An intelligent polynomial chaos expansion method based upon features selection

Wei Zhang¹, Qiang Wang² and Chao Yan*¹

¹ School of Aeronautic Science and Engineering, Beihang University, China
² China Academy of Aerospace Aerodynamics, China

E-mail: yanchao@buaa.edu.cn

Abstract. Polynomial chaos expansion (PCE) method is a common tool for uncertainty quantification (UQ) in fluid mechanics. However, there exists ‘Dimensional Curse’ when the parameters dimension is very high, and large samples are required to solve the PCE function. This would hinder the application of PCE in high dimensions. An intelligent PCE method based on the idea of features selection in machine learning is proposed in this paper. Therefore, only several important features will be selected to construct the PCE function, then fewer samples will be needed to solve the model, and it will be more efficient. Several benchmark functions and an RAE2822 airfoil flow case are utilized to verify the UQ capability of the intelligent PCE. It is proved to be more efficient than the original PCE, with nearly same accuracy.

1. Introduction
Uncertainties (i.e. aerodynamic environment, shape, engine thrust, etc.) are common throughout the entire life cycle of the aircraft. These uncertainties are vital to the aircraft quality. Nowadays, the main uncertainty quantification (UQ) methods consist of Polynomial Chaos Expansion (PCE) method[1]-[2], Monte Carlo Simulation (MCS) method[3], metamodeling method[4], etc. Among them, the PCE method has been widely used in UQ of many fields (i.e. finite deformation[5], environmental acoustics[6], heat transfer[7], fluid mechanics[8], etc.). There are two categories of PCE, including the intrusive PCE and the non-intrusive PCE. The readers can refer to Ref[9] for more information. All the mentioned PCE methods in this paper are the non-intrusive PCE without explanation. The Sobol indices method[10], designed for parameters sensitivity analysis, is also developed based upon the PCE method.

However, not each feature term in the PCE model is efficient and important. Here, as exhibited in Table 1, an example dataset from Ref[2] will be borrowed here, and utilized to view the phenomenon. The Sobol indices of the geometric parameters (αc, Rn and Rs) are analysed to observe the parameters sensitivity on the laminar maximum surface heat flux (QL) of the Mars entry vehicles. The Sobol process is shown in Table 2. ‘Vterm’ and ‘S_i,term’ are the variance and the Sobol index of each feature in the PCE model. It is easy to draw the following summaries. Some features of the model are crucial. For example, the Sobol index of the second order term of αc reaches nearly 52%. However, there are still some features not important and could be neglected. For instance, the Sobol indices of the second interaction terms of αc & Rn, Rn & Rs, and R, & Rs are very small, as low as 0.07% to 1.58%. When if these terms are removed from the PCE model, the total variance predicted will changed from 5.2514 to 5.1573, with only about 1.8% relative error. In this case, the accuracy of the uncertainty quantified via
PCE without the non-important terms, still can be guaranteed. Moreover, the required samples number would be reduced due to the PCE features reduction, which is more efficient. Therefore, this article will focus on how to construct the intelligent PCE model based on features selection.

| num | $a_c$ | $R_n$ | $R_s$ | $Q_L$ | num | $a_c$ | $R_n$ | $R_s$ | $Q_L$
|-----|-------|-------|-------|-------|-----|-------|-------|-------|-------|
| 1   | 68.16 | 1130.92 | 116.45 | 65.32 | 11  | 77.00 | 1071.71 | 124.34 | 72.03 |
| 2   | 63.74 | 1119.08 | 132.24 | 69.86 | 12  | 74.79 | 1178.29 | 112.50 | 70.70 |
| 3   | 69.63 | 1190.13 | 136.18 | 64.85 | 13  | 66.68 | 1036.18 | 115.13 | 68.81 |
| 4   | 71.11 | 1225.66 | 119.08 | 67.44 | 14  | 68.89 | 1095.40 | 137.50 | 64.61 |
| 5   | 73.32 | 1237.50 | 129.61 | 69.12 | 15  | 70.37 | 1142.76 | 126.97 | 66.00 |
| 6   | 71.84 | 1012.50 | 120.39 | 67.99 | 16  | 65.95 | 1213.82 | 128.29 | 65.04 |
| 7   | 72.58 | 1048.03 | 130.92 | 68.58 | 17  | 67.42 | 1059.87 | 125.66 | 67.39 |
| 8   | 65.21 | 1024.34 | 134.87 | 70.75 | 18  | 63.00 | 1107.24 | 121.71 | 71.14 |
| 9   | 64.47 | 1201.97 | 117.76 | 66.86 | 19  | 75.53 | 1154.61 | 133.55 | 69.89 |
| 10  | 76.26 | 1166.45 | 123.03 | 70.96 | 20  | 74.05 | 1083.55 | 113.82 | 69.95 |

Table 1. Example data for PCE model [2]

Table 2. Specific Sobol indices of the polynomial chaos model

| Features | Total | $a_c$ | $R_n$ | $R_s$ | $a_c$ & $a_s$ | $R_n$ & $R_s$ | $a_c$ & $R_n$ | $R_s$ | $a_c$ & $R_s$ | $R_n$ & $R_s$ | $R_n$ & $R_s$ & $R_s$ & $R_s$
|-----------|-------|-------|-------|-------|---------------|---------------|---------------|-------|---------------|---------------|---------------|---------------|-------|
| $V_{err}$ | 5.2514 | 0.5163 | 0.3685 | 0.1131 | 2.7499 | 0.1714 | 0.0832 | 0.0071 | 0.0038 |
| $S_{err}$ | 100% | 9.83% | 7.02% | 2.15% | 52.37% | 3.26% | 1.58% | 0.14% | 0.07% |

The paper will be arranged as below. Section 1 gives an overview of the PCE method, and introduces the importance of features selection in the PCE construction. In Section 2, the principles of the original and the intelligent PCE models are briefly demonstrated. After that, several benchmark functions are utilized to verify the intelligent PCE model in Section 3. Section 4 applies this proposed model to the UQ of RAE2822 airfoil. Finally, summaries are concluded in Section 0.

2. Fundamental principles

2.1. The Original PCE Model
The PCE model can be treated as a linear combination of series of orthogonal basis $\psi(\xi)$, as shown in Equation (1). $c_i$ is the relative coefficients to be solved. $\xi = (\xi_1, ..., \xi_n)$ is an n-dimensional standard random vector, and $\psi(\xi)$ is the product of orthogonal basis functions of related stochastic parameters.

$$Y^* \approx \sum_{i=0}^{p} c_i\psi_i(\xi) \quad (1)$$

The selection of orthogonal basis function is related to the probability distribution of corresponding stochastic parameter, and can be observed in Table 3.

Table 3. Relationship between the orthogonal basis function and the probability distribution

| probability distribution | Orthogonal basis function |
|--------------------------|---------------------------|
| Normal                   | Hermite                   |
| Uniform                  | Legendre                  |

It is infinite theoretically for series of Equation (1), whereas the PCE model is truncated for implementation. The total number of samples $N_{SC}$ will be obtained from Equation (2). $D$ is the PCE truncated order, $n_D$ is the oversampling ratio, and is recommended as 2 in Ref[11].

$$N_{SC} = n_D \cdot (P + 1) = n_D \left[ \frac{(n + D)!}{n!D!} \right] \quad (2)$$

The samples set will be obtained from Latin hypercube design (LHD)[12]. The PCE model is rewritten in matrix form in Equation (3), and the least quadratic regression is utilized for functions solving.
The mean $\mu_{y^*}$ and standard deviation $\sigma_{y^*}$ of the stochastic response $Y^*$ can be achieved in Equations (4)-(5), and the readers can refer to Ref[1] for specific derivation process.

$$
\mu_{y^*} = E[Y^*(\xi)] = c_0
$$

$$
\sigma_{y^*}^2 = D[Y^*(\xi)] = \sum_{i=1}^{p} c_i^2 \langle \psi_i^2 \rangle
$$

2.2. The Intelligent PCE Model

The novel proposed PCE model should be intelligent to find which features are crucial, and are suitable to be added into the model. In machine learning, there are three different categories of features selection, including the filter selection, the wrapper selection, and the embedding selection. The idea of the wrapper selection is utilized in the proposed intelligent PCE model in this paper, and the fundamental principles of model construction are refined as below.

**Step 1.** Introduce the feature terms into the model step by step, and only the tested most significant feature will be introduced at a time.

Assuming it is the $k^{th}$ introduction. The activated feature terms in model are $\psi_1, \psi_2, \ldots, \psi_r$, and the features to be selected are $\psi_{r+1}, \ldots, \psi_p$, the temporary model $y_k$ can be written as in Equation (6). $C$ is the model coefficients vector, and $\varepsilon$ is the approximation error. $N_t$ in Equation (7) is the number of samples utilized for the intelligent PCE construction, to cover all the possible features in model.

$$
\boldsymbol{\Psi}' = \begin{bmatrix}
1 & \psi_1 & \cdots & \psi_r
\end{bmatrix}
\boldsymbol{\Psi}'_k = \begin{bmatrix}
\psi_{r+1} & \cdots & \psi_p
\end{bmatrix}
\boldsymbol{y}_k = \boldsymbol{\Psi}' \boldsymbol{C} + \varepsilon = \boldsymbol{\Psi}' \boldsymbol{C}
$$

$$
N_t = (P + 1) = \frac{(n + D)!}{n!D!}
$$

The sum of squares of deviation $S_T$ of this temporary model can be decomposed as follows. $S_r$ is the sum of squares of regression, and $S_e$ is the sum of squares of residuals.

$$
S_T = S_r(\psi_1, \psi_2, \ldots, \psi_r) + S_e(\psi_{r+1}, \psi_{r+2}, \ldots, \psi_p)
$$

After that, add the remaining features for selection into the temporary model in Equation (6), sequentially and respectively. Taking the term $\psi_{r+1}$ as an example, the new model will be:

$$
y_{k+1}' \approx (\boldsymbol{\Psi}_{k+1}', \psi_{r+1}) (\boldsymbol{C}^T c_{r+1}^T)^T
$$

And the sum of squares of deviation of this newly temporary model can be re-divided as:

$$
S_T = S_r(\psi_1, \psi_2, \ldots, \psi_r, \psi_{r+1}) + S_e(\psi_{r+2}, \ldots, \psi_p, \psi_{r+1})
$$

Afterwards, F-test will be performed to test whether the feature $\psi_{r+1}$ is significant enough to be added into the model. As shown in Equation (11), the statistical value of this F-test is recorded as $F_{r+1}$. 

$$
F_{r+1} = \frac{S_r(\psi_1, \psi_2, \ldots, \psi_{r+1})}{S_e(\psi_{r+2}, \ldots, \psi_p, \psi_{r+1})}
$$
Similarly, the statistical values for these alternative features can be obtained and recorded as \( F_{r+1} \), \( F_{r+2} \), \ldots, \( F_r \). \( F_{r+1} \) is assumed as the maximum among them for convenience. \( \psi_{r+1} \) is the corresponding feature. An activating threshold is marked as \( F_{act}(1, n-r-2) \) at the activating significance factor \( \alpha_{act} \) and can be obtained from existing statistical tables. Thereafter, \( F_{r+1} \) is compared with \( F_{act}(1, n-r-2) \) to decide if the most significant feature \( \psi_{r+1} \) will be added or not. When \( F_{r+1} \) is greater, \( \psi_{r+1} \) will be introduced into the new model. Otherwise, no features will be added, and the construction will be ended.

\[
\psi_{r+1} \text{ activated, if } F_{r+1} > F_{act}(1, n-r-2) \\
\text{end if } F_{r+1} \leq F_{act}(1, n-r-2)
\]  

(12)

**Step 2.** Test all the features in the model to remove the least crucial feature, after each feature added.

Assume that a new feature is just imported into the model, and we mark these temporarily activated features as \( \psi_1, \psi_2, \ldots, \psi_{k+1} \). Afterwards, these features need to be checked via \( F \)-test to find if they could be excluded or not. The relative statistical values are recorded as \( F_1, F_2, \ldots, F_{k+1} \). Among them, \( F_{k+1} \) is supposed as the minimum one. \( \psi_{k+1} \) is the corresponding feature. Another threshold is marked as \( F_{act}(1, n-k-2) \) at an excluding significance factor \( \alpha_{act} \). By comparison, the feature \( \psi_{k+1} \) will be excluded from the temporary PCE model, if \( F_{k+1} \) is smaller than the excluding threshold.

\[
\psi_{k+1} \text{ excluded, if } F_{k+1} \leq F_{act}(1, n-k-2)
\]  

(13)

**Step 3.** Repeat steps 1-2, until no features can be added, and no features can be eliminated.

**Step 4.** After construction of the intelligent PCE model, the uncertainty characteristics of the stochastic response can be acquired by using Equations (4)-(5).

### 3. Benchmark testing

Several benchmark functions are used to test the property of the intelligent PCE model on UQ.

#### 3.1. Test 1: Ackley function

The first benchmark is the Ackley function, as shown in Equation (14). The stochastic parameters \( x_1 \) and \( x_2 \) are assumed to be uniformly distributed in interval \([-2, 2]\), as exhibited in Table 4. The truncated order of the two PCE models (the original and the intelligent ones) are both set as 4. Thus, the numbers of the samples set are severally 30 and 15 for the original and the intelligent PCE models. These two sample sets are arranged from the LHD design. A Monte Carlo Simulation (MCS) carried for 100000 times is treated as standard. These three sample sets are demonstrated in Figure 1.

\[
A(x) = -20 \exp \left( -0.2 \sqrt{0.5(x_1^2 + x_2^2)} \right) - \exp \left( 0.5 \left( \cos(2\pi x_1) + \cos(2\pi x_2) \right) \right) + 20 + e
\]  

(14)

**Table 4. Parameters variation for Ackley function**

| Parameters Interval range Distribution | Parameters Interval range Distribution |
|--------------------------------------|--------------------------------------|
| \( x_1, x_2 \) \([-2,2]\) Uniform   | \( x_1, x_2 \) \([-2,2]\) Uniform   |

**Table 5. Comparison between the PCE models in Test I**

| Features Num. | Mean Value | Standard Deviation |
|---------------|------------|--------------------|
|               | Responses  | Error (%)          | Responses  | Error (%)          |
| MCS (Standard)| /          | /                  | 1.419      | /                  |
| Original PCE  | 15         | +0.39%             | 1.321      | -6.88%             |
| Intelligent PCE| 8          | -1.79%             | 1.488      | +4.90%             |

In the intelligent model, the activating and excluding thresholds are severally set as 0.20 and 0.30. As shown in Table 5, uncertainty quantified from the original and the intelligent PCE models are...
compared to that of MCS. The number of features in the PCE models are also given. Obviously, only half features, which are crucial, of the original model, are selected in the intelligent model. Besides, the UQ capabilities of two models are comparable. Therefore, the intelligent model is more efficient.

![Figure 1. Samples Set for Test I.](image)

3.2. Test II: Rosenbrock function (2-dim)

The second benchmark is the 2-dimensional Rosenbrock function, as shown in Equation (15). The stochastic parameters are exhibited in Table 6 with their distributions. The PCE models are still truncated to the fourth order. The arithmetic results of another MCS with 100000 random samples is assumed as standard. As shown in Table 7, uncertainty quantified from the two PCE models are compared to the standard values of MCS. Similar conclusions as above could be drawn from Table 7.

\[
y = (x_i - 1)^2 + 100(x_i^2 - x_{i+1})^2
\]

(15)

| Parameters variation for Rosenbrock function (2-dim) |
|-----------------------------------------------|
| Parameters | Interval range | Distribution |
|-------------|----------------|--------------|
| $x_1 \sim x_2$ | [0, 2] | Uniform |

| Features Num. | Mean Value | Standard Deviation |
|----------------|------------|-------------------|
| MCS (Standard) | 188.579 | 260.326 |
| Original PCE | 15 | 187.000 | 255.549 |
| Intelligent PCE | 9 | 187.865 | 258.096 |

3.3. Test III: Rosenbrock function (6-dim)

The third benchmark is the 6-dimensional Rosenbrock function, as shown in Equation (16). The stochastic parameters are exhibited in Table 8 with their distributions. The PCE models are truncated to the third order. The numbers of samples for the original and the intelligent PCE models are respectively 168 and 84. The arithmetic results of a third MCS with 100000 random samples is treated as standard. As shown in Table 9, still similar conclusions with the above tests could be drawn.

\[
y = \sum_{i=1}^{5} \left[ (x_i - 1)^2 + 100(x_i^2 - x_{i+1})^2 \right]
\]

(16)

| Parameters variation for Rosenbrock function (6-dim) |
|-----------------------------------------------|
| Parameters | Interval range | Distribution |
|-------------|----------------|--------------|
| $x_1 \sim x_6$ | [0, 2] | Uniform |

| Features Num. | Mean Value | Standard Deviation |
|----------------|------------|-------------------|
| MCS (Standard) | 935.1975 | 564.9914 |
| Original PCE | 84 | 936.3557 | 548.6624 |
| Intelligent PCE | 33 | 938.7748 | 544.8881 |
4. Application in Fluid Mechanics
An experiment of RAE2822 airfoil[13] was selected to preliminarily verify the UQ capability of the proposed intelligent PCE model in aerodynamic problems. The baseline condition for RAE2822 is shown in Table 10. The pressure distribution on airfoil surface is exhibited in Figure 2. It can be drawn that the simulation agrees well with the experimental data.

Table 10. Baseline condition for RAE2822 airfoil

| $Ma_\infty$ | $\alpha^\circ$ | $Re/m$ |
|-------------|---------------|--------|
| 0.73        | 2.79          | 6.5$\times10^6$ |

Table 11. Perturbations for RAE2822 airfoil

| Perturbation Sources | Interval Range | Distribution |
|----------------------|----------------|--------------|
| $Ma_\infty$          | [0.70, 0.76]   | Uniform      |
| $\alpha^\circ$       | [2.59, 2.99]   | Uniform      |

Table 12. Comparison between lift uncertainty

| $C_l$ Features Num. | Mean value | Standard deviation |
|---------------------|------------|--------------------|
| Original PCE 15     | 0.7632     | 0.03273            |
| Intelligent PCE 8   | 0.7634     | 0.03303            |
| Relative error /     | +0.03%     | +0.92%             |

Table 13. Comparison between drag uncertainty

| $C_d$ Features Num. | Mean value | Standard deviation |
|--------------------|------------|--------------------|
| Original PCE 15    | 0.1732E-1  | 0.5552E-2          |
| Intelligent PCE 8  | 0.1732E-1  | 0.5554E-2          |
| Relative error /    | +0.00%     | +0.04%             |

Perturbations are assumed at the freestream Mach number and angle of attack, as demonstrated in Table 11. The original and the intelligent PCE models are utilized to quantify the aerodynamic uncertainty (i.e. lift uncertainty, and drag uncertainty) of RAE2822 under perturbations. The sample numbers are 30 and 15 respectively for the two models. The activating and excluding significance factors are still set as 0.20 and 0.30 for the intelligent PCE model. Results are compared in Table 12-Table 13. It can be viewed that, although with fewer features and fewer samples, both the uncertainty responses of lift and drag quantified via the intelligent PCE model are similar and comparable to those of the original model. Therefore, the application of the intelligent PCE method based on features selection in aerodynamic uncertainty quantification has been preliminarily verified.

![Figure 2. Comparisons of the pressure coefficient distribution on the airfoil surface.](image)

5. Conclusions
An intelligent polynomial chaos expansion method based upon features selection has been presented in this paper. Several benchmark functions are utilized to verify the uncertainty quantification (UQ) property of this method. After that, this intelligent PCE method is applied to quantify the aerodynamic uncertainty of RAE2822 airfoil. The original PCE method is also performed for comparison. Conclusions are summarized as below.
a) The novel proposed PCE method is intelligent to grasp the crucial features in its model construction.
b) The UQ capability of the intelligent PCE model is similar and comparable with the original PCE model.
c) The intelligent PCE model is more efficient than the original model. This intelligent model only needs half samples of the original one to solve the model function.
d) The application of the intelligent PCE model in the aerodynamic uncertainty quantification has been preliminarily validated.

In the future works, attentions will be paid to application of multiple features selection strategies, and realization of the sequential sampling to further the performance of the intelligent PCE model.

6. Acknowledgments
This work was supported by grants from the National Natural Science Foundation of China (No. 11721202).

References
[1] Zhang W, Wang S, Wang Q, Ju S and Yan C, 2020, Reverse jet parameters study on aerodynamic thermal uncertainty of a blunt body, Aerosp. Sci. Technol. 107: 106260.
[2] Ju S, Yan Y, Wang X, Qin Y and Ye Z, 2018, Sensitivity analysis of geometric parameters upon the aerothermodynamic performances of mars entry vehicle, Int. J. Heat Mass Transf. 120: 597–607.
[3] Matala A, 2008, Sample size requirement for monte carlo simulations using latin hypercube sampling, Mat-2.4108 Independent Research Projects in Applied Mathematics, Helsinki Univ. of Technology.
[4] Martinelli M and Ducigneau R, 2010, On the use of second-order derivatives and metamodel-based monte-carlo for uncertainty estimation in aerodynamics, Computers & Fluids, 399: 953-64.
[5] Acharjee S and Zabaras N, 2007, A non-intrusive stochastic galerkin approach for modeling uncertainty propagation in deformation processes, Computer & Structures, V85: 244-54.
[6] Finette S, 2006, A stochastic representation of environmental uncertainty and its coupling to acoustic wave propagation in ocean wave guides, J ACOUST SOC AM, 120: 2567-79.
[7] Xiu D and Karniadakis G, 2003, A new stochastic approach to transient heat conduction modeling with uncertainty, Int. J. Heat Mass Transf, 46: 4681-93.
[8] Walters R, 2003, Towards stochastic fluid mechanics via polynomial chaos, 41st AIAA aerospace sciences meeting & exhibit, Reno, Neveda.
[9] Xiong F, Yang S, Liu Y and Chen S, 2015, Engineering probabilistic uncertainty analysis method, Beijing, Beijing Science Press. (in Chinese)
[10] Zhao Y, Yan C, Wang X and Liu H, 2019, Uncertainty and sensitivity analysis of sst turbulence model on hypersonic flow heat transfer, Int. J. Heat Mass Transf, 136: 808-20.
[11] Hosder S, Walters R and Balch M, 2007, Efficient sampling for non-intrusive polynomial chaos applications with multiple uncertain input variables, 48th AIAA/ASME/ASCE/AHS/ASC Structures, Structural Dynamics, and Materials Conference, Honolulu, Hawaii.
[12] Ju S, Yan C and Ye Z, 2017, Multi-parameter sensitivity analysis of air-breathing hypersonic vehicle, J. Beijing Univ. Aeronautics Astronautics, 43: 927-35. (in Chinese)
[13] Wang X, 2000, Research on genetic algorithm and its application in aerodynamic optimization design, PhD thesis, Northwestern Polytechnical University, Xi’an. (in Chinese)