Model Reduction by PCA and Kriging

G. Aversano[1,2,3,b)], Z. Li[1,2,c)], O. Gicquel[3,d)] and A. Parente[1,2,e)]

[1]*Université Libre de Bruxelles, Aero-Thermo-Mechanics Departement, Avenue F.D. Roosevelt 51, CP 165/41, 1050 Brussels, Belgium*
[2]*Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*
[3]*Laboratoire EM2C, CNRS, Centrale-Supélec, Université ParisSaclay, 92295 Chatenay-Malabry Cedex, France*

[a)]Corresponding author: Gianmarco.Aversano@ulb.ac.be
[b)]Gianmarco.Aversano@ulb.ac.be
[c)]Zhiyi.Li@ulb.ac.be
[d)]Olivier.Gicquel@centralesupelec.fr
[e)]Alessandro.Parente@ulb.ac.be

Abstract. Combustion systems are characterized by very complex physical interactions, between chemistry, fluid dynamics and heat transfer. Detailed numerical simulations of such systems require substantial computational resources that are unfortunately restricted in the industry, limiting their use during the optimization processes or uncertainty quantification studies.

In this study we consider the combination of Principal Component Analysis (PCA) with the Kriging interpolation method to identify accurate low-order models. PCA is used to identify and separate invariants of the system from the variables that are related to the characteristic operating conditions. The Kriging interpolation method is then used to find a response surface for these variables. The methodology has been applied to 2D flames produced by OpenFOAM. The inlet velocity and fuel composition were the input parameters of the system. A set of simulations, for different values of the input parameters, were carried out and a the available solutions were divided into training and validation data. The test data was predicted with an error of approximately 5% for the prediction of the temperature field.

INTRODUCTION

In many engineering applications, complex physical systems can only be described by high-fidelity expensive simulations. Due to the non-linearity of these problems, changing the operating conditions, namely the model's input parameters, can drastically change the state of the considered system.

Our objective is to develop advanced Surrogate Models (SM) that can accurately represent the behavior of complex reacting systems in a wide range of conditions, without the need for expensive CFD simulations. This is particularly attracting for the development of digital counterparts of real systems, with application in monitoring, diagnostics and prognostics [1, 2]. Examples of SMs used in combustion applications can be found in [3] and in [4].

In our approach a specific computationally-expensive CFD simulation or computer code, referred to as Full-Order Model (FOM) [5, 6], is treated as a black box that generates a certain output $\mathbf{y}$ (e.g. the temperature field) given a set of input parameters $\mathbf{x}$ (e.g. the equivalence ratio) and indicated by $\mathbf{y} = \mathcal{F}(\mathbf{x})$. The evaluation of the function $\mathcal{F}(\cdot)$ usually requires many hours of computational time. After enough observations of the FOM's output are available, $\mathbf{y}(\mathbf{x}_i) \quad \forall i = 1, ..., M$, a Surrogate Model (SM) can be trained and the output $\mathbf{y}^*$ for a particular set of unexplored inputs $\mathbf{x}^*$ can be predicted without the need to evaluate $\mathcal{F}(\mathbf{x}^*)$ and, thus, no simulation is run. The function $\mathcal{F}(\cdot)$ is therefore approximated by a new function $\mathcal{M}(\cdot)$ whose evaluation is very cheap compared to $\mathcal{F}(\cdot)$ and predictions are possible as follows: $\mathbf{y}^* = \mathcal{M}(\mathbf{x}^*) \approx \mathcal{F}(\mathbf{x}^*)$.

SMs are generally constructed directly on the analyzed system's output, i.e. directly on the variables of interest like the velocity and temperature fields. For each individual output variable a SM is trained and a response surface is found, indicating the relationship between the variable and the input parameters. If the number of variables of interest is high, many SMs need to be trained. Besides, any correlation between these variables of interest might be lost in the process of training individual SMs: the information about the physics of the phenomena involved is lost. Reducing

the number of SMs to train is possible if the original set of variables can be represented by a new set of fewer scalars. This corresponds to the idea that the original variables are actually realization of unknown *latent* variables [7].

Principal Component Analysis (PCA) [8] and related techniques offer the potential of preserving the physics of the system while reducing the size of the problem. PCA can be used to find a new, smaller set of uncorrelated variables, often referred to as *PCA scores*, which is representative of the original variables of interest. Once these PCA scores are found, a SM can be built for each one of them, giving birth to Reduced-Order (Surrogate) Models (ROMs). SMs usually include interpolation or regression techniques which depend on the choice of some particular design functions. As shown in [6], ROMs are less sensitive to the particular design functions chosen for their construction, which is desirable. These features are what makes PCA-based ROMs very attractive candidates for the development of physics-preserving SMs.

In the present work, the developed SMs were based on the combination of PCA [9] and Kriging [10, 11]. PCA was used to extract the invariant (w.r.t. the input parameters) physics-related information of an investigated combustion system and identify the system's coefficients which instead depend on the operating conditions, the PCA scores. The Kriging interpolation method was then able to find a response surface for these scores. With this strategy it was possible to build a ROM that granted the possibility of parameter exploration with reduced computational cost. The Kriging interpolation method was chosen over other regression techniques because of the encouraging results shown in [6, 12, 13, 14]. However, the application was limited to non-reacting flow problems.

In the paper, the Kriging-PCA approach is extended to combustion applications, to develop a ROM that can faithfully reproduce the temperature and chemical species fields in a reacting flow simulation. The methodology is demonstrated on a methane laminar premixed flame. The configuration of the simulated flame is described in [15] and in [16]. The input parameters were two, namely the inlet velocity and the molar fraction of CH_4 in the inlet stream, which was a mixture of CH_4 and N_2. OpenFOAM was employed to produce a set of training observations (spanning the two input parameters in the range $24 \div 55$ cm/s and $40 \div 100$ % for the inlet velocity and inlet molar fraction of CH_4, respectively) and test data. The chemical mechanism was GRI3.0 with no NO_x. A total of 36 physical variables was considered: 35 chemical species and temperature. The test data was predicted with an error of approximately 5% for the prediction of the temperature field.

THEORY

PCA

Principal Component Analysis (PCA) is a statistical technique that reduces a large number of interdependent variables (i.e. independent up to the second-order statistical moments) to a smaller number of uncorrelated variables, while retaining as much of the original data variance as possible [17, 8, 18, 19, 20]. For a data-set $\mathbf{Y}(M \times N)$, containing M observations of N original variables, PCA finds a set of Principal Components (PCs), collected into a matrix $\mathbf{A} = \{\mathbf{a_1}, \mathbf{a_2}, ...\}$. The PCs are the eigenvectors of the data correlation matrix. Their eigenvalues represent the portion of the data variance that they account for. Thus, the PCs can be sorted in descent order of importance. Only the first few $q < N$ PCs can be retained if they already provide a good approximation of the total data variance. The PCs can be viewed as vectors of weights that map the original variables onto a lower-dimensional manifold. One particular observation or row of $\mathbf{Y}$ can be approximated as $\mathbf{y} \approx \tilde{\mathbf{y}} = \mathbf{z}\mathbf{A}^T$, where $\mathbf{z} \in \mathbb{R}^{q<N}$ is the projection of $\mathbf{y} \in \mathbb{R}^N$ on PCs: $\mathbf{z} = \mathbf{y}\mathbf{A}$.

Kriging

Accurate prediction of the PCA scores at unexplored points $\mathbf{x}^*$ translate into accurate estimation of the original variables as the mapping from $\mathbf{z}(\mathbf{x}^*)$ to $\mathbf{y}(\mathbf{x}^*)$ is known and explained in section e). The data-set $\mathbf{Z} = \{\mathbf{z}(\mathbf{x}_1),\ \mathbf{z}(\mathbf{x}_2),\ ...,\ \mathbf{z}(\mathbf{x}_M)\}$ of PCA scores evaluated at different training points is used to build a response surface in the region spanned by $\mathbf{X}_M$.

Kriging is an interpolation method in which every realization $y(\mathbf{x})$ is expressed as a combination of a trend function and a residual [21]: $y(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon(\mathbf{x}) = \sum_{i=0}^{p} b_i f_i(\mathbf{x}) + \epsilon(\mathbf{x}) = \mathbf{f}^T(\mathbf{x})\mathbf{b} + \epsilon(\mathbf{x})$. The trend function $\mu(\mathbf{x})$ is expressed as a weighted linear combination of $p + 1$ polynomials $\mathbf{f}(\mathbf{x}) = [f_0(\mathbf{x}),\ ...,\ f_p(\mathbf{x})]^T$ with the weights $\mathbf{b} = \left[b_0,\ ...,\ b_p\right]^T$ determined by generalized least squares (GLS). The subscript p also indicates the degree of the polynomial. The residuals $\epsilon(\mathbf{x})$ are modeled by a Gaussian process with a kernel or correlation function that depends on a set of hyper-parameters $\mathbf{h}$ to be evaluated by Maximum Likelihood Estimation (MLE) [21, 11, 10].

The final form of the Kriging predictor for any realization $y(\mathbf{x}^*)$ is $y(\mathbf{x}^*) = \mathbf{f}(\mathbf{x})^T\mathbf{b} + \mathbf{r}(\mathbf{x}^*)^T\mathbf{R}^{-1}(\mathbf{y} - \mathbf{F}\mathbf{b}) = \mathbf{f}(\mathbf{x})^T\mathbf{b} + \mathbf{r}(\mathbf{x}^*)^T\mathbf{g}$, where $\mathbf{F}$ is the matrix of polynomials evaluated at the training locations; $\mathbf{R}$ is the kernel matrix or

matrix of correlations between the training points; and **r** is the vector or correlations between the training points and the point $\mathbf{x}^*$ for which we wish to make a prediction.

RESULTS

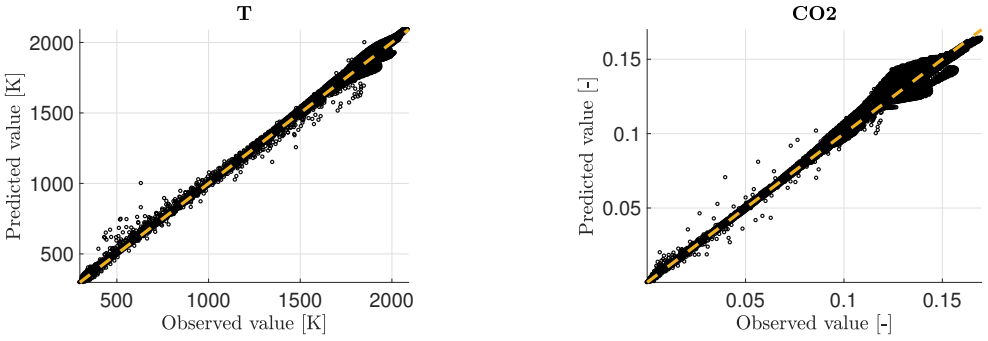

FIGURE 1: (*left*) Parity plot for the predictions of the temperature. (*right*) Parity plot for the predictions of the CO_2. ROM developed by means of PCA and Kriging: quadratic trend function, Matern52 kernel.

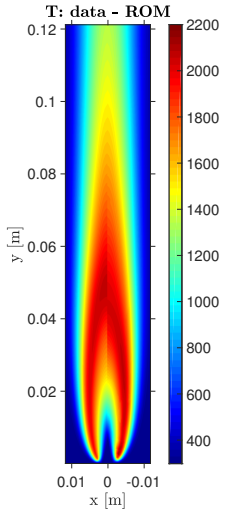

FIGURE 2: Prediction of the temperature field by the developed ROM based on PCA+Kriging: quadratic trend function, Matern52 kernel.

The configuration of the simulated flame is described in [15] and in [16]. The input parameters were two, namely the inlet velocity and the molar fraction of CH_4 in the inlet stream, which was a mixture of CH_4 and N_2. Samples were produced by OpenFOAM, spanning the two input parameters in the range $24 \div 55$ cm/s and $40 \div 100$ % for the inlet velocity and inlet molar fraction of CH_4, respectively. A total of 23 observations were used for the training of the ROM in this region. The chemical mechanism was GRI3.0 with no NO_x. A total of 36 physical variables was considered: 35 chemical species and temperature. Predictions were made for the values 45 % - 35 m/s, 65 % - 30 m/s, 75 % - 40 m/s, 95 % - 40 m/s.

The overall performance of the model for the reconstruction of the original data can be deduced in the parity plots shown in figures 1. These Figures report the parity plots for the reconstructed temperature and CO_2 mass fraction fields, respectively. The prediction of the PCA+Kriging ROM for the temperature field had a mean error of 4%. Figure

2 reports the temperature field predicted by the developed ROM as well as the actual field. These results indicate that in spite the high dimensionality of the problem, the combination of a dimension reduction technique such as PCA with the Kriging interpolation method is a valid candidate for the development of reduced-order surrogate models. PCA provides the identification of the invariant structures of the system as well as a smaller set of coefficients (w.r.t. the original number of variables) that can be interpolated for parameter exploration.

CONCLUSIONS

In this work, a reduced-order surrogate model was trained by combining a dimensionality-reduction technique such as PCA with the Kriging interpolation method. The approach was tested on a 2D flame, described in [15] and in [16], by varying two input parameters, namely the inlet velocity (in the range $24 \div 55$ cm/s) and the molar fraction of CH_4 in the inlet stream (in the range $40 \div 100$ %), which was a mixture of CH_4 and N_2. Results have shown that the ROM developed on PCA+Kriging possessed the required predicting capabilities for the prediction of the considered variables fields (temperature and 35 chemical species). The prediction of this ROM for the temperature field had a mean error of 4%. The present work represents the first application of the Kriging+PCA methodology to combustion problems. As such, it is intended to be a proof of concept that will pave the way for more complex applications where the computational savings will be clearly evident. The ROM's predictions can also be used as starting solutions for a CFD-combustion solver, thus reducing the time needed for convergence.

ACKNOWLEDGMENTS

This project has received funding from the European Unions Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement No 643134. The research was sponsored by the European Research Council, Starting Grant No 714605.

REFERENCES

[1] B. Schleich, N. Anwer, L. Mathieu, and S. Wartzack, CIRP Annals - Manufacturing Technology **66**, 141–144 (2017).
[2] T. H. Uhlemann, C. Schock, C. Lehmann, S. Freiberger, and R. Steinhilper, Procedia Manufacturing **9**, 113–120 (2017).
[3] T. Lancien, N. Dumont, K. Prieur, D. Durox, S. Candel, O. Gicquel, and R. Vicquelin, (2016).
[4] M. Guenot, I. Lepot, C. Sainvitu, J. Goblet, and R. F. Coelho, Engineering Computations **30**, 521–547 (2013).
[5] K. Bizon and G. Continillo, Computers and Chemical Engineering **39**, 22–32 (2012).
[6] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, M. Sidorkiewicz, and P. Villon, Structural and Multidisciplinary Optimization **41**, 555–574 (2010).
[7] C. M. Bishop, *Journal of Chemical Information and Modeling*, Vol. 53 (2013), pp. 1689–1699, arXiv:arXiv:1011.1669v3 .
[8] I. T. Jolliffe, (2002).
[9] S. Karamizadeh, S. M. Abdullah, A. A. Manaf, M. Zamani, and A. Hooman, Journal of Signal and Information Processing **4**, 173–175 (2013).
[10] M. Seeger, *International journal of neural systems*, Vol. 14 (2004), pp. 69–106, arXiv:026218253X .
[11] S. N. Lophaven, J. Søndergaard, and H. B. Nielsen, 1–28 (2002).
[12] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, and P. Villon, Structural and Multidisciplinary Optimization **46**, 129–136 (2012).
[13] M. Xiao, P. Breitkopf, R. Filomeno Coelho, C. Knopf-Lenoir, P. Villon, and W. Zhang, Applied Mathematics and Computation **223**, 254–263 (2013).
[14] M. Xiao, P. Breitkopf, R. F. Coelho, P. Villon, and W. Zhang, Applied Mathematics and Computation **247**, 1096–1112 (2014).
[15] S. Cao, B. Bennett, B. Ma, and D. Giassi, 1–9 (2013).
[16] S. Cao, B. Ma, B. A. Bennett, D. Giassi, D. P. Stocker, F. Takahashi, M. B. Long, and M. D. Smooke, Proceedings of the Combustion Institute **35**, 897–903 (2015).
[17] A. Parente and J. C. Sutherland, Combustion and Flame **160**, 340–350 (2013).
[18] A. Coussement, B. J. Isaac, O. Gicquel, and A. Parente, Combustion and Flame **168**, 83–97 (2016).
[19] K. Bizon, G. Continillo, E. Mancaruso, S. S. Merola, and B. M. Vaglieco, Combustion and Flame **157**, 632–640 (2010).
[20] N. Kambhatla and T. Leen, Neural Computation **9**, 1493–1516 (1997).
[21] P. G. Constantine, E. Dow, and Q. Wang, SIAM Journal of Scientific Computation **36**, 1500–1524 (2014).