Feature Extraction by means of Unsupervised Learning Techniques on LES Combustion Data

G. D'Alessio[1,2,3,b)], G. Aversano[1,2,c)], Z. Li[1,2,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]*Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, P.zza Leonardo da Vinci 32, 20131 Milan, Italy*

[a)]Corresponding author: Giuseppe.Dalessio@ulb.ac.be
[b)]Giuseppe.Dalessio@ulb.ac.be
[c)]Gianmarco.Aversano@ulb.ac.be
[d)]Zhiyi.Li@ulb.ac.be
[e)]Alessandro.Parente@ulb.ac.be

Abstract. Principal Component Analysis (PCA) has shown very promising capabilities from the perspective of reduced-order model development as it provides optimal progress variables that can be used for the parametrization of the reacting system as well as to gain insight on the features of combustion processes. Local PCA is able to find lower-dimensional clusters in the data-space, i.e. regions where a subset of the original variables account for most of the original variance, and thus overcomes the limits of global PCA in dealing with the recurrent non-linearities of combustion problems. Self-Organizing Maps (SOMs) are a class of Artificial Neural Network (ANN) which are used to map an ensemble of high dimensional observations onto a non-linear, lower-dimensional grid. Both methodologies are useful in finding structures in the data manifold and in extracting information that can also be used for mechanism reduction by means of an adaptive kinetic scheme. In this work, Local PCA and SOMs are applied to a data-set obtained from a LES simulation of two co-flow jets: the central jet is an equimolar mixture of CH_4 and H_2; the annulus jet has an oxygen content of 3%. Results show that Local PCA is able to efficiently cluster the data, while also offering a measure for the quality of the clustering process itself. The methodology also extracts information about the dominant variables in each clustered region. On the contrary, SOMs do not provide a measure for the quality of their clustering solution nor do they indicate a subset of dominant variables. Furthermore, the method wrongly detected features in the data manifold.

INTRODUCTION

The increasing detail level of the chemical kinetic schemes developed for the modeling of combustion problems, coupled with the strong non-linearities of Fluid-Dynamics equations, requires huge efforts in terms of computational resources [1]. The high dimensionality of data produced by the implementation of such detailed schemes is also challenging in terms of interpretation. Data mining and feature extraction methodologies have become popular as solutions to this problem in the field of industrial engineering [2, 3]. These techniques can be exploited to reduce computational costs [4] and for easier data interpretation as they grant a lower-dimensional representation of high-dimensional data. Principal Component Analysis (PCA) is one of the most employed techniques for dimension reduction and data encoding [5]. PCA finds a new, reduced set of progress variables, usually referred to as Principal Components (PCs), which are linear combinations of the original variables, that is able to describe the total variation of the data. For inherently or strongly non-linear systems, e.g. in combustion applications, PCA still needs a relevant number of PCs to correctly describe the system [6, 7]. Local PCA [8] can overcome the limits of PCA, which derive from its linear nature, by partitioning the data into regions where a local linear approximation is more suitable. Another unsupervised clustering technique is provided by Self-Organizing Maps (SOMs) [9]. SOMs are a type of artificial neural network (ANN) that produces a low-dimensional representation of the data, which is achieved by spanning a lower-dimensional map of

nodes or neurons in the original data-space. SOMs are non-linear and they are employed for data clustering. These techniques can be employed to find structures in the data-space for a certain system. A structure can be a region of the state-space where a subset of the thermo-chemical variables, such as temperature and species mass fractions, are more relevant in transport and chemical phenomena involved. This information can be utilized for the creation of a reduced adaptive kinetic mechanism as well as to aid data interpretation.

In this paper, both Local PCA and SOMs are applied on a data-set obtained from an LES simulation of two co-flow jets: the central jet is composed of an equimolar mixture of CH_4 and H_2; the annulus jet has an oxygen content of 3% by mass. It is shown how the different spatial locations of the mesh have been separated into clusters according to the thermo-chemical activities which occur in that region. Results show that the Local PCA approach can efficiently partition the state-space as 5 PCs can explain over 99% of the variance in each cluster. SOMs led to the separation of regions that share the same structures. Besides, SOMs do not provide a measure for the quality of the data partitioning solution, neither do they indicate the dominant actors in a certain cluster. On the contrary, Local PCA can quantify the recovered variance and reconstruction error while also being able to find the PVs. Finally, the clustering solutions are also visualized.

THEORY

Local 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 [10, 5, 11, 12]. 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 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. A set of Principal Variables (PVs) can be identified as a subset of the original variables that explain the same amount of variance of the PCs. They are usually found as the ones with higher weights on the first few PCs. This strategy can be used to detect a subset of the original variables One particular observation or row of $\mathbf{Y}$ can be approximated as $\mathbf{y} \approx \tilde{\mathbf{y}} = \mathbf{z}\mathbf{A}^T$, where $\mathbf{z}$ is the projection of $\mathbf{y}$ on PCs: $\mathbf{z} = \mathbf{y}\mathbf{A}$. The difference between $\mathbf{y}$ and its projection on the PCs $\tilde{\mathbf{y}}$ is the reconstruction error. Thus, the reconstruction error is the squared Euclidean distance to the linear manifold that is found by applying PCA. Data are usually centered by subtracting their mean and scaled by their standard deviations before PCA. Centering represents all observations as fluctuations, leaving only the relevant variation for analysis. Scaling is fundamental in multivariate data-sets as variables come in different units. Because PCA is a linear combination of basis functions, a large number of PCs may be required when applying PCA on highly non-linear systems. Local PCA constructs local models, each pertaining to a different disjoint region of the data space [8]. Within each region, the model complexity is limited, thus the construction of linear models by PCA is more suitable. The partition in local clusters is accomplished using a Vector Quantization (VQ) algorithm that minimizes the reconstruction error [8]. In each cluster, PCA is carried out and a set of local PCs are found: $\mathbf{A}_i$, with i being the index of the cluster. One observation $\mathbf{y}$ is assigned to cluster i if the Euclidean distance from the linear manifold indicated by $\mathbf{A}_i$ is minimum.

SOM

Self Organizing Maps (SOMs) belong to the class of artificial neural network (ANN) [13]. Each row in the data matrix can be seen as one point in a N-dimensional space. The data matrix is usually standardized beforehand. The main idea of SOMs is to map a high-dimensional data-set onto a lower dimensional, non-linear grid. SOMs consist of a grid of neurons in competition to be activated. A neuron is activated after the inner product between the weight vector of a node (representing its location in the N-dimensional data space) and a point from the data-set is maximized, namely when the Euclidean distance between this point and the neuron is minimized. During the training of the algorithm, each neuron can move across the N-dimensional data space, and the map is geometrically organized. After this procedure, a high-dimensional data-set can be mapped onto a lower dimensional grid [9].

RESULTS

Large Eddy Simulation (LES) was applied on the selected test case. A transient combustion solver based on the open source software OpenFOAM is used in the simulation. The simulation domain starts from the jet exit and is discretized with a 3D cylinder structured mesh containing 1.5 million cells. Experimentally measured profiles (H_2O,

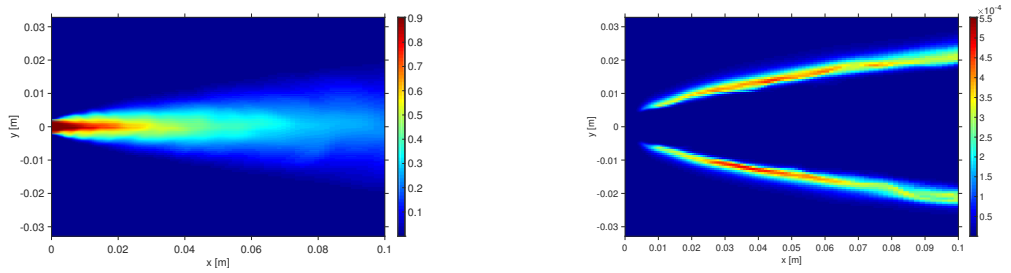

FIGURE 1: (*left*) Original field of CH_4 mass fraction; (*right*) Original field of OH mass fraction.

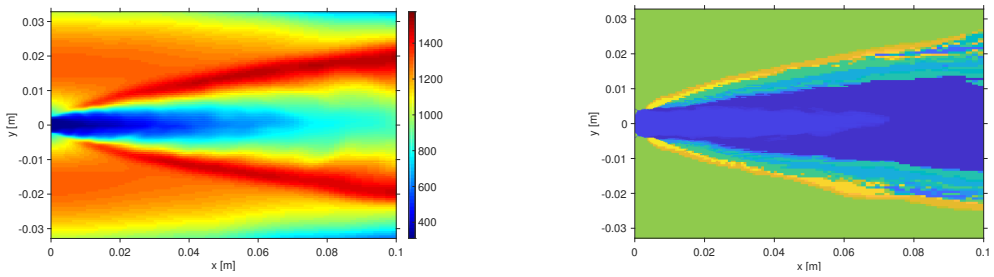

FIGURE 2: (*left*) Original field of temperature; (*right*) clusters found by Local PCA with 20 clusters and 5 PCs. Colors represent cluster assignments.

CO_2, O_2 species mass fractions and temperature) are directly used as inflow data for the simulation. A reduced skeletal mechanism KEE58 containing 17 species and 58 reactions [14] are considered. The combustion model of Partially Stirred Reactor [15] and the LES turbulent model of one equation eddy viscosity [16] are employed.

The dataset is a matrix of size $1,555,200 \times 19$. Rows represent observations and columns correspond to variables. Thus, one column of this matrix corresponds to the spatial profile of one variable, e.g. temperature, pressure or species mass fraction.

Figure 1 reports the fields of CH_4 (left) and OH radical (right). Figure 2 reports the temperature field (left) and the clustered spatial domain (right) by means of Local PCA. The auto-scaling criterion was applied. The dimension of the local manifolds was set to 5 as a number of 5 PCs was necessary to explain over 99% of the variance in each cluster. With the Global PCA approach, a number of 10 PCs was necessary to do so. Thus, this can be considered to be the intrinsic dimensionality of the problem. Each colored portion of the geometrical domain corresponds to a region of the data-space where a subset of the original variables account for most of the data variation. Interestingly, the Local PCA approach could identify the different flame zones despite being based on a purely mathematical principle, as visible in Figure 2 (right). This allows for feature extraction with limited user expertise and thus can aid data interpretation. By looking at the PVs in each cluster, it is possible to characterize the clusters in terms of dominant thermo-chemical actors. This is paramount for the implementation of local adaptive kinetic schemes that reduce the computational cost of an LES simulation with negligible loss of information. Another unsupervised learning methodology has been employed, namely SOMs. Results are reported for the SOM methodology with a two-dimensional map in Figure 3. The use of SOMs as clustering procedure led to a different solution. In particular, in spite of the intrinsic axial-symmetry of the problem, the regions outside the flame have been grouped into different clusters. In order to compare the two methodologies, the Local PCA clustering has been run again with the same number of clusters and local dimensionality. Results for this are reported in Figure 3 (right). The first emerging difference is the classification of the outer regions of the domain, which have been grouped again into the same cluster by Local PCA. Thus, this is explained by the fact that SOMs with a small number of nodes behave in a way that is similar to k-means as it uses Euclidean distances, and should not be attributed to the low-dimensionality of the SOM. In fact, Local PCA looks for local manifolds such that the projection of each clustered object onto it is minimized. Furthermore, this process allows for data encoding and thus the reconstruction error for the clustered objects can be used a measure for the quality of the clustering itself. Because Local PCA finds a set of local directions where most of the data variation is explained, this information is paramount for the development of reduced mechanisms, such as DGR.

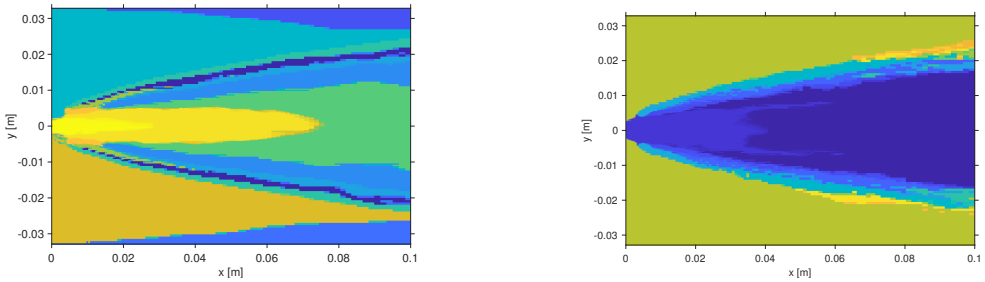

FIGURE 3: (*left*) Clusters found by a two-dimensional 4×4 SOM (auto-scaling); (*right*) clusters found by Local PCA with 16 clusters and 2 PCs. Colors represent cluster assignments.

CONCLUSIONS

In this work two unsupervised learning methodologies, namely SOMs and Local PCA, have been applied to an LES data-set of two co-flow jets: the central jet is an equimolar mixture of CH_4 and H_2 and the annulus jet has an oxygen content of 3% by mass. Local PCA was run with a number of 20 clusters. Results show that Local PCA could partition the data into clusters with low reconstruction errors by retaining with only 5 PCs in each cluster. Global PCA needed a number of 10 PCs to have comparable reconstruction errors. A two-dimensional SOM was also trained on the same data-set, with a 4×4 net. Regions which were previously proved to have very similar structures by Local PCA were separated into different clusters. The Local PCA methodology was run again with a number of 2 PCs and 16 clusters, in order to reproduce the same conditions and compare the two methodologies. Results showed that Local PCA still did not separate the aforementioned regions. Thus, this misclassification can be attributed to the Euclidean nature of the SOM more than to the low dimensionality of the map itself. Furthermore, Local PCA provides information about the PVs which is fundamental for the development of reduced adaptive kinetic schemes, such as DRG.

ACKNOWLEDGMENTS

The research was sponsored by the European Research Council, Starting Grant No 714605.

REFERENCES

[1] T. Lu and C. K. Law, Progress in Energy and Combustion Science **35**, 192–215 (2009).
[2] Z. Song and A. Kusiak, IEEE Transactions on Industrial Informatics **2**, 176 – 184 (2016).
[3] W. Li, Z. Zhu, F. Jiang, G. Zhou, and G. Chen, Mechanical Systems and Signal Processing **50-51**, 414–426 (2015).
[4] K. Bizon, G. Continillo, L. Russo, and J. Smula, Computers and Chemical Engineering **32**, 1313–1323 (2008).
[5] I. T. Jolliffe, (2002).
[6] B. J. Isaac, A. Coussement, O. Gicquel, P. J. Smith, and A. Parente, Combustion and Flame **161**, 2785–2800 (2014).
[7] A. Parente, J. C. Sutherland, B. B. Dally, L. Tognotti, and P. J. Smith, Proceedings of the Combustion Institute **33**, 3333–3341 (2011).
[8] N. Kambhatla and T. Leen, Neural Computation **9**, 1493–1516 (1997).
[9] T. Kohonen, Neurocomputing **21**, 1–6 (1998).
[10] A. Parente and J. C. Sutherland, Combustion and Flame **160**, 340–350 (2013).
[11] A. Coussement, B. J. Isaac, O. Gicquel, and A. Parente, Combustion and Flame **168**, 83–97 (2016).
[12] K. Bizon, G. Continillo, E. Mancaruso, S. S. Merola, and B. M. Vaglieco, Combustion and Flame **157**, 632–640 (2010).
[13] M. Kubat, Neural networks: a comprehensive foundation by Simon Haykin, Macmillan, 1994, ISBN 0-02-352781-7. 1999, arXiv:arXiv:1312.6199v4 .
[14] R. W. Bilger, S. H. Stårner, and R. J. Kee, Combustion and Flame **80**, 135–149 (1990).
[15] J. Chomiak, *Combustion: A study in theory, fact and application* (Abacus Press, Philadelphia, PA, 1987).
[16] K. H. Akira Yoshizawa, Journal of the Physical Society of Japan **54**, 2834–2839 (1985).