Principal Component Analysis of Turbulent Combustion Data: data pre-processing and manifold sensitivity

Alessandro Parente[a,*], James C. Sutherland[b]

[a]*Service d'Aéro-Thermo-Mécanique, Université Libre de Bruxelles, Bruxelles, Belgium*
[b]*Department of Chemical Engineering, University of Utah, Salt Lake City, UT, 84112, USA*

Abstract

Principal Component Analysis has demonstrated promise in its ability to identify the low-dimensional chemical manifolds of turbulent reacting systems by providing a basis for the *a priori* parameterization of such systems based on a reduced number of parameterizing variables. Previous studies on PCA have only mentioned the importance of data pre-processing and scaling on the PCA analysis, without detailed consideration. This paper assesses the influence of data-preprocessing techniques on the size-reduction process accomplished through PCA. In particular, a methodology is proposed to identify and remove outlier observation from the datasets on which PCA is performed. Moreover, the effect of centering and scaling techniques on the PCA manifold is assessed and discussed in detail, to investigate how different scalings affect the size of the manifold and the accuracy in the reconstruction of the state-space. Finally, the sensitivity of the chemical manifold to flow characteristics is considered, to investigate the invariance of the manifold with respect to the Reynolds number. Several high-fidelity experimental datasets fro the TNF workshop database are considered in the present work, to demonstrate the effectiveness of the proposed methodologies.

[*]Corresponding author. Phone + 32 2 650 26 80 Fax +32 2 650 27 10 Address: Avenue F. D. Roosevelt 50, 1050 Bruxelles, Belgium.

Email address: Alessandro.Parente@ulb.ac.be (Alessandro Parente)

1. Introduction

Recently, principal component analysis (PCA) was introduced as a method of identifying manifolds in turbulent combustion [?]. PCA has also been used by others to analyze combustion data [? ? ?], but for different purposes - see [?] for a discussion. The merits of PCA in the context of modeling turbulent reacting flows have been demonstrated for identifying low-dimensional manifolds underlying the thermo-chemical state [? ?] and toward the development of PCA-based combustion models [? ?]. A particularly noteworthy feature of PCA-based models is the possibility of obtaining low-dimensional parameterizations satisfying well-defined error bounds. Previous studies on PCA [? ?] have mentioned the importance of pre-processing data prior to applying PCA, but the effects of pre-processing strategies have not been assessed in detail. In particular, the effect of potential outlier observations as well as the role of centering and scaling on the principal component structure has not been addressed. The objective of the present paper is to review the PCA procedure and highlight the role of the available pre-processing techniques on the robustness of PCA and its ability to identify a low-dimensional representation of a thermo-chemical manifold. The sensitivity of PCA to modifications of the database from which the low-dimensional basis is extracted is also considered, to investigate the universality of the PCA method.

Section 2 provides a review of PCA as well as a discussion on outlier removal (2.1), data centering and scaling (2.2), and dimension reduction (2.3). Section 3, applies PCA to several experimental datasets from the Sandia non-premixed flame datasets, to illustrate the effect of pre-processing and scaling on the PCA reduction. Finally, the invariance of the chemical manifold with respect to the Reynolds number is demonstrated for a set of piloted flames at different Reynolds number.

2. Principal Component Analysis

Principal Component Analysis (PCA) provides a rigorous mathematical formalism for the identification of the most active directions in multivariate datasets. PCA identifies correlations among the variables defining the state space. As a result, a new coordinate system is identified in the directions of maximal data variance, which allows less important dimensions to be eliminated while maintaining the primary structure of the original data. Details of the PCA reduction have been already provided [**?**]. Here, the PCA concept will be reviewed briefly whereas the impact of pre-processing and post-processing on PCA results will be discussed in detail.

In PCA, n observations of Q variables are assigned to an $(n \times Q)$ matrix $\mathbf{X}$ whose rows represent individual observations of all Q variables $\boldsymbol{x}$. For the combustion applications considered in this paper, the Q columns in $\mathbf{X}$ are taken to be the temperature and species mass fractions[1]. PCA projects $\boldsymbol{x}$ onto a rotated basis obtained from the eigenvalue decomposition of the $(Q \times Q)$ covariance matrix,

$$\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X} = \mathbf{A}\mathbf{L}\mathbf{A}^T, \tag{1}$$

where $\mathbf{A}$ and $\mathbf{L}$ are the eigenvectors and eigenvalues of $\mathbf{S}$. The rotated basis, defined by the eigenvectors $\mathbf{A}$, may be truncated to retain the most energetic directions (those columns of $\mathbf{A}$ associated with the largest eigenvalues of $\mathbf{L}$), providing the non-square matrix $\mathbf{A}_q$ on which the original data are projected to obtain the principal components (PC), $\mathbf{Z}_q$,

$$\mathbf{Z}_q = \mathbf{X}\mathbf{A}_q. \tag{2}$$

Eq. (2) can be inverted to obtain an approximate reconstruction of the original $(n \times Q)$ dimensional sample:

$$\mathbf{X}_q = \mathbf{Z}_q\mathbf{A}_q^T. \tag{3}$$

Note that (3) is a linear reconstruction. Nonlinear reconstructions can provide more accurate mappings from $\mathbf{Z}_q$ to $\mathbf{X}_q$ [**?**]. The PCA reduction process is represented schematically in Figure 1.

Several procedures are required prior to performing the PCA reduction process (Figure 1):

[1] Formally, pressure should also be included, but for low mach number flows in open domains, it is safely neglected.

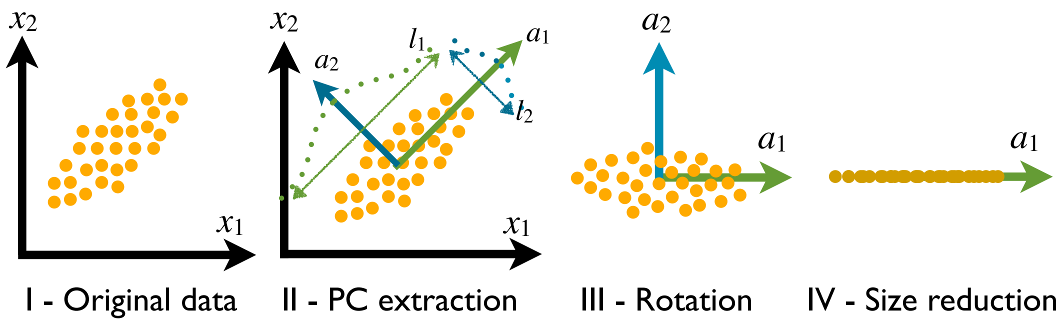


Figure 1: PCA reduction process.

1. *Outlier removal*

   Experimental datasets usually contain a few unusual observations which can strongly affect the data covariance structure and, therefore, the structure for the principal components. If we refer to a one-dimensional problem, the outliers can be classified as those observations which are either very large or very small with respect to the others. In high dimensions, there can be outliers that do not appear as outlying observations when considering each dimension separately and, therefore, they will not be detected using univariate criteria. Thus, a multivariate approach must be pursued. PCA itself represents an ideal tool for the identification and removal of outlier observations.

2. *Centering and scaling*

   Data are usually *centered* and *scaled* before PCA is carried out. Centering represents all observations as fluctuations, leaving only the relevant variation for analysis. Scaling is a crucial operation when analyzing the thermochemical state of a reacting system since temperature and species concentrations have different units and vary over different scales. The choice of scaling significantly affects the subsequent PCA analysis: different scalings allow to emphasize correlations among different groups of state variables, providing an effective tool for targeting the PCA analysis on the variables which are most relevant for an investigated application.

Section 2.1 presents a technique to identify outliers, while §2.2 addresses centering and scaling.

2.1. Outlier Detection and Removal with PCA

The usual procedure for outlier detection in multivariate data analysis is to measure the distance of each realization i of the Q observed variables, from the data center, using the so called Mahalanobis distance:

$$D_M = \left(\mathbf{X} - \overline{\mathbf{X}}\right)^T \mathbf{S}^{-1} \left(\mathbf{X} - \overline{\mathbf{X}}\right), \tag{4}$$

where $\overline{\mathbf{X}}$ is a matrix containing the average values, $\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, of the original variables. The observations associated with large values of D_M are classified as outliers and then discarded. The Mahalanobis distance can be related to the principal components: it can be shown, in fact, that the sum of squares of the PC, standardized by the eigenvalue size, equals the Mahalanobis distance for observation i:

$$\sum_{k=1}^{Q} \frac{z_{ik}^2}{l_k} = \frac{z_{i1}^2}{l_1} + \frac{z_{i2}^2}{l_2} + \ldots + \frac{z_{iQ}^2}{l_Q} = D_M. \tag{5}$$

This realization can be exploited for building a robust methodology based on PCA for outlier identification and removal. As mentioned previously, the first few principal components have large variances and explain most of the variation in $\mathbf{X}$. Therefore, those components are strongly affected by variables with relatively large variances and covariances. Consequently, the observations that are outliers with respect to the first few components usually correspond to outliers on one or more of the original variables. On the other hand, the last few principal components represent linear functions of the original variables with minimal variance. These components are sensitive to the observations that are inconsistent with the correlation structure of the data but are not outliers with respect to the original individual variables. Based on the above considerations, the following detection scheme can be proposed, as suggested by [**?**]:

1. *Multivariate trimming.* A fraction γ (typically 0.01%-0.1%) of the data points characterized by the largest value of D_M are classified as outliers and removed. $\overline{\mathbf{X}}$ and $\mathbf{S}$ are then computed from the remaining observations. The trimming process can be iterated to ensure that $\overline{\mathbf{X}}$ and $\mathbf{S}$ are resistant to outliers.

2. *Principal components classifier* (PCC). The PCC consists of two functions, one from the major, $\sum_{k=1}^{q} \frac{z_{ik}^2}{l_k}$, and one from the minor principal component, $\sum_{k=Q-r+1}^{Q} \frac{z_{ik}^2}{l_k}$. The first function can easily detect observations with large values on some of the original variables; in addition, the second function helps detect the observations that do not conform to the correlation structure of the sample. The number of major components, q, is determined by retaining the minimum number of PC required to account for at least 50% of the original data variance, while r is chosen so that the minor components used for the definition of the PCC are those whose variance is less than $0.2 \cdot \bar{l}$, where $\bar{l}$ is the average value of the eigenvalues of $\mathbf{S}$. This ensures that the selected minor components account for a very marginal variance and they only represent linear relations among the variables. Based on the PCC definition, an observation $\mathbf{X}_i$ is classified as an outlier if:

$$\sum_{k=1}^{q} \frac{z_{ik}^2}{l_k} > c_1 \qquad \text{or} \qquad \sum_{k=Q-r+1}^{Q} \frac{z_{ik}^2}{l_k} > c_2, \tag{6}$$

   where c_1 and c_2 are chosen as the 99^{th} quantile of the empirical distributions of $\sum_{k=1}^{q} \frac{z_{ik}^2}{l_k}$ and $\sum_{k=Q-r+1}^{Q} \frac{z_{ik}^2}{l_k}$.

The convergence of the algorithm is verified by looking at the third and fourth order moments of the major principal components. Since the structure of the data is frequently non-normal, the skewness and kurtosis are monitored from one iteration to the other and convergence is achieved when the rate of change of such quantities falls below an a priori defined tolerance, (e.g., 10^{-6}) or a maximum number of iterations is reached.A schematic representation of the outlier removal process is shown in Figure 2.

An example of the outlier detection scheme applied to a dataset consisting of 62,766 observations of 10 state variables [**?**] is shown in Figures 3a and 3b. Outliers were artificially introduced in the experimental data: specifically, 1000 observations have been generated from a matrix (1000 × 10) of random numbers between 0 and 1 and scaled using the standard deviation, s_j, of the variables $\boldsymbol{x_j}$. The effect of the outliers on the PCs is very clear from Figure 3a. The introduced outliers (black circles) are mostly outliers with respect to the original variables and they are visible in the plot of first two

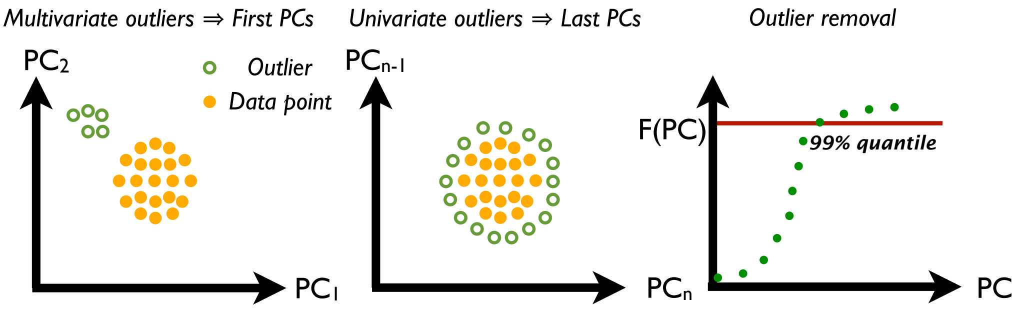


Figure 2: Outlier identification and removal.

PCs: a small cluster of points, separated from the majority of observations, appears in the plot of the first and second PC scores. They are also apparent (although less so) in the plot of the last and second-last scores as observations scattered around the main cloud of points. If the outlier detection scheme is applied (Figure 3b), the introduced outliers are completely removed; in addition, outliers present in the original experimental dataset, affecting the first and last PC scores (univariate and multivariate outliers), are also detected with the procedure described. A closer look at Figure 3b also indicates that the elimination of the outliers results in a slight modification of the first two PC scores, which are rotated counter-clockwise and compressed (especially in the $\mathbf{z}_2$ direction).

Outliers must be treated with care as they can strongly affect the covariance matrix, thus leading to the identification of false PCs.

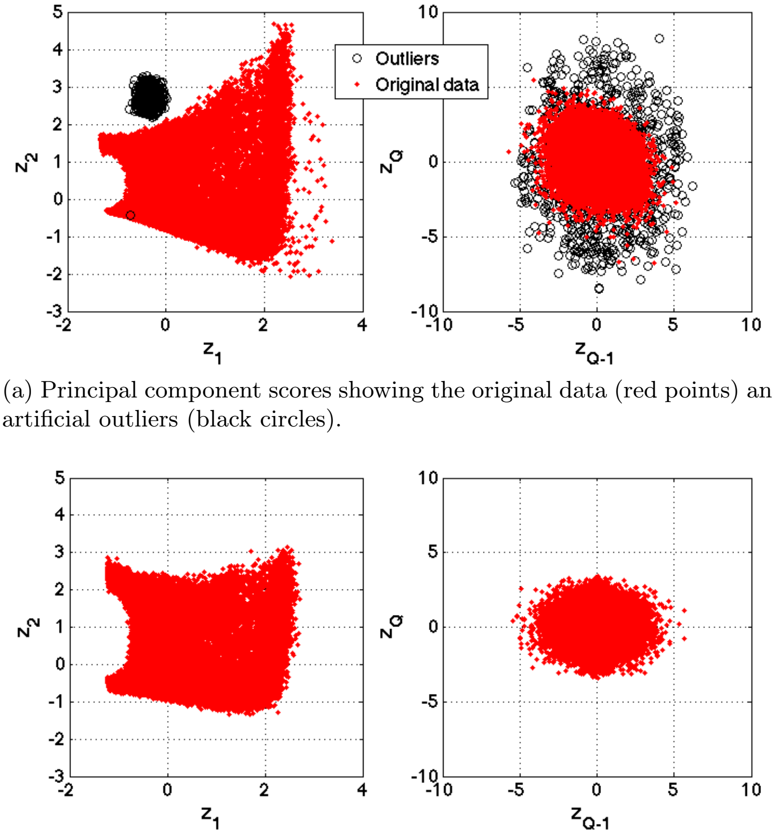
2.2. Centering and Scaling

When the variables are centered and scaled, a reduced variable can be defined as:

$$\widetilde{\boldsymbol{x}}_j = \frac{(\boldsymbol{x}_j - \overline{\boldsymbol{x}}_j)}{d_j}, \tag{7}$$

where d_j is the scaling parameter for variable $\boldsymbol{x}_j$. PCA, as discussed above, is applied on $\widetilde{\boldsymbol{x}}$ rather than $\boldsymbol{x}$. Centering is always applied in conjunction with scaling. We consider the following scaling methods:

1. *Auto scaling.* Also called unit variance scaling, auto scaling uses the standard deviation, s_j, as the scaling factor for each $\boldsymbol{x}_j$. After auto scaling, all the elements



(a) Principal component scores showing the original data (red points) and artificial outliers (black circles).

(b) Principal component scores after outlier removal.

Figure 3: Demonstration of removal of outliers artificially inserted into a dataset and the effect on the resulting PCA structure.

of $\mathbf{X}$ have a standard deviation equal to one and therefore the data is analyzed on the basis of correlations instead of covariances.

2. *Range scaling.* Range scaling adopts the difference between the minimal and the maximal value, $(\max(\boldsymbol{x}_j) - \min(\boldsymbol{x}_j))$, as scaling factor. A disadvantage of range scaling with respect to other scaling methods is that only two values are used to estimate the range, while for the standard deviation all measurements are taken into account. This makes range scaling more sensitive to outliers (see §2.1). To increase the robustness of range scaling, the range could also be determined by using robust estimators for maximum and minimum sample values, or after outliers have been removed.

3. *Level scaling.* The mean values of the variables, $\overline{x}_j$, are used as scaling factors. As with range scaling, level scaling can be affected by outliers. Therefore, a more robust estimator of the mean (the median) could be used or the mean could be determined after outlier removal. Level scaling can be used when large relative changes are of specific interest. However, in the case of the thermochemical state of a system, this could exaggerate the role of chemical species which appear in very small concentrations (*e.g.* radicals).

4. *Max scaling.* The variables are normalized by their maximum values, $\max(\boldsymbol{x}_j)$, so that they are all bounded between zero and one. As for the range and level scaling, a robust estimator of maximum values or a procedure for outliers removal should be employed.

5. *VAST scaling* [**?**]. *VAST* is an acronym for variable stability scaling and it is an extension of auto scaling. It focuses on variables which do not show strong variation, using the product between the standard deviation and the so-called coefficient of variation, defined as ${}^{s_j}/_{\overline{x}_j}$. Such scaling results in a higher importance for variables with a small relative standard deviation.

6. *PARETO scaling* [**?**]. *PARETO* scales each variable by the square root of its standard deviation. As a consequence, PARETO gives the variable under evaluation a variance equal to its standard deviation instead of unit variance, which is used for

auto scaling.

The impact of the different scaling methods will be discussed for several datasets in §3.

2.3. Choosing a Subset of Principal Components

The major objective of PCA is to replace the Q elements of $\overline{\mathbf{X}}$ with $q < Q$ principal components, while minimizing information loss. The most obvious criterion for choosing q is to select a cumulative fraction of the total variance that the PCs have to account for. The required number of PCs, q, is then the smallest value of q for which this chosen percentage is exceeded. The cumulative variance in the data can be obtained as

$$\sum_{k=1}^{Q} l_k = \sum_{j=1}^{Q} \mathrm{var}\left(\boldsymbol{x}_j\right). \tag{8}$$

Then the fraction of the total variance accounted for by retaining q of the Q eigenvectors can be defined as:

$$t_q = \frac{\sum_{k=1}^{q} l_k}{\sum_{k=1}^{Q} l_k}. \tag{9}$$

Importantly, it can be shown that the definition of t_q is equivalent to the so-called R^2 value,

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(x_{q,ij} - x_{ij})^2}{\sum_{i=1}^{n}(x_{ij} - \overline{x}_j)^2}, \tag{10}$$

where $x_{q,ij}$ is the reconstructed i^{th} observation of $\boldsymbol{x}_j$. Following the derivation of t_q, an appropriate measure of lack-of-fit of the rank q linear approximation of $\mathbf{X}$ can be related to the size of the discarded eigenvalues, *i.e.*

$$\epsilon_j = \sum_{k=q+1}^{Q} l_k = \sum_{i=1}^{n}\sum_{j=1}^{Q}\left(x_{q,ij} - x_{ij}\right)^2. \tag{11}$$

For a given number (q) of retained components, it is also possible to determine the variance accounted for each variable by the retained eigenvectors as:

$$t_{q,j} = \sum_{k=1}^{q}\left(\frac{a_{jk}\sqrt{l_k}}{s_j}\right)^2, \tag{12}$$

where a_{jk} is the weight of the j^{th} variable on the k^{th} eigenvector of $\mathbf{S}$.

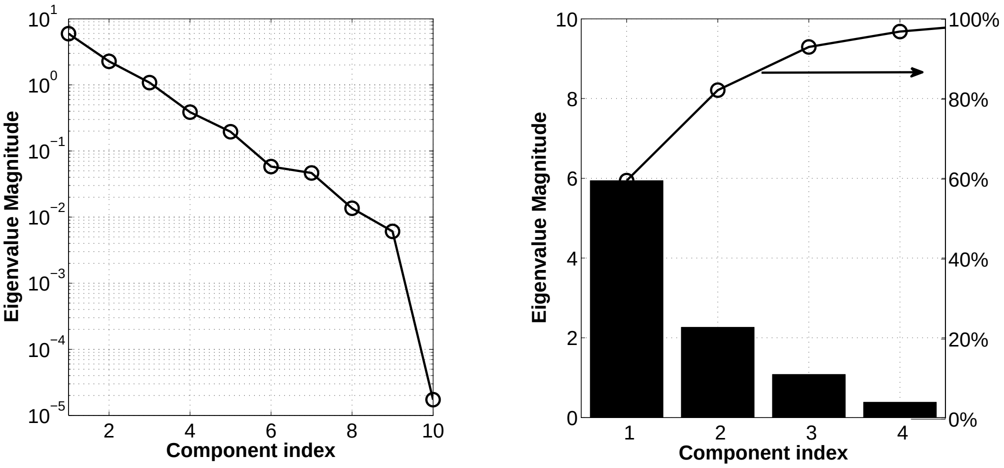


Figure 4: Scree plot for the determination of the number of principal components. The left frame shows the eigenvalue magnitudes (log scale) while the right frame shows eigenvalue magnitudes (linear scale) along with the associated variance in the data recovered from retaining the given number of eigenvalues.

A less rigorous method to identify the number of retained PCs uses a *scree plot*, as shown in Figure 4 for the jet in hot co-flow dataset, presented in section 3. This is a simple plot of the eigenvalue magnitudes sorted in descending order against their indexes, and provides a graphical interpretation of the information encoded in each dimension. As previously observed [? ?], there is an exponential decay in the information encoded in each succeeding dimension.

3. Results

High fidelity experimental data provided under the framework of the Workshop on Measurement and Computation of Turbulent Non-premixed Flames (TNF workshop) [?] are analyzed in the present paper. In particular, the following TNF datasets are employed:

- Turbulent non-premixed $CO/H_2/N_2$ (0.4/0.3/0.3 by vol.) jet flame [?]. This flame represents an ideal test-case due to its simplicity in terms of turbulence/chemistry interactions.

- Flames C, D E and F, a set of four piloted CH_4 jet flames [?], are characterized by an increasing Reynolds number and exhibit increasing non-equilibrium phenomena, including local extinction and re-ignition.

- The jet in hot co-flow (JHC) burner [?], designed to emulate MILD conditions. It consists of a central fuel jet (80% CH_4 and 20% H_2) within an annular co-flow of hot exhaust products from a secondary burner mounted upstream of the jet exit plane. O_2 in the co-flow is controlled at three different levels, 3, 6 and 9 mol%, while the temperature and exit velocity are kept constant.

- A bluff-body stabilized flame [? ?]. The experimental data used in this paper, designated as HM1, refer to an equimolar mixture of CH_4/H_2 with a fuel velocity of 118 m/s and coflow air velocity of 40 m/s.

"Instantaneous" (as opposed to ensemble-averaged) measurements were used for all analyses presented here.

It should be emphasized that experimental data are "incomplete" in the sense that we do not have simultaneous measurements of all species and temperature, as is possible from computationally obtained data (from, e.g., DNS). In previous studies that employed computational data, we employ all species and temperature in the analysis. However, for the purposes of this paper, this is not an issue, since we are focused on data preprocessing strategies.

In the following, the effect of data pre-processing on PCA results will be discussed in detail, pointing out the possible impact of outliers on PC determination and the impact of scaling methods on the PC structure. Moreover, the PCA structure will be analyzed and processed with PC rotation, to provide a physical interpretation for the extracted components.

3.1. Effect of Outliers on the PCA Structure

Outlier detection and removal is particularly important when using PCA with experimental data. If outliers are not removed, the resulting PCA can show significant sensitivity to their existence, thereby complicating interpretation of the PCA.

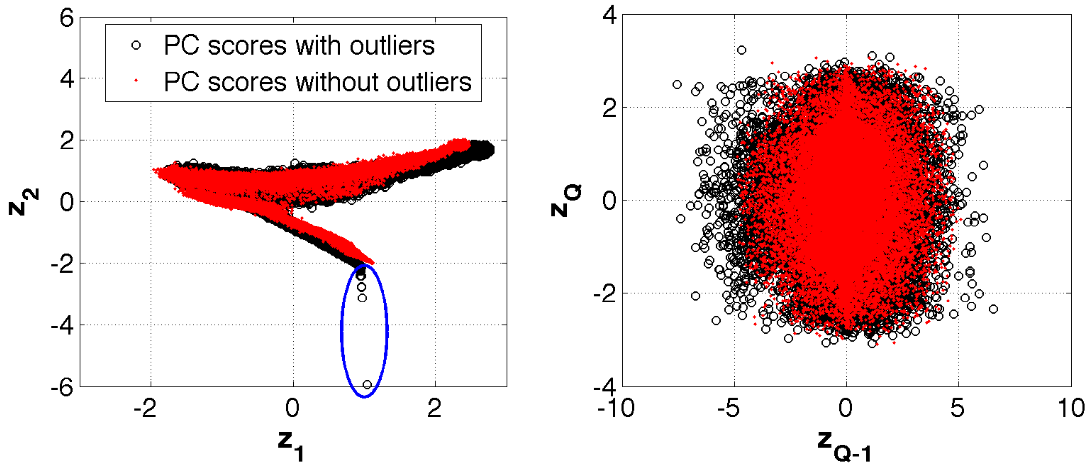
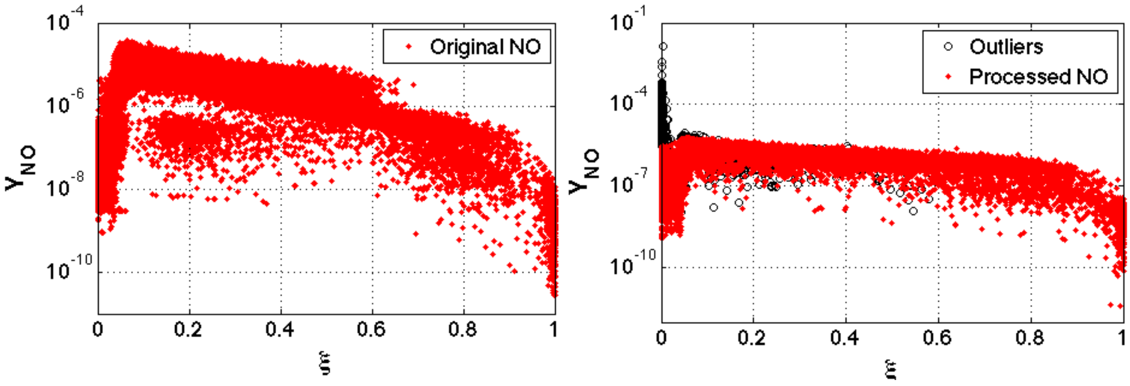


Figure 5: Scatter plot of the first two (left frame) and last two (right frame) principal components from the JHC dataset before (black circles) and after (red circles) outlier removal. Scaling method: auto scaling. Trimming fraction, γ: 0.1%. Outliers are indicated by the blue circle.

Figure 5 shows the effect of the outlier removal process on the principal component structure for the JHC dataset, scaled using auto scaling and a trimming fraction γ 0.5%. In contrast to the example discussed in §2.1, the dataset is not augmented with artificial outliers, but processed to identify the existence of experimental measurements inconsistent with the primary structure of the data.

It can be observed that, using the original data without any pre-processing, the scatter plot of the first two principal component scores show the existence of observations which strongly differ from the main multi-variate structure of the data. Those can be classified as *univariate outliers* (see §3.1), as they correspond primarily to the components associated with the largest eigenvalues. Mathematically, those observations are flagged as outliers because, as explained in §2.1, the PC classifier related to the first few PCs is larger than the 99^{th} quantile of the actual PC distribution, indicating that the scores associated with those observations largely deviate from the main data structure. Multivariate outliers are also present in the original dataset, as indicated by the plot of the last two principal components.

To confirm the existence of univariate outliers, NO mass fraction is plotted against



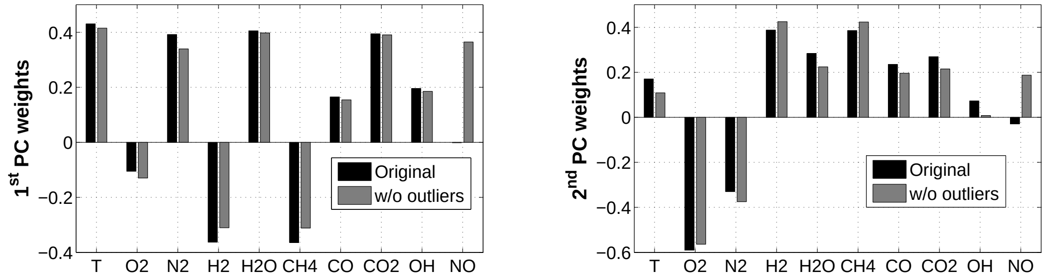
(a) Original data from 6% O_2 case.

(b) Retained data (red points) and removed outliers (black circles) for 3% O_2 case. Trimming fraction, γ: 0.1%.

Figure 6: Scatter plot of NO mass fraction as a function of the mixture fraction. Effect of outlier removal for the 3% O_2 dilution case. JHC dataset [**?**].

the mixture fraction, ξ, for the original data, at 3% and 6% O_2 mass fraction in the co-flow. Figure 6 shows, for the 3% O_2 case, unphysically high concentrations of NO on the oxidizer side ($\xi = 0$), which are not observed for the other dilution cases (*e.g.,* 6% O_2, Figure 6a) and that determine the extreme score values observed in Figure 5. Figure 6b points out that some "feasible" observations are also removed during the outlier detection process (black circles behind red dots). This does not affect the statistical value of the analysis since only approximately 1500 out of more than 60,000 observations are removed, and only a few hundred of those are in the feasible NO range.

The eigenvectors ($\mathbf{A}$) of the covariance matrix provide insight into the effect of outliers on the principal components. Figure 7 shows a comparison between the first two PCs obtained for the JHC dataset with (gray) and without (black) outlier observations. In particular, the figures graphically indicate the weight of the original variables on the first two components. From Figure 7, it is clear that, for the JHC dataset, outlier removal results in the chemical species NO being eliminated from the first two PCs, while the remaining weights in $\mathbf{A}$ remain largely unaffected. This confirms, as indicated by Figure 6, that the outliers identified in Figure 5 are related to NO measurements, leading to the overestimation of such species in the PC structure when the dataset is not pre-processed.



(a) (b)

Figure 7: Weights of the original variables on the first (a) and second (b) principal components before (black bars) and after (gray bars) outlier removal for the JHC data. Scaling method: auto scaling. Trimming fraction, γ: 0.1%.

The results shown above indicate the strong relevance that outlier observations, caused by measurement errors, may have on the covariance structure of the data, confirming the need for effective outlier removal tool, as the one employed here and based on PCA. The importance of an effective pre-precessing of the data is not limited to the application of PCA but it should aways be considered, whenever PCA is used to extract information about the system behavior.

3.1.1. Effect of the Threshold Parameter, γ

The trimming fraction γ (see §2.1) plays a critical role in the outlier removal process: large values of γ may result too many samples being eliminated, resulting in an unphysical modification of the PC structure. Figure 8 shows the number of removed points as a function of γ for the HM1, 3% O_2 JHC and flame F datasets. Figure 9 shows the effect of γ on the structure of the first PC for the 3% O_2 JHC (9a), flame F (9b) and HM1 (9c) datasets. The PC structure is relatively constant for $\gamma < 0.05\%$, but begins to change noticably for $\gamma > 0.1\%$. Flame F and the HM1 were inluded in such analysis with the JHC dataset as they show specific features which can help identifying appropriate ranges for γ, to avoid over-agressive observation removal during the outlier identification process. Indeed, those systems show singificant extinction: flame F is close to global extinction [?] and the HM1 bluff-body stabilized flame is known to show intermittent local extinction being at 50% blow-of conditions [? ?] Figure 9c indicates $\gamma < 0.05\%$ leads, in all cases,

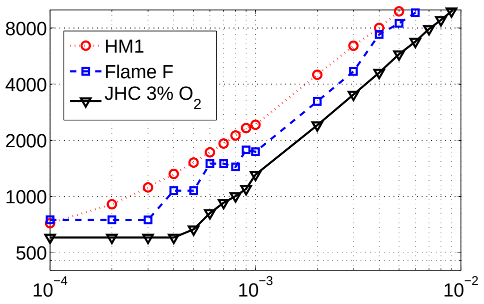


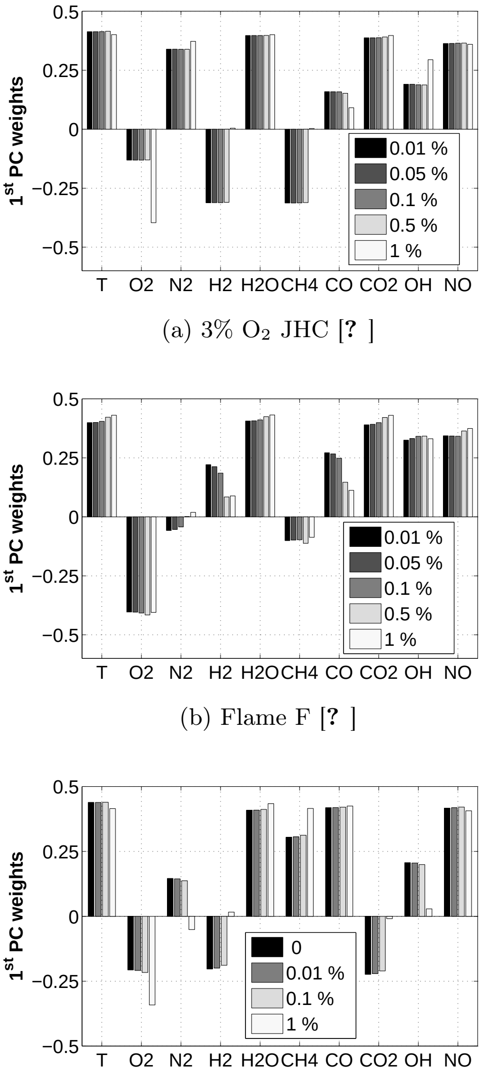
Figure 8: Effect of γ on the number of points removed during the outlier identification process for the 3% O_2 JHC, flame F and HM1 datasets. Scaling method: auto scaling.

to a PC structure which is unaffected by outlier removal, for all analyzed datasets.

Figure 10 illustrates the effect of over-agressive outlier removal ($\gamma = 1\%$) on the temperature distribution for the JHC and flame F datasets. Over-agressive outlier removal eliminates observations corresponding to extinction for the JHC dataset while eliminating fully-burning regions for the flame F datset. A similar effect (not shown here) is observed for the HM1 bluff-body dataset, where large γ also results in removal of points corresponding to extinction. Figure 9a indicates that the corresponding PC structure is significantly altered in both cases when the choice of γ is too large.

Figures 9b and 11 indicate that, for flame F, the outlier removal process with an appropriate choice for γ does not significantly impact the PC structure, although some of the realizations inside the main "data cloud" are removed because they exceed the 99^{th} quantile of the experimental distribution of the first and last eigenvectors. As a consequence, the PC determined before and after the outlier removal procedure show very minor differences among the weights, as shown by Figure 9b.

Based on the observations above, we recommend $0.01\% < \gamma < 0.05\%$, which effectively removes outliers but does not remove enough physically meaningful datapoints to alter the PC structure.

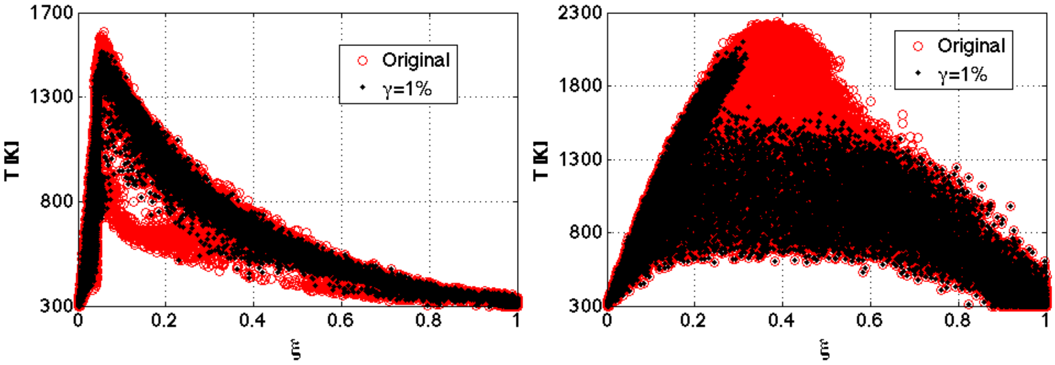


(a) 3% O_2 JHC [?]

(b) Flame F [?]

(c) Bluff body HM1 [? ?]

Figure 9: Effect of the trimming fraction γ on the structure of the first PC. Scaling method: auto scaling.



(a) JHC dataset

(b) HM1 dataset

Figure 10: Effect of the trimming fraction γ on the temperature distribution (plotted against the mixture fraction, ξ) for the JHC and flame F datasets.

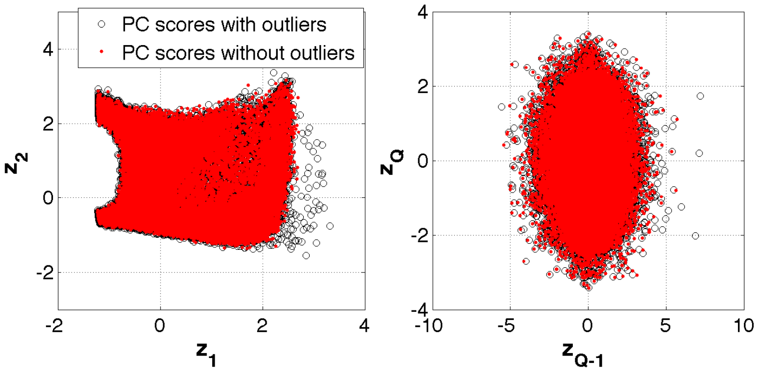


Figure 11: Scatter plot of the first and last two principal components from flame F before (black circles) and after (red circles) outlier removal. Scaling method: auto scaling. Trimming fraction, γ: 0.5%.

3.2. Effect of Scaling

We now consider the effect of scaling strategies outlined in §2.2 on the PCA reduction, focusing on the Sandia CO/H_2 jet flame dataset. Figure 12 shows the normalized[2] eigenvalue size distribution obtained by applying the different scaling options. It indicates that the VAST and PARETO scaling methods result in larger weights for the first few eigenvalues while the other scaling options are all very similar in their eigenvalue size distribution. This is a consequence of the high importance given by the VAST and PARETO scaling methods to temperature over the chemical species mass fractions. This effect is accentuated for the PARETO scaling, where the square root of the standard deviation is used to scale variables: this enhances the relevance of temperature with respect to the other variables defining the state-space. Indeed, the application of PARETO scaling results in temperature being the variable carrying most of the data variance and is, therefore, equivalent to forcing the first principal component to align with the temperature. Such behavior is a consequence of the size-dependency of PCA for non-homogeneous datasets (where the variables have very different scales) as is characteristic of combustion. Therefore, the choice of such scaling does not appear very useful for the analysis carried out in the present paper, as it is equivalent to an *a priori* choice

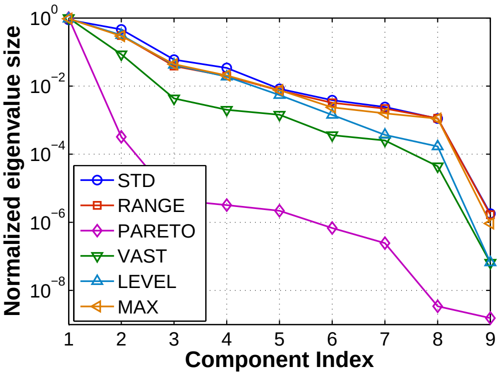


Figure 12: Ordered normalized eigenvalue magnitudes for the CO/H_2 jet flame dataset for various scaling strategies.

[2] The eigenvalues obtained using different scaling methods are normalized between 0 and 1, to allow comparison of different pre-processing techniques.

of the PC. However, PARETO scaling can be extremely appealing for the definition of reduced-order combustion models, as the choice of temperature within the set of PC has a dramatic influence on the model's accuracy [?].

Table 1 shows t_q and $t_{q,j}$ (see Eqs. (9) and (12)) obtained by applying range, max, VAST and level scaling to the CO/H_2 dataset. Results indicate that auto scaling is the only scaling technique that provides a uniform reconstruction of the state variables (for $q = 3$), as evidenced by relatively high values of $t_{q,j}$ for all variables. Range and max scaling, whose behavior is very similar (as expected), perform slightly better than auto scaling for most of the main species and temperature. However, they cannot properly capture NO variation, even with $q = 3$. Similarly, VAST scaling concentrates on extremely stable variables such as N_2, but fails to recover minor species such as OH properly. This effect is accentuated in PARETO scaling, which clearly emphasizes main species and temperature. The higher values of t_q given by range, max, VAST and PARETO scaling, compared to auto scaling, are due to the higher variance explained for the major variables. However, these scaling approaches do not preserve features related to minor species such as NO and OH. The variance accounted for OH and NO by auto scaling is up to 16% and 25% higher, respectively, than that explained by the other scaling methods. On the other hand, level scaling focuses on variables characterized by large changes (relative to their mean) and leads to an overestimation of the role of minor species in the PCA reduction. Therefore, the reconstruction of minor species such as OH and NO is very accurate, but major species such as H_2O are poorly recovered. On the basis of the described sensitivity, it becomes clear how scaling can be constructively employed to target the desired accuracy of different subsets of state variables. In particular, auto scaling appears very well-suited when an exploratory analysis on the chemical manifold should be performed, whereas range, max and vast scaling are useful for capturing the principal features of the systems and the behavior of the main chemical species. This appears very appealing for building reduced-order models of combustion systems to be used in optimization studies.

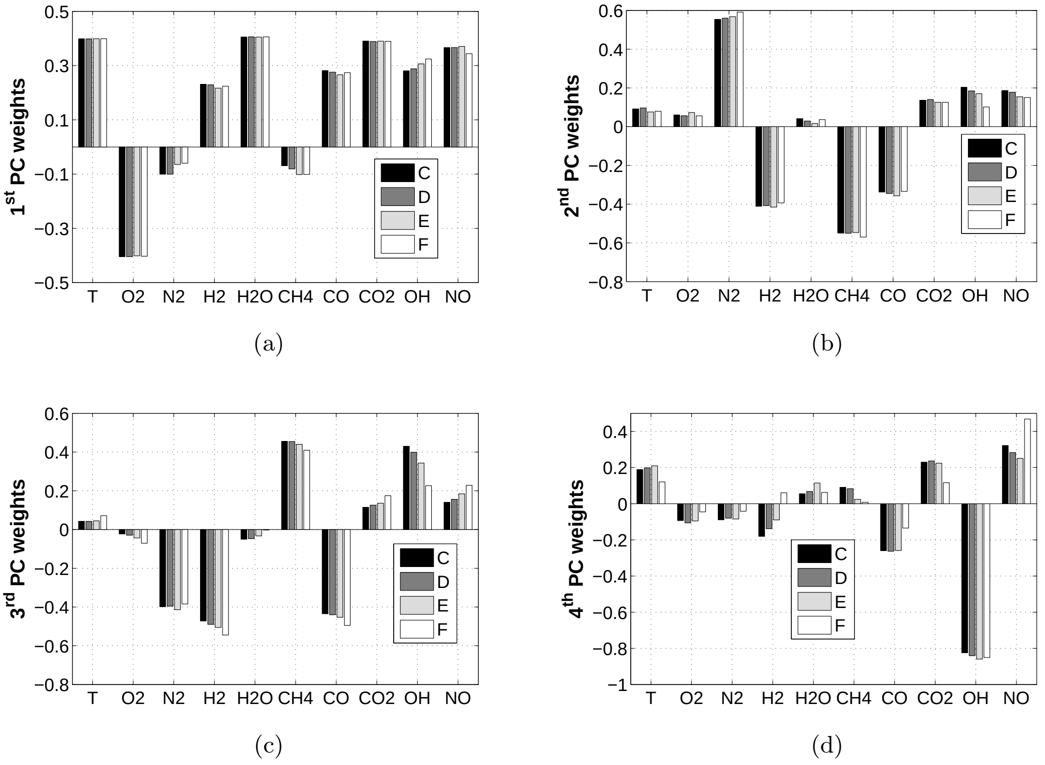
Table 1: Total, t_q, and individual variance, $t_{q,j}$, (see Eqs. (9) and (12)) accounted for the CO/H_2 jet flame dataset, as a function of the number of retained PC, q, and the scaling option used.

	Auto (std)		Range		Max		VAST		Level		PARETO	
	$q = 2$	$q = 3$	$q = 2$	$q = 3$	$q = 2$	$q = 3$	$q = 2$	$q = 3$	$q = 2$	$q = 3$	$q = 2$	$q = 3$
T	0.971	0.973	0.983	0.991	0.979	0.990	0.992	0.992	0.896	0.943	1.000	1.000
Y_{O_2}	0.986	0.986	0.994	0.994	0.997	0.997	0.975	0.978	0.942	0.961	0.990	0.991
Y_{N_2}	0.986	0.986	0.981	0.981	0.971	0.971	1.000	1.000	0.965	0.970	0.989	0.994
Y_{H_2}	0.968	0.969	0.962	0.963	0.957	0.960	0.945	0.947	0.991	0.991	0.965	0.967
Y_{H_2O}	0.930	0.936	0.945	0.945	0.944	0.944	0.940	0.978	0.870	0.884	0.917	0.968
Y_{CO}	0.994	0.994	0.995	0.997	0.990	0.994	0.979	0.980	0.987	0.987	0.999	0.999
Y_{CO_2}	0.973	0.977	0.979	0.987	0.977	0.988	0.981	0.985	0.908	0.959	0.967	0.993
Y_{OH}	0.738	0.940	0.731	0.991	0.745	0.992	0.660	0.687	0.870	0.993	0.554	0.567
Y_{NO}	0.772	0.930	0.728	0.795	0.729	0.802	0.744	0.970	0.701	0.926	0.759	0.813
t_q	0.924	0.966	0.946	0.975	0.942	0.975	0.992	0.996	0.949	0.973	0.999	0.999

3.3. PC sensitivity to system variability

We now consider the question of how sensitive a PCA is to the characteristics of a system such as Reynolds number. To investigate this, the four piloted jet flames (Sandia flames C to F) are considered. These flames have increasing Reynolds numbers that lead to significant extinction in flames E and F, which is near blow-out. Given that significantly different regions of state space are realized in these flames (e.g. extinction), one may not think that the PCA structure should be consistent across all flames. Figure 13 shows the weights of the original variables on the first four PCs (columns of **A**) for Sandia flames C, D, E and flame F. The PC structure remains very similar for the first four PCs. The possible exception is weights on intermediate species such as CO, H_2 and OH, which show some variation in their contributions to the eigenvectors across the range of Reynolds numbers. This is a consequence of the increasing degree of extinction which characterizes flames C to F: the OH distribution shows a larger scatter as the Reynolds number is increased. As a result, OH contribution to the covariance matrix is decreased (OH is less correlated with the other state parameters) and the corresponding weights on the PC is reduced.

Although Figure 13 indicates that the PCA structure is largely unchanged, the ques-



(a)

(b)

(c)

(d)

Figure 13: Structure of the first four PC for PCA applied to flames C, D, E and F in the TNF series [?]. Scaling criterion adopted: auto scaling.

tion remains of how sensitive the PCA reconstruction of state variables is to slight variation in the PCA structure. In other words, can the low-dimensional representation obtained via PCA for one of the four systems can be exploited for the others, without performing a new decomposition? This question is crucial to assess the *universality* of the PCA approach for identifying manifolds in reacting systems. To answer this question, a PCA was performed for flames C and F and employed to reconstruct the other datasets. This implies projecting the scores of each system onto a single PC basis (C or F):

$$\mathbf{X}_{q,i} = \mathbf{Z}_{q,i}\mathbf{A}_{q,k}^{T} \qquad k = C,\ F. \tag{13}$$

Table 2 lists $t_{q,i}$ (the R^2 values for a linear reconstruction) for

1. flames C-F using PCA on each dataset (labeled as $t_{q,i}$)
2. flames C-E using the PCA obtained from flame F (labeled as $t_{q,i}^{F}$).
3. flames D-F using the PCA obtained from flame C (labeled as $t_{q,i}^{C}$)

Results indicate that the low-dimensional representation found for flame F provides a very satisfactory representation of the other systems. In all cases, the relative error with respect to an optimal reconstruction ($t_{q,i}$ versus $t_{q,i}^{F}$) is less than 1%. When the basis found for flame C is employed, a very interesting result is observed: the reconstruction of most state variables slightly improves and the accuracy in NO reconstruction decreases. This is probably due to the increasing degree of extinction determined by the increase of Re, which leads to a large variability of NO species, as shown in Figure 14. As a consequence, the basis identified directly from the flames reflect the major variability of NO, leading to larger weights on the first components. This is not the case when the basis is extracted from flame C, leading to less accurate NO predictions.

3.3.1. Effect of Scaling on Manifold Invariance

For completeness, we also consider the effect of scaling on the manifold invariance. Figure 15 shows the standard deviation of the first PC weights on each original variable considering PCA performed on flames C, D, E and F independently with different scaling methods. Large standard deviations indicate an alteration in the PCA structure across

Table 2: Individual variance, $t_{q,j}$, accounted for the Sandia flame C, D, E and F datasets by the PCA reduction, as a function of the number of retained PCs, q. Note that $t^F_{c,i}$ and $t^C_{q,i}$ refer to the accuracy by which variables are reconstructed using the PCA obtained for flame F and flame C, respectively.

	$t_{q,i}$ (%)				$t^F_{q,i}$ (%)			$t^C_{q,i}$ (%)		
	C	D	E	F	C	D	E	D	E	F
T	0.985	0.985	0.976	0.971	0.982	0.981	0.974	0.984	0.977	0.974
Y_{O_2}	0.987	0.986	0.980	0.979	0.983	0.983	0.979	0.986	0.980	0.977
Y_{N_2}	0.982	0.982	0.980	0.980	0.983	0.981	0.980	0.981	0.979	0.979
Y_{H_2}	0.975	0.969	0.964	0.970	0.973	0.966	0.966	0.966	0.964	0.965
Y_{H_2O}	0.989	0.989	0.986	0.984	0.987	0.986	0.984	0.988	0.985	0.984
Y_{CH_4}	0.987	0.987	0.984	0.984	0.986	0.985	0.984	0.986	0.985	0.985
Y_{CO}	0.972	0.968	0.962	0.969	0.970	0.963	0.962	0.964	0.962	0.970
Y_{CO_2}	0.987	0.986	0.976	0.974	0.983	0.980	0.975	0.985	0.977	0.975
Y_{OH}	0.999	0.999	0.995	0.978	0.990	0.990	0.984	0.999	0.999	0.998
Y_{NO}	0.945	0.932	0.887	0.892	0.942	0.931	0.895	0.933	0.877	0.850

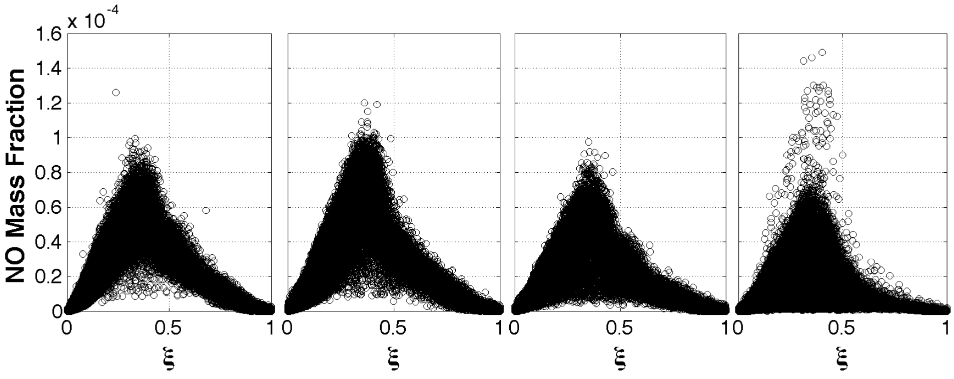


Figure 14: NO distribution with increasing Reynolds number (from left to right) for flames C-F.

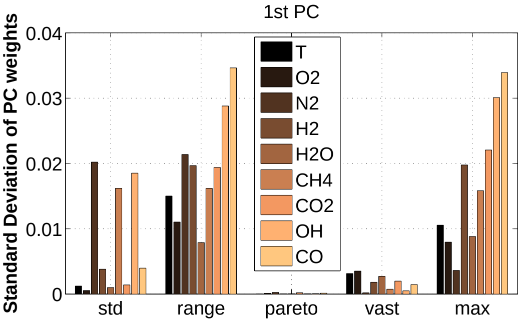


Figure 15: Standard deviation of the weights on the first principal component across flames C-F when applying different scaling methods.

flames C-E when using the given scaling method. In all cases, there is not a significant variation in the PC structure (with the standard deviation remaining below 0.04 in all cases), but they appear extremely stable for VAST and PARETO scaling methods, which emphasize major and stable variables of the state-space.

The results shown in this section indicate the potential of exploiting a PCA-reduced representation even when the system characteristics are modified. In particular, the relative independence of the basis on the Reynolds number indicate the invariability of the manifold in a range of operating conditions. Nevertheless, further study considering more systems over wider ranges of Re is warranted before concluding that PCA is entirely independent of Re.

4. Conclusions

PCA has recently been proposed as a technique to identify correlations among the multivariate datasets ubiquitous to turbulent combustion. These correlations imply the existence of manifolds in the chemically reactive systems, and PCA has shown promise in identifying these manifolds [? ? ? ?]. This paper has explored the details of data pre-processing for use in PCA. Specifically, scaling and centering the data as well as outlier removal have been discussed.

The existence of outliers in the dataset can significantly alter the determination of the PC structure and this can lead to the overestimation of the role of specific variables, or sets of variables, for which outlier observations exist. A method based on PCA has proved very satisfactory for the elimination of the observations which differ from the main multi-variate structure, based on PC classifier built from the first and last few PC, respectively. The effectiveness of the approach was proven for a jet in hot co-flow dataset, and results indicate that outlier removal does not alter the PC structure of outlier free datasets.

The choice of scaling in particular has a significant impact on the resulting PCA structure by altering the relative importance of various species and temperature. Indeed, different scaling choices may be made depending on the goal of the resulting PCA to optimize the reconstruction of specific classes of state variables. In particular, auto-scaling appears the best option where a balanced reconstruction of the state-space is required for exploratory analysis, whereas level scaling enhances the role of minor species. All the other tested scalings (range, max, VAST) appear ideal for the optimization of stable and major species.

Finally, for the TNF flame datasets, we have demonstrated that the PCA structure remains nearly invariant with Reynolds number across the range from flame C to flame F. This observation is further substantiated by the fact that reconstructing flame C data from a PCA obtained on flame F (or vice-versa) is nearly as accurate as reconstructing data from a PCA obtained directly on that dataset.

Acknowledgments

The reserach was sponsored by the National Nuclear Security Administration under the Accelerating Development of Retrotable CO2 Capture Technologies through Predictivity program through DOE Cooperative Agreement DE-NA0000740 and by National Science Foundation PetaApps award 0904631.

References