This is a pre-copyedited version of a contribution published in Pitsch, H. and Attili, A. (eds) "Data Analysis for Direct Numerical Simulations of Turbulent Combustion", published by Springer, Cham. The definitive authenticated version is available online via https://doi.org/10.1007/978-3-030-44718-2_12

Analysis of turbulent reacting jets via Principal Component Analysis

G. D'Alessio, A. Attili, A. Cuoci, H. Pitsch and A. Parente

1 Abstract

The interpretation of high-dimensional data, like those obtained from Direct Numerical Simulations (DNS) of turbulent reacting flows, constitutes one of the biggest challenges in science and engineering. Although these simulations are a source of key information to advance the knowledge of turbulent combustion, as well as to develop and validate modeling approaches, the dimensionality of the data often limits the full opportunity to leverage the detailed and comprehensive information stored in data-sets.

G. D'Alessio
Université Libre de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium
CRECK Modeling Lab, Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20131 Milano, Italy
e-mail: giuseppe.dalessio@ulb.be

A. Attili
Institute for Combustion Technology, RWTH Aachen University, 52056 Aachen, Germany
e-mail: a.attili@itv.rwth-aachen.de

A. Cuoci
CRECK Modeling Lab, Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20131 Milano, Italy
e-mail: alberto.cuoci@polimi.it

H. Pitsch
Institute for Combustion Technology, RWTH Aachen University, 52056 Aachen, Germany
e-mail: h.pitsch@itv.rwth-aachen.de

A. Parente
Université Libre de Bruxelles, Aero-Thermo-Mechanics Laboratory, Bruxelles, Belgium
Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium
e-mail: alessandro.parente@ulb.be

The Principal Component Analysis (PCA), and its local formulation (LPCA), are widely used in many fields, including combustion. During the last 20 years, they have been used in combustion for the identification of low-dimensional manifolds, data analysis, and development of reduced-order models. Lower-dimensional structures, either global or local, can provide better insights on the underlying physical phenomena, and lead to the formulation of high-fidelity models.

This chapter aims to offer to the reader a comprehensive introduction of the PCA potential for data analysis, firstly introducing the main theoretical concepts, and then going through all the required computational steps by means of a `MATLAB®` code. Finally, the methodology is applied to data obtained from a DNS of a turbulent reacting non-premixed n-heptane jet in air. The latter can be regarded as an optimal case for data-analysis because of the complex physics characterized by turbulence-chemistry interaction and soot formation.

2 Theory

2.1 Building the data-set and data-set preprocessing

In order to apply any kind of statistical tool, data must be organized as matrices. The matrix $\mathbf{X}$, representing the original dataset, consists of n rows, which represent the statistically equivalent observations of a phenomenon, i.e., the different samples of an experiment, or the grid points of a numerical simulation, and p columns, which represent the variables of the problem, i.e., chemical species, velocity, temperature, pressure. Since the variables are characterized by different units and ranges, preprocessing in the form of centering and scaling is a mandatory operation [1, 2]. Data centering consists of subtracting the mean value of each variable to all data-set observations: in this way, all the observations can be seen as fluctuations from a mean value. Scaling is achieved by dividing each variable by a given scaling factor, which can be different depending on the adopted scaling criterion. Therefore, the *i-th* observation of the *j-th* variable, $x_{i,j}$, from the original data-set matrix $\mathbf{X}$, can be centered and scaled by means of Equation 1, where $\bar{x}_j$ and d_j are the centering and scaling factor for the considered j-th variable, respectively.

$$\tilde{x}_{i,j} = \frac{x_{i,j} - \bar{x}_j}{d_j} \tag{1}$$

The way data are preprocessed can have a strong influence on the data analysis and the reduced-order modeling for combustion applications [3], as the scaling technique can be more or less sensitive to the presence of outliers or it can highlight a specific pattern in the data:

1. **Auto scaling**: the standard deviation of each variable, $\sigma(\mathbf{x})$, is used as a scaling factor. After Auto scaling, all the variables are characterized by a standard devi-

ation equal to one. It is also one of the most used scaling criteria as it allows to consider evenly all the variables.
2. **Pareto scaling**: it uses the square root of the standard deviations for each variable, $\sqrt{\sigma(\mathbf{x})}$, as a scaling factor. More importance is given to variables with a very high standard deviation and with high numerical values. Importantly, the variables do not become dimensionless after scaling.
3. **Range scaling**: the difference between the minimum and the maximum value is adopted as scaling factor. It results to be more sensitive, if compared to the other scalings, to outliers, which can significantly change the numerical values of minimum and maximum.
4. **Vast scaling**: the scaling factor is the product between a variable's standard deviation and the coefficient of variation, i.e., the ratio: $\frac{\sigma(\mathbf{x})}{\text{mean}(\mathbf{x})}$. It has been proven to focus on the variables which do not show strong variation.

Auto scaling is the optimal option for combustion applications if the main objective is the reconstruction of the overall state space, with no major differences between the major and minor state variables [3]. The other scalings, such as Range scaling and Vast scaling, on the other hand, are more focused on the stable and major species [3]. In Table 1, the aforementioned scaling criteria, as well as their associated scaling factors, are summarized.

Scaling criterion	*Scaling factor (**d**)*
Auto	$\sigma(\mathbf{x})$
Pareto	$\sqrt{\sigma(\mathbf{x})}$
Range	$\max(\mathbf{x}) - \min(\mathbf{x})$
Vast	$\sigma(\mathbf{x})\frac{\sigma(\mathbf{x})}{\text{mean}(\mathbf{x})}$

Table 1 Scaling criteria and scaling factors for multivariate data-sets.

2.2 Principal Component Analysis

The Principal Component Analysis (PCA) is a statistical technique used to find a reduced set of uncorrelated variables, starting from a larger set of interdependent variables, losing only a small amount of information [4, 5]. Starting from a centered and scaled data matrix $\tilde{\mathbf{X}}$, consisting of n observations and p variables, it is possible to compute the associated covariance matrix $\mathbf{S}$ and decompose it by means of an eigenvalue decomposition:

$$\mathbf{S} = \frac{1}{n-1}\tilde{\mathbf{X}}^T\tilde{\mathbf{X}} \tag{2}$$

$$\mathbf{S} = \mathbf{A}\mathbf{L}\mathbf{A}^T \tag{3}$$

The columns of the matrix $\mathbf{A}$, whose size is $p \times p$, are an orthonormal basis of eigenvectors (Principal Components), while the diagonal elements of $\mathbf{L}$ correspond to their associated eigenvalues. Each eigenvalue represents a fixed percentage of information, in terms of variance of the original data-set, accounted by the associated Principal Component (PC). As the eigenvalues are ordered in descending order of magnitude $l_1 > l_2 > ... > l_p$, the PCs are also ordered in descending order of importance. The matrix $\tilde{\mathbf{X}}$ can be expressed as a function of the Principal Components by means of the scores matrix, $\mathbf{Z}$:

$$\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{A}. \tag{4}$$

With the linear transformation described in Equation 4, the original variables are recasted into a new set of uncorrelated variables. From a geometrical point of view, the axes of the new variables are represented by the columns of the matrix $\mathbf{A}$. Moreover, given the orthonormality of the latter, it results that: $\mathbf{A^T} = \mathbf{A^{-1}}$.
Thus, it is possible to uniquely recover the values of the original variables from $\mathbf{Z}$:

$$\tilde{\mathbf{X}} = \mathbf{Z}\mathbf{A^T}. \tag{5}$$

The dimensionality reduction comes by considering only a q-dimensional subset, $\mathbf{A}_q$, from the original p-dimensional full set of PCs, $\mathbf{A}$. If the cumulative variance, t_q, for the truncated q-dimensional basis of eigenvectors is within a desired accuracy, the basis can be considered as representative of the problem, and the original dataset $\tilde{\mathbf{X}}$ can be correctly compressed to the chosen reduced dimensionality finding the matrix of the scores, $\mathbf{Z_q}$. Equations 6 and 7 report the expressions for the cumulative variance, t_q, and the representation of the original data by means of the truncated basis of PCs, $\mathbf{A_q}$:

$$t_q = \frac{\sum_{j=1}^{q} \lambda_j}{\sum_{j=1}^{p} \lambda_j}, \tag{6}$$

$$\mathbf{Z_q} = \tilde{\mathbf{X}}\mathbf{A_q}. \tag{7}$$

The data matrix $\tilde{\mathbf{X}}$ can be reconstructed from the reduced-dimensionality space by means of Equation 8:

$$\tilde{\mathbf{X}} \approx \tilde{\mathbf{X}}_q = \mathbf{Z}_q\mathbf{A}_q^T. \tag{8}$$

The difference between the data matrix $\tilde{\mathbf{X}}$ and the reconstructed $\tilde{\mathbf{X}}_\mathbf{q}$ is defined as the low-rank approximation error, and it can be used to evaluate the quality of the dimensionality reduction. The low-rank approximation error, ϵ, can be defined as:

$$\epsilon = \sum_{j=q+1}^{p} \lambda_j = \sum_{i=1}^{n}\sum_{j=1}^{p}(\tilde{\mathbf{x}}_{q,ij} - \tilde{\mathbf{x}}_{ij})^2, \tag{9}$$

where $\tilde{\mathbf{x}}_{q,ij}$ and $\tilde{\mathbf{x}}_{ij}$ correspond to the lower-dimensional and the original observation, respectively.

The possibility to use PCA as a data-analysis tool comes by considering that each PC is a linear combination of the original variables of the data-set. The *j-th* variable will be characterized by a weight $w_{i,j}$, indicating how much it is represented by the *i-th* PC [6, 7]. Thus, analyzing the distribution of the weights on the retained PCs of $\mathbf{A_q}$, it is possible to gain a better insight about the features of the system. As PCA is particularly sensitive to the presence of outliers, an outlier removal procedure is recommended in the preprocessing step, before applying the algorithm, in case of analysis on data obtained from experimental set-ups [3].

In the case of data-sets with a large number of variables, it could be sometimes difficult to perform the analysis via visual inspection of the weights, because many variables could have comparable weights. Thus, the PCs physical interpretation can be aided by *rotation* methods, a class of statistical tools often coupled to PCA and other similar techniques such as Factor Analysis [7, 8]. The Varimax rotation, firstly developed by Kaiser [9], is an orthogonal rotation method which rigidly rotates the PCs over a fixed angle, while keeping the components orthogonal. When rotated, the subset of PCs spanning the lower dimensional space accounts for the same amount of cumulative variance as the unrotated, but it is redistributed within the components. Therefore, the information regarding the relative importance of the PCs, if the latter are rotated, is lost [4].

2.3 Local Principal Component Analysis

PCA is a linear technique, so the dimensionality reduction is limited when dealing with data-sets obtained from non-linear systems such as those in combustion, since large reconstruction errors are obtained.

One option to overcome the intrinsic limitation of the PCA is to adopt a piecewise linear, local formulation for the dimensionality reduction (LPCA). Partitioning the data in k groups (clusters) and then performing the dimensionality reduction in each of them separately, can lead to a drastic decrease of the reconstruction error.

Two methods [10, 11] are available to perform the data-set partitioning: an iterative unsupervised algorithm based on the minimization of the reconstruction error, the Vector Quantization Principal Component Analysis (VQPCA), or a supervised partitioning based on an *a-priori* conditioning, by means of a selected variable which is known to be important for the process (FPCA). As the latter is not an iterative algorithm, it allows for a faster clustering in comparison with VQPCA, even if the choice of the optimal variable could constitute a difficult task for some applications, as it requires prior knowledge on the process, and the choice must be assessed case-by-case. For non-premixed, turbulent combustion applications, the mixture fraction Z is an optimal variable for the data conditioning, leading to excellent results both for data compression and interpretation tasks [11]. In the present approach, the FPCA algorithm groups the data in k bins: $k/2$ are allocated for all the observations under the condition of Z being lower than the stoichiometric mixture fraction Z_{st}, and the remaining $k/2$ for the observations at $Z > Z_{st}$.

The iterative VQPCA algorithm, instead, is based on the following steps:

1. *Initialization*. The cluster centroids $\tilde{\mathbf{r}}^{(\mathbf{k})}$ are initializated: a random allocation, a uniform distribution between all the observations of the data-set or a previous clustering solution can be chosen to compute the $\tilde{\mathbf{r}}^{(\mathbf{k})}$ initial values. The eigenvectors in each cluster, $\mathbf{A}^{(\mathbf{k})}$, are initialized as identity matrices.
2. *Partition*. Each observation is assigned to a cluster k such that the local reconstruction error is minimized:

$$\epsilon(\tilde{\mathbf{x}}_\mathbf{i}, \tilde{\mathbf{r}}^{(\mathbf{k})}) = (\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{r}}^{(\mathbf{k})})^T \mathbf{A}_q^{(k)T} \mathbf{A}_q^{(k)} (\tilde{\mathbf{x}}_\mathbf{i} - \tilde{\mathbf{r}}^{(\mathbf{k})}) \tag{10}$$

3. *PCA*. The Principal Component Analysis is performed in each of the clusters found in the previous step. A new set of centroids is computed after the new partitioning step: their coordinates are calculated as the mean of all the observations in each cluster.
4. *Iteration*. All the previous steps are iterated until convergence is reached.

The available convergence criteria are the following:

i. The global mean reconstruction error, i.e., the averaged reconstruction error taking into account all the clusters, is below a fixed threshold.
ii. All the clusters centroids positions are not changing between two consecutive iterations.
iii. The variation of the global mean reconstruction error between two consecutive iterations is below a fixed threshold.

3 Methods and MATLAB® code for data analysis with PCA

In this Section, a simple procedure for data analysis via PCA with the corresponding MATLAB® code is presented, in order to show in details the application of the theory explained in Section 2. To perform data analysis with PCA, the following steps must be followed:

i. Standardize the initial data-set by means of centering and scaling.
ii. Perform PCA (global or local algorithm).
iii. Evaluate how many PCs are necessary to properly describe the system using Equation 6.
iv. Evaluate the variables' weights on the retained PCs, and possibly apply a rotation to increase interpretability.

After the data have been organized as a matrix and loaded in MATLAB®, the first operation to accomplish is the standardization. The code for centering and scaling a generic $\mathbf{X}$ matrix to obtain $\tilde{\mathbf{X}}$ is reported below:

Data pre-processing

```
[n_obs, n_var] = size(X);
tolerance = 1e-08;

% First of all, the mean value for each variable
% is calculated.
xbar = mean(X);

% Depending on the scaling criterion, the
% scaling factor is calculated:
switch upper(scaling)
  case {'NONE' ''}
    d = ones(1,nvar);
  case {'AUTO' 'STD'}
    d = zeros(1,n_var);
    for i=1:n_var
      d(i) = std(X(:,i),1);
    end
  case 'VAST'
    d = zeros(1,n_var);
    for i=1:n_var
      d(i) = std(X(:,i),1).^2 ./ (xbar(i));
    end
  case 'RANGE'
    d = max(X)-min(X);
  case 'PARETO'
    d = zeros(1,n_var);
    for i=1:n_var
      d(i) = sqrt(std(X(:,i),1));
    end
  otherwise
    error('Unsupported scaling option');
end

% All the observations can be now centered and scaled.
% In case of big matrices it is convenient to preallocate
% the memory.
X_tilde = zeros(size(X));
for i = 1: n_var
X_tilde(:,i) = (X(:,i) - xbar(i)) / (d(i) + tolerance);
end
```

After these operations, it is possible to perform PCA.

Principal Component Analysis

```
% The covariance matrix of the centered and scaled data
% must be computed.
cov_data = cov(X_tilde, 1);

% The eigenvectors and the eigenvalues are calculated
% from the covariance matrix.
% The eigenvalues, originally returned as a diagonal matrix,
% are stored into a vector (lambda).
[eigenvectors, eigenvalues] = eig(cov_data);
lambda = diag(eigenvalues);

% The eigenvalues must be now sorted in descending order.
% Their original indeces (sort_index) are also stored, as
% they are later used to arrange the eigenvectors in order
% of descending importance, thus building the matrix of
% the Principal Components (PCs).
[sort_eigval, sort_index] = sort(lambda, 'descend');
PCs = zeros(n_var, n_var);
for i = 1 : n_var
    PCs(:,i) = eigenvectors(:, sort_index(i));
end

% Two different kinds of principal component scores can be
% now computed.
% 1) U-scores: obtained by projecting the matrix X_tilde of
% the centered and scaled data on the PCs.
% The resulting U-scores are uncorrelated and have variances
% equal to the corresponding eigenvalues.
% 2) W-scores: obtained by projecting the matrix X_tilde of
% the centered and scaled data on the PCs previously
% scaled by the inverse of the eigenvalues square root.
% The W-scores are still uncorrelated and have variances
% equal to 1.
U_vec = PCs;
W_vec = zeros(n_var, n_var);
for j = 1 : n_var
    W_vec(:, j) = (PCs(:, j)/sort_eigval(j)^0.5);
end
U_scores = X_tilde*U_vec;
W_scores = X_tilde*W_vec;
```

In alternative, a built-in function is also already available in MATLAB®:

Principal Component Analysis

```
[PCs, U_scores, Eigenvalues, ~, Explained] = ...
pca(X_tilde, 'Centered', false);
```

In this function, as well as providing the $n \times p$ standardized matrix in input, it is specified that the data have already been centered and scaled. In the output, the function returns:

i. **PCs**: the PCs matrix $\mathbf{A}$ containing all the principal components, whose size is $p \times p$.
ii. **U_scores**: the scores matrix, $\mathbf{Z}$, which consist of the projection of the input matrix on the full set of PCs: $\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{A}$.
iii. **Eigenvalues**: the eigenvalues vector ($p \times 1$), containing the eigenvalues associated to each PC.
iv. **Explained**: the explained variance vector ($p \times 1$), which consists of the percentage of explained variance by each PC depending on the magnitude of the associated eigenvalue.

This last output is important for the next step, the choice of the number of PCs. As already explained in Section 2, a good approximation of the original problem requires selecting a basis that can explain a large amount of cumulative variance, i.e, from 95% to 100%. In this way, it is possible to have an initial guess for the dimensionality needed by the reduced basis to be representative. In the following code, it is required that more than 99% of the global variance is explained by the retained Principal Components:

Choice of the reduced dimensionality

```
cumulative_explained = cumsum(Explained)/sum(Explained);
variance_cut = find(cumulative_explained > 0.99);
required_number = variance_cut(1);
```

Once the required number of PCs is calculated, the reduced eigenvectors basis, $\mathbf{A_q}$, can be built as:

Selecting the PCs

```
A_q = PCs(:, 1:required_number);
```

Sometimes the explained variance criterion might not be enough to assess the number of required PCs, especially if the data are not standardized with the Auto scaling criterion. Thus, another method must be taken into account to verify if the choice of the number of PCs is appropriate, checking the reconstruction error for the single variables. If the number of PCs is correctly determined, the variables' reconstruction from the reduced dimensionality is characterized by a low error, otherwise the number of the retained PCs must be increased. A large reconstruction error can have a negative impact on the analysis, as the feature extraction process from the data could also be compromised. In fact, the distribution of the weights on the modes could be too noisy or some important processes might not be extracted. The code to reconstruct the original matrix from the reduced dimensionality space is reported below, assessing the quality of the reconstruction using parity plots of the original and the reconstructed variables, after uncentering and unscaling the data.

Reconstruction error for the variables

```
% The original matrix must be reconstructed using the set of
% truncated modes calculated in the previous steps, first.
recovered_X_fromPCA = X_tilde*A_q*A_q';

% This matrix must be then unscaled and uncentered
% (in this order), with the same scaling and centering
% factors used in the previous  steps, to make a proper
% comparison with the original matrix, X.

% Unscaling
for i = 1:n_var
X_unscal(:, i) = recovered_X_fromPCA(:,i)*d(i);
end

% Uncentering
for i = 1:n_var
X_recovered(:,i) = X_unscal(:,i) + xbar(i);
end

% The parity plots between the original and the reconstructed
% variables can then be drawn to evaluate the reconstruction:
% the more the scatter points are aligned with the red solid
% line, the more precise is the PCA reconstruction. In order
% to have  a quantitative indication, it is also possible to
% use error metrics such as the Mean Square Error or the
```

```
% Root Mean Square Error.
for i = 1:n_var
figure, plot(X(:,i), X(:,i), 'r', 'LineWidth', 2);
hold on
scatter(X(:,i), X_recovered(:,i), 15, 'filled');
xlabel('Original variable');
ylabel('Reconstructed variable');
end
```

For a local, supervised, partitioning in k clusters, using mixture fraction as a conditioning variable, the following code can be implemented:

Local partitioning via FPCA

```
% Initialization of bin data matrices
bin_data = cell(k, 1);
idx_clust = cell(k, 1);
idx = zeros(size(X_tilde,1), 1);

% Number of intervals
n = k + 1;
min_z = min(Z);
max_z = max(Z);

ints_1 = linspace(min_z, z_stoich, ceil(n/2));
ints_2 = linspace(z_stoich, max_z, ceil((n+1)/2));
ints = [ints_1(1:ceil(n/2-1)) ints_2];

%  Partition
for bin = 1 : k
   idx_clust{bin} = find((Z>=ints(bin))&(Z<=ints(bin+1)));
   bin_data{bin} = X_tilde(idx_clust{bin}, :);
   idx(idx_clust{bin}) = bin;
end

% Perform PCA in each bin after the centroid has been removed
PCs_clusters = cell(k,1);
for i = 1:k
[rows, columns] = size(bin_data{i});
mean_var = mean(bin_data{i}, 1);
X_ave = repmat(mean_var, rows, 1);
X0 = bin_data{i} - X_ave;
[PCs, Scores, Eigenvalues, ~, Explained] = ...
```

```
pca(X0, 'Centered', false);
PCs_clusters{i} = PCs;
end

% Plot the weights on the PCs to analyze the data
for j = 1:k
A_q = PCs_clusters{j}(:, 1:required_number);
for i = 1: required_number
figure,bar(A_q(:,i),'FaceColor',[0, 0, 0],'LineWidth',1.5)
xlabel('Variables');
ylabel('Weights on the PC');
end
end
```

---

4 Application: PCA of a non-premixed sooting flame DNS

4.1 Data description

Data obtained from a 2D slice of a 3D temporally evolving DNS simulation of an n-heptane turbulent jet [12] are here considered for the analysis by means of PCA. The jet is non-premixed with a Reynolds number of 15,000. The fuel has an initial temperature of 400 K, while the oxidizer stream (air) is at 800 K. The kinetic mechanism used for the n-heptane flame consists of 47 species, including naphthalene and other soot precursors, with 290 reactions in total. Additional information on the mechanism and the gas phase hydrodynamics can be found in [12, 13, 14, 15]. This can be considered as an optimal case for a data-analysis task since it includes a large number of available observations and chemical species in the mechanism, and the physics is characterized by many complex phenomena such as turbulence-chemistry interaction and soot formation. The data-set considered here consists of the full thermo-chemical space and it is organized as a matrix of 1,048,576 observations (grid points of the 2D slice) and 48 variables (temperature and mass fractions of all the chemical species).

4.2 Analysis

4.2.1 Principal Component Analysis

The matrix **X** containing the input data was scaled with the four scalings methods discussed in Section 2 (Auto, Vast, Pareto, Range) to test their effect on the analysis. A different relative importance given to the PCs is observed in the four cases, due to the different eigenvalues magnitude distribution. As it is shown in Figure 3, for Auto and Range scaling criteria, the curve representing the cumulative explained variance has a more moderate slope, starting from small values (i.e. about 0.4 and 0.65, respectively) and then asymptotes to one for a relatively large number of PCs, while for Pareto and Vast scaling criteria, the first principal component already explains an almost unitary cumulative variance. With Auto scaling, the number of eigenvectors to retain in order to explain at least 99% of the data cumulative variance is 19, while with Pareto scaling 1 PC is already enough to explain 99% of the total data variation. The examination of the cumulative explained variance is not always optimal for the assessment of the number of PCs to analyze, so the reconstruction error for the variables must be also investigated.

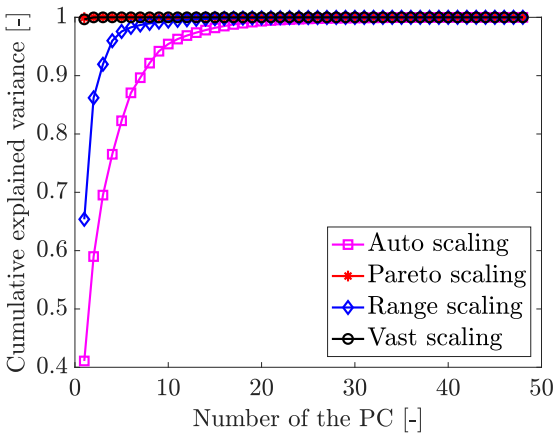

Fig. 1 Cumulative explained variance for the PCs for different scaling criteria.

If the data are reconstructed from nineteen-dimensional compressed space and from the one-dimensional compressed space, respectively, the results in terms of accuracy of the reconstruction are totally different. Even if the amount of explained cumulative variance is the same, the coefficient of determination, R^2, for the parity-plots obtained from the variables' reconstruction are completely different: for the reconstruction of the oxygen radical, in case of auto-scaling with 19 PCs, the R^2 amounts to 0.996, while in case of pareto-scaling with 1 PC it amounts to 0.218. The parity plots for the reconstructed variable using the two different scalings are

shown in Figure 2. The only variable which has an acceptable reconstruction error with only 1 PC, in case of Pareto scaling, is the temperature with a R^2 equal to 1.

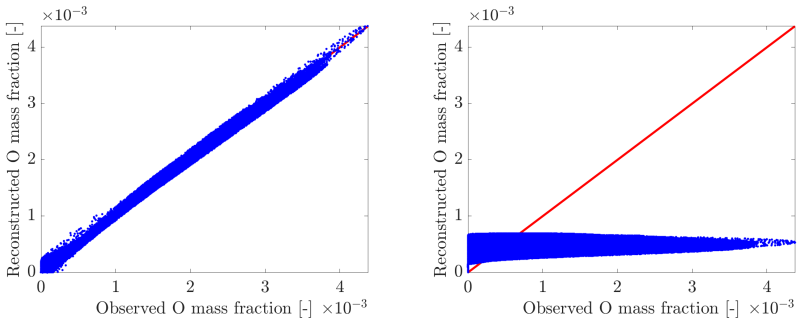

Fig. 2 Left: Parity plot for the reconstruction of the oxygen radical from the compressed space adopting Auto scaling with 19 retained PCs; Right: Parity plot for the reconstruction of the oxygen radical from the compressed space adopting Pareto scaling with 1 retained PC; on equal terms of cumulative explained variance ($t_q > 0.99$).

As highlighted in Section 2, the analysis is performed via examination of the weights' distribution on the PCs, and in many cases they tend to represent one or more physical quantities. For example, examining the sixth rotated PC obtained from the data scaled with Auto scaling reported in Figure 3, it is possible to notice that nitrogen and oxygen have the largest negative weights, while n-heptane has the largest positive weight. This PC clearly represents the mixture fraction, as also confirmed by its correlation coefficient with the mixture fraction itself, which is equal to 0.72. This high correlation is particularly interesting because the mixture fraction was not included in the variables of the data-set, which consisted only of temperature and species mass fractions.

Other global important features extracted via PCA were, for example, the most important radicals involved in the branching reactions (O, OH, H on one rotated PC and HO_2 on another) and in the soot formation mechanism (C_6H_5, C_7H_8, C_3H_3 with highest weights on the third rotated PC).

4.2.2 Local Principal Component Analysis

The quality of the data analysis can be enhanced if a local formulation is considered. A piecewise linear local formulation for PCA has several advantages with respect to a global analysis: a lower reconstruction error, a lower intrinsic reduced dimensionality and the possibility to highlight local processes. The algorithm has only one hyperparameter, namely the choice of the number of bins of mixture fraction (the number of clusters) to use to partition the data, k. Despite the fact that a defined way to properly set the number of variables does not exist in literature, the total number of clusters can be retrieved from a trade-off between the accuracy of the reconstruction

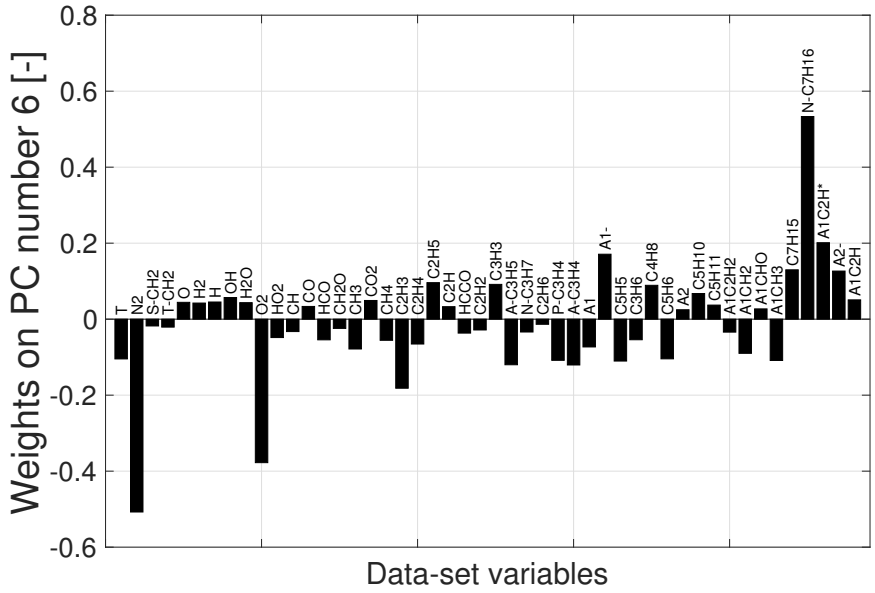

Fig. 3 Weights distribution on the sixth Principal Component, Auto scaling criterion.

and the feasibility of the data-analysis. In fact, as the number of clusters grows, the reconstruction error decreases, but the analysis with an excessively large k could be infeasible.

In Figure 4, the reconstruction error obtained with FPCA for an increasing number of clusters, starting from the global PCA ($k = 1$), is reported. The error decreases in total of one order of magnitude, and it starts asymptoting from $k = 16$. The latter could also be ideally chosen as a good k to perform the analysis as it is a reasonable number of clusters to examine, not being too large for the manual weights inspection and interpretation, as $k = 64$ or $k = 128$ could be. LPCA is also robust to underfitting and overfitting, as the reconstruction error does not depend on the data-set dimensions, as also shown in the aforementioned figure where the errors using the full dataset and only the 50% of the observations are compared.

As discussed in Section 2, for the FPCA algorithm the mixture fraction space is divided in k/2 bins for all the points below the stoichiometric mixture fraction and in k/2 above Z_{st}, while VQPCA assigns the cluster index to one particular observation on the basis of the reconstruction error minimization criterion. In Figures 5 and 6, the flame partitionings for $k = 16$ for the two different LPCA algorithms are reported. As expected, the results are different; even if the VQPCA partitioning is completely unsupervised, it gives better results in terms of data compression, as its average reconstruction error for $k = 16$ amounts to 0.0121, while with FPCA, the reconstruction error for the variables is 0.0256 on equal terms of number of clusters and number of retained principal components. In each of the sixteen clusters, the

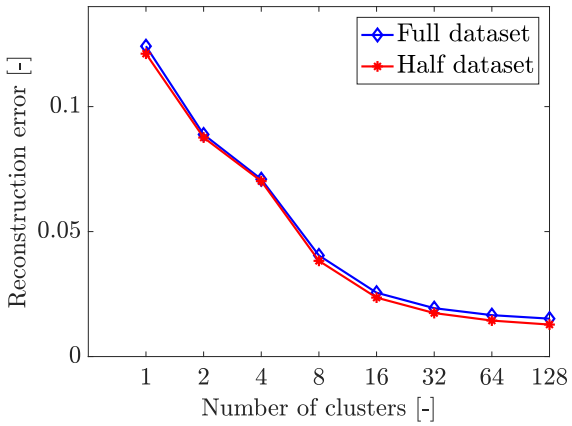

Fig. 4 Reconstruction error for an increasing number of bins of mixture fraction for the full data-set (all the observations) and a data-set consisting only of half of the total observations. The value for number of clusters equal to 1 is the error with global PCA.

main features of the flame can be visualized plotting the weights of the variables on the modes as done in the previous paragraph.

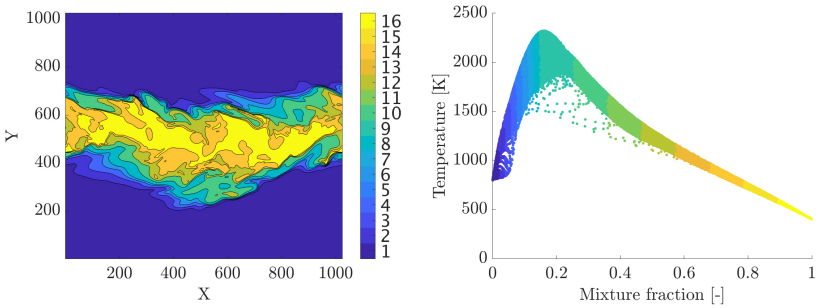

Fig. 5 Left: DNS flame partitioning via FPCA with $k = 16$; Right: mixture fraction partitioning via FPCA with $k = 16$. The colorbar indicates the index of the cluster assignment.

As shown in the mixture fraction - temperature plot, coloured by means of the LPCA partitioning, the unsupervised algorithm allocates only two clusters for the lean conditions (from zero up to stoichiometric), while with the supervised approach, eight clusters were intentionally allocated for low values of Z. One of these two clusters found via VQPCA for the lean conditions contains all the points below the curve (in light green, cluster number 11), and it groups all the points of local extinction, characterized by quasi-stoichiometric and stoichiometric mixture fraction, low temperature, and OH concentration almost equal to zero. In the right branch of the

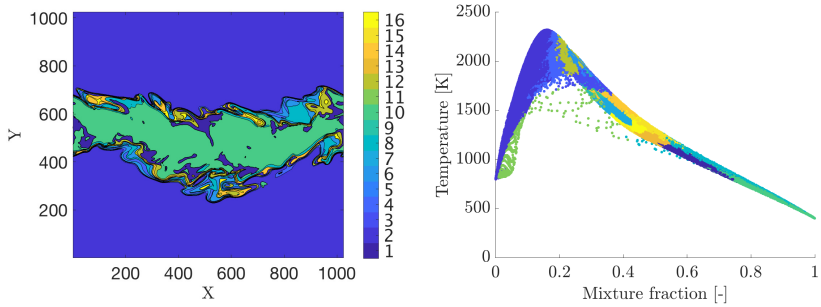

Fig. 6 Left: DNS flame partitioning via VQPCA with $k = 16$; Right: mixture fraction partitioning via VQPCA with $k = 16$. The colorbar indicates the index of the cluster assignment.

mixture fraction - temperature plot, corresponding to rich conditions, the clusters found via the unsupervised algorithm have a larger extension in terms of mixture fraction range if compared to the supervised ones. In this case, separate clusters are assigned to points at the same mixture fraction range, but different temperature. Examining the plot on the right of Figure 6, indeed, it is possible to see that cluster number 9 goes from $Z \sim 0.5$ to $Z \sim 0.95$, the same mixture fraction covered by clusters number 1 and number 10, but with the latter being at a lower temperature. This partitioning is totally in line with the physics of the flame, as these clusters are representative of different chemical features. In fact, while clusters number 1 and 10 are representative for the fuel jet and its decomposition, as the highest weights on the first modes are representative for species such as n-C_7H_{16}, C_7H_{15}, C_5H_{10}, C_4H_8, C_2H_5, CH_2O, cluster number 9 is representative of soot precursors, as in the first two modes the species characterized by highest weights are all aromatics involved in the soot formation such as naphtalene and its naphtyl radical $C_{10}H_8$, $C_{10}H_7^-$, benzyl radical and ethynyl benzene C_7H_7, C_8H_7. Thus, the FPCA algorithm, despite the conditioning variable proven to be optimal for turbulent non-premixed reacting jets, was not capable to be competitive with VQPCA for data analysis purposes as the physics of the jet were too complex, involving local extinction phenomena and the dynamics of soot precursors.

These data-analysis methods can also be coupled to other techniques, such as Principal Variables (PV) [4, 16, 17]. Its purpose is to find a relationship between the Principal Components and a subset of the original variables by means of the maximization of the variance of the original data. Many strategies to find the PVs are available in the literature: in the present work, the B2 backward method [4] was tested. This technique has been successfully used for reduced order modeling in combustion applications [17, 18, 19, 20]. The variables extracted by the latter can be grouped in three main categories: radicals involved in branching reactions, such as H, HO_2, CH_3, C_7H_{15}, stable species, such as O_2, CO_2, and species involved in the soot formation mechanism, like aromatic compounds $C_{10}H_8$, C_8H_7, C_7H_7, C_7H_6O and propargyl C_3H_3. Many of these species were also extracted by means of a direct analysis of the clusters' weights found via VQPCA, as a proof of the

effectiveness of the local PCA for data analysis tasks. Moreover, the possibility to know the spatial position of the cluster where these features are important constitutes a relevant property of the partitioning algorithm. On the other hand, the PV algorithm offers the possibility to identify features in an automated fashion, without having to visually inspect the weights of the LPCA modes, which could be unfeasible in case of a high number of clusters or retained modes for each cluster.

An optimal solution for data analysis could be represented by a hybrid method VQPCA-PV. The first algorithm can be used to partition the original data-set in k groups according to the reconstruction error minimization, and then the principal variables can be found in each cluster. In this case, applying the PV algorithm in each cluster found with the previous VQPCA partitioning led to similar results in terms of extracted features, but in a totally unsupervised fashion, without a visual inspection of the weights to be required. In Figure 7, the contours of the mass fractions of naphthalene and propargyl, two species identified as principal variables in cluster number 9, are reported, and it can be observed that the highest concentrations for these species are observed in the region of the geometrical domain corresponding to that cluster, which is reported, colored in yellow, in Figure 8. In fact, this cluster resulted again to be associated to soot formation, with the chemical species C_6H_5, C_6H_7, C_7H_8, $C_{10}H_7$, C_2H_3 identified as *local* principal variables.

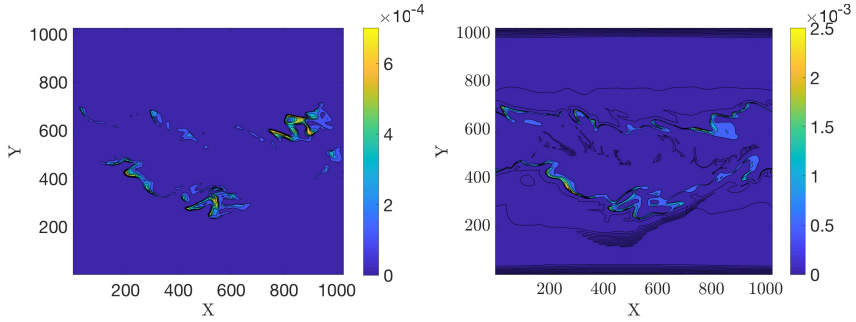

Fig. 7 Left: mass fraction contours of naphtalene; Right: mass fraction contours of propargyl.

5 Conclusions

The present work investigates the potential of the Principal Component Analysis to analyze data obtained from Direct Numerical Simulations of turbulent reacting flows.

The Principal Component Analysis is widely used in many fields for dimensionality reduction, but it can also be exploited for data analysis tasks. In fact, if the original variables' weights distribution on the principal components is examined, it is possi-

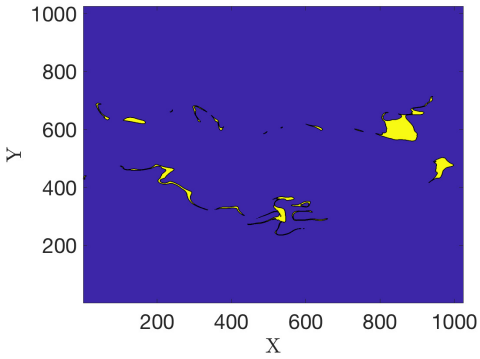

Fig. 8 Cluster number 9 obtained via VQPCA partitioning (in yellow).

ble to obtain a physical interpretation for the latter, and an insight about the system features can be gained. Moreover, two local formulations of the PCA are available to overcome the limitations due to the linearity of the method: an iterative unsupervised algorithm, based on the minimization of the reconstruction error (VQPCA), and a supervised partitioning algorithm, based on an a-priori conditioning by means of a selected variable which is known to be important for the process (FPCA). With the last two algorithms, the local phenomena, which could have been overlooked by a global analysis, can be highlighted.

The aforementioned techniques were tested on the analysis of data obtained in a 3D temporally evolving DNS of an n-heptane turbulent jet in air. The data-set consisted of the full thermo-chemical space, organized as a matrix of 1,048,576 observations (grid points of the simulation) and 48 variables (temperature and mass fractions of all the chemical species).

The global PCA was able to recognize the key-role covered by the mixture fraction in the process even if it was not included in the variables of the dataset, as well as to highlight the most important radicals involved in the branching reactions and in the soot formation mechanism. The analysis done using the local algorithms was more effective both in terms of reconstruction error and feature extraction, as more physical processes such as: decomposition of the fuel jet, local extinction phenomena, branching reactions and soot formation were highlighted. In particular, VQPCA was more effective than FPCA because the conditioning variable, the mixture fraction, although proven to be optimal for turbulent non-premixed reacting jets, was not capable to deal with the complex physics of the system, characterized by local extinction phenomena and the dynamics of soot precursors. Finally, an hybrid algorithm coupling VQPCA with the Principal Variables method was proposed. This led to similar results as with VQPCA in terms of extracted features, but in a totally automated fashion, without requiring a visual inspection of the weights.

Acknowledgements The first author acknowledges the support of the Fonds National de la Recherche Scientifique (FRS-FNRS) through a FRIA fellowship.
A.A. and H.P. acknowledge funding from the European Research Council (ERC) under the European Unions Horizon 2020 research and innovation program under grant agreement No 695747.
A.P. acknowledges funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program, grant agreement No 714605.

References

1. Bro R, Smilde AK. Centering and scaling in component analysis. Journal of Chemometrics. 2003 Jan;17(1):16-33.
2. van den Berg RA, Hoefsloot HC, Westerhuis JA, Smilde AK, van der Werf MJ. Centering, scaling, and transformations: improving the biological information content of metabolomics data. BMC Genomics. 2006 Dec;7(1):142.
3. Parente A, Sutherland JC. Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity. Combustion and Flame. 2013 Feb 1;160(2):340-50.
4. Jolliffe I. Principal component analysis. Springer Berlin Heidelberg; 2011.
5. Bishop CM. Pattern recognition and machine learning. Springer; 2006.
6. Parente A, Sutherland JC, Dally BB, Tognotti L, Smith PJ. Investigation of the MILD combustion regime via principal component analysis. Proceedings of the Combustion Institute. 2011 Jan 1;33(2):3333-41.
7. Bellemans A, Aversano G, Coussement A, Parente A. Feature extraction and reduced-order modelling of nitrogen plasma models using principal component analysis. Computers & Chemical engineering. 2018 Jul 12;115:504-14.
8. Richman MB. Rotation of principal components. Journal of climatology. 1986;6(3):293-335.
9. Kaiser HF. The varimax criterion for analytic rotation in factor analysis. Psychometrika. 1958 Sep 1;23(3):187-200.
10. Kambhatla N, Leen TK. Dimension reduction by local principal component analysis. Neural computation. 1997 Jul 10;9(7):1493-516.
11. Parente A, Sutherland JC, Tognotti L, Smith PJ. Identification of low-dimensional manifolds in turbulent flames. Proceedings of the Combustion Institute. 2009 Jan 1;32(1):1579-86.
12. Attili A, Bisetti F, Mueller ME, Pitsch H. Formation, growth, and transport of soot in a three-dimensional turbulent non-premixed jet flame. Combustion and Flame. 2014 Jul 1;161(7):1849-65.
13. Attili A, Bisetti F, Mueller ME, Pitsch H. Effects of non-unity Lewis number of gas-phase species in turbulent nonpremixed sooting flames. Combustion and Flame. 2016 Apr 1;166:192-202.
14. Attili A, Bisetti F, Mueller ME, Pitsch H. Damkohler number effects on soot formation and growth in turbulent nonpremixed flames. Proceedings of the Combustion Institute. 2015 Jan 1;35(2):1215-23.
15. Attili A, Bisetti F. Application of a robust and efficient Lagrangian particle scheme to soot transport in turbulent flames. Computers & Fluids. 2013 Sep 15;84:164-75.
16. McCabe GP. Principal variables. Technometrics. 1984 May 1;26(2):137-44.
17. Isaac BJ, Coussement A, Gicquel O, Smith PJ, Parente A. Reduced-order PCA models for chemical reacting flows. Combustion and Flame. 2014 Nov 1;161(11):2785-800.
18. Isaac BJ, Parente A, Galletti C, Thornock JN, Smith PJ, Tognotti L. A novel methodology for chemical time scale evaluation with detailed chemical reaction kinetics. Energy & Fuels. 2013 Mar 27;27(4):2255-65.
19. Coussement A, Isaac BJ, Gicquel O, Parente A. Assessment of different chemistry reduction methods based on principal component analysis: Comparison of the MG-PCA and score-PCA approaches. Combustion and Flame. 2016 Jun 1;168:83-97.

20. Coussement A, Gicquel O, Parente A. MG-local-PCA method for reduced order combustion modeling. Proceedings of the Combustion Institute. 2013 Jan 1;34(1):1117-23.