{ with } \nu \ll N.
\]

Here, \( \nu \) is called the truncation (or effective) dimension representing the highest dimension of the ANOVA component functions.

**Active dimension \( DT \).** For problems featuring a high dimensionality \( N \), ANOVA decomposition method is still very expensive even when we only choose a truncation dimension \( \nu = 2 \). An efficient way to solve this problem is to use the adaptive ANOVA decomposition. To this purpose, we replace the approximation (27) by

\[
f(\mathbf{x}) = f_0 + \sum_{T=1}^{\nu} \sum_{i_1 < \cdots < i_T} f_{i_1, i_2, \ldots, i_T}(x_{i_1}, x_{i_2}, \cdots, x_{i_T}), \text{ with } \nu \ll N, \ D_T \leq N.
\]

Here \( DT \) is the active dimension for ANOVA component functions of order \( T \). For problems considered in this work, we take \( D_1 = N \). The active dimension for higher order component functions will be determined by the criterion presented below.

In this work we use a variance-based criterion for choosing the active dimension \( D_2 \) and further selecting the most important second- and higher-order terms. It is assumed \( \text{Var}(f_i) (i \in [1, N]) \) of first-order terms are monotonically decreasing with respect to \( i \). \( D_2 \) is evaluated using the sum of variances of first-order terms:

\[2\text{ This criterion (29) is the same as the Criterion 1 already presented in [15] in choosing the active dimension \( D_2 \). However, the way we further add significant second- and higher-order components and remove those of less importance is different from [15].]
\[
\sum_{i=1}^{D_2} \text{Var}(f_i) \geq p \sum_{i=1}^{N} \text{Var}(f_i),
\]

where \( p \) is a proportionality constant in \((0, 1)\), and is very close to 1.

For simplicity in our applications, we set

\[ D_T = D_2, \text{ for } T \geq 3 \]

in this work. Note however \( D_T \) can certainly be further reduced depending on problems and objectives.

We point out that the truncation dimension \( \nu \) and active dimension \( D_T \) used in this work are different from the definitions of the “effective dimension” (or “dimension distribution”) appeared in \([47, 48]\). We recall their effective dimension, in the superposition sense, is the smallest integer \( d_S \) such that \( \sum_{0<|u|\leq d_S} \text{Var}(f_u) \geq 0.99 \text{Var}(f) \). Here \( u \) is multi-index of a component function, and 0.99 is an arbitrary choice of threshold in their paper. In other words, component functions in ANOVA with their interaction order less or equal to \( d_S \) are retained, while their variance contributions represent at least 99% of the total output variance. We remind the truncation dimension \( \nu \) set in this paper results from the assumption of low order input interactions have more important effects than high order ones. However, the value of \( \nu \) used in this work does not guarantee any quantitative weight of the retained component variance contributions over the total variance. On the other hand, their effective dimension, in the truncation sense, is the smallest integer \( d_T \) such that \( \sum_{u \leq 1, \ldots, d_T} \text{Var}(f_u) \geq 0.99 \text{Var}(f) \). That is, the dimensionality \( N \) of the original model is reduced to \( d_T \), while the reduced model output variance is at least 99% of the original output variance. If we impose the active dimension employed in this work to be constant for all interaction orders, this dimension \( D_T \) shares the same idea with the effective dimension \( d_T \) in the truncation sense in \([47]\). In the framework of this work, since we do not know a priori the total output variance, it is not straightforward to employ the effective dimension introduced in \([47, 48]\). We remind also, in Equation (29), the active dimension for second-order interaction is determined over the total variances of all first-order component functions.

### 4.2. Adaptive PDD algorithm – eliminating non-important polynomials

In section 4.1, we have talked about how to reduce the size of ANOVA expansion. However, even with a sparse ANOVA expansion, if applying the classical PDD expansion to the component function (see the formulation (16) and (17)), the required computational cost still remains very high. Indeed, it reveals that, for a large number of engineering problems, the contribution of many polynomial terms is negligible when regarding the accuracy of the constructed meta-model \([40, 13]\). In this work on the other hand, as will be shown in our numerical results, if we eliminate those polynomials whose variance is negligible, we can also build a very sparse PDD representation without compromising the accuracy of the meta-model.

Considering the adaptive ANOVA technique presented in Section 4.1, let us describe our adaptive algorithm for stepwise regression from a practical point of view as follows.

1. **Determination of active dimension.** First of all, we construct a full set of PDD representation (given \( m \)) for all first-order ANOVA component functions, namely

\[
f(x) \approx f_m(x) = f_0 + \sum_{i=1}^{N} f_i(x_i) = f_0 + \sum_{i=1}^{N} \left( \sum_{j=1}^{m} C^i_j \psi_j(x_i) \right).
\]

We then compute the total first-order variance by

\[
\text{Var}(f_m(X)) = \sum_{i=1}^{N} \sum_{j=1}^{m} (C^i_j)^2 \gamma^i_j.
\]

The first-order global sensitivity indices can be obtained by a simple rearrangement:

\[
S_i = \sum_{j=1}^{m} (C^i_j)^2 \gamma^i_j/\text{Var}(f_m(X)).
\]

Let us assume that the sensitivity indices \( S_i \) are monotonically decreasing with respect to \( i \) (thus, a re-ordering task is generally required), so we choose the active dimension \( D_2 \) for second-order ANOVA functions in such a way that

\[
\sum_{i=1}^{D_2} S_i \geq p,
\]

with \( p \) a constant close to the unity (e.g. \( p = 0.999 \)). Note for the sake of simplicity in this work, we emphasize again the same active dimension is employed for second- and also higher-order component functions if applicable. However, one can further reduce this dimension if necessary.
2. Building a concise (multivariate) set of univariate polynomials of first-order ANOVA components \( \{ \Phi_{a1} \} \). The objective of this step is to reduce the size of the first-order PDD expansion as expressed in (30). The principle remains similar as in the previous step: we eliminate those non-important polynomial terms by measuring their variance contribution. The selection criterion is further explained in next step. The resultant first-order model representation contains only significant components, and thus is more concise. Let us denote this first-order PDD bases by \( \{ \Phi_{a1} \} \).

3. Enriching the existing set \( \{ \Phi_{a1} \} \) by selecting efficiently from multivariate polynomials of higher order ANOVA components \( \{ \Phi_{a2+} \} \). Starting from the concise first-order PDD representation, the task of this step is to enrich the model representation by significant second and higher order PDD polynomials. After choosing a truncation dimension \( \nu \) and an active dimension for each order \( (D_2, \ldots, D_\nu) \), the full set of second and higher order PDD polynomial bases \( \{ \Phi_{a2+} \} \) can be constructed easily from the tensor product rule (see the formulation (17) or (18)).

Algorithm 1 Adaptive ANOVA and PDD featuring stepwise regression with Criterion 1.

1. Determine active dimension (variance-based)
2. Initialize concise first-order PDD bases (variance-based): \( \{ \Phi^w \} \leftarrow \{ \Phi_{a1} \} \)
3. Construct high order bases \( \{ \Phi_{a2+} \} \) using truncation dimension \( \nu \) and active dimensions
4. for all \( \Phi_{a_i} \in \{ \Phi_{a2+} \} \) do
5. Add \( \Phi_{a_i} \) into \( \{ \Phi^w \} \), i.e. \( \{ \Phi^w \} \leftarrow \{ \Phi^w, \Phi_{a_i} \} \)
6. Depending on size \( P^w \) of \( \{ \Phi^w \} \), adjust, if necessary, the size \( Q^w \) of experimental design (see formulation (22))
7. Solve the regression system (24) to determine PDD expansion coefficients \( C^w \)
8. Evaluate the model accuracy by a leave-one-out cross validation estimator \( Q^2_i \)
9. if \( Q^2_i \geq Q^2_{tgt} \) then Exit
10. end if
11. Compute total variance: \( \text{Var}(f^w(X)) \leftarrow \sum_k(C_{a_k})^2 \gamma_{a_k} \)
12. for all \( \Phi_{a_j} \in \{ \Phi^w \} \) do
13. if \( \frac{(C_{a_j})^2 \gamma_{a_j}}{\text{Var}(f^w(X))} < \theta \) then
14. Eliminate this polynomial: \( \{ \Phi^w \} \leftarrow \{ \Phi^w \} \setminus \{ \Phi_{a_j} \} \)
15. end if
16. end for
17. end for
18. Solve final regression system based on the constructed bases \( \{ \Phi^f \} \) – an accurate surrogate model is thus built
19. Compute final total and partial variances by post-processing of surrogate coefficients
20. Compute global (total) sensitivity indices
21. Surrogate model evaluations using MC/QMC importance sampling from input distributions can provide reliable output pdf

The selection process of important polynomials from \( \{ \Phi_{a2+} \} \) can be explained by the simple Algorithm 1. The essential idea is that we add to our existing polynomial set by adding recursively one new polynomial candidate from the set of higher-order ANOVA functions. During this recursive procedure, we eliminate those polynomial terms that are quantified as “unimportant”. In order to measure this relative importance between multivariate polynomials that form the current surrogate model, we propose the following variance-based Criterion 1.

**Criterion 1.** Supposing \( \{ \Phi \} \) is the basis multivariate polynomials currently built, and \( \{ C \} \) the corresponding coefficients, the current total variance of our surrogate model is thus given by manipulating \( \{ C \} \) and the normalization constants \( \{ \gamma \} \): \( \text{Var}(f^w(X)) = \sum_k(C_{a_k})^2 \gamma_{a_k} \). We eliminate any multivariate polynomial \( \Phi_{a_j} \in \{ \Phi \} \) that satisfies the following condition

\[
\frac{(C_{a_j})^2 \gamma_{a_j}}{\text{Var}(f^w(X))} < \theta.
\]

(31)

Here \( \theta \) is a positive predefined threshold (e.g. \( \theta = 10^{-5} \)).

**Criterion 1** is applied as Operation 13–14 in Algorithm 1 as a critical step whose objective is to eliminate non-important polynomial terms. This is achieved by discarding any component polynomial whose variance is negligible compared to the total variance.

Once all possible polynomials have been examined by either being included or neglected, the computation of surrogate model output statistics as well as the global sensitivity indices is a simple post-processing task of expansion coefficients. Monte Carlo type importance sampling from the surrogate model can provide reliable probability density function (pdf) and cumulative distribution function (cdf).
Let us mention in this algorithm the cost of the recursive resolution of the regression linear system is negligible compared to the one of realistic deterministic model evaluations.

If denoting the size of the finally obtained sparse PDD representation by $P_{\text{sparse}}$, we define the sparsity of our adaptive PDD approach as follows

$$\text{Sparsity} = \frac{P_{\text{sparse}}}{(1 + m)^N}.$$  \hfill (32)

Note that the formulation (32) will be employed in our numerical applications for assessing the efficiency of the proposed approach.

4. Evaluation of surrogate model accuracy by a leave-one-out cross validation estimator $Q^2$. Several polynomial chaos error estimators are presented in [40,13]. These estimators can be directly used for the sparse PDD expansions in this work. We present the leave-one-out cross validation estimator $Q^2$ used in this work in Appendix A. Note that $Q^2 = 1$ indicates a perfect fit between surrogate and original model, while $Q^2 \approx 0$ or $Q^2 < 0$ reflects a poor model accuracy. Note the lines 8 to 10 of Algorithm 1 is an optional procedure. $Q_{\text{tgt}}^2$ is a predefined target accuracy (e.g. 0.999). Once the algorithm reaches this target accuracy, we can choose to stop and use the constructed bases.

Operation 13–14 in Algorithm 1 can be replaced by a second criterion (already used in [13]) that relies on the estimator $Q^2$. We recall this criterion as follows

**Criterion 2.** Supposing $\Phi$ is the basis multivariate polynomials currently built, and $Q^2$ the corresponding estimator of accuracy. We eliminate any multivariate polynomial $\Phi_{\alpha_j} \in \Phi$ that satisfies the following condition

$$Q^2 - Q_{\alpha_j}^2 < \varepsilon Q^2.$$  \hfill (33)

Here $Q_{\alpha_j}^2$ is the accuracy estimator obtained by solving a regression system excluding $\Phi_{\alpha_j}$ in the current bases, and $\varepsilon Q^2$ a positive predefined threshold (e.g. $\varepsilon Q^2 = 10^{-5}$).

Briefly speaking, one eliminates those polynomials whose contribution to the model accuracy can be negligible. The technical details of the estimator of accuracy and how to implement this criterion can be found in Appendix A (see Algorithm 2).

**Remark.** We point out a disadvantage of Criterion 2, compared to Criterion 1, consists in the fact that we need to solve a considerably larger number of regression systems. For instance, in order to select significant member polynomials from the set $\Phi$ whose cardinality denoted by $|\Phi|$, we need to solve $|\Phi| + 1$ least-squares systems, while the Criterion 1 requires only once. Although the cost of resolving a linear system is negligible compared to the one of a deterministic fluid/multi-physics simulation, this disadvantage of Criterion 2 is “visible” when we treat high-dimensional problems that usually involve large-scale regression systems.

4.3. A synopsis of arbitrary methodology parameters

The overall methodology presented above involves a couple of arbitrary parameters necessary for the proposed procedure, that we list explicitly as follows.

**Parameters related to the PDD expansion and the least-squares regression**

- **PDD polynomial degree $m$.** This parameter is linked to the order of the PDD expansion and characterizes the size of model representation prior to the adaptivity. On the one hand, when it is given too small, it is hard to obtain accurate results. On the other hand, the well-known overfitting phenomenon [49] would appear if $m$ is too big letting the polynomial expansion size tend to the size of experimental design $Q$. In our examples, we select $m$ from 2 up to 11.

- **Experimental design size $Q$.** This parameter must satisfy that the least-squares matrix is well-conditioned, so it should never be smaller than the polynomial expansion size. In this respect, when enriching the polynomial bases during the adaptivity procedure, one may need to increase $Q$ in order to insure this condition, while of course insuring the absence of overfitting phenomenon. However, it is obvious this parameter represents the number of deterministic simulation runs, so it should be kept as small as possible, and meanwhile must insure an acceptable accuracy.

**Parameters related to the adaptivity procedure**

- **Truncation dimension $v$ in ANOVA.** This parameter accounts for the maximum interaction order of inputs that is considered in ANOVA computation. Under the assumption of low-order interactions are dominant in many engineering applications, $v$ is usually taken between 2 and 4 in our examples.

- **$p$ constant (which determines active dimension $D_T$).** This parameter close to unity is used to select active dimension for higher order ANOVA terms. We usually choose a $p$ inside the range of [0.9, 1]. This mainly has objective to reduce effectively the computational cost. In order to make it possible to solve very high dimensional problems, we expect a
low active dimension. We emphasize this is an assumption, and is not guaranteed. The user is supposed to verify the number of retained active dimensions, and adjust p if necessary. For low-dimensional problems, p can be set to unity, and we take all dimensions into account.

- **Thresholds $\theta$ (Criterion 1) & $\epsilon_{q_2}$ (Criterion 2).** These two parameters are used to reduce the size of the model representation during the adaptivity procedure. A sensitivity study on these parameters will be carried out for numerical examples. In our examples, we usually choose them between $10^{-2}$ and $10^{-6}$.

- **Target accuracy $Q^2_{tg}$.** This is an optional target accuracy that can be set in order to terminate the algorithm. When the entire adaptive algorithm is efficient, this parameter is not necessary.

We point out these arbitrary parameters are case-dependent. A systematic study of the parameter’s sensitivity will be performed in the section devoted to numerical examples.

4.4. CPU cost of the methodology

The CPU cost of the proposed technique depends, on one hand “slightly” upon the methodology parameters listed above, and on the other hand “strongly” upon the stochastic problem under consideration (dimensionality $N$ of inputs, number of active dimensions, etc). The successive resolution of least-squares problems is the main cost of the methodology, apart from the global cost of the whole study, i.e. the cost of one deterministic run multiplied by the number of samples. As already stated earlier, Criterion 1 has a considerable advantage compared to Criterion 2 in this regard, since a much fewer number of linear systems are required to be resolved. In the section devoted to numerical examples, we will provide a picture of required CPU time for certain problems. In general it only requires seconds to minutes for problems featuring a moderate number of uncertain inputs. When compared to the sparse Polynomial Chaos Expansion (PC) combined with ANOVA [13], our approach can be advantageous for high-dimensional problems if a relatively low active dimensionality can be retained. We explain this statement in a quantitative way as follows:

- The conventional PC expansion (1) truncated at degree $p$ for a $N$-dimensional problem has its size $P_{PC}$ equaling

$$P_{PC} = \binom{N + p}{p}.$$  

(34)

- The PDD expansion featuring ANOVA’s active dimension truncated at degree $m$ has its size $P_{PDD}$:

$$P_{PDD} = 1 + Nm + \left(\frac{D_2}{2}\right)m^2 + \left(\frac{D_3}{3}\right)m^3 + \cdots$$

(35)

For example if we set $N = 10$ and $p = 6$, the PC size is $P_{PC} = 8008$. Regardless of the target accuracy, the technique proposed in [13] should select important polynomials from a set that contains 8008 members. Even with Criterion 1, the forward step would require repeatedly to solve linear systems about eight thousand times. In case of $N = 50$, $p = 6$, we have a huge number of $P_{PC} = 32,468,436$. It becomes challenging if we do not adopt any further adaptive strategy.

While the proposed approach is considered with a truncation dimension $v = 3$ and $m = 6$, we have $P_{PDD} = 27,601$ in case of $N = 10$, and $D_2 = D_3 = 10$. This is less advantageous than using the PC expansion, since the multivariate polynomials have their degree up to 18. However, if we can manage to choose a lower active dimension for second- and third-order ANOVA functions, the PDD size would be much smaller. For instance, if active dimensions are retained as $D_2 = D_3 = 5$, $P_{PDD} = 2581$. This is more than three times smaller than the PC expansion. This difference is even more significant when considering the 50-dimensional case. When the “full” PDD expansion is truncated at $v = 3$ and active dimensions are all taken as 50, we will have $P_{PDD} = 4,278,001$. Although this is about 7 times smaller than in the PC case, it is still a huge number to work with. However, if active dimensions can be retained to $D_2 = D_3 = 10$, we have $P_{PDD} = 27,841$. If further reducing to $v = 2$, we will only have $P_{PDD} = 1921$ polynomials to choose from. This is much easier than for the PC case. Moreover, we can potentially explore higher order multivariate polynomials than in the PC case.

We emphasize again the above comparison can only make sense for high-dimensional problems under two assumptions:

1. The variance contribution of low-order input interactions is significantly more important than the one of high-order interactions.
2. All model inputs do not have equal importance, and active dimensionality is significantly lower than original problem’s dimensionality $N$.

In next section, we will demonstrate the method’s capability of solving problems of dimensionality up to 100.

5. Numerical results

This section is devoted to the presentation of our numerical results. Three academic functions will be studied, and we will also investigate one CFD application example.
Table 1
Analytical variance and variance-based sensitivity indices for Ishigami function.

<table>
<thead>
<tr>
<th>Variable</th>
<th>( S_0 )</th>
<th>( S_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>0.3138</td>
<td>0.5574</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.4424</td>
<td>0.4424</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0</td>
<td>0.2436</td>
</tr>
<tr>
<td>( X_1, X_2 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( X_1, X_3 )</td>
<td>0.2436</td>
<td></td>
</tr>
<tr>
<td>( X_2, X_3 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( X_1, X_2, X_3 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( V(Y) )</td>
<td>13.845</td>
<td></td>
</tr>
</tbody>
</table>

Table 2
Numerical results for the Ishigami test case. Criterion 1, \( \theta = 10^{-5} \).

<table>
<thead>
<tr>
<th>Variable</th>
<th>( S_0 )</th>
<th>( S_1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X_1 )</td>
<td>0.3139</td>
<td>0.5573</td>
</tr>
<tr>
<td>( X_2 )</td>
<td>0.4427</td>
<td>0.4427</td>
</tr>
<tr>
<td>( X_3 )</td>
<td>0</td>
<td>0.2434</td>
</tr>
<tr>
<td>( X_1, X_2 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( X_1, X_3 )</td>
<td>0.2434</td>
<td></td>
</tr>
<tr>
<td>( X_2, X_3 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( X_1, X_2, X_3 )</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>( V(Y) )</td>
<td>13.8304</td>
<td></td>
</tr>
</tbody>
</table>

| Model evaluations | 200 |
| Sparse polynomial bases | 15 |
| PDD order \( m \) | 10 |
| Sparsity | 15/1331 \approx 0.01127 |
| Model accuracy \( Q^2 \) (Appendix A) | 0.999965 |

5.1. 3-dimensional Ishigami function

Let us consider the Ishigami function [50] which has been thoroughly studied in our previous work [28] by making use of the Covariance-based Sensitivity Analysis:

\[
Y = \sin X_1 + a \sin^2 X_2 + bX_3^4 \sin X_1, 
\]

where the input random variables \( X = (X_1, X_2, X_3) \) are uniformly distributed over \([-\pi, \pi]\). The constants are set to \( a = 7 \), \( b = 0.1 \), as done in [11,51].

As presented in [50,11,28], the total output variance and the component variances based on standard ANOVA expansion can be obtained analytically:

\[
\begin{align*}
V(Y) &= \frac{a^2}{8} + \frac{b\pi^4}{5} + \frac{b^2\pi^8}{18} + \frac{1}{2}, \\
V_1 &= \frac{b\pi^4}{5} + \frac{b^2\pi^8}{50} + \frac{1}{2}, \quad V_2 = \frac{a^2}{8}, \quad V_3 = 0, \\
V_{12} &= V_{23} = 0, \quad V_{13} = \frac{8b^2\pi^8}{225}, \\
V_{123} &= 0.
\end{align*}
\]

Thus, the variance-based sensitivity indices can be gathered in Table 1.

We set the truncation dimension \( v = 3 \); i.e. all interactions are considered. We use the quasi-random sampling design based on Sobol’ sequences for solving the regression system. For the sake of comparison with [13], let us choose the experimental design size \( Q = 200 \), and set the PDD order \( m = 10 \). The active dimension selection technique in Section 4.1 is not adopted for this relatively low dimensional test case. The numerical results with the proposed approach are reported in Table 2.

Comparing our results with the ones reported in [13, Table 2, last column, p. 1223], sensitivity indices are obtained with a similar accuracy. With our approach, only 15 polynomial terms are necessary to obtain the model accuracy of 0.999965, while, in [13], 77 terms are needed to have an accuracy of 0.9999. We write the constructed surrogate polynomial model as follows to approximate the Ishigami function:
\[ Y = 3.50078 + 2.81152\tilde{L}_1(X_1) - 3.41737\tilde{L}_3(X_1) + 0.649461\tilde{L}_5(X_1) \]
\[ - 1.33092\tilde{L}_2(X_2) - 5.85130\tilde{L}_4(X_2) + 4.90081\tilde{L}_6(X_2) \]
\[ - 1.39774\tilde{L}_8(X_2) + 0.212954\tilde{L}_10(X_2) \]
\[ + 5.29778\tilde{L}_1(X_1)\tilde{L}_2(X_2) + 2.09215\tilde{L}_1(X_1)\tilde{L}_4(X_3) \]
\[ - 6.46581\tilde{L}_3(X_1)\tilde{L}_2(X_3) - 2.60564\tilde{L}_3(X_1)\tilde{L}_5(X_3) \]
\[ + 1.23633\tilde{L}_5(X_1)\tilde{L}_2(X_3) + 0.495326\tilde{L}_5(X_1)\tilde{L}_4(X_3). \] (38)

where $\tilde{L}_j(x_i)$ represents the $j$-th order shifted Legendre polynomial for the variable $x_i$ with respect to the weight function $w(x_i) = 1/(2\pi)$ for $x_i \in [-\pi, \pi]$.

Because the underlying function in (36) is even with respect to the variable $X_2$ and $X_3$, the odd polynomials related to these variables are found to be zero in (38) as expected. For the same reason, the even polynomials linked to the variable $X_1$ have zero coefficients.

As it is known, by the analytical analysis using the standard ANOVA (see e.g. [28]), the maximum interaction order for Ishigami function is 2, as is also found in the surrogate model (38). Note thus if setting the truncation dimension $v = 2$ before constructing the surrogate model, one finds the same results as in (38).

**Sensitivity to the PDD order $m$ with the variance-based Criterion 1**

In this section, we study the sensitivity of the proposed approach with respect to the univariate polynomial order $m$, by reporting the sparsity of the surrogate representation and showing the convergence of the method in terms of model accuracy.

Note we fix the size of our experimental design to $Q = 200$. The dimension reduction technique presented in Section 4.1 is not adopted ($v = 3$ and $D_{1,2,3} = 3$); only the adaptive PDD algorithm in Section 4.2 is employed here. Note the variance selection threshold is set to $\theta = 10^{-5}$ for Criterion 1. The polynomial order $m$ varies from 5 to 11. We report the results in Table 3.

From Table 3, we notice, for this specific case, the adaptive PDD approach with the even number of $m$ provides better results in general than with the odd $m$. The convergence tendency can be observed when we increment $m$ by 2 (even or odd $m$).

Let us fix $m = 10$ and still using $Q = 200$ in order to test the method’s sensitivity to the threshold $\theta$. Results are presented in Table 4 by decreasing $\theta$. We observe all sensitivity indices tend to be more accurate with a lower threshold $\theta$ until the value of $10^{-4}$, by means of including significant polynomial terms up to size of 15. The same results have been obtained when setting $\theta$ to $10^{-4}$ and $10^{-5}$, since the finally built polynomial bases are identical. Further decreasing $\theta$ to $10^{-6}$ results in a polynomial representation with a much heavier size of 162. This makes the least-squares matrix ill-conditioned and sensitivity indices start to diverge. In order to avoid this problem, we need to add more sampling points. However, this is unnecessary for Ishigami function, since with a sparse representation containing 15 polynomials, we already obtain very accurate results. Also, using more than 200 model evaluations for a 3-dimensional case in real life is not the purpose of the proposed approach. Thus, we do not further analyze the method’s sensitivity to $\theta$ using more sampling points here.

**Sensitivity to the target accuracy $Q_{tgt}^2$ with the accuracy-based Criterion 2**

In order to evaluate the model accuracy by using the Criterion 2, we vary the target accuracy $Q_{tgt}^2$ to compute the sensitivity indices. The results are reported in Table 5.
Table 4
Ishigami test. Numerical results using Criterion 1 by varying the threshold $\theta$. The experimental design size $Q = 200$. $m = 10$. The model accuracy is estimated by the cross validation method.

<table>
<thead>
<tr>
<th>SI</th>
<th>Exact</th>
<th>$\theta = 10^{-1}$</th>
<th>$\theta = 10^{-2}$</th>
<th>$\theta = 10^{-3}$</th>
<th>$\theta = 10^{-4}$</th>
<th>$\theta = 10^{-5}$</th>
<th>$\theta = 10^{-6}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.3138</td>
<td>0.3593</td>
<td>0.3153</td>
<td>0.3131</td>
<td>0.3139</td>
<td>0.3139</td>
<td>0.3021</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.4424</td>
<td>0.4731</td>
<td>0.4265</td>
<td>0.4376</td>
<td>0.4427</td>
<td>0.4427</td>
<td>0.4227</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.5693E-4</td>
</tr>
<tr>
<td>$S_{12}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.1541E-2</td>
</tr>
<tr>
<td>$S_{13}$</td>
<td>0.2436</td>
<td>0.1675</td>
<td>0.2582</td>
<td>0.2492</td>
<td>0.2434</td>
<td>0.2434</td>
<td>0.2374</td>
</tr>
<tr>
<td>$S_{23}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.2928E-1</td>
</tr>
<tr>
<td>$S_{123}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0.6889E-3</td>
</tr>
<tr>
<td>$S_1^2$</td>
<td>0.5574</td>
<td>0.5269</td>
<td>0.5735</td>
<td>0.5624</td>
<td>0.5573</td>
<td>0.5573</td>
<td>0.5479</td>
</tr>
<tr>
<td>$S_2^2$</td>
<td>0.4424</td>
<td>0.4731</td>
<td>0.4265</td>
<td>0.4376</td>
<td>0.4427</td>
<td>0.4427</td>
<td>0.4604</td>
</tr>
<tr>
<td>$S_3^2$</td>
<td>0.2436</td>
<td>0.1675</td>
<td>0.2582</td>
<td>0.2492</td>
<td>0.2434</td>
<td>0.2434</td>
<td>0.2736</td>
</tr>
<tr>
<td>$P_{qane}$</td>
<td></td>
<td>6</td>
<td>9</td>
<td>12</td>
<td>15</td>
<td>15</td>
<td>162</td>
</tr>
<tr>
<td>Accuracy $Q^2$</td>
<td>0.8551</td>
<td>0.9812</td>
<td>0.9963</td>
<td>0.999965</td>
<td>0.999965</td>
<td>0.99993</td>
<td></td>
</tr>
</tbody>
</table>

Table 5
Ishigami test. Numerical results using Criterion 2 by varying the target accuracy $Q_{tgt}^2$. The model accuracy is estimated by the cross validation method. $\epsilon_{Q^2} = 10^{-7}$.

<table>
<thead>
<tr>
<th>SI</th>
<th>Exact</th>
<th>$Q_{tgt}^2 = 0.9$</th>
<th>$Q_{tgt}^2 = 0.99$</th>
<th>$Q_{tgt}^2 = 0.999$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.2138</td>
<td>0.2796</td>
<td>0.3171</td>
<td>0.3129</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.4424</td>
<td>0.4607</td>
<td>0.4381</td>
<td>0.4424</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_{12}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_{13}$</td>
<td>0.2436</td>
<td>0.2597</td>
<td>0.2448</td>
<td>0.2437</td>
</tr>
<tr>
<td>$S_{23}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_{123}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$S_1^2$</td>
<td>0.5574</td>
<td>0.5393</td>
<td>0.5619</td>
<td>0.5576</td>
</tr>
<tr>
<td>$S_2^2$</td>
<td>0.4424</td>
<td>0.4607</td>
<td>0.4381</td>
<td>0.4424</td>
</tr>
<tr>
<td>$S_3^2$</td>
<td>0.2436</td>
<td>0.2597</td>
<td>0.2448</td>
<td>0.2437</td>
</tr>
<tr>
<td>Model evaluations</td>
<td>75</td>
<td>390</td>
<td>430</td>
<td></td>
</tr>
<tr>
<td>PDD degree $m$</td>
<td>6</td>
<td>7</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>$P_{qane}$</td>
<td>10</td>
<td>14</td>
<td>15</td>
<td></td>
</tr>
<tr>
<td>Sparsity</td>
<td>$\frac{100}{123} \approx 0.029$</td>
<td>$\frac{14}{123} \approx 0.027$</td>
<td>$\frac{15}{123} \approx 0.021$</td>
<td></td>
</tr>
<tr>
<td>Model accuracy $Q^2$</td>
<td>0.9428</td>
<td>0.9908</td>
<td>0.9996</td>
<td></td>
</tr>
</tbody>
</table>

The convergence of the sensitivity indices to the reference results is clearly verified when we increase the target model accuracy.

In the authors' experience, we observe it is relatively difficult to obtain an accuracy superior to 0.9999 for this case when using the Criterion 2. However, the numerical results are found to be sufficiently accurate with the conditions in the last column in Table 5.

By comparing these results with the ones in Table 3, we notice that, with Criterion 2, a larger number of model evaluations are usually required to obtain the sensitivity indices with a comparable accuracy. On the other hand, it shows, with the Criterion 2, we have sparser polynomial bases than with the Criterion 1.

5.2. 8- and 100-dimensional Sobol’ function

The second test case is devoted to the Sobol’ function. We consider first an 8-dimensional case, and discuss a 100-dimensional variant later in the same section.

5.2.1. 8-dimensional case

The 8-dimensional Sobol’ function has been considered in [13, Section 6.2]:

$$Y = f(X) = \prod_{k=1}^{N} f^{(k)}(X_k), \quad X_k \sim \mathcal{U}(0, 1),$$ (39)
Table 6
8-dimensional Sobol’ test case. \( D_2 = 8, \theta = 2 \times 10^{-4} \) as the predefined threshold for the variance Criterion 1. Numerical results with the proposed sparse PDD approach are compared to the ones obtained in [13] (sparse PC and Crude MC).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>( S_1 )</td>
<td>0.60</td>
<td>0.60</td>
<td>0.56</td>
<td>0.57</td>
</tr>
<tr>
<td>( S_2 )</td>
<td>0.27</td>
<td>0.25</td>
<td>0.22</td>
<td>0.29</td>
</tr>
<tr>
<td>( S_3 )</td>
<td>0.07</td>
<td>0.06</td>
<td>0.05</td>
<td>0.06</td>
</tr>
<tr>
<td>( S_4 )</td>
<td>0.02</td>
<td>0.02</td>
<td>0.02</td>
<td>0.03</td>
</tr>
<tr>
<td>( S_5 )</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>( S_6 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( S_7 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( S_8 )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>( S_9 )</td>
<td>0.63</td>
<td>0.62</td>
<td>0.59</td>
<td>0.66</td>
</tr>
<tr>
<td>( S_{10} )</td>
<td>0.29</td>
<td>0.27</td>
<td>0.26</td>
<td>0.27</td>
</tr>
<tr>
<td>( S_{11} )</td>
<td>0.08</td>
<td>0.08</td>
<td>0.10</td>
<td>0.08</td>
</tr>
<tr>
<td>( S_{12} )</td>
<td>0.02</td>
<td>0.03</td>
<td>0.05</td>
<td>0.01</td>
</tr>
<tr>
<td>( S_{13} )</td>
<td>0.01</td>
<td>0.01</td>
<td>0.03</td>
<td>0.01</td>
</tr>
<tr>
<td>( S_{14} )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.04</td>
<td>0.00</td>
</tr>
<tr>
<td>( S_{15} )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>( S_{16} )</td>
<td>0.00</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>Model evaluations</td>
<td>150</td>
<td>150</td>
<td>100,000</td>
<td></td>
</tr>
<tr>
<td>Sparse polynomial bases</td>
<td>38</td>
<td>76</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ANOVA order</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max polynomial degree</td>
<td>( \frac{10}{m-5} )</td>
<td>6</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Sparsity</td>
<td>( \approx 2.3e^{-5} )</td>
<td>( \approx 0.02 )</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Model accuracy ( Q^2 )</td>
<td>0.993</td>
<td>0.99</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

where

\[
N = 8, \\
f^{(k)}(X_k) = \frac{|4X_k - 2| + a_k}{1 + a_k}, \\
\mathbf{a} = \{1, 2, 5, 10, 20, 50, 100, 500\}^T.
\] (40)

Note the members of the input random vector \( \mathbf{X} \) are uniformly distributed over \([0, 1]\). \( \{a_k\}_{k=1}^N \) are positive coefficients whose values are gathered in the vector \( \mathbf{a} \).

In the computation by the proposed approach, let us first set the ANOVA interaction order equal to \( \nu = 2 \) (a sensitivity study on \( \nu \) is given later in the same section). We first set the active dimension \( D_1 = D_2 = 8 \) here (by imposing a large enough \( p \)). The experimental design size is set to 150 by a quasi-random Sobol’ sequence, for the sake of comparison with the results obtained with the adaptive sparse polynomial chaos method and the Monte Carlo method reported in [13]. Our numerical results with the variance Criterion 1 (the predefined threshold \( \theta = 2 \times 10^{-4} \)) are shown in Table 6 compared to the ones in the reference [13].

It is noticed, by using a half number of polynomial terms with the highest degree equal to 10, the corresponding model representation by the present approach is able to provide more accurate sensitivity indices than the method presented in [13] using the sparse PC of degree 6. On the other hand, we clearly observe the advantage of our method compared to the classical MC method when looking at the corresponding number of model evaluations.

**Sensitivity to the \( p \) constant (Criterion 1 vs Criterion 2)**

In this section, we test the method sensitivity to the \( p \) constant (see Section 4.1) which allows the efficient selection of the active dimension, and thus reducing the order of modeling difficulty.

The results by varying \( p \) from 0.9 to 0.999 and using Criterion 1 are reported in Table 7. We can conclude, from this analysis, that the using of only the 3 most important variables (by setting \( p = 0.9 \)) responsible for the second-order interactions are sufficient to provide very accurate results regarding both the sensitivity indices and the meta-model representation whose error is measured by \( 1 - Q^2 \).

We can also verify the quality of the built meta-model representation by comparing their output pdf and cumulative distribution function (cdf) with the ones of the original mathematical function. In Fig. 1, we show this comparison by varying \( p \) from 0.9 to 0.999 as in Table 7. All the distribution curves are obtained by carrying out histograms with \( 10^6 \) samples using a QMC Sobol’ sequence either through the built surrogate polynomial model or based on the original Sobol’ function. It is observed that the three pdf/cdf by the surrogates are all very close to the reference distribution curve, which matches the analysis in Table 7. We also notice the pdf/cdf with \( p = 0.9 \) is the most accurate, which is an opposite conclusion with respect to Table 7.
Table 7
8-dimensional Sobol' test case. Experimental design size $Q = 150$, $\theta = 2 \times 10^{-4}$ as the predefined threshold for the variance Criterion 1. Numerical results with the proposed sparse PDD approach by varying $p$.

<table>
<thead>
<tr>
<th>Sl</th>
<th>Exact</th>
<th>$p = 0.9$</th>
<th>$p = 0.99$</th>
<th>$p = 0.999$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S_1$</td>
<td>0.60</td>
<td>0.61</td>
<td>0.61</td>
<td>0.60</td>
</tr>
<tr>
<td>$S_2$</td>
<td>0.27</td>
<td>0.27</td>
<td>0.25</td>
<td>0.24</td>
</tr>
<tr>
<td>$S_3$</td>
<td>0.07</td>
<td>0.06</td>
<td>0.06</td>
<td>0.07</td>
</tr>
<tr>
<td>$S_4$</td>
<td>0.02</td>
<td>0.03</td>
<td>0.02</td>
<td>0.02</td>
</tr>
<tr>
<td>$S_5$</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
<td>0.01</td>
</tr>
<tr>
<td>$S_6$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$S_7$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$S_8$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Active dimension $D_2$ 3 6 7
Sparse polynomial bases 25 40 43
ANOVA order 2 2 2
Max polynomial degree $9 (m = 5)$ $9 (m = 5)$ $10 (m = 5)$
Sparsity $25 \times 1.5e-5$ $40 \times 2.4e-5$ $43 \times 2.6e-5$
Model accuracy $Q^2$ 0.982 0.987 0.992

Fig. 1. 8-dimensional Sobol' test case. Probability density function (pdf) and cumulative distribution function (cdf) of the model output $y$ obtained by carrying out histograms with $10^6$ samples using a QMC Sobol' sequence. The reference pdf/cdf is obtained by sampling from original mathematical function, while the others from the polynomial surrogates built with the proposed approach by varying $p$. Criterion 1. For other method parameters, see Table 7, Section 6.

Next we provide a picture of CPU time comparison of the proposed approach using two different criteria. In addition to the computation results shown in Table 7, we carry out the same analysis but using Criterion 2 with its threshold $\varepsilon_{Q^2} = 10^{-6}$. The meta-model's error and size, as well as the CPU time, is given in Fig. 2 by varying $p$ for both criteria. By increasing $p$, Criterion 1 is able to give continuously a lower error, while Criterion 2 is not. In our system, about 7 seconds were required to carry out this computation with Criterion 2 and $p = 0.999$. By retaining a considerably larger number of polynomial bases, Criterion 1 required however less than 1 second for this calculation. This gain in terms of CPU time is significant, and would be greater if more uncertain parameters should be considered in reality.

Sensitivity to Criterion 2's threshold $\varepsilon_{Q^2}$ and PDD degree $m$
Let us study the method sensitivity to the threshold $\varepsilon_{Q^2}$ of Criterion 2. The same parameters in Table 7 are used, and we choose $p = 0.999$. The active dimension is found to be 7 for the case of $m = 5$, and 8 for $m = 6$ and $m = 7$. Fig. 3 shows the representation error $1 - Q^2$ decreases quickly when we put a smaller $\varepsilon_{Q^2}$ until $10^{-7}$, while the size of constructed
Fig. 2. 8-dimensional Sobol’ test case. Numerical results by varying $p$ and Criterion. $\theta = 2 \times 10^{-4}$ for Criterion 1. $\epsilon_{Q^2} = 10^{-6}$ for Criterion 2.

The smaller $\epsilon_{Q^2}$, the more strongly the method is sensitive to $m$. Also, the even degree $m = 6$ is found to be the most accurate with a slightly larger number of polynomial bases.

Sensitivity to truncation dimension $\nu$ and Criterion 1’s threshold $\theta$

A sensitivity on the truncation dimension $\nu$ is carried out here. We use the same size of experimental design $Q = 150$. The PDD degree $m = 5$, and $p = 0.99$. Using Criterion 1’s threshold $\theta = 10^{-3}$ and $10^{-4}$, the model approximation error and the size of sparse representation are shown in Fig. 4 by varying $\nu$ from 2 to 5. The retained active dimension for all interaction order is 6. We observe the model error decreases by a small amount of about 0.5% when increasing $\nu$ from 2 to 5, while the gain between $\nu = 4$ and 5 is negligible for both curves. This indicates that low order ANOVA component functions help slightly improving the surrogate modeling, while high order components can be neglected. We also observe the method’s sensitivity to $\theta$ depends weakly on $\nu$. We show the (total) sensitivity indices in Fig. 5 with $\nu = 2$ and $\nu = 5$. The input importance decays as expected, and both curves are in perfect agreement with respect to the exact values shown in Table 6.
Fig. 4. 8-dimensional Sobol’ test case. Numerical results by varying $\nu$ and $\theta$. Criterion 1.

Fig. 5. 8-dimensional Sobol’ test case. The (total) sensitivity indices. $\nu = 2$ vs $\nu = 5$, $\theta = 10^{-4}$.

Table 8
100-dimensional Sobol’ test. The sensitivity studies on methodology parameters, $\nu = 2$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$p$</th>
<th>$m$</th>
<th>$Q$</th>
<th>$\theta$ (Criterion 1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Varying $m, \theta$</td>
<td>0.99</td>
<td>(2, 3, 4, 5, 6)</td>
<td>2000</td>
<td>$[10^{-4}, 10^{-5}]$</td>
</tr>
<tr>
<td>Varying $Q, m$</td>
<td>0.99</td>
<td>(3, 4)</td>
<td>[1000, 1250, 1500, 1750, 2000]</td>
<td>$10^{-4}$</td>
</tr>
<tr>
<td>Varying $\theta, p$</td>
<td>0.99</td>
<td>4</td>
<td>2000</td>
<td>$[10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}, 10^{-7}]$</td>
</tr>
</tbody>
</table>

5.2.2. 100-dimensional case

In order to demonstrate the proposed approach’s capability of solving high-dimensional stochastic problems, we consider here a 100-dimensional Sobol’ test [6] that has been similarly studied in [27,15] (with different numbers of inputs). The model is the same as Equation (39), but with different parameters:

$$N = 100,$$

$$f^{(k)}(X_k) = \frac{4X_k - 2| + a_k}{1 + a_k},$$

$$a_k^2 = k^2. \quad (41)$$

The output variance can be easily computed: $V(Y) = 0.103949$. We have demonstrated earlier in this section the advantage of Criterion 1 in terms of CPU time compared to Criterion 2. Thus, for this high-dimensional case, we only assess our approach using Criterion 1 (to obtain a similar accuracy, hundreds of times more CPU time should be required if using Criterion 2 for this specific case). The truncation dimension will be set to second-order ($\nu = 2$). The methodology parameters are gathered in Table 8 for detailed sensitivity studies below.

Sensitivity to PDD degree $m$ and Criterion 1’s threshold $\theta$

Let us first look at the sensitivity to the methodology parameter $m$. The parameters are gathered in the first row in Table 8. The scatter plots of model outputs using the experimental design of size $Q = 2000$ are given in Fig. 6. We can observe clearly the first variable $x_1$ gives the most variation on model output $y$. The influence decays quickly with respect to the input index. By increasing $m$, the numerical results of model error $1 - Q^2$, number of active dimensions, the size of sparse representation $P_{\text{sparse}}$, and the output variance $V(Y)$ are shown in Fig. 7. We observe the retained active dimensions
increase from 5 to 9, and the convergence tendency is confirmed by the error profile and output variance value. The size \( P_{\text{sparse}} \) of the finally constructed polynomial bases with \( \theta = 10^{-5} \) varies between 49 and 162, and this confirms the approach’s capability of an efficient selection of significant polynomials, since the size of a full PDD expansion for \( m = 6 \) is \( 7^{100} \). Using a bigger \( \theta = 10^{-4} \) in this case provides a considerably sparser representation while we still have a surrogate with a similar accuracy. This observation indicates, for some cases, after a certain threshold, further decreasing \( \theta \) does not provide a much better accuracy, while the computational costs increases considerably because of the size of involved linear systems of regression problems. Fig. 8 shows, for different \( m \), the progress of the adaptive approach regarding the stepwise enriching iteration from high-order polynomials of second-order ANOVA components. The number of total iterations is equal to \( (D_2)_m^2 \). It is shown \( P_{\text{sparse}} \) decreases at the early stage and increases later, while the representation model error always decreases. This demonstrates the method is able to correctly choose important polynomials to ameliorate the surrogate accuracy. We further plot the output pdf/cdf in Fig. 9, by increasing \( m \), where we show the convergence tendency towards the reference profile that is drawn from the original 100-dimensional mathematical Sobol’ function. Finally, the sensitivity indices (SI) are shown in Fig. 10. The total sensitivity indices (TSI) are slightly above the SI. Both SI and TSI decay with respect to input index as expected.

### Sensitivity to experimental design size \( Q \) and PDD degree \( m \)

When we vary the experimental design size \( Q \) and \( m \) as shown in Table 8 (second row), results are presented in Fig. 11 for model error \( 1 - Q^2 \), active dimension \( D_2 \), size of sparse bases \( P_{\text{sparse}} \), and output variance \( V(Y) \), and in Fig. 12 for pdf/cdf. It is shown a larger \( Q \) provides a smaller \( D_2 \), and a sparser polynomial representation for this specific case. This means the importance of input variables decays more quickly when imposing more evaluations of deterministic simulations, and thus results in a smaller number of active dimensions. Also, the variance contributions are more focused within a smaller group of polynomials with a larger \( Q \), which makes the polynomial surrogate sparser. We observe the error \( 1 - Q^2 \) increases with a larger \( Q \) in general, while the variance \( V(Y) \) is meanwhile closer to the exact value for the case of \( m = 4 \). We point out it is not an appropriate way to compare model accuracy by only looking at \( Q^2 \) between two cases where the number of evaluations \( Q \) differs, since the estimator \( Q^2 \) itself strongly depends on \( Q \) (see Appendix A). We observe when setting a smaller \( m = 3 \) in this case, the variance result becomes less accurate when increasing \( Q \). However, Fig. 7 demonstrates the variance’s accuracy can be improved quickly by increasing \( m \). On the other hand, the pdf/cdf presented in Fig. 12 shows that increasing \( Q \) is not able to provide a more accurate distribution function for the Sobol’ function case.
Fig. 7. 100-dimensional Sobol’ test case. Numerical results by varying $m$ and $\theta$.

Fig. 8. 100-dimensional Sobol’ test case. The evolution of the number of sparse polynomial bases and model error $1 - Q^2$ with respect to the iterations concerning adaptive selection of polynomials from second-order ANOVA components, $\theta = 10^{-5}$.
Fig. 9. 100-dimensional Sobol' test case. Probability density function (pdf) and cumulative distribution function (cdf) of the model output \( y \) obtained by carrying out histograms with \( 10^5 \) samples using a QMC Sobol sequence. The reference pdf/cdf is obtained by sampling from original 100-dimensional mathematical function, while the others from the polynomial surrogates built with the proposed approach by varying \( m, \theta = 10^{-5} \). For other method parameters, see Section 5.2.2.

Fig. 10. 100-dimensional Sobol' test case. The sensitivity indices of the first 10 inputs. \( m = 6, \theta = 10^{-5} \).

**Sensitivity to Criterion 1’s threshold \( \theta \) and \( p \) constant**

Finally, the sensitivity study is performed on the methodology parameter \( \theta \) of Criterion 1 by setting \( p = 0.9 \) and 0.99. Other parameters are gathered in the last row of Table 8. Since in the proposed approach, the active dimension \( D_2 \) is independent with respect to \( \theta \), we have selected a constant active dimension of \( D_2 = 2 \) for \( p = 0.9 \) and \( D_2 = 6 \) for \( p = 0.99 \). Fig. 13 shows the model error \( 1 - Q^2 \) does not decrease monotonically when decreasing \( \theta \). With \( \theta = 10^{-5} \), the error reaches its minimum value. While the sparse PDD representation obviously increases its size when further decreasing \( \theta \), the accuracy of variance \( V(Y) \) is slightly improved. This means below a certain value, further decreasing \( \theta \) will add less important polynomials and the method becomes less efficient without a significant gain in accuracy. On the other hand, we notice the gain when increasing \( p \) from 0.9 to 0.99 by incorporating 4 more dimensions in second-order ANOVA components is considerable for this case, while the size of polynomial representation remains quasi-constant until \( \theta = 10^{-5} \). Fig. 14 shows the surrogate output pdf/cdf is not sensitive to Criterion 1’s threshold when \( \theta \) is smaller than \( 10^{-5} \).

5.3. 21-dimensional product model [52, Plischke et al. 2013]

In last section, we have studied a 100-dimensional test problem where the input importance decays fast, and thus the active dimension is much lower than 100. In this section, we consider a numerically more challenging example. The following product model function has been presented in Plischke et al. 2013 [52] and also analyzed in Borgonovo et al. 2014 [53].

\[
Y = \prod_{k=1}^{N} X_k^{\theta}, \quad N = 21, 
\]
where the random variables $X = (X_1, \ldots, X_N)$ are all lognormally distributed with unitary mean and variance, i.e.

$$\ln(X_k) \sim \mathcal{N}(1, 1), \ k = 1, \ldots, N.$$ 

The weights are imposed as

$$a_1, \ldots, a_7 = 4, \ a_8, \ldots, a_{14} = 2, \ a_{15}, \ldots, a_{21} = 1,$$

and in this way the 21 inputs are divided into three groups of relatively high, moderate and low importance. As indicated in [52, 53], the analytical moments of this test function are made extremely high:
Fig. 13. 100-dimensional Sobol’ test case. Numerical results by varying $\theta$ and $p$.

Fig. 14. 100-dimensional Sobol’ test case. Probability density function (pdf) and cumulative distribution function (cdf) of the model output $y$ obtained by carrying out histograms with $10^5$ samples using a QMC Sobol sequence. The reference pdf/cdf is obtained by sampling from original 100-dimensional mathematical function, while the others from the polynomial surrogates built with the proposed approach by varying $\theta$, $p = 0.99$.

$$E(Y) = \exp \left( \sum_{k=1}^{N} (\alpha_k^2/2 + a_k) \right) = 1.589 \cdot 10^{53},$$

$$V(Y) = \exp \left( \sum_{k=1}^{N} (2a_k^2 + 2a_k) \right) - [E(Y)]^2 = 1.752 \cdot 10^{170}.$$ 

As a result of large value of the variance, the first-order sensitivity indices are almost all zero by analytical calculation, which implies the total variance are contributed entirely by input interaction terms. This is due to the fact that the model function is highly non-additive [52]. As also stated in these two references, when using 100,000 importance samples by a
QMC Sobol’ sequence generator, we obtain model evaluation outputs that have an extremely large range from a minimum of 0.8283 to a maximum of $5.995 \cdot 10^{42}$. Note the expectation and variance obtained by a simple MC method using the 100,000 samples are,

$$E_{MC}(Y) = 1.075 \cdot 10^{38},$$

$$V_{MC}(Y) = 4.144 \cdot 10^{80}.$$  

We remark that these values are not close to the exact ones. This is because the model output values are sparse due to the extremely large value of $V(Y)$. This test case is thus made very difficult, on one hand, because of this high variability in $Y$, and on the other hand, due to the high number of effective dimension of the model. In [52], it is mentioned these difficulties impair a direct application of any global sensitivity estimation procedure, unless one resorts to monotonic transformations to reduce numerical noise. [53] confirms this statement by showing the global sensitivity measures are covered by numerical noise and one cannot distinguish the three groups of inputs.

We recall the objective of the proposed approach is to compute sensitivity measures by using as few samples as possible. To this, let us test our approach by using the fewest 512 and 1,024 as the sample size as employed in [53]. Methodology parameters used and the obtained surrogate size, accuracy and moments are reported in Table 9. The fact that a lower $Q^2$ is obtained when using a higher experimental design size $Q$ is not surprising, and has also been observed in 100-dimensional Sobol’ test case. Note that the convergence is not really expected, especially with such a small number of model evaluations for this case with an extremely large value of variance. Indeed, we have mentioned that even with 100,000 samples using Sobol’ sequence, the maximum model output we have obtained is $5.995 \cdot 10^{42}$, which is still ten orders-of-magnitude smaller than the exact output expected value of $1.589 \cdot 10^{53}$. The numerical results of total sensitivity indices (TSI)$^4$ using the proposed method are presented in Fig. 15. It is shown clearly the results are covered by numerical noises for both sample sizes, as the TSI of input variables in a same importance group ($X_1, \ldots, X_7, X_8, \ldots, 14$, or $X_{15}, \ldots, 21$) have very different values. This has also been reported in [52] using a density-based sensitivity measure. However, the average values of these indices for each group allow us to distinguish the three groups of inputs. We notice the cases with an experimental design size of 512 and 1024 give very similar average measures. On the other hand, the sensitivity indices (SI) of each ANOVA component function are given in Fig. 16. Note the first 21 indices correspond to first-order SI, and the remaining 210 indices are second-order ones (separated by a dash line). Although noised, we observe, for both sample sizes in general, that the variance contribution of first-order components is significantly less noticed than its second-order counterpart (in particular

---

3 According to [53], the mean effective dimension of this model is $D = 18.30$.
4 No logarithmic transformation is used for model output, since this type of transformation modifies the model structure, and also changes the value of variance-based sensitivity measures.
individual component indexed between 22 and around 40 in Fig. 16 is found of greatest importance, which is expected because the first-order SI are close to zero analytically. This confirms that the model is highly non-additive.

The above results show that the proposed approach, with a reasonable sample size, is able to distinguish the three groups of different importance, and at the same time can capture correctly the principal source of output variance.

5.4. 7-dimensional application example

The method proposed in this paper is applied to the uncertainty quantification of an atmospheric re-entry spacecraft case. Numerical simulation solves a set of governing equations including modelization of rarefied gas effects, aerothermo-chemistry, radiation, and the response of thermal protection materials to extreme conditions. A global overview over this problem has been provided in [54].

Here, the focus is in predicting stagnation-point pressure and heat flux from freestream conditions, and is described by a physico-chemical model and solved by suitable numerical methods proposed by Barbante [55,64].

We use a set of physico-chemical models to simulate high temperature reacting flows, including 2D axisymmetric Navier Stokes equations and gas/surface interaction equations (see Ref. [55]). Indeed, the wall of the spacecraft acts as a catalyzer and promotes recombination of atoms. This phenomenon is modeled by a catalytic wall at radiative equilibrium, where the so-called effective catalytic recombination coefficient $\gamma$ represents the proportion of gas impinging the body that will recombine. A mixture of 5 species of air is used, namely N, O, N$_2$, O$_2$, and NO, with chemical mechanism due to Park [57]. Input data for the forward model are the freestream pressure $p_\infty$ and Mach number $M_\infty$, the effective catalytic recombination coefficient $\gamma$, and the gas reaction rate coefficients $k_r$ of the chemical reactions $r$.

The code COSMIC developed by Barbante [55] is used, which was designed to approximate hypersonic flow models where chemical non-equilibrium effects need to be accounted for. It includes a Hybrid Upwind Splitting (HUS) scheme [58], which is an interesting attempt of combining, in a mathematically rigorous way, Flux Vector Splitting (FVS) and Flux Difference Splitting (FDS) schemes. The design principle combines the robustness of FVS schemes in the capture of nonlinear waves and the accuracy of some FDS schemes in the resolution of linear waves. In particular, COSMIC uses the hybridization of the Van Leer scheme [59] and the Osher scheme [60] and includes a carhuncle fix.

The boundary conditions are illustrated in the left panel of Fig. 17: an axisymmetric condition is imposed on the $y$ axis (horizontal axis on Fig. 17), while the wall of the body is modeled by a partially catalytic wall at radiative equilibrium. The mesh used for the computations is given in the right panel of Fig. 17. Pressure and temperature iso-contours of the flow around the European EXPerimental Reentry Test-bed (EXPERT) vehicle obtained with COSMIC for input data mean values are shown in Fig. 18. Note that a specific point of the trajectory of EXPERT is considered [61]. The trajectory point corresponds roughly to the chemical non-equilibrium flow conditions of Table 10. Uncertainties are considered on $p_\infty$, $M_\infty$, and $\gamma$, with uniform distributions detailed in Table 11. Concerning $p_\infty$ and $M_\infty$, only a priori ranges of plausible values

Fig. 16. 21-dimensional product model case. 21 first-order and 210 second-order sensitivity indices (SI) using 512 and 1024 as sample size. The vertical dash line separates these two groups.

Table 10

<table>
<thead>
<tr>
<th>Application test. Freestream conditions for one trajectory point of the EXPERT vehicle.</th>
</tr>
</thead>
<tbody>
<tr>
<td>-----------------</td>
</tr>
<tr>
<td>Chemical non-equilibrium</td>
</tr>
</tbody>
</table>
are known. Concerning \( \gamma \), the mean value corresponds roughly to the EXPERT material, while the 33% error have been previously determined [62].

Uncertainty is also considered on the gas reaction rate coefficients \( k_i \) of four chemical reactions of the dissociation reaction. For the trajectory point investigated, the dissociation reaction of molecular oxygen and nitric oxide was found important. Following the suggestion of Bose et al. [63], the uncertainty concerns only the pre-exponential factor \( A_i \) of the Arrhenius rate equation:

\[
k_i = A_i T^{b_i} \exp(-E_i/RT)
\]

Since the uncertainties on \( k_i \) can be quite large, it is appropriate to
consider them on a logarithmic scale; in particular, \( \log_{10}(k_t/k_{t_0}) \), with \( k_{t_0} \) the recommended rate constant, is commonly assumed to vary following a normal distribution \([63]\),

\[
P(k_t) \propto \exp \left[ -\frac{1}{2} \left( \frac{\log_{10}(k_t/k_{t_0})}{\sigma_t} \right)^2 \right]
\]

where \( \pm 2\sigma_t \) (reported in Table 12) defines the 95% confidence limits symmetrically bounding \( k_{t_0} \).

Note that the quantities of interest are the pressure \( p_{st} \) and heat flux \( q_{st} \) at the stagnation point.

**Uncertainty quantification results**

As far as this application test is concerned, 1000 resolutions of the deterministic code are used to recursively solve linear regression systems. Let us first consider the pressure \( p_{st} \) as the quantity of interest and set the highest ANOVA interaction order equal to \( v = 2 \) (the truncation dimension). We further set \( p = 0.99999 \) for the selection of the active dimensions for the second-order ANOVA interaction terms. For the sake of comparison with the UQ results presented in the reference \([54]\), let us vary the PDD polynomial order from \( m = 2 \) to \( m = 4 \). Both criteria for the selection of the most important polynomial terms are tested. The computed mean, variance and sensitivity indices are reported in Table 13. Note that all first-order and total indices are presented here; concerning the second-order sensitivity indices, for conciseness, we only present the most important one measuring the interaction between \( p_\infty \) and \( M_\infty \). The moment results from Table 13 show the method is convergent when increasing the PDD order \( m \) for both of the two selection criteria. Meanwhile, for all cases considered here, the obtained model accuracy is very high (all superior to 0.9995), \( p_\infty \) and \( M_\infty \) are found to be the two most important parameters whose first-order sensitivity measures are largely superior to those of other parameters. The second-order interaction between these two parameters is found to be non-negligible compared to other first-order sensitivity estimates. \( y \) reveals its negligible influence on \( p_{st} \). We further observe the two criteria provide similar sensitivity estimates and moments. Following the numerical conditions used for this case, the **Criterion 2** generally produces a sparser
Table 14
Application test. Uncertainty quantification results for \( q_{st} \) using Criterion 1 and Criterion 2 by varying the PDD polynomial order \( m \). The truncation dimension \( \nu = 3 \). The experimental design size \( Q = 1000 \). The model accuracy is estimated by the cross validation method (see Appendix A). For Criterion 1, the threshold \( \theta = 10^{-4} \); for Criterion 2, the threshold \( \epsilon_{\text{Q2}} = 10^{-7} \).

<table>
<thead>
<tr>
<th>( q_{st} ) (Criterion 1)</th>
<th>( m = 2 )</th>
<th>( m = 3 )</th>
<th>( m = 4 )</th>
<th>( q_{st} ) (Criterion 2)</th>
<th>( m = 2 )</th>
<th>( m = 3 )</th>
<th>( m = 4 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( E(q_{st}) )</td>
<td>2869270</td>
<td>2869450</td>
<td>2859070</td>
<td>2868980</td>
<td>2869470</td>
<td>2872330</td>
<td></td>
</tr>
<tr>
<td>( V(q_{st}) )</td>
<td>0.296806E+10</td>
<td>0.303000E+10</td>
<td>0.387263E+10</td>
<td>0.296506E+10</td>
<td>0.301310E+10</td>
<td>0.311091E+10</td>
<td></td>
</tr>
<tr>
<td>( S_1 )</td>
<td>0.110434E+00</td>
<td>0.102941E+00</td>
<td>0.833866E-01</td>
<td>0.104603E+00</td>
<td>0.103322E+00</td>
<td>0.101710E+00</td>
<td></td>
</tr>
<tr>
<td>( S_2 )</td>
<td>0.874650E+00</td>
<td>0.872848E+00</td>
<td>0.712564E-00</td>
<td>0.876280E+00</td>
<td>0.877072E+00</td>
<td>0.872107E+00</td>
<td></td>
</tr>
<tr>
<td>( S_3 )</td>
<td>0.946742E-02</td>
<td>0.961702E-02</td>
<td>0.656856E-02</td>
<td>0.908636E-02</td>
<td>0.876599E-02</td>
<td>0.877659E-02</td>
<td></td>
</tr>
<tr>
<td>( S_4 )</td>
<td>0.366610E-03</td>
<td>0.367807E-03</td>
<td>0.611833E-02</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_5 )</td>
<td>0.219254E-02</td>
<td>0.192399E-02</td>
<td>0.189310E-02</td>
<td>0.225662E-02</td>
<td>0.195543E-02</td>
<td>0.923366E-03</td>
<td></td>
</tr>
<tr>
<td>( S_6 )</td>
<td>0.200434E-03</td>
<td>0.139166E-02</td>
<td>0.151648E-02</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_7 )</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_8 )</td>
<td>0.588836E-02</td>
<td>0.616165E-02</td>
<td>0.624935E-02</td>
<td>0.588886E-02</td>
<td>0.597824E-02</td>
<td>0.642160E-02</td>
<td></td>
</tr>
<tr>
<td>( S_9 )</td>
<td>0.140176E-02</td>
<td>0.151489E-02</td>
<td>0.805465E-03</td>
<td>0.133857E-02</td>
<td>0.135844E-02</td>
<td>0.104616E-02</td>
<td></td>
</tr>
<tr>
<td>( S_{10} )</td>
<td>/</td>
<td>/</td>
<td>0.121303E-02</td>
<td>/</td>
<td>/</td>
<td>0.117491E-02</td>
<td></td>
</tr>
<tr>
<td>( S_{11} )</td>
<td>0.176827E-03</td>
<td>0.945794E-03</td>
<td>0.331797E-03</td>
<td>0.000000E+00</td>
<td>0.532325E-03</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_{12} )</td>
<td>/</td>
<td>/</td>
<td>0.486300E-02</td>
<td>/</td>
<td>/</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_{13} )</td>
<td>/</td>
<td>/</td>
<td>0.218688E-01</td>
<td>/</td>
<td>/</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_{14} )</td>
<td>0.161953E-02</td>
<td>0.225210E-02</td>
<td>0.734882E-02</td>
<td>0.392357E-03</td>
<td>0.699826E-03</td>
<td>0.178231E-02</td>
<td></td>
</tr>
<tr>
<td>( S_{15} )</td>
<td>0.146464E-02</td>
<td>0.293194E-01</td>
<td>0.597520E-01</td>
<td>/</td>
<td>/</td>
<td>0.000000E+00</td>
<td></td>
</tr>
<tr>
<td>( S_{16} )</td>
<td>/</td>
<td>/</td>
<td>0.557466E-01</td>
<td>/</td>
<td>/</td>
<td>0.146205E-02</td>
<td></td>
</tr>
<tr>
<td>( D_2, D_3 )</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td>3</td>
<td>3</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>( P_{\text{gase}} )</td>
<td>19</td>
<td>31</td>
<td>122</td>
<td>11</td>
<td>14</td>
<td>18</td>
<td></td>
</tr>
<tr>
<td>Sparsity</td>
<td>( 10^{-9} \approx 8.7e-3 )</td>
<td>( 10^{-9} \approx 1.9e-3 )</td>
<td>( 10^{-9} \approx 1.6e-3 )</td>
<td>( 10^{-10} \approx 5.0e-3 )</td>
<td>( 10^{-14} \approx 8.5e-4 )</td>
<td>( 10^{-18} \approx 2.3e-4 )</td>
<td></td>
</tr>
<tr>
<td>Accuracy ( Q^2 )</td>
<td>0.783770</td>
<td>0.793848</td>
<td>0.805251</td>
<td>0.785050</td>
<td>0.796990</td>
<td>0.811363</td>
<td></td>
</tr>
</tbody>
</table>

The polynomial basis for a similar model accuracy. Note nevertheless the Criterion 2, based essentially upon the model error estimates for the selection of polynomial terms, requires more resolutions of regression problems than the Criterion 1.

When the heat flux quantity \( q_{st} \) is considered, the same type of UQ analysis is realized and the results are reported in Table 14. We set the maximum interaction order \( \nu = 3 \). When the \( p \) constant is modified to \( p = 0.995 \) for both criteria, 3 active dimensions are retained for second- and third-order interactions when the PDD order is set as \( m = 2 \) and \( m = 3 \), whilst 4 active dimensions are retained for \( m = 4 \). Concerning the computed expectation and variance, the convergence is less good than for the quantity of \( P_{\text{gase}} \) regarding both criteria. The poor convergence for \( q_{st} \) can also be observed by viewing the modeling accuracy estimate \( Q^2 \) whose value is around 0.8, while for \( p_{\text{gase}} \) we have obtained \( Q^2 \) above 0.999. We think this is due to the high variability and non-linearity in the response of the stagnation-point heat-flux. For Criterion 1, \( P_{\text{gase}} \) and \( M_{\infty} \) are still found to be the most significant parameters, while the gas reaction \( O_2 + O \rightarrow 2O + O \) (\( X_7 \)) is negligible. Indeed, this reaction uncertainty is not taken into consideration in the final meta-model polynomial representation, since its total sensitivity effect is also zero. Moreover, Table 14 shows that all second- and third-order interactions are non-negligible: the orders of magnitude of these sensitivity indices are comparable to those of first-order contributions (\( X_3, X_4, X_5, X_6 \)). In particular, in the case of \( m = 4 \), the third-order interactions are found more important than the second-order ones. It is also shown that the sensitivity measure of a parameter or a group of parameters can vary significantly when a different number of active dimensions is employed. For instance, the total sensitivity index of \( X_4 \) is about 500 times bigger with \( D_2, D_3 = 4 \) than with \( D_2, D_3 = 3 \). Finally, the accuracy of the PDD model representation for \( q_{st} \) is found to be less good than for \( p_{\text{gase}} \), as also mentioned in the reference [54]. As far as the Criterion 2 is concerned, in order to obtain a similar model accuracy, a smaller number of polynomial terms are required. However, the model representation includes fewer uncertain parameters. For instance, \( X_4 \) and \( X_7 \) are excluded when using \( m = 4 \), and \( X_4 \) is additionally neglected if we set \( m = 2 \) or \( m = 3 \).

6. Conclusions and perspectives

This paper aims to deal with Uncertainty Quantification of engineering and physical problems featuring a moderate to large number of uncertain input parameters. The purpose is to identify the relative importance of these uncertainties onto a given quantity of interest. This is achieved in this work by performing global sensitivity analysis, and in particular by
combining the Analysis of Variance technique (ANOVA) and the polynomial dimensional decomposition approach (PDD). Complexities present in practical problems usually make the global methods prohibitive due to the large uncertainty in the model output. In this paper, we have employed three levels of adaptivity to reduce the meta-modeling difficulty which read as follows:

1. We set a truncation dimension (the maximum interaction order) in the ANOVA expansion.
2. Resulting from the solution of the regression system including only the PDD terms of the first-order ANOVA component functions, a rank of importances can be established quantitatively for all the input parameters. Hence, we can retain so-called active dimensions (the most influential parameters) for the PDD terms of the second- and higher-order ANOVA components.
3. Starting from the PDD polynomials of the first-order components, we enrich our surrogate model representation by adding only significant polynomials of second- and higher-order interaction terms by keeping only the corresponding active dimensions. Two selection criteria have been utilized in this work for this purpose. We emphasize that recursive resolutions of regression problems are required for this task.

The resulting surrogate polynomial approximation is a very sparse representation of the deterministic model. Since the surrogate model size is updated recursively with respect to the resolutions of regression problems subject to the enrichment of the polynomial basis, the number of the required deterministic model evaluations is well controlled, and its final value is significantly smaller than when employing a standard Monte Carlo or quasi-Monte Carlo method. The computation of the global sensitivity indices simply requires a post-processing of the polynomial coefficients.

Concerning the selection criterion for retaining the significant polynomial terms, the Criterion 1 based on the variance contribution is found to be very efficient, but can result in a slightly larger basis than the Criterion 2 that based on the model error comparison, for a similar meta-model accuracy. As explained in Section 4.2 and emphasized in the numerical examples section, Criterion 1 is more appropriate when one considers a high-dimensional case because Criterion 2 can suffer from CPU cost in solving a significantly larger number of least-squares problems. As far as a low-dimensional case is concerned, both criteria can be used. The satisfactory uncertainty quantification results for the application example show that our approach is capable of treating complex engineering problems. Note in particular for this application case, the finally obtained surrogate model does not include all the input parameters, which means throughout the meta-modeling approach, non-important parameters are well identified, which reduces the modeling difficulty.

Only independent inputs have been taken into account in this work. Note indeed that the independence is important to insure the uniqueness of the functional ANOVA expansion. Bedford [64] shows that, in the case correlations are present, the uniqueness of the representation is lost. However, for many realistic problems, model parameters are calibrated using a Bayesian approach (see e.g. [65]), and the resulting joint probability density function (pdf) is in general correlated. Saltelli & Tarantola [66] and Oakley & O’Hagan [67] show that one can obtain spurious terms in the decomposition when correlations are present. Most recently, Kucherenko et al. [68] shows that total order indices can be smaller than individual indices if correlations are present. The readers are also referred to Borgonovo & Plischke [21] for a detailed discussion on the current status of the study of variance-based sensitivity measures in the presence of correlations among the model inputs. We point out one possible way to consider correlations when using the proposed PDD approach in this work is to employ the covariance decomposition of the unconditional variance (see for instance [51]). In this approach, three sensitivity indices should be defined, instead of a single index, to take account of input correlations. This topic remains our ongoing research effort.

The density-based (moment-independent) sensitivity measures (see [18,52], also [21]) have been proposed by considering the entire distribution without reference to a particular moment. This measure does not require independence between the model inputs. We point out that it is not prohibitive to compute this type of density-based measure if an accurate PDD surrogate is available. Indeed, since surrogate model evaluations can be carried out “almost” immediately, performing the Monte Carlo type importance sampling from the joint input pdf can allow us to propagate the uncertainty into the model output in a very efficient way. We can then approximate the unconditional and/or conditional output pdf by drawing a histogram based on the model evaluations. Note however, how to approximate efficiently the expectation of a functional featuring a conditional output pdf, necessary to compute the density-based measure, is still one of our current research efforts.

Acknowledgements

The authors are grateful to the reviewers whose constructive comments have helped us greatly in improving the paper.

Appendix A. Estimator of accuracy $Q^2$

The Leave-one-out cross validation estimator presented in [13] is directly used in this work of which we summarize the main ingredients as follows. We keep the same notations as in [13].
Algorithm 2 Adaptive ANOVA and PDD featuring stepwise regression with Criterion 2.

1. Determine active dimension (variance-based)
2. Initialize concise first-order PDD bases (variance-based): \( \{\Phi^w\} \leftarrow \{\Phi_{a1}\} \)
3. Construct high order bases \( \{\Phi_{a2+}\} \) using truncation dimension \( \nu \) and active dimensions
4. **for all** \( \Phi_{ai} \in \{\Phi_{a2+}\} \) **do**
5. Add \( \Phi_{ai} \) into \( \{\Phi^w\} \), i.e. \( \{\Phi^w\} \leftarrow \{\Phi^w, \Phi_{ai}\} \)
6. Depending on size \( P^w \) of \( \{\Phi^w\} \), adjust, if necessary, the size \( Q^w \) of experimental design (see formulation (22))
7. Solve the regression system (24) to determine PDD expansion coefficients \( C^w \)
8. Evaluate the model accuracy by a leave-one-out cross validation estimator \( Q_1^2 \)
9. **if** \( Q_1^2 \geq Q_{tgt} \) **then** Exit **# Optional target accuracy**
10. **end if**
11. **for all** \( \Phi_{aj} \in \{\Phi^w\} \) **do**
12. Solve the regression system with the polynomial bases \( \{\Phi^w\} \setminus \Phi_{aj} \), i.e. by excluding \( \Phi_{aj} \)
13. Evaluate the model accuracy \( Q_i^2 \)
14. **if** \( Q_i^2 - Q_{1i}^2 \leq \varepsilon_Q \) **then** **# Criterion 2**
15. **end if**
16. **end if**
17. **end for**
18. **end for**
19. Solve final regression system based on the constructed bases \( \{\Phi^f\} \) --- an accurate surrogate model is thus built
20. Compute final total and partial variances by post-processing of surrogate coefficients
21. Compute global (total) sensitivity indices
22. Surrogate model evaluations using MC/QMC importance sampling from input distributions can provide reliable output pdf

Let \( f_X(x) \) denote the surrogate model representation obtained by adaptive sparse polynomial dimensional decomposition (PDD) with the coefficients determined by the regression approach with the experimental design \( X = (x^1, x^2, \ldots, x^Q)^T \) (see Sections 3 and 4). We remind that \( Y = (y^1, y^2, \ldots, y^Q)^T \) represents the corresponding experimental design outputs.

Let \( f_{X\setminus i}(x) \) denote the sparse PDD representation constructed from the experimental design \( X \setminus [x^i] \).

The following empirical mean-square predicted residual [13] is computed to estimate the approximation error:

\[
I_X^*[f_X] = \frac{1}{Q} \sum_{i=1}^{Q} \left( y^i - f_{X\setminus i}(x^i) \right)^2.
\]  
(A.1)

In the case of linearly parameterized regression, we have the following result [13] useful for the computation of (A.1):

\[
y^i - f_{X\setminus i}(x^i) = \frac{y^i - f_X(x^i)}{1 - h_i}
\]  
(A.2)

where \( h_i \) is the \( i \)-th diagonal term of the projection matrix

\[
\mathcal{M}(AM^TAM)^{-1}AM^T
\]

where \( \mathcal{M} \) is expressed as in (25). Hence, (A.1) can be expressed as

\[
I_X^*[f_X] = \frac{1}{Q} \sum_{i=1}^{Q} \left( y^i - f_X(x^i) \right)^2 / (1 - h_i).
\]  
(A.3)

Note the use of the formulation (A.3) avoids any additional resolution of regression problems concerning \( f_{X\setminus i}(x) \).

The determination coefficient then writes as follows:

\[
Q^2[f_X] = 1 - \frac{I_X^*[f_X]}{\hat{\psi}[Y]},
\]  
(A.4)

where

\[
\hat{\psi}[Y] = \frac{1}{Q-1} \sum_{i=1}^{Q} (y^i - \bar{y})^2, \quad \bar{y} = \frac{1}{Q} \sum_{i=1}^{Q} y^i.
\]  
(A.5)

The quantity \( Q^2 \) is exclusively used in this work to estimate the accuracy of the surrogate model representation. \( Q^2 = 1 \) indicates a perfect fit, while \( Q^2 \approx 0 \) or \( Q^2 < 0 \) reflects a poor model accuracy. We also use \( 1 - Q^2 \) to represent the modeling error.
Algorithm 1 can be modified to Algorithm 2, by considering Criterion 2 based on the model accuracy $Q^2$ (in particular, see Operation 14–15 in Algorithm 2). $Q^2_{gr}$ in Operation 9 of Algorithm 2 is an optional predefined target accuracy (e.g., 0.999). Once the algorithm reaches this target accuracy, we can choose to stop and use the constructed bases. $Q^2$ in Operation 14 is a predefined threshold (e.g., $10^{-5}$). If the decrease of the model accuracy is smaller than $e_{Q^2}$ when excluding the polynomial term in consideration, we then eliminate this term from our working bases.

Note that the initialization of multivariate PDD bases (Operation 2) can be realized by Criterion 2.

Operation 12 in Algorithm 2 involves additional resolutions of regression systems compared to Algorithm 1. Thus, Algorithm 2 can be in general more expensive than Algorithm 1.

References
