ECCOMAS Congress 2016
VII European Congress on Computational Methods in Applied Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou, V. Plevris (eds.)
Crete Island, Greece, 5–10 June 2016

A STATISTICAL APPROACH FOR BUILDING SPARSE POLYNOMIAL CHAOS EXPANSIONS

Simon Abraham, Ghader Ghorbaniasl, and Chris Lacor

Vrije Universiteit Brussel (VUB), Department of Mechanical Engineering Research Group Fluid Mechanics and Thermodynamics
Pleinlaan 2, 1050 Brussels
e-mail: Simon.Abraham@ulb.ac.be, Ghader.Ghorbaniasl@vub.ac.be, Chris.Lacor@vub.ac.be

Keywords: Sparse polynomial chaos expansion, Uncertainty quantification, Regression, Statistical inference, Curse of dimensionality, Compressive sampling

Abstract. *Over the last years, a lot of effort has been made to make existing uncertainty quantification techniques more efficient in high dimensions. An important class of methods relies on the assumption that the polynomial chaos representation of the model response is sparse. This paper contributes to the validation and assessment of an innovative basis selection technique for building sparse polynomial chaos expansions. A regression approach is used for computing the polynomial chaos coefficients. The technique is based on statistical inference theory which provides information about the true regression model from an estimated regression model based on samples. The latter information is used to build iteratively the sparse polynomial chaos expansion. Using the developed methodology, a more robust and efficient basis selection technique is obtained. For validation purpose, the methodology is applied to high dimensional analytical test cases, including the Oakley & O'Hagan function (d=15) and the Morris function (d=20). The results are compared with those obtained from two state-of-the-art techniques, namely the LARS-based algorithm and compressive sampling. As compared to previous work, more comparisons with the LARS-based method are provided, through the use of UQLab, a MATLAB-based uncertainty quantification framework developed by Sudret and Marelli [1]. It is shown that, with equal settings, the developed methodology results in a more accurate polynomial chaos expansion compared to the aforementioned technique. In addition, a new criterion for building an optimal polynomial chaos expansion is further investigated. The conclusions are in-line with previous findings, i.e. the present criterion always builds a sparser polynomial chaos expansion which is, in addition, at least as accurate as compared to the optimal polynomial chaos expansion obtained from the classical cross validation technique.*

1 INTRODUCTION

In recent years, due to the increase in computational power, the interest in uncertainty quantification (UQ) in computational fluid dynamics (CFD) has drastically increased. This has led to an extensive use of polynomial chaos (PC) methods, which are known for their ability to propagate efficiently uncertainties through complex engineering models.

Initially, the PC applications were highly intrusive in the sense that the PC expansion was inserted in the partial differential equations describing the problem. Some applications of intrusive PC are available in [2, 3]. Though, it turned out that, in addition to being error prone, a lot of effort was required to modify the CFD code. This made intrusive PC less attractive for industry who are relying on their own well-validated CFD code. It is for this reason that a focus was made on non-intrusive techniques, where no change to the CFD software is required.

In non-intrusive PC, the PC coefficients are computed using either a projection or a regression approach [4]. In both cases, the stochastic solution is calculated by running a series of deterministic simulations for different realization of the uncertain input parameters. The exact number of calls to the CFD software, also referred to as the number of samples, depends on the PC expansion order but also, and most importantly, on the number of input random variables. For a given PC order, the number of samples required grows exponentially with the number of dimension. In the literature, this exponential increase of computational cost with the number of random dimensions is often referred to as the *curse-of-dimensionality*. The latter issue constitutes a serious brake in the application of non-intrusive PC to relevant industrial applications, which are inherently characterized by many uncertainties, e.g. uncertainties on operational conditions, geometrical uncertainties, etc. Hence, in order to handle this issue, efficient non-intrusive techniques have been developed in the last few years.

To takle the curse of dimensionality, an important class of methods, which has been proven particularly effective, relies on the assumption that the PC solution is sparse. This means that only a limited number of features will contribute significantly to the modeling of the model response. The most famous techniques relying on such assumption are the sparse regression technique of [5, 6] and the compressive sampling technique [7, 8]. In the former approach, the PC coefficients are calculated using the LARS method while in the later approach, an underdetermined system of equations is solved using ℓ_1 regularization [7].

In this paper, an innovative basis selection technique for building sparse polynomial chaos expansion is further investigated. The methodology combines statistical inference theory with regression-based PC. The use of statistical inference will provide information about the true regression given an estimated regression model based on samples. The latter information is then used to build the sparse PC metamodel iteratively, following a forward-backward strategy. It follows that the developed methodology has two important features, i.e. (i) *robustness* as the dependency with respect to the sampling strategy is removed and (ii) *effectiveness* as most of the terms that are captured will contribute to the true regression model. An extensive comparison with state-of-the-art techniques such as the LARS-based method and the compressive sampling technique is provided.

2 REGRESSION-BASED POLYNOMIAL CHAOS

Suppose Y is the exact model response. Y is function of a set of random variables $\boldsymbol{\xi} = (\xi_1, ..., \xi_d)$, where d is the dimension of the random space. The polynomial chaos theory consists in expanding the exact model response into a series of orthogonal polynomials,

$$Y(\boldsymbol{\xi}) = \sum_{i=0}^{P} u_i \psi_i(\boldsymbol{\xi}) \tag{1}$$

where u_i are the PC coefficients and ψ_i are the PC basis, chosen in accordance with the probability density function of the input random variables following the so-called Askey scheme of polynomials [9]. As an example, if the random variables are uniformly distributed, then the Legendre polynomials are used.

The PC coefficients can be calculated using either a quadrature (projection) or a regression method. In the present work, we focused on the regression approach. The regression method consists in evaluating Equation (1) at different location in the stochastic space and solving the resulting system of equations, i.e.

$$Y\left(\boldsymbol{\xi}^{(j)}\right) = \sum_{i=0}^{P} u_i \psi_i\left(\boldsymbol{\xi}^{(j)}\right), \quad j = 1, ..., n \tag{2}$$

In practice, an overdetermined system of equations is built to avoid the overfitting phenomenon [10]. As a rule of thumb, the number of samples is often chosen as twice the number of PC terms and the system is solved in a least squared sense. The statistical moments are then derived by post-processing the estimated PC coefficients, as detailed in [11].

3 STATISTICAL INFERENCE IN REGRESSION ANALYSIS

In this work, a distinction is made between the population regression model, which is built based on an infinite number of samples, and the estimated regression model, based on a limited number of samples. A convention commonly used is to denote estimated parameters with a "hat" superscript. This convention will be followed throughout this paper. The population regression model is written as

$$Y = \sum_{i=0}^{P} u_i \psi_i + \varepsilon \tag{3}$$

where ε denote the error made by approximating the exact model by the population regression model. In the sequel, it will be assumed that the error is iid. with zero mean and constant variance σ^2. On the other hand, an estimated regression model is given by

$$Y = \sum_{i=0}^{P} \hat{u}_i \psi_i + \hat{\varepsilon} \tag{4}$$

where $\hat{\varepsilon}$, the difference between the exact model and an estimated regression model, is called residual. The primary goal of statistical inference is to derive information about the parameters of the true regression model given the parameters of the estimated regression model. In the following, statistical inference will be used for (i) building confidence intervals on the regression coefficients, (ii) testing the dependency between the exact model response and one specific predictor.

3.1 Confidence intervals

In this section, we build confidence intervals (CIs) on the regression coefficients. A CI always takes the same form, i.e. an estimate $\pm$ critical value $\times$ standard deviation of the estimate. The estimate $\hat{u}_j$ is provided by ordinary least squares (OLS) while it is possible to show that the variance of the regression coefficients can be calculated as [12]

$$\mathbb{V}[\hat{u}] = \sigma^2(\mathbf{\Psi}^T\mathbf{\Psi})^{-1} \tag{5}$$

where $\mathbf{\Psi}$ denotes the design matrix (matrix containing all the features). The variance σ^2 is a population parameter which can be estimated as follows [12]

$$\hat{\sigma}^2 = \frac{1}{\text{DOF}}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \tag{6}$$

which is nothing else but the residual sum of squares (RSS) divided by the number of degrees of freedom (DOF) [12]. It turns out that the CIs for the regression coefficients can be built as

$$u_j \in \left[\hat{u}_j \pm t_{\text{DOF};\alpha/2}\sqrt{\mathbb{V}[\hat{u}_j]}\right] \tag{7}$$

where $t_{\text{DOF};\alpha/2}$ is the critical value.

3.2 Hypothesis testing

The goal of hypothesis testing in regression analysis is to test the dependency between the exact model response Y and a given predictor. Let consider the following estimated regression model

$$\hat{Y} = \sum_{i=1}^{P}\hat{u}_i\psi_i \tag{8}$$

In order to test the dependency between the response and one specific predictor ψ_i, the null and alternative hypotheses are stated as follows

$$\text{H}_0 : u_i = 0 \tag{9}$$
$$\text{H}_1 : u_i \neq 0 \tag{10}$$

To assess the veracity of H_0 a test statistic is defined

$$t_{\hat{u}_i} = \frac{\hat{u}_i}{\sqrt{\mathbb{V}[\hat{u}_i]}} \sim t_{\text{DOF}} \tag{11}$$

which means that the test statistic follows a Student-t-distribution with DOF degrees of freedom. The decision rule is

$$\text{RH}_0 \text{ : if } |t_{\hat{u}_i}| \geq t_{\text{DOF};1-\alpha/2} \tag{12}$$
$$\cancel{\text{RH}}_0 \text{ : if } |t_{\hat{u}_i}| < t_{\text{DOF};\alpha/2} \tag{13}$$

where RH_0 means the null hypothesis is rejected in favour of the alternative hypothesis and $\cancel{\text{RH}}_0$ the null hypothesis is not rejected. In other words, the test statistic is compared with a critical value. If the test statistic lies inside the critical region, then H_0 is rejected.

4 RESULTS AND DISCUSSION

4.1 Oakley & O'Hagan function (d=15)

First the Oakley & O'Hagan function [13] is considered:

$$Y = \mathbf{a_1}^T \boldsymbol{\xi} + \mathbf{a_2}^T \sin(\boldsymbol{\xi}) + \mathbf{a_3}^T \cos(\boldsymbol{\xi}) + \boldsymbol{\xi}^T \mathbf{M} \boldsymbol{\xi} \tag{14}$$

where $\boldsymbol{\xi} = \{\xi_1, ..., \xi_{15}\}$ are independent random variables, normally distributed with zero mean and variance equals one, i.e. $\xi_i \sim \mathcal{N}(0,1)$, $i = 1, ..., 15$. The Oakley & O'Hagan function consists of 15 random dimensions, among which 5 variables contribute significantly to the variability of the output, 5 have a smaller effect and the remaining 5 input variables have almost no effect on the output of interest [14]. The vectors $\boldsymbol{a_j}$, $j = \{1, 2, 3\}$ and the matrix $\mathbf{M}$ are reported at `http://www.jeremy-oakley.staff.shef.ac.uk/psa_example.txt`.

The Oakley & O'Hagan function is seen as the expensive model, that will be modeled by a sparse PC expansion using Hermite polynomials. In the sequel, for the sake of comparison, two sparse PC metamodels will be built. The first one is built using the LARS-based method and the second one using the statistic-based approach. On the one hand, the LARS-based calculation of the PC coefficients is provided by UQLab, a MATLAB-based UQ framework developed by Sudret and Marelli [1]. Regarding the settings, the degree-adaptivity is activated, i.e. the calculation of the PCE coefficients is performed for a range of PC degree and the degree with the lowest leave-one-out (LOO) error is selected [REF]. In addition, no truncation to the set of candidates basis has been applied ($q = 1$). On the other hand, the STAT method is also run for a range of PC degree, among which the best PC metamodel is selected. A cut-off level of 20% is applied for the STAT-CI method. In all cases, the PCE-based methods are run with two quasi-random experimental designs of size $n = 500$ and $n = 750$ respectively. The relative ℓ_2 error is used to measure the accuracy of the resulting PC expansions.

Results are reported in Table 1. It turns out that, using the same experimental design, the STAT method results in a metamodel which is almost one order of magnitude more accurate as compared to the accuracy of the best metamodel calculated with the LARS method. Moreover, the level of sparsity of the resulting PCE is significantly much smaller when the STAT-CI method is used (almost 50% less terms are captured). This confirms that the statistic-based method selects the PC terms in a smarter way, which can result in large savings in terms of computational cost.

	LARS-MCV		**STAT-MCV**		**STAT-CI**	
	n=500	n=750	n=500	n=750	n=500	n=750
Relative ℓ_2 error	$2.0 \cdot 10^{-2}$	$4.6 \cdot 10^{-3}$	$3.8 \cdot 10^{-3}$	$7.3 \cdot 10^{-4}$	$3.4 \cdot 10^{-3}$	$4.9 \cdot 10^{-4}$
Level of sparsity	204	341	234	374	135	189
PC order	3	4	4	5	4	5

Table 1: Oakley function - Comparison of different adaptive methods for building sparse PC expansions.

More detailed comparisons are reported in Figure 1. In that Figure, the magnitude of the estimated PC coefficients is plotted against a reference solution. The reference solution is given by a full PCE of order 4 where PC coefficients are computed with regression (requiring 7752 samples). For the sake of comparison only, the LARS algorithm is run using a PC degree of 4 (which is not optimal as shown in Table 1). It is shown that, using only 500 samples, the STAT

method is able to capture exactly the most important contributions while the LARS method also captures lots of irrelevant contributions. It is therefore not surprizing that the STAT method results in better performance as compared to the LARS method.

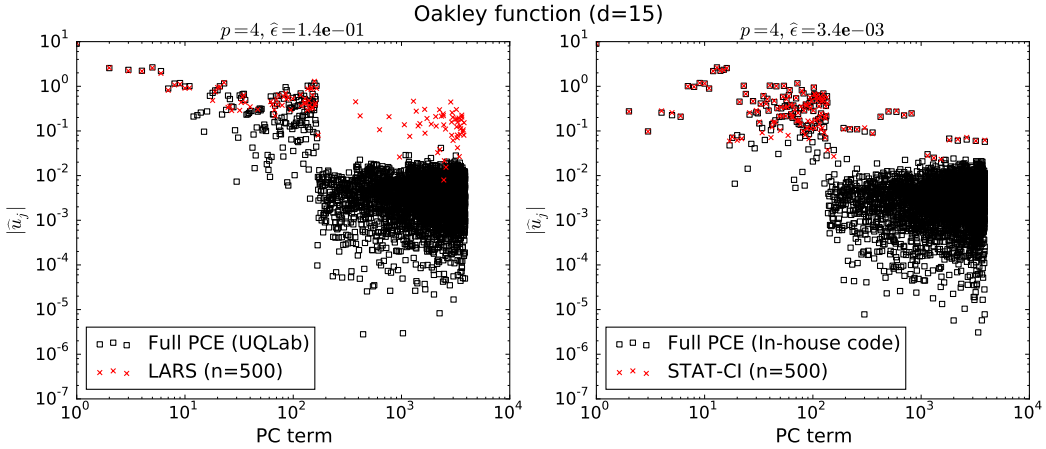

Figure 1: Oakley function - Magnitude of estimated regression coefficients versus a full PCE of order 4 (Left: LARS, Right: STAT)

Eventually, a comparison is made between the STAT method and compressive sampling, using the same experimental design (n=500). Results are reported in Figure 2. It is shown that the compressive sampling technique captures some important contributions. Roughly, the coefficients whose order of magnitude lies in the range $1-10$ are correctly captured. Though CS completely fails in capturing the PC coefficients with lower order of magnitude, resulting in a metamodel of poor quality compared to the one computed with the STAT-CI method.

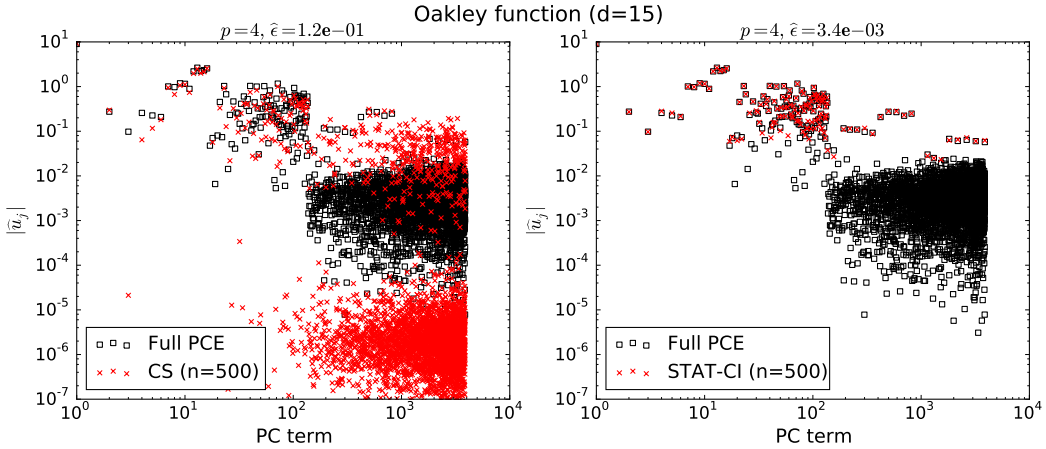

Figure 2: Oakley function - Comparison between the STAT-CI method and compressive sampling

4.2 Morris function (d=20)

The second function considered is another high dimensional function, namely the so-called Morris function [14]:

$$Y = \beta_0 + \sum_{i=1}^{20} \beta_i w_i + \sum_{i<j}^{20} \beta_{ij} w_i w_j + \sum_{i<j<k}^{20} \beta_{ijk} w_i w_j w_k + \sum_{i<j<k<\ell}^{20} \beta_{ijk\ell} w_i w_j w_k w_\ell \tag{15}$$

where

$$w_i = \begin{cases} 2\left(1.1\frac{\xi_i}{\xi_i+0.1} - 0.5\right) & \text{if } i = 3,5,7 \\ 2(\xi_i - 0.5) & \text{otherwise.} \end{cases} \qquad \xi_i \sim \mathcal{U}(0,1) \tag{16}$$

and

$$\begin{cases} \beta_i = 20 & \text{for } i = 1,...,10 & \text{and } \beta_i = (-1)^i & \text{otherwise} \\ \beta_{ij} = -15 & \text{for } i = 1,...,6 & \text{and } \beta_{ij} = (-1)^{i+j} & \text{otherwise} \\ \beta_{ijk} = -10 & \text{for } i = 1,...,5 & \text{and } \beta_{ijk} = 0 & \text{otherwise} \\ \beta_{ijk\ell} = 5 & \text{for } i = 1,...,4 & \text{and } \beta_{ijk\ell} = 0 & \text{otherwise} \end{cases} \tag{17}$$

The Morris function consists of 20 random dimensions, uniformly distributed over [0,1]. A detailed sensitivity analysis of the Morris function is available in [14]. The Morris function is seen as the expensive model which will be replaced by a sparse PC-based metamodel using Legendre polynomials. A similar study is performed as compared to the previous test case. The PC-based methods are run with two quasi random designs of experiments ($n = 500$, $n = 750$). Again, a cut-off level of 20% is applied for the STAT-CI method.

Results are shown in Table 2. In that case, the STAT-CI method still outperforms the LARS method but the gap between both approaches is less significant. The gain in terms of accuracy reaches 27% (resp. 34%) using a design of experiments made of 500 (resp. 750) individuals.

	LARS-MCV		**STAT-MCV**		**STAT-CI**	
	n=500	n=750	n=500	n=750	n=500	n=750
Relative ℓ_2 error	$8.4 \cdot 10^{-2}$	$5.9 \cdot 10^{-2}$	$1.1 \cdot 10^{-1}$	$5.7 \cdot 10^{-2}$	$6.1 \cdot 10^{-2}$	$3.9 \cdot 10^{-2}$
Level of sparsity	52	93	192	195	53	62
PC order	3	3	3	4	3	4

Table 2: Morris function - Comparison of different adaptive methods for building sparse PC expansions.

The resulting sparse PC expansions ($n = 500$) are now faced with a reference solution, namely a full PCE of order 3 whose PC coefficients were calculated with regression (see Figure 3). Those results are in-line with previous findings, i.e. in contrast with the STAT-based method, the terms captured by the LARS method do not exactly correspond to those calculated by the full PC solution. This will inevitably result in a deterioration in the estimation of the statistical moments.

Eventually, the performance of the STAT-CI method is compared with the compressive sampling technique (see Figure 4). Using the same design of experiments ($n = 500$), the performance of compressive sampling is comparable to the performance of the statistic-based method. The main difference with previous test case is that the order of magnitude of the PC coefficients

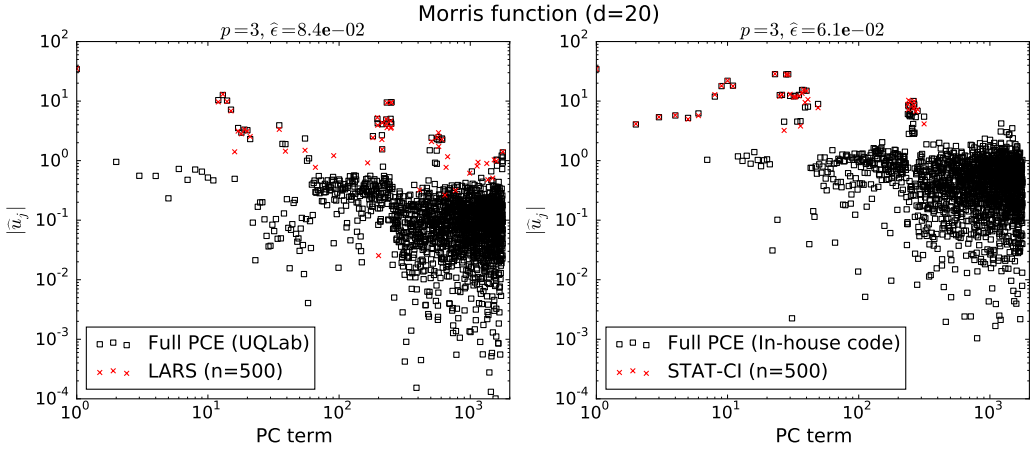

Figure 3: Morris function - Magnitude of estimated regression coefficients versus a full PCE of order 3 (Left: LARS, Right: STAT)

is greater than in the previous test case. Those terms are perfectly capture by the compressive sampling technique.

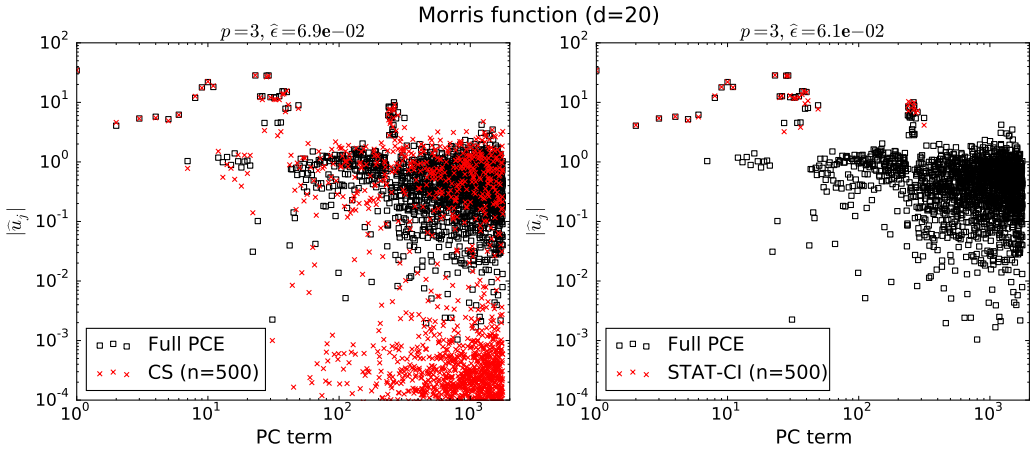

Figure 4: Morris function - Comparison between the STAT-CI method and compressive sampling

5 CONCLUSIONS

In this paper, the performance of a novel basis selection technique are further investigated. The developed methodology shows good potential in building sparse PC expansions at a reduced computational expense. Moreover, it has two essential features, i.e. (i) robustness with respect to the experimental design and (ii) effectiveness in the sense that only terms that truly contribute to the true regression model are captured.

In order to show the effectiveness of the proposed method, two high-dimensional analytical test cases are considered, namely the Oakley & O'Hagan function ($d = 15$) and the more challenging Morris function ($d = 20$). In each case, at equal settings, the accuracy of the sparse PC metamodel shows significant improvement as compared to existing state-of-the-art techniques. Particularly, the results confirm the superiority of the statistic-based approach over the

LARS-based approach. In addition, it is confirmed that the terms captured by the statistic-based method correspond exactly to the most important features calculated by a full PC expansion.

In the future, the method will be applied to relevant industrial applications.

REFERENCES

[1] S. Marelli, B. Sudret. UQLab, The framework for Uncertainty Quantification. Retrieved February 26, 2016, from `http://www.uqlab.com`

[2] C. Lacor, S. Smirnov. Uncertainty propagation in the solution of compressible Navier-Stokes equations using polynomial chaos decomposition. In *CD Rom Proc. of NATO AVT symposium*, page 13, Athens, December 2007.

[3] C. Lacor, S. Smirnov. Non-deterministic compressible Navier-Stokes simulations using polynomial chaos. In *Proc. ECCOMAS Conf*, Venice, July 2008.

[4] M.P. Petterssen, G. Iaccarino, J. Nordstrm. *Polynomial Chaos Methods for Hyperbolic Partial Differential Equations*. Springer 2015. ISBN 978-3-319-10714-1

[5] G. Blatman, B. Sudret. Sparse polynomial chaos expansions and adaptive stochastic finite elements using a regression approach. *C.R. Mecanique*, **336**:518523, 2008.

[6] G. Blatman, B. Sudret. Adaptive sparse polynomial chaos expansion based on Least Angle Regression. *J. Comput. Phys.*, **230**:23452367, 2011.

[7] A. Doostan, H. Owhadi. A non-adapted sparse approximation of PDEs with stochastic inputs. *J. of Comput. Phys.*, **230**(8):30153034, April 2011.

[8] J. Hampton, A. Doostan. Compressive sampling of polynomial chaos expansions: Convergence analysis and sampling strategies. *J. Comput. Phys.*, **280**:363386, 2015.

[9] R. Askey and J. Wilson. *Some Basic Hypergeometric Orthogonal Polynomials That Generalize Jacobi Polynomials*. AMS, 1985.

[10] G. Blatman. *Adaptive sparse polynomial chaos expansions for uncertainty propagation and sensitivity analysis*. PhD thesis, Universite Blaise Pascal, Clermont Ferrand, 2009.

[11] B. Sudret. Global sensitivity analysis using polynomial chaos expansions. *Reliab. Eng. Syst. Safe.*, **93**(7):964979, 2008.

[12] J.O. Rawlings, S.G. Pantula, D.A. Dickey. *Applied Regression Analysis: A Research Tool*. Springer Texts in Statistics. Springer, 2nd edition, 2001. ISBN: 978-0-387-98454-4.

[13] J. Oakley, A. OHagan. Probabilistic sensitivity analysis of complex models: a Bayesian approach. *J Royal Stat Soc*, Series B 2004;**66**:751 769.

[14] B. Sudret, C.V. Mai. Computing derivative-based global sensitivity measures using polynomial chaos expansions. *Reliab. Eng. Syst. Safe.*, **134**:241250, 2015.