Assessment of different chemistry reduction methods based on principal component analysis: comparison of the MG-PCA and score-PCA approaches.

Axel Coussement^{a,*}, Benjamin J. Isaac^{a,d}, Olivier Gicquel^{b,c}, Alessandro Parente^a

^aUniversité Libre de Bruxelles, Ecole Polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium

^bEcole Centrale Paris, Grande Voie des Vignes, 92295 Chatenay-Malabry. France

^cCNRS, UPR 288 Laboratoire d'Energétique moléculaire et macroscopique, combustion, Grande Voie des Vignes, 92295 Chatenay-Malabry. France.

^dDepartment of Chemical Engineering, University of Utah, Salt Lake City, UT, 84112, USA

Abstract

Two families of PCA based combustion model were recently developed: score-PCA and MG-PCA. This paper presents the first comparative study of the two aftermentioned methods. Both methods are first benchmarked on 1-D laminar flames of hydrogen-air and syngas-air, to verify the consistency of the approaches. The sensitivity to differential diffusion is carefully addressed and a rotation technique is proposed, to allow for using score-PCA with differential diffusion. Then, 2-D flame-turbulence interactions are considered, spanning a wide range of velocity and length scale ratios, to demonstrate the ability of PC-based combustion models to efficiently capture the behaviour of different combustion regimes, from flamelet-like flames to distributed reaction regimes.

Keywords: Low-dimensional manifolds, Principal Component Analysis, Reacting Flows, Reduced-order models

1. Introduction

The simulation of turbulent reacting systems using detailed (and reduced) kinetic mechanisms tends to be extremely expensive. For instance, the Direct Numerical Simulation of a reacting jet at high Reynolds number using detailed chemistry can already require millions of CPU hours, leading to unacceptable costs for more realistic configurations (see [?]). Therefore, there is a need for low order models such as low-dimensional manifold in the composition space that parameterize the full system with a reduced number of variables. In recent years, the potential for Principal Component Analysis (PCA) [? ? ?] towards the identification of these manifold

*Corresponding author. Phone + 32 2 650 26 80 Fax +32 2 650 27 10

Email address: axcousse@ulb.ac.be (Axel Coussement)

Preprint submitted to Elsevier

March 22, 2016

has been investigated thoroughly, [? ? ? ? ? ? ?], with the main objective of defining a new family of reduced-order combustion models.

Starting from a thermo-chemical state space consisting of Q variables, PCA can provide a truncated set of q ($q < Q$) principal components, providing a low-dimensional representation of the manifold. Therefore, only q variables need to be taken into account instead of Q , accelerating the computation accordingly. Contrary to other reduction techniques, PCA models offer the advantage of not being bound to the original dataset used to train the model [?].

Two families of PCA models were proposed. The first approach is based on the use of the transport equations for the scores (i.e the projection of the thermo-chemical state space on the PCA basis), and originates from the work of Sutherland and Parente [? ?]. It was further developed by [?] and Pope [? ?], using PCA in conjunction with non-linear regression techniques, to maximize the dimensionality reduction. [? ?] proposed the use of neural networks and investigated the potential of kernel PCA [?], to transform the initial problem into a non-linear feature space where the linear PCA is carried out. Finally [?] and [? ?] provided the first *a posteriori* studies using the score-PCA approach.

[?] proposed an alternative method based on PCA, the so-called Manifold Generated from PCA (MG-PCA). MG-PCA relies on classical transport equations for a reduced set of principal variables (PV), which are used to reconstruct the full thermo-chemical state using the information contained in the PC matrix. [?] studied the behavior of MG-PCA on both 0-D auto-ignition and flame-vortex interactions, showing the capacity of MG-PCA based model trained on simple systems, to reproduce the behavior of more complex ones.

Despite recent research on PCA methods, only a few multidimensional cases were investigated, namely flame-turbulence and flame-vortex interaction cases with MG-PCA in DNS [?], and the Sandia flame F [?] with score-PCA method in one dimensional turbulence (ODT) simulations. The aim of the present paper is to study the behavior of both score-PCA and MG-PCA models, for a wide range of combustion regimes in the framework of premixed turbulent combustion. First, 1-D laminar flames are used to generate the database for model training and verify the consistency of the method. Several equivalence ratios are considered, using syngas and hydrogen, paying particular attention to the effect of differential diffusion on the applicability of the methods. Various 2-D flame-turbulence interaction cases are then considered for both models and benchmarked against DNS, to investigate the ability of the methods to be applied outside of the original training manifold boundaries. The different cases investigated in the present study are shown in the Borghi diagram in Figure 1

Figure 1: Borghi diagram on which is reported the different flame-turbulence interactions studied in this work (×).

2. PCA method

The thermo-chemical state-space can be represented by the following variables: $[Y_1, \dots, Y_{N_{sp}}, T, p]$, where Y_i is the mass fraction of species i , N_{sp} is the number of species, T is the temperature and p is the pressure. PCA is concerned with finding the best $q < Q$ dimensional representation of the original state-space. Then, it is sufficient to transport q variables to recover the missing $Q - q$ ones.

Considering a dataset $\mathbf{Y} (n \times Q)$, containing n samples of Q variables, the first operation to be considered is a centering and scaling of the dataset to give each variable the same importance during the analysis. [?] and [?] showed that the scaling of the dataset has great influence on the quality of PCA. This operation reads:

$$\tilde{\mathbf{Y}} = (\mathbf{Y} - \bar{\mathbf{Y}})\mathbf{D}^{-1}, \quad (1)$$

where $\bar{\mathbf{Y}}$ is a $n \times Q$ matrix containing the mean of each variable and $\mathbf{D}$ is a $n \times Q$ matrix containing the standard deviation of each variable [?]. In the reminder of this paper, all the datasets are assumed to be centered and scaled for the sake of clarity

In order to perform PCA the covariance matrix has to be evaluated,

$$\mathbf{S} = \frac{1}{n-1} \mathbf{Y}^t \mathbf{Y}, \quad (2)$$

$\mathbf{S}$ can be decomposed as:

$$\mathbf{S} = \mathbf{A} \mathbf{L} \mathbf{A}^t, \quad (3)$$

where $\mathbf{A}$ ($Q \times Q$) is the matrix of eigenvectors or Principal Components (PCs), associated to the eigenvalues of $\mathbf{S}$, $\mathbf{L}$ ($Q \times Q$), arranged in decreasing order of magnitude. Also, from equation (2) and (3), it follows that $\mathbf{A}^{-1} = \mathbf{A}^t$. The matrix $\mathbf{A}$ defines a new basis composed of uncorrelated vectors.

The projection of the original sample onto the new basis defines the scores called $\mathbf{Z}$ ($n \times Q$):

$$\mathbf{Z} = \mathbf{Y}\mathbf{A}. \quad (4)$$

Since the variance accounted by each PC eigenvector decreasing with its index it is possible to form a truncated basis retaining only the first q PCs. Thus, the original dataset can be approximated as:

$$\mathbf{Y} \approx \mathbf{Y}_q = \mathbf{Z}_q \mathbf{A}_q^t, \quad (5)$$

where $\mathbf{Z}_q$ ($n \times q$) and $\mathbf{A}_q$ ($Q \times q$) are computed from $\mathbf{Z}$ and $\mathbf{A}$ by retaining only the first q PCs (see [?] for details). Note that $\mathbf{A}_q^t$ is of size ($q \times Q$) and is the transpose of $\mathbf{A}_q$. It is equivalent to the pseudo-inverse of $\mathbf{A}_q$, since the columns of $\mathbf{A}_q$ are made of orthonormal vectors.

3. Score-PCA approach

The score-PCA approach proposed by [?] is a direct application of PCA theory presented in the previous section. If the original dataset includes n samples, PCA allows to compute an approximation of this dataset using equation (5).

The Q species transport equations are thus replaced by q scores equations, reducing the number of variables to be transported and therefore the CPU cost of the computation. The score transport equations reads [?]:

$$\frac{\partial \rho \mathbf{z}}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{z}) = \nabla \cdot \rho D_{\mathbf{z}} \nabla \mathbf{z} + \Omega_{\mathbf{z}}, \quad (6)$$

where ρ is the density and $\mathbf{z} = \mathbf{Z}_i^t$ represent an individual score realization. Although the model in [?] applies to the full thermo-chemical state, for practical applications of score-PCA, it is recommended to limit the dataset $\mathbf{Y}$ to the chemical species [?]. In such a case, the diffusion matrix $D_{\mathbf{z}}$ and the reaction terms matrix Ω can easily be derived from the species matrix. The species conservation equations can be expressed as:

$$\frac{\partial \rho \mathbf{y}}{\partial t} + \nabla \cdot (\rho \mathbf{u} \otimes \mathbf{y}) = \nabla \cdot \rho D_{\mathbf{y}} \nabla \mathbf{y} + \Omega_{\mathbf{y}}, \quad (7)$$

with $\mathbf{y}$ is the chemical species vector. As for the scores $\mathbf{y} = \mathbf{Y}_i^t$ is an individual realization of the chemical species (i.e. the species values at a particular point in space where equation (7) is

applied). The dataset $\mathbf{Y}$ is obtained by sampling $\mathbf{y}$ at different locations in space. Using the score definition, one can write:

$$\mathbf{Z} = \mathbf{Y} \mathbf{A}_q. \quad (8)$$

Using the definition in equation (8) in equation (7) and noticing that:

$$\nabla \mathbf{Y} = \nabla \mathbf{Z} \mathbf{A}_q^t, \quad (9)$$

it is fairly easy to derive an expression for the score diffusion coefficient,

$$\mathbf{D}_z = \mathbf{A}_q^t D_y \mathbf{A}_q \quad (10)$$

and for the score source term,

$$\Omega_z = \Omega_y \mathbf{A}_q, \quad (11)$$

where D_y is the diffusion matrix of the species $\mathbf{y}$ and Ω_y the corresponding source terms per unit of mass.

Equation (6) is easy to compute in a reactive flow solver but still requires closure at the boundaries. This closure is achieved using:

$$\mathbf{z}|_{boundary} = \mathbf{y}|_{boundary} \mathbf{A}_q, \quad (12)$$

where $\mathbf{y}|_{boundary}$ are the species mass fraction values to be enforced at the boundaries. In the case of the Navier-Stokes equations, the viscous fluxes to be enforced on the boundary can be derived easily from [? ?]

4. MG-PCA approach

The MG-PCA approach also relies on PCA but is based on the resolution of classic transport equations for a selected number of transported species, while the non transported ones are reconstructed using PCA. Considering equation (5):

$$\mathbf{Z} \approx \mathbf{Y}_q \mathbf{A}_q. \quad (13)$$

However, $\mathbf{Z}$ can also be found by using a subset of species $\mathbf{Y}(\mathbf{q})$ and a subset of the $\mathbf{A}_q$ matrix, $\mathbf{A}(\mathbf{q})_q$:

$$\mathbf{Z} = \mathbf{Y}(\mathbf{q}) \mathbf{A}(\mathbf{q})_q. \quad (14)$$

Introducing equation (14) in equation (5) we get:

$$\mathbf{Y} \approx \mathbf{Y}_q = \mathbf{Y}(\mathbf{q}) \mathbf{B}_q, \quad (15)$$

where $\mathbf{B}_q$ is a $(q \times Q)$ matrix derived from $\mathbf{A}_q$ [?], allowing the reconstruction of $(Q - q)$ non-transported species from the q transported ones.

Similarly to score-PCA, MG-PCA allows reducing the number of transport equation to $q < Q$. Additional analysis is, however, required to determine the variables to transport. Several methods exist to this purpose and a full sensitivity to them is shown in [?]. Here the B2 backward selection method is used. The choice of the variables is also very important from the numerical perspective, to ensure stability of the solution.

5. Test cases and reduced-order model generation

The following section briefly discusses the test cases selected for the assessment of the models. First the database used to generate PCA has to be chosen. Here, a database consist of a converged one-dimensional freely propagating laminar flame. To avoid any bias, the same explicit solver is used to generate the database and test the models. Convergence is reached when the relative change in the manifold is less then 10^{-6} on species and temperature.

Eight different databases are used corresponding to two different fuels: syngas and hydrogen each at two different equivalence ratios with and without differential diffusion. In each instance both PCA-based methods will be compared, they will be trained with the same database.

For the hydrogen cases the following equivalence ratios are considered: $\phi = 0.75$ and $\phi = 1$, with and without differential diffusion. For each case, the grid is composed of 5,121 points with a spacing of $6.25 \cdot 10^{-6}$ m. Pressure is set to 101,325 Pa and inlet temperature is 300 K. Inlet velocity is different for each equivalence ratio: $u = 1.25$ m/s at $\phi = 0.75$ and $u = 1.5$ m/s at $\phi = 1$. The chemical scheme is the one from [?], containing 9 species: Y_H , Y_{H_2} , Y_O , Y_{O_2} , Y_{OH} , Y_{H_2O} , Y_{N_2} , Y_{HO_2} and $Y_{H_2O_2}$.

For the syngas (50% H_2 /50% CO by vol.) cases, two different equivalence ratios are considered: $\phi = 0.7$ and $\phi = 1$, with and without differential diffusion. The grid, pressure and inlet temperature are exactly the same as for the hydrogen cases. Inlet velocities are set to $u = 0.5$ m/s at $\phi = 0.7$, and $u = 0.65$ m/s at $\phi = 1$. The chemical scheme is taken from [?], containing 12 species: Y_{N_2} , Y_{O_2} , Y_{H_2} , Y_{H_2O} , $Y_{H_2O_2}$, Y_{CO} , Y_{CO_2} , Y_O , Y_H , Y_{OH} , Y_{HO_2} and Y_{HCO} .

To generate the reduced models, PCA is carried out on one of the eight database, depending on the computation to be performed. The number of variables (either scores or species) to be transported is chosen to allow a for quasi-exact reconstruction of the original dataset. If not, the method could introduce large errors in the calculation of thermodynamic data and source terms, leading to numerical instabilities. The reconstruction is assumed to be quasi-exact when

a coefficient of determination, R^2 , greater than 0.999 is obtained for all variables (see [?]). By doing so on each database, one can see that 7 PCs must be retained for all the syngas cases and 6 for the hydrogen cases.

For MG-PCA, one must still choose the species to be transported. The species are selected using the B2 criterion [?]:

- $Y_H, Y_O, Y_{OH}, Y_{HO_2}, Y_{H_2O}$ and $Y_{H_2O_2}$ for the hydrogen cases.
- $Y_{H_2O}, Y_{H_2O_2}, Y_O, Y_H, Y_{OH}, Y_{HO_2}$ and Y_{HCO} for the syngas cases.

It is worth noting that the number of variables and the species transported depends only on the fuel. This yields a database which contains a 9×6 matrix and two 9×1 vectors (centering and scaling) for the hydrogen scheme, and a 12×7 matrix and two 12×1 vectors for syngas. It is also interesting to note that, using different equivalence ratios or different diffusion law assumptions, the number of components or variables to be transported does not change.

From a manifold reduction point of view, using 6 PCs with hydrogen test case do not lead to a reduction in the degrees of freedom of the problem when $Le = 1$. However when the differential diffusion is taken into account, a reduction of 1 degree of freedom is achieved (15%). In the syngas case, using 7 PCs leads to a reduction of 1 degree of freedom out of 8 (12.5%) for $Le = 1$, and 4 degrees of freedom when $Le \neq 1$, providing a reduction of 50%. The reduction obtained in this paper is in line with previous investigations [? ? ? ?]

Although coupling PCA with non-linear regression [? ?] has the potential of ensuring very strong compression, the approach strongly relies on the ability of the chosen non-linear model to reproduce the system physics. On the other hand, keeping the full original mechanism results in lower reduction but allows, at the same time, to increase the range of applicability and the generality of the reduced models. The present paper aims to focus on the latter aspect, to allow a throughout comparison of the two methods.

Once both reduced methods are trained on the laminar flame simulations, they are benchmarked against each other and with respect to DNS. First the eight one-dimensional freely propagating laminar flames used to train the model are considered. The relative simplicity of these cases is ideal to benchmark the different models. In all cases, the database used to perform PCA is the flame itself, therefore both methods are *a priori* expected to yield perfect results.

Then, eight-flame turbulence interactions cases are considered covering various regimes (see Figure 1). The objective is to investigate the range of applicability of the reduced models, trained with the same single freely propagating laminar premixed flame, when changing the

Reynolds and Damkohler numbers. The wide range of regimes spanned by our investigation can be appreciated in Figure 2, which shows Y_{HCO} distribution for four of the eight cases, namely cases 1 to 4. Not only flamelet-like regimes are considered but also distributed reaction zones, which show significant deviations with respect to the original training dataset.

All simulations are performed using the compressible solver YWC developed at the EM2C Laboratory, using a CHEMKIN-like [?] formalism for the computation of thermo-chemical properties (see [?] for details).

5.1. 1-D flame cases

The present section discusses the results obtained using the score-PCA and MG-PCA methods on the 1-D flames. In order to compare them, each 1-D flame was simulated from the DNS solution until convergence was once again reached. The criterion retained for assessing the quality of the models is the maximum error relative to the maximum value for each species $\epsilon_{k,\text{model}}$:

$$\epsilon_{k,\text{model}} = \max \frac{|Y_{k,\text{model}} - Y_{k,\text{DNS}}|}{\max Y_{k,\text{DNS}}}, \quad (16)$$

where $Y_{k,\text{model}}$ is the mass fraction of species k computed with the reduced model and $Y_{k,\text{DNS}}$ the mass fraction of species k computed with DNS. As only the maximum is taken into account, the criterion is very severe. To have a sound comparison between MG-PCA and score-PCA, results were post-processed so that the point corresponding to $T = 1000 \text{ K}$ is at the same location with respect to the DNS solution. The results are presented in Tables 1 to 4.

The error in the laminar flame speed predictions, S_l , is presented in Tables 5 - 6 for hydrogen and syngas, respectively. The deviation is estimated by computing the relative error on the laminar flame speed value as provided by the PCA models with respect to the DNS solution. The S_l value is computed as:

$$S_l = -\frac{1}{\rho Y_F} \int_{\text{mesh}} \dot{\omega}_F dx, \quad (17)$$

where the subscript F indicates the fuel and $\dot{\omega}_F$ its reaction rate. The syngas case with $Le \neq 1$ and $\phi = 1$ was unstable when computed using the score-PCA approach, so that data is not available for this case. This is linked to the reconstruction of the diffusive fluxes as explained below.

In terms of maximum errors (ϵ_k), when the $Le = 1$, results are in favor of MG-PCA, where errors tends to be one order of magnitude smaller than for score-PCA. However, even the latter errors are still very low (1-2 %). Both models are able to accurately reproduce the flame structure,

Figure 2: HCO field for case 1 to 4.

Species	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$
	$\phi = 0.75$	$\phi = 0.75$	$\phi = 1$	$\phi = 1$
Y_{N_2}	≈ 0	≈ 0	≈ 0	≈ 0
Y_{H}	0.0036	0.3262	0.0495	2.5381
Y_{H_2}	0.0018	0.1783	0.0259	1.3237
Y_{O}	0.0031	0.3483	0.0464	2.3372
Y_{O_2}	0.0013	0.1331	0.0237	1.2063
Y_{OH}	0.0023	0.3056	0.0321	1.5732
$Y_{\text{H}_2\text{O}}$	0.0016	0.1569	0.0230	1.1783
Y_{HO_2}	0.0052	0.2461	0.0815	4.4077
$Y_{\text{H}_2\text{O}_2}$	0.0043	0.2726	0.0671	3.6362

Table 1: Maximum errors, $\epsilon_{k,\text{model}}$ for MG-PCA and score-PCA approaches for the hydrogen test cases with $Le = 1$.

Species	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$
	VARIMAX			VARIMAX		
	$\phi = 0.75$	$\phi = 0.75$	$\phi = 0.75$	$\phi = 1$	$\phi = 1$	$\phi = 1$
Y_{N_2}	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0	≈ 0
Y_{H}	3.9219	1.7359	2.1356	1.5592	0.6226	1.2135
Y_{H_2}	8.1131	3.2091	6.1245	2.6318	1.2782	1.8902
Y_{O}	2.3025	0.7016	1.2432	2.0334	1.2435	2.1356
Y_{O_2}	1.8564	3.0198	2.1466	0.2751	1.1857	0.6721
Y_{OH}	2.7657	1.0816	5.1245	1.0733	0.7062	3.1245
$Y_{\text{H}_2\text{O}}$	2.2144	3.2597	5.1246	0.4388	1.1190	1.3456
Y_{HO_2}	17.5078	7.0385	6.2211	3.8679	4.9692	4.5311
$Y_{\text{H}_2\text{O}_2}$	11.6380	5.3595	7.1241	2.0933	1.1518	3.1345

Table 2: Maximum errors, $\epsilon_{k,\text{model}}$ for MG-PCA and score-PCA approaches for the hydrogen test cases with $Le \neq 1$.

Species	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$
$\phi = 0.7$	$\phi = 0.7$	$\phi = 1$	$\phi = 1$	
$Y_{\text{N}_2} \approx 0$	≈ 0	≈ 0	≈ 0	
Y_{O_2}	0.2679	0.3651	$5.7009 \cdot 10^{-4}$	0.6804
Y_{H_2}	0.6005	0.2771	$8.5543 \cdot 10^{-4}$	1.0311
$Y_{\text{H}_2\text{O}}$	0.7011	0.7757	$7.8693 \cdot 10^{-4}$	0.9444
$Y_{\text{H}_2\text{O}_2}$	1.1133	0.3086	0.0021	2.5878
Y_{CO}	0.2092	0.2093	$3.3805 \cdot 10^{-4}$	0.4034
Y_{CO_2}	0.1994	0.2361	$3.9419 \cdot 10^{-4}$	0.4677
Y_{O}	1.3036	0.9571	0.0015	1.7437
Y_{H}	2.7675	2.1277	0.0016	1.8974
Y_{OH}	2.8536	1.5922	$7.9389 \cdot 10^{-4}$	0.9402
Y_{HO_2}	1.2945	0.4642	0.0026	3.1401
Y_{HCO}	1.7357	0.6363	0.0032	3.9520

Table 3: Maximum errors, $\epsilon_{k,\text{model}}$ for MG-PCA and score-PCA approaches for the syngas test cases with $Le = 1$.

indicating that the reconstruction error on the species are kept small enough to avoid biasing the source terms and transport properties.

When considering the $Le \neq 1$ cases, Table 2 and 4 shows that both methods are comparable. However when looking at the laminar flame speed error, it is clear that the MG-PCA approach is able to capture the correct flame speed, while score-PCA displays large errors. To emphasize this, Figure 3 shows a comparison of the two methods for the syngas case at $\phi = 0.7$. We can observe the score-PCA approach cannot deal with cases where differential diffusion is taken into account.

The overestimations of S_l by the score-PCA can be due to an error in the evaluation of the diffusion fluxes and/or source terms. As indicated above, the latter can be attributed to the non linear propagation of the error on the reconstructed species, whereas the former is due to the appearance of negative off-diagonal source terms in the diffusion matrix. As the errors are observed only on the predictions of S_l with differential diffusion, the origin of the errors lies in the diffusion modeling. Indeed, if the score diffusion matrix is computed (Table 7) for hydrogen at $\phi = 0.75$ and $Le \neq 1$, negative cross-diffusion terms appear. The diagonal terms do not appear to be dominant. Some negative off-diagonal terms are also reported in other PCA studies [? ?

Species	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{MG-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$	$\epsilon_{k,\text{S-PCA}} (\%)$
	VARIMAX			VARIMAX		
	$\phi = 0.7$	$\phi = 0.7$	$\phi = 0.7$	$\phi = 1$	$\phi = 1$	$\phi = 1$
Y_{N_2}	≈ 0	≈ 0	≈ 0	≈ 0	N/A	≈ 0
Y_{O_2}	1.1515	1.3284	1.2134	0.4905	N/A	2.6531
Y_{H_2}	10.4504	1.0814	3.2464	6.7174	N/A	2.8506
$Y_{\text{H}_2\text{O}}$	1.0337	4.0001	1.3461	0.7402	N/A	4.2862
$Y_{\text{H}_2\text{O}_2}$	4.8962	6.1756	5.6799	3.3443	N/A	10.6046
Y_{CO}	1.0899	2.0329	0.9235	0.4527	N/A	0.9798
Y_{CO_2}	2.6966	2.1660	2.3461	0.5863	N/A	1.1895
Y_{O}	4.3250	2.9284	3.1344	2.4246	N/A	3.9111
Y_{H}	2.1114	2.7048	4.1211	1.1407	N/A	2.9958
Y_{OH}	3.6944	3.3695	4.1241	1.1500	N/A	2.6153
Y_{HO_2}	5.9161	7.1765	10.1351	4.5448	N/A	19.2864
Y_{HCO}	4.9373	6.7068	7.8912	3.0235	N/A	9.0309

Table 4: Maximum errors, $\epsilon_{k,\text{model}}$ for MG-PCA and score-PCA approaches for the syngas test cases with $Le \neq 1$.

Condition	S_l error in % MG-PCA	S_l error in % Score-PCA	S_l error in % Score-PCA and VARIMAX
$\phi = 0.75$ $Le = 1$	$< 10^{-4}$	$< 10^{-4}$	-
$\phi = 1$ $Le = 1$	$< 10^{-4}$	$< 10^{-4}$	-
$\phi = 0.75$ $Le \neq 1$	1.2183	7.823	0.3139
$\phi = 1$ $Le \neq 1$	1.5597	8.125	0.667

Table 5: Relative error on the laminar flame speed for hydrogen test cases.

Condition	S_l error in % MG-PCA	S_l error in % Score-PCA	S_l error in % Score-PCA and VARIMAX
$\phi = 0.75 \ Le = 1$	$< 10^{-4}$	$< 10^{-4}$	-
$\phi = 1 \ Le = 1$	$< 10^{-4}$	$< 10^{-4}$	-
$\phi = 0.75 \ Le \neq 1$	0.620	12.039	0.145
$\phi = 1 \ Le \neq 1$	0.401	-	0.243

Table 6: Relative error on the laminar flame speed for syngas test cases.

Figure 3: Comparison of Y_{CO_2} and Y_{HCO} for the syngas test case.

	z_1	z_2	z_3	z_4	z_5	z_6
z_1	$1.2303 \cdot 10^{-3}$	$1.6157 \cdot 10^{-4}$	$-2.8651 \cdot 10^{-4}$	$-4.7042 \cdot 10^{-4}$	$-4.0757 \cdot 10^{-5}$	$2.9592 \cdot 10^{-4}$
z_2	$1.6157 \cdot 10^{-4}$	$1.2096 \cdot 10^{-3}$	$-7.5123 \cdot 10^{-4}$	$-6.1785 \cdot 10^{-4}$	$-7.7428 \cdot 10^{-5}$	$4.8711 \cdot 10^{-4}$
z_3	$-2.8651 \cdot 10^{-4}$	$-7.5123 \cdot 10^{-4}$	$1.5325 \cdot 10^{-3}$	$7.4382 \cdot 10^{-4}$	$9.4947 \cdot 10^{-5}$	$-5.9447 \cdot 10^{-4}$
z_4	$-4.7042 \cdot 10^{-4}$	$-6.1785 \cdot 10^{-4}$	$7.4382 \cdot 10^{-4}$	$1.8451 \cdot 10^{-3}$	$5.1270 \cdot 10^{-5}$	$-7.1834 \cdot 10^{-4}$
z_5	$-4.0757 \cdot 10^{-5}$	$-7.7428 \cdot 10^{-5}$	$9.4947 \cdot 10^{-5}$	$5.1270 \cdot 10^{-5}$	$6.6198 \cdot 10^{-4}$	$-1.8693 \cdot 10^{-4}$
z_6	$2.9592 \cdot 10^{-4}$	$4.8711 \cdot 10^{-4}$	$-5.9447 \cdot 10^{-4}$	$-7.1834 \cdot 10^{-4}$	$-1.8693 \cdot 10^{-4}$	$1.3889 \cdot 10^{-3}$

Table 7: Score diffusion matrix, for the hydrogen case at $\phi = 0.75$ and $Le \neq 1$.

]. When differential diffusion is not taken into account, the off-diagonal terms vanish. As the diffusion coefficients, D_i , are equal for all the species, $D_i = D$, equation (10) becomes:

$$D_{\mathbf{z}} = D \mathbf{A}_q^t \mathbf{A}_q \quad (18)$$

Considering that, by definition, $\mathbf{A}_q^t \mathbf{A}_q = \mathbf{I}$, we obtain $D_{\mathbf{z}} = D \mathbf{I}$, which implies that no off-diagonal terms are present.

Handling differential diffusion with the score approach requires the transformation of the diffusion matrix, to make it diagonally dominant and allow the use of Fick's law. This can be accomplished using the PC rotation. Through rotation, the weights can be redefined to meet an alternative criteria. In particular, rotation is aimed at attaining a simple structure for $\mathbf{A}_q$, so that weights on a PC are either close to unity or close to zero and, thus, variables have large weights on only few or (ideally) one PC. The most common orthogonal rotation is based on the maximization of the VARIMAX criterion [?]. Verbally, such rotation concentrates the amount of variance explained for any of the original variables on single PCs. After VARIMAX rotation, $\mathbf{A}_q$ will generally have fewer large coefficients in its columns. Using VARIMAX implies that the error associated to the reduction is also the same after the rotation. The diffusion matrix presented in Table 8 displays more dominant diagonal terms. Also this method conserves the global variance accounted for by the scores [?].

In their work, [?] also proposed a rotation of the diffusion matrix to attenuate the issue of the off-diagonal terms, using a minimization algorithm. While interesting, such an approach does not *a priori* guarantee that the global variance accounted for by the score will be kept, while this is ensured by VARIMAX rotation.

All cases with $Le \neq 1$ were recomputed using these rotated PCs. Results are given in Tables

	z_1	z_2	z_3	z_4	z_5	z_6
z_1	$1.1778 \cdot 10^{-3}$	$-5.0800 \cdot 10^{-6}$	$-3.8897 \cdot 10^{-5}$	$2.1682 \cdot 10^{-5}$	$-2.1537 \cdot 10^{-5}$	$4.4452 \cdot 10^{-5}$
z_2	$-5.0800 \cdot 10^{-6}$	$5.9260 \cdot 10^{-4}$	$-2.3508 \cdot 10^{-7}$	$-2.2615 \cdot 10^{-7}$	$2.4600 \cdot 10^{-7}$	$-4.0640 \cdot 10^{-7}$
z_3	$-3.8897 \cdot 10^{-5}$	$-2.3508 \cdot 10^{-7}$	$3.6705 \cdot 10^{-3}$	$9.9370 \cdot 10^{-7}$	$-9.2803 \cdot 10^{-7}$	$2.7471 \cdot 10^{-6}$
z_4	$2.1682 \cdot 10^{-5}$	$-2.2615 \cdot 10^{-7}$	$9.9370 \cdot 10^{-7}$	$9.0556 \cdot 10^{-4}$	$-9.7649 \cdot 10^{-7}$	$1.6148 \cdot 10^{-6}$
z_5	$-2.1537 \cdot 10^{-5}$	$2.4600 \cdot 10^{-7}$	$-9.2803 \cdot 10^{-7}$	$-9.7649 \cdot 10^{-7}$	$5.9742 \cdot 10^{-4}$	$-1.7610 \cdot 10^{-6}$
z_6	$4.4452 \cdot 10^{-5}$	$-4.0640 \cdot 10^{-7}$	$2.7471 \cdot 10^{-6}$	$1.6148 \cdot 10^{-6}$	$-1.7610 \cdot 10^{-6}$	$9.2465 \cdot 10^{-4}$

Table 8: Score diffusion matrix, for the hydrogen case at $\phi = 0.75$ and $Le \neq 1$ using a varimax rotation.

2, 4 for ϵ_k and in Tables 5 and 6 for the laminar flame speed error. Moreover, the syngas case at $\phi = 1$ using rotation is now stable. Using this method, results show that score-PCA outperforms MG-PCA both in terms of species mass fraction distribution and laminar flame speed predictions. Such result can be directly linked to the condition number of matrices $\mathbf{A}_q$ and $\mathbf{B}_q$, used for score-PCA and MG-PCA, respectively. In the case of score-PCA, $\mathbf{A}_q$ is composed of orthogonal vectors by construction and since the species are centered and scaled (i.e. normalized) it results in a perfectly conditioned matrix (condition number equal to one). On the other hand, the $\mathbf{B}_q$ used in MG-PCA displays a condition number ranging from 1.54 to 3.53, depending on the cases. This is explained by partial correlation left in the transported species, even when selecting the most uncorrelated ones of the sample [?]. Adding to this the fact that the transported species can span different order of magnitudes (while the scores are scaled), the numerical reconstruction error will be eventually higher for MG-PCA.

Finally it should be stressed that off-diagonal terms, while neglectable, are still present in the diffusion matrix. They could generate numerical instabilities, due to error amplification, for very long simulations. However, we did not find any evidence of this in the present study.

5.2. Flame Turbulence interaction

The present Section aims to assess the behavior of the score-PCA (using rotation) and MG-PCA models when outside the bounds defined by the training dataset. As showed in [?], PCA methods are able to predict the state-space of a stretched flame, using a model built with an undisturbed laminar flame. To further this enquiry, eight flame-turbulence interaction cases are simulated using four different eddy sizes and two maximum eddy velocities, to cover several combustion regimes (see Figure 1). Each case is computed 3 times, using DNS, MG-PCA and

score-PCA with VARIMAX rotation.

The baseline for all cases is the syngas laminar flame at $\phi = 0.7$ which is extended on the y -axis on a 2-D domain of 640×640 points, with a mesh spacing of $1.25 \cdot 10^{-6}$ m. At $x = x_{\min}$ the boundary condition fixes a constant speed with a superimposed turbulent field at a temperature of 300 K. At $x = x_{\max}$, an outlet boundary is imposed with $p = 101,325$ Pa. Remaining boundary conditions ensure periodicity along the y -axis. The turbulent field is obtained using the Passot-Pouquet spectrum [?]. The numerical parameters for the different cases are summarized in Table 9. Figure 2 shows the Y_{HCO} field for cases 5 to 8, emphasizing the wide spectrum of conditions considered here. All numerical simulations are performed with differential diffusion. All simulation were run for 0.5 ms, ensuring that manifold statistics do not evolve anymore with time.

In order to assess the quality of the results, [?] and [?] used a simple comparison of species scatter plots. While this provides valuable general information of the performance of the method, this type of comparison cannot provide a quantitative assessment of the agreement between reduced models and DNS.

In this study, the probability density function (PDF) of the species mass fraction are used to benchmark the reduced-order model performances to DNS. Indeed, if the models behaved perfectly, their PDFs should overlap with those provided by DNS. Thus, difference between model and DNS PDFs will indicate sources of error from the reduced models. In order to discard fresh and burned gases (over-represented in the computational domain [?]), only the nodes where $Y_{\text{HCO}} > 10^{-8}$ are used to compute the PDFs. The PDFs of Y_{H} , $Y_{\text{H}_2\text{O}}$ and Y_{OH} are considered, along with the PDF of the laminar flame used to train the database. Figures 4 to 19 provide the PDFs and the scatter plots of Y_{H} and Y_{OH} as a function of Y_{CO_2} .

It is possible to observe that case 5 displays oscillations in the PDF, especially for Y_{H} (Figure 12). This is due to the flame discretization on the mesh. Indeed the case is laminar (see Figure 1) and initialized by extending the 1-D flame on the y -axis, leading to an oversampling of specific species values. Since the flame is propagating along the x -axis in a laminar manner, such over-representation results in higher probability for the repeated values, thus explaining the observed peaks in the PDF. Although less obvious, this can be observed on case 1 as well. This is confirmed by the scatter plots in Figures 5 and 13.

Comparing the Y_{H} scatter plots as well as the conditioned PDFs for the different investigated cases, it is possible to conclude that turbulence strongly affects the original manifold. While case 1 and 5 display PDFs close to the original training data (Figures 4 and 12), case 3 on

Figure 8 differs completely. Regarding the performances of the reduced models, it is clear that the score-PCA method (using the VARIMAX rotation) is able to follow more closely the DNS results, while the MG-PCA appears constrained to the laminar conditioned PDF. When the flame regime is the flamelet one, as for case 1 and 5, MG-PCA is able to provide very accurate results. However, for cases where turbulence plays a major role on the flame structure, MG-PCA is not able to correctly capture the deviation from the original manifold, especially for high value of Y_H .

When considering the scatter plots for Y_H (Figures 5, 7, 9, 11, 13, 15, 17 and 19), similar observations can be drawn. Score-PCA well mimics the distributions obtained with DNS for all cases, except for case 3 (Figure 9), while MG-PCA results in a flame structure which is very close to the laminar one, as indicated by the small amount of scatter in Y_H distribution. Concerning case 3, it is interesting to point out that DNS displays a *two-branches* configuration, one around the laminar manifold and one above (Figure 9). While MG-PCA captures only a small deviation from the training dataset, score-PCA is able to predict the *two branches* behavior, although without the desired accuracy.

The above observations can be directly extended to the analysis of Y_{OH} distribution. Score-PCA predicts correctly the deviation from the training dataset, while MG-PCA tends to overlap with it. Indeed when considering a well stirred reactor case like case 2, Figure 6 indicates that MG-PCA provides a PDF close to the laminar flame one, while score-PCA follows more closely the DNS statistics (Figure 8). MG-PCA can provide results superior to score-PCA only when operating very close to the flamelet regime, as it happens for cases 1, 5 and 8 (Figures 4, 5, 12, 13, 18, 19).

As far as case 3 is concerned (Figures 8, 9), the two branches are also observed for Y_{OH} . While the scatter plots (Figure 9) indicates a significant difference between score-PCA and DNS, the analysis of the PDFs suggest that the statistics provided by the two simulations do not differ strongly. The analysis of Y_H and Y_{OH} distribution indicate that score-PCA is able to capture the two branches behavior of case 3, using the information contained in the 1-D flame only. However, since the combustion regime for case 3 is significantly different than a laminar flame.

From the analysis of Y_{H_2O} distribution, it appears that both methods behave well. However, MG-PCA is able to capture the first peak around 0.07, providing better results for cases 1, 5, 6 and 8 (see Figures 4, 12, 14, 18 respectively), which are all close to the flamelet regime in the Borghi diagram (Figure 1). This confirms, as indicated previously, that MG-PCA results in

	$l_e(\text{m})$	$u_{up}(\text{m/s})$	$u_{conv}(\text{m/s})$	Re_t	l_t
Case 1	$5 \cdot 10^{-5}$	10	10	14.231	$1.899 \cdot 10^{-5}$
Case 2	$5 \cdot 10^{-4}$	10	10	1423	$1.899 \cdot 10^{-3}$
Case 3	$2 \cdot 10^{-3}$	10	10	22,781	$3.038 \cdot 10^{-2}$
Case 4	$1 \cdot 10^{-2}$	10	10	569,544	$7.596 \cdot 10^{-1}$
Case 5	$5 \cdot 10^{-5}$	2	1	0.565	$3.798 \cdot 10^{-6}$
Case 6	$5 \cdot 10^{-4}$	2	1	56.952	$3.798 \cdot 10^{-4}$
Case 7	$2 \cdot 10^{-3}$	2	1	911	$6.077 \cdot 10^{-3}$
Case 8	$1 \cdot 10^{-2}$	2	1	22,781	$1.519 \cdot 10^{-1}$

Table 9: Parameter used for the different cases : l_e and u_{up} are the Passot-Pouquet Parameters [?], u_{conv} is the convection speed superimposed on the turbulence spectrum, Re_t and l_t are respectively the turbulent Reynolds number and integral length scale.

better predictions for cases close to the original training dataset. On the other hand, for regimes far from the original one, score-PCA is able to better capture the deviation from the original manifolds, while MG-PCA appears constrained to it. This is especially obvious for case 3 and 7, which are quite far from the laminar flame used to train the model (see Figures 8 and 16, respectively).

In conclusion, both reduced models can provide very accurate results when the departure from the flamelet regime is not severe, representing an attractive alternative to tabulation techniques as FPI [?] or SLFM [?]. Far from the flamelet regime, the score-PCA approach is able to correctly reproduce the departure of the system from the original manifold, while MG-PCA appears to be constrained to that manifold. Interestingly, none of the methods produce numerical instabilities, which could yield to the simulation failure. Let's recall here that no artificial stabilization (like a tabulation of the source terms [? ?]) is used here, emphasizing the numerical stability of the methods. Finally, in terms of CPU time, both score-PCA and MG-PCA result in a speed-up of $\approx 31\%$ with respect to DNS.

6. Conclusions

This study presented an in-depth analysis of method for the parametrization of the thermo-chemical state using two formulations of PCA-based models variations of PCA: score-PCA and

Figure 4: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 1.

Figure 5: Scatter plots of Y_H vs Y_{CO_2} for case 1, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 6: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 2.

Figure 7: Scatter plots of Y_H vs Y_{CO_2} for case 2, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 8: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 3.

Figure 9: Scatter plots of Y_H vs Y_{CO_2} for case 3, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 10: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 4.

Figure 11: Scatter plots of Y_H vs Y_{CO_2} for case 4, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 12: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 5.

Figure 13: Scatter plots of Y_H vs Y_{CO_2} for case 5, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 14: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 6.

Figure 15: Scatter plots of Y_H vs Y_{CO_2} for case 6, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 16: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 7.

Figure 17: Scatter plots of Y_H vs Y_{CO_2} for case 7, for DNS (left), MG-PCA (center), score-PCA (right).

Figure 18: Conditional Pdf of Y_H , Y_{H_2O} and Y_{OH} for case 8.

Figure 19: Scatter plots of Y_H vs Y_{CO_2} for case 8, for DNS (left), MG-PCA (center), score-PCA (right).

MG-PCA. This study demonstrated that both methods accurately treat cases in which differential diffusion is not taken into account. When differential diffusion is considered, the score-PCA requires a rotation in order to diagonalize its diffusion matrix. Here the VARIMAX rotation was chosen in order to preserve the variance (i.e. the information) transported by the scores.

Section 5.2 proved that the score-PCA approach is able to better predict the deviation from the training dataset when flame-turbulence interactions are considered. On the other hand, MG-PCA performs well only when the investigated configurations do not significantly differ from the one characterizing the training data-set, i.e. flamelet regime. From a modeling point of view the score-PCA method is then superior to MG-PCA. Furthermore, score-PCA does not require any selection of transported species, which is a crucial step for the quality of MG-PCA results (see ?).

MG-PCA offers two advantages over the score-PCA approach. It requires minor code modifications and it can be very easily formulated for clusters [?] thus reducing the number of transported species. The local formulation appears challenging for the score approach, since the base change between clusters leads to a modification of PCs transport equations, thus multiplying the problem dimensions.

Finally, for both methods, a combination of both fast and slow chemical species is used in the set of transported variables, thus indicating that the reduced model do not guarantee a reduction of the problem stiffness. In the context of score-PCA a stiffness reduction is observed only when a reduced number of PCs is considered, in combination with non-linear regression methods.

7. Acknowledgments

The first author was supported by a fellowship from the Fonds National de la Recherche Scientifique, FRS-FNRS (Communauté Française de Belgique). This material is based on work supported by the Department of Energy, National Nuclear Security Administration, under Award Number(s) DE-NA0002375. The last author would like to acknowledge the support of Fédération Wallonie-Bruxelles, via Les Actions de Recherche Concertée (ARC) call for 2014-2019, to support fundamental research.

References