
Coussement Axel$^{a,b,c,*}$, Gicquel Olivier$^{a,b,c}$, Parente Alessandro$^{a}$

$^{a}$Université Libre de Bruxelles, Aero-Thermo-Mechanics Department, Avenue F.D. Roosevelt 51, CP 165/11. 1050 Bruxelles. Belgium.


Abstract

Principal component analysis (PCA) has been successfully applied to the analysis of combustion data-sets. However using PCA on a raw direct numerical simulation or an experimental data-set is not straightforward. Indeed, those data-sets usually show non homogenous data density, hot and cold zones being generally over represented. This can introduce bias in the PCA reconstruction, especially when strong non-linear relationships characterize the data sample. To tackle this problem, a combination of the kernel density method and PCA is introduced here. This new PCA algorithm, called Temperature BAseD KErnel Density weighted PCA (T-BAKED PCA) allows to enhance the PCA accuracy especially in the flame front zone, which is the principal zone of interest. The performance of this new approach is benchmarked against classical PCA. Moreover, a new method called Hybrid T-BAKED PCA or HT-BAKED PCA, combining both classical and T-BAKED PCA, is proposed to provide an optimal representation of all flame regions.

Keywords: Principal Component Analysis, Combustion, Tabulated Chemistry

1. Introduction

Principal component analysis has been applied, since its introduction [1], to a wide range of problem requiring dimension reduction. Recently, the application of PCA for size-reduction in turbulent-reacting systems was investigated by Frouzakis et al. [2] and Dandy et al. [3]. Indeed, a primary challenge in turbulent combustion modeling lies in the broad range of scales which are inherently coupled, in space and time, through thermo-chemical and fluid dynamic interactions. However, in many applications [4], several chemical time scales are smaller than the fluid dynamic time scales, allowing to reduce the dimensionality of the problem. This reduction is of critical importance in Computational Fluid Dynamics (CFD): if one reduces the number of equations needed by detailed chemical mechanisms, the computing power needed will decrease accordingly. Indeed, even simple fuels such as methane are characterized by chemical mechanisms involving a large number of species and chemical reactions: 53 species and 325 reaction for the GRI-3.0 mechanism [5].

Two methods can be used to perform a dimensionality reduction:

- mechanism reduction: the chemical mechanism is modified to reduce the number of species and its stiffness.

- state-space parameterization: where it is assumed that the combustion processes can be parametrized by a small number of variables defining a reduced manifold in the state space.

Numerous models have so been proposed to identify this manifold: Steady Laminar Flamelet Method (SLFM) [6, 7, 8], Flamelet-Generated Manifold (FGM) [9] and Flamelet-Prolongation of ILDM model (FPI) [10, 11, 12] are just few examples. However, Sutherland and Parente [13] and Parente et al. [14] showed that the selection of the parameters defining the low dimensional manifold is not straightforward, even though some parameters are known to be optimal for specific combustion regimes, such as the mixture fraction ($\phi$) for nonpremixed combustion.

Another method which allows identifying a chemical manifold is Principal Component Analysis (PCA). In particular, PCA offers the possibility to automate the parameter identification process [13, 14], controlling at the same time the error induced by the reduction. PCA can provide an optimal representation of a system based on $q$ Principal Components (PC) which are linear correlations of the $Q$ original variables. The error can be controlled by choosing the number of PCs ensuring an $a$ priori defined accuracy. Sutherland and Parente [13] showed that a transport equation can be solved for the score of each PC leading to a low dimensional representation of the combustion process. Moreover, Parente et al. [14] have demonstrated an improvement of the PCA-based model if local PCA is used. However, the actual implementation of a model based on local PCA is still to be established. Finally, PCA was also

*Corresponding author. Phone + 32 2 650 26 71 Fax +32 2 650 27 10

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

Preprint submitted to Elsevier February 15, 2012
exploited to investigate the main features of combustion regimes [15].

However, the application of PCA to reacting flows still requires further developments. The present work focuses on the influence of data sampling on the derivation of the principal components and on the parameterization error. As it will be discussed below, the PCA results and the PCs are strongly dependent on the data-set set conditioning. In particular, without an appropriate conditioning of the data-set, the extracted PCs can be biased.

After a brief description of the PCA reduction process, the problem of data sampling for PCA in combustion process will be emphasized and a weighting function based on the kernel density method [16, 17] will be introduced. The improvement of the PCs in the flame front description using the kernel density method with respect to classic PCA will be demonstrated using a 1-D premixed laminar flame, a 2-D flame flame-vortex interaction and a diffusion jet flame as data-sets. Finally, a hybrid PCA method that combines the advantages of the two methods will be presented.

2. Principal component analysis, centering and scaling

The dimension of the chemical state space associated to a reaction mechanism including $N_{sp}$ species is equal to $Q = N_{sp} + 2$. This state space can be parameterized for instance by the $N_{sp}$ species mass fractions, $Y_i$, temperature, $T$ and pressure $p$. This means that $Q$ transport equations have to be solved to characterize the combustion processes leading to a large CPU cost.

To tackle this problem, previous studies [13] used the Principal Component Analysis (PCA) to identify $q$ ($q < Q$) linear correlations between the state space variables to reduce the degree of freedom characterizing the system. Employing those correlations, the approximated combustion system reads at each point $c$:

$$[Y_1, \ldots, Y_{n_{sp}}, T, p]_{c} \approx [Y_1, \ldots, Y_{n_{sp}}, T, p]_{c,q} = \sum_{i=1}^{q} z_{c,i} a_i$$

(1)

where $a_i = [a_{i,1}, \ldots, a_{i,Q}]$ and the $a_{i,j}$ are coefficients of linear correlation $a_i$. If $q = Q$, equation 1 describes the full state space, while when $q = 1$ chemistry evolutions are limited to a simple linear subspace in the state space.

To identify the most accurate linear correlations, $a_i$, a representative sample of $n$ observations, $[Y_1, \ldots, Y_{n_{sp}}, T, p]_{c}$, $c \in [1, n]$ are used. This sample could result from 1D laminar flame simulation, DNS of a 3D turbulent flame or experimental results [13, 14], defining the matrix $X$ ($n \times Q$).

Then, the covariance matrix $S$ is computed:

$$S = \frac{1}{n-1} X' X$$

(2)

where the superscript $t$ indicates the transpose matrix.

If the covariance matrix is non singular, it can be decomposed:

$$S = A \Lambda A'$$

(3)

where $A$ ($Q \times Q$) and $\Lambda$ ($Q \times Q$) contain respectively the $Q$ eigenvectors called principal component (PC) and the eigenvalues of $S$, in decreasing order of magnitude.

The principal component scores, $Z$ ($n \times Q$), are then computed using the eigenvectors:

$$Z = X A$$

(4)

One of the main advantage of PCA is that the original set of data, $X$, can be uniquely recovered using the PCs and their associated scores:

$$X \approx X_q = Z_q A_q'$$

(5)

where it should be noted that $A^{-1} = A'$. The main feature of PCA is dimension reduction. Indeed, if one only uses the first $q$ PCs ($q < Q$) in equation 5, an approximation of the original data-set, $X_q$ ($n \times P$), is obtained:

$$X_q \approx X_q = Z_q A_q'$$

(6)

It should be stressed that PCA is affected by the relative difference of magnitude between variables contained in the data-set [14]: if, for example, the magnitude of one variable in the data-set is one (or more) order of magnitude higher than the other for all observations, the first principal component will be aligned with this particular variable. Therefore, the data should be centered and scaled before applying PCA [14]. Centering is achieved by subtracting the mean value of each $Q$ original variable to their respective observation:

$$X' = X - \bar{X}$$

(9)
where $\mathbf{X} (n \times Q)$ is a matrix containing the mean of each variable over the $n$ observations:

$$
\mathbf{X} = \begin{pmatrix}
\bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_Q \\
\bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_Q \\
\vdots & \vdots & \ddots & \vdots \\
\bar{x}_1 & \bar{x}_2 & \ldots & \bar{x}_Q
\end{pmatrix}
$$

(10)

where $\bar{x}_j$ is the mean of the $j^{th}$ variable over the $n$ observations.

Scaled data can be defined as:

$$
\tilde{\mathbf{X}} = \mathbf{X}' \mathbf{D}^{-1}
$$

(11)

with $\mathbf{D} (Q \times Q)$ being a matrix containing the scaling parameter for each variable:

$$
\mathbf{D} = \begin{pmatrix}
d_1 & 0 & \ldots & 0 \\
0 & d_2 & \ldots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \ldots & d_Q
\end{pmatrix}
$$

(12)

While various scaling methods can be applied [18], the classic choice is to use the "auto-scaling" [14]: the standard deviation of the $j^{th}$ variable computed from the data-set, $\hat{\sigma}_j$, is used as $d_j$. After this scaling the variance of each original variable is one, meaning that the analysis is based on correlations among the state variables. Applying centering and scaling, the data-set used to perform the PCA is thus:

$$
\tilde{\mathbf{X}} = (\mathbf{X} - \bar{\mathbf{X}}) \mathbf{D}^{-1}
$$

(13)

This method is used throughout the present work.

3. Problem statement

Centering and scaling are crucial for the effectiveness of PCA. However, this does not address the problem of the data sampling. Indeed, if one considers a combustion data-set to be analyzed with PCA, special attention should be paid to the over-representation of the hot and cold zone, especially when dealing with numerical data. For example, if one considers the temperature field from a vortex-flame interaction (see section 7.1 for details) represented on Figure 1, the relevant zone to be considered for PCA is the grey highlighted region, corresponding to the flame front. It is clear that if one considers the whole domain as a data-set the hot and cold zone, where no reaction takes place, will be over represented. This over-representation is emphasized in Figure 2 where the density of the normalized temperature is shown.

To characterize the problem of over-sampling and to show its influence on PCA, let’s consider the following
data-set:

\[
\begin{align*}
  x_{1,n} & = -1 + 20R \\
  x_{2,n} & = 5x_{1,n} + 100R \\
\end{align*}
\]

for \( n \in [1, 1000] \)

\[
\begin{align*}
  x_{1,n} & = 420 + 8(n - 1001) \\
  x_{2,n} & = \frac{5000}{200} \times (x_{1,n} - 400) + 500R \\
\end{align*}
\]

for \( n \in [1001, 1021] \)

\[
\begin{align*}
  x_{1,n} & = 1000 + 20R \\
  x_{2,n} & = 5x_{1,n} + 100R \\
\end{align*}
\]

for \( n \in [1021, 2021] \)

with \( R \) being an uniformly distributed random noise which varies between 0 and 1. This data-set, plotted on Figure 3, consists in two linear functions, with variable data densities. Equation 14 and 16 define a linear relationship for low and high values of \( x_1 \), with a large density of points, and equation 15 provides another linear relationship for intermediate \( x_1 \) values, with a lower density of points. The point density along \( x_1 \) and \( x_2 \) was chosen to mimic the behavior of the temperature field shown in Figure 2.

Processing this data-set using the PCA, yields the results plotted on Figure 4, where the recovered \( x_2 \) considering one PC is plotted as a function of the observed \( x_2 \). As expected, the first PC is "attracted" by the over represented extreme values. If the relevant information is contained in the over represented zone, this does not yield any problem. However, if most of the relevant information is contained in the intermediate region, the PCA fails to capture that information.

Note that, if data-set had contained only one linear function the over-representation would not have yielded any problem. It is therefore not only the sample density that raise an issue, but the combination of non-linearity with the uneven sampling.

Applying PCA to combustions data-sets will suffer from these issue, as cold and hot zones are generally over represented, especially in numerical data-sets. PCA will thus be able to represent the hot and cold zone but not the flame front zone which is the main zone of interest. Indeed if one want to accurately reproduce the source terms or the flame speed using the results of PCA, the set of PCs chosen should be able to accurately reconstruct the flame front.

To solve this problem, a pretreatment should be applied to the data-set, to remove the effects of uneven sampling. The most obvious approach is to suppress the over represented zone prior to the PCA. While simple, this process has two main disadvantages: first, one should find a criterion to distinguish the relevant zones from the non-relevant ones, which is not an easy task mainly because this criterion should be robust. Second, even if such a criterion can be found, the PCA process will no longer be fully automatic, and as stated in the work of Sutherland and Parente [13], the main advantage of the PCA is the automatic identification of relevant variables.

In the following section, the kernel density method is proposed to solve this problem: the sample density can be computed and used to lower the influence of the over represented zone in the covariance matrix.

### 4. Weighting based on kernel density method

The kernel density method was introduced by Rosenblatt [16] and Parzen [17]. The general idea of this method is to compute the density of the statistical sample in a distribution using a presumed normal distribution.

#### 4.1. Single-variable case:

For each observation the distance, \( d_{c,c'} \), between the current observation \( c \) and an observation \( c' \) is computed:

\[
d_{c,c'} = |x_{c'} - x_c| \tag{17}
\]
and the global weighting
\[ W_c = \frac{1}{\sqrt{n}} \text{max}(\frac{1}{K_c}) \] (25)

4.3. Choice of \( h \)

The density, and so the weighting, is strongly dependent of the bandwidth value, \( h_k \), as it will be shown in section 7.1. However, for a given data-set an optimal value for \( h_k \) can be found (see [19]). This optimum value can be either found by an iterative process, that is changing \( h_k \) until the error is smaller than a given threshold, or from the data-set characteristics. Using the latter method the bandwidth can be computed for a mono-variable weighting as [19]:
\[ h = \left( \frac{4\hat{\sigma}}{\sqrt{3n}} \right)^\frac{1}{5} \] (26)

where \( \hat{\sigma} \) is the standard deviation of the considered variable within the data-set.

For multi-variable kernel density a different bandwidth can be used for each variable [19], but since the data used to compute the weighting are centered and scaled by the standard deviation of all the variables distributions, \( \hat{\sigma}_k \), is therefore be equal to one, so that:
\[ \hat{\sigma}_k = \hat{\sigma} \quad \forall k \] (27)

Using a global bandwidth or a variable dependent bandwidth is therefore equivalent:
\[ h_k = \left( \frac{4\hat{\sigma}_k}{\sqrt{3n}} \right)^\frac{1}{5} = h \] (28)

where \( \hat{\sigma} \) is the mean of the standard deviations \( \sigma_k \).

4.4. Application

To demonstrate the ability of the proposed weighting to tackle the problem exposed in section 3, the method will be applied on the data-set defined by equations 14 to 16. Each centered-scaled observation is multiplied by its associated weighting before applying the PCA, that is:
\[ \tilde{X} = \mathbf{W} \tilde{X} \] (29)

where \( \mathbf{W} \) is the matrix containing the weighting for each observation. Note that the weightings are computed using the centered-scaled data-set, \( \tilde{X} \), and that \( \tilde{X} \) is only used for the computation of the covariance matrix.

Figure 5 shows the recovered \( x_2 \) as of function of the observed \( x_2 \) using one PC. The weighting was performed using the multi-variable kernel density method, and \( h \) was computed using equation 26. Figure 5 emphasizes that the weighting behaves as expected: it emphasizes the importance of the region originally under-represented in the data-set, yielding a much accurate representation of this region but a worse representation of the originally over-represented regions. It can be concluded that sampling is crucial for the analysis of non-linear data set, and an appropriate choice of the weighting function can emphasize forward the main features of interest for the investigated system.

The characteristics kernel density weighted PCA will now be studied on the basis of a 1-D premixed hydrogen/air flame. Note that, in section 5.3, a method to solve the issue related to the reconstruction of the originally over-represented regions will be proposed.
5. Characteristics of the kernel density weighted PCA

The performances of the kernel density weighted PCA are studied for a 1-D premixed hydrogen/air flame.

This data-set was obtained from a DNS computation using the compressible Navier-Stokes solver YWC developed at the EM2C laboratory. The thermochemical properties are computed using the REGATH package developed at the EM2C laboratory which is based on the CHEMKIN [20] formalism (i.e differential diffusion effects are taken into account).

The case is a H₂-Air 1-D laminar flame at an equivalence ratio of $\phi = 0.88$ and a temperature of 300 K. The pressure is set to 101,325 Pa, velocity of the unburned gases is $1.25 \text{ m/s}$, and the chemical reaction mechanism by Katta et al. [21] was employed. The domain consists of 1,281 points with a constant grid spacing of $\Delta x = 1.25 \times 10^{-5} \text{ m}$. A view of the temperature field of the data-set is shown in Figure 6. Nine variables are included in the data-set: $Y_H$, $Y_{H_2}$, $Y_O$, $Y_{O_2}$, $Y_{OH}$, $Y_N$, $Y_{H_2}O$, $Y_{H_2}O_2$ and $Y_{H_2}O_3$.

5.1. T-BAKED PCA

The multi-variable kernel density computation, presented in section 4.2, yields two issues when applied to a combustion data-set. First computing cost can be high when large dataset are considered. To reduce this cost a tree-based algorithm could be used but a mono-variable kernel density can lead to erroneous evaluation of the weights due to the very nature of combustion processes data-sets. This is why it is of great interest to investigate if a mono-variable kernel density can be used.

If a mono-variable kernel density is computed, a variable must be chosen. The most efficient constraint is to require that this variable must describe the progress of the combustion process. The most relevant choice is therefore the temperature, yielding a Temperature BAsed KErnel Density weighted PCA or T-BAKED PCA.

Figure 7 A and B show the density and the weighting function for the multi-variable kernel density method and the T-BAKED method, respectively. The bandwidths were computed using equation 26. The comparison of Figure 7 A and B, for $Y_{H_2}O$, $Y_{H_2}O_2$ and $Y_{OH}$. Those reconstructions are achieved using 4 PCs. Those species were chosen because $Y_{H_2}O$ and $Y_{OH}$ are present in the flame front and in the post flame zone and $Y_{H_2}O_2$ is only present in the flame front. From this, one can conclude that the T-BAKED PCA, yields more accurate results than the multi-variable
Figure 7: Density (grey) and weighting (black) functions as a function of the temperature, T in K for: (A) multi-variable kernel density, (B) mono-variable kernel density based on T.

Figure 8: Recovered $Y_{\text{H}_2\text{O}}$, $Y_{\text{H}_2\text{O}_2}$, $Y_{\text{OH}}$ as a function of their observed counterparts with 4 PCs, using a multi-variable kernel density (A) and a mono-variable kernel density (B).
The T-BAKED PCA approach allows improving the representation of the flame front region, while the representation of the hot and cold regions is worsened and left in the minor PCs. A hybrid PCA relying on the following idea can be adopted: if p PCs should be retrieved from the data-set, qw PCs will be obtained using the T-BAKED PCA and qnw PCs with the classical, non weighted (nw) PCA (q = qw + qnw). In our combustion problem, the qw PCs will account for the flame front and qnw PCs for the hot and cold regions.

The resulting approach, called n Hybrid T-BAKED PCA or HT-BAKED PCA, can be summarized as follows: from the centered-scaled data-set, \( \tilde{X} \), qw PCs are computed using the T-BAKED PCA and the data-set, \( \tilde{X}_w \) is reconstructed with those PCs:

\[
\tilde{X}_w = Z_{qw}A_{qw}^t
\]  

(30)

The data-set \( \tilde{X}_w \) containing the information relative to the low sample density regions is then subtracted from the original data-set, yielding \( \tilde{X}^* \):

\[
\tilde{X}^* = \tilde{X} - \tilde{X}_w
\]  

(31)

Thus, \( \tilde{X}^* \) will only contains information relevant to the over represented zones, on which a classical PCA can be performed to extract the remaining qnw PCs. The data-set recovered using those qnw PCs is:

\[
\tilde{X}_{nw}^* = Z_{nw}A_{nw}^t
\]  

(32)

So that the data-set recovered using the q PCs computes as:

\[
\tilde{X} = \tilde{X}_{nw}^* + \tilde{X}_w = Z_{qw}A_{qw}^t + Z_{nw}A_{nw}^t
\]  

(33)

From a numerical study performed on the datasets presented below, the optimal choice for retaining the PCs of each type is the following: first the classical PCA is performed retaining one component, then the T-BAKED PCA is performed retaining one component. This process is then repeated until the total number of PCs retained is reached. Centering and scaling and the inverse operations are performed at the beginning and the end of the algorithm to avoid introducing a bias in the second PCA.

Results of the T-BAKED and HT-BAKED PCA will now be benchmarked against the classical PCA for the 1-D flame data-set.

6. Results for the 1-D flame data-set

The reconstruction results for a classical T-BAKED and HT-BAKED PCA are plotted on Figure 10 A, B and C respectively for \( Y_{H_2O} \), \( Y_{H_2O_2} \), \( Y_{OH} \). The weighting was computed using the bandwidth from equation 26 and for each reconstruction 4 PCs were used. This number of PCs allows to recover more than 98% of the variance contained in the original data-set.

For the HT-BAKED, 2 PCs of each type were used and the bandwidth was also computed using equation 26.

Table 2 shows the \( R^2 \) statistic of the species reconstruction for the three methods. Before any interpretation is made, it should be stressed that the \( R^2 \) square

<table>
<thead>
<tr>
<th>Specie</th>
<th>( h ) from equation 26</th>
<th>Optimal ( h )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( Y_{H_2O} )</td>
<td>237 K</td>
<td>280 K</td>
</tr>
<tr>
<td>( Y_{H_2O_2} )</td>
<td>237 K</td>
<td>220 K</td>
</tr>
<tr>
<td>( Y_{OH} )</td>
<td>237 K</td>
<td>240 K</td>
</tr>
</tbody>
</table>

Table 1: Value of the bandwidth for \( Y_{H_2O} \), \( Y_{H_2O_2} \), and \( Y_{OH} \) allowing an optimal reconstruction using a T-BAKED PCA with 4 PCs.
Table 2: $R^2$ statistics for the reconstruction of 1-D premixed flame using 4PCs using HT-BAKED PCA and T-BAKED PCA.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$R^2$ Classical</th>
<th>$R^2$ T-BAKED</th>
<th>$R^2$ HT-BAKED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_H$</td>
<td>0.98612</td>
<td>0.9561</td>
<td>0.99861</td>
</tr>
<tr>
<td>$Y_{H_2}$</td>
<td>0.99841</td>
<td>0.5095</td>
<td>0.99722</td>
</tr>
<tr>
<td>$Y_O$</td>
<td>0.98478</td>
<td>0.7722</td>
<td>0.99804</td>
</tr>
<tr>
<td>$Y_{O_2}$</td>
<td>0.99940</td>
<td>0.5318</td>
<td>0.99756</td>
</tr>
<tr>
<td>$Y_{OH}$</td>
<td>0.95047</td>
<td>0.9494</td>
<td>0.96956</td>
</tr>
<tr>
<td>$Y_{H_2O}$</td>
<td>0.9514</td>
<td>0.4938</td>
<td>0.99693</td>
</tr>
<tr>
<td>$Y_{HO_2}$</td>
<td>0.95477</td>
<td>0.9977</td>
<td>0.99062</td>
</tr>
<tr>
<td>$Y_{H_2O_2}$</td>
<td>0.96671</td>
<td>0.9729</td>
<td>0.98635</td>
</tr>
</tbody>
</table>

The configuration of the computational case used to generate the data-set is presented in Figure 11. It is a premixed hydrogen flame-vortex interaction at a temperature of 300 K and a pressure of 101,325 Pa. The unburned gases are composed of hydrogen, nitrogen and oxygen with the following mass fractions: $Y_H = 0.05$, $Y_{O_2} = 0.4$ and $Y_{N_2} = 0.55$. The grid consist of 161 x 321 nodes with a constant grid spacing in all directions $\Delta x = \Delta y = \Delta z = 2.5 \times 10^{-3}$ m which results in a box of $2 \times 10^{-3}$ m by $4 \times 10^{-3}$ m. The premixed hydrogen flame chemistry is described using the mechanism of Katta et al. [21].

As in the work of Renard et al. [22] the flow is initialized by a one dimensional premixed flame, extended along the $y$-axis on which two counter-rotating vortices are superimposed in the unburnt gases. The vortices are initialized using equations 34, 35 and 36 with $R_c = 0.2 \times 10^{-3}$ m, $\tau = 20 \times 10^{-3}$, $u_{10} = 2$ m/s and $p_0 = 101,325$ Pa. The distance between the vortices center is equal to $R_c$. This result in a maximum $x$-axis velocity of around $122$ m/s and so a convection speed of the ring of $61$ m/s.

\[
\begin{align*}
  u_1(x,y) &= u_{10} + \frac{\tau}{R_c} y \exp \left( -\frac{(x^2 + y^2)}{2 R_c^2} \right) \\
  u_2(x,y) &= \frac{\tau}{R_c} x \exp \left( -\frac{(x^2 + y^2)}{2 R_c^2} \right) \\
  p(x,y) &= p_0 \exp \left( -\frac{\gamma}{2} \left( \frac{\tau}{c R_c} \right)^2 \exp \left( -\frac{(x^2 + y^2)}{R_c^2} \right) \right)
\end{align*}
\]

The temperature field and the distribution of $Y_H$ for the data-set are shown in Figure 1 and 12, respectively. For the sake of clarity, the $y$-axis extent has been limited to $2 \times 10^{-3}$ m in the two Figures.
Figure 10: Recovered $Y_{\text{H}_2\text{O}}$, $Y_{\text{H}_2\text{O}_2}$, $Y_{\text{OH}}$ as a function of their observed counterparts with 4 PCs, using a classical PCA (A), a T-BAKED PCA (B) and a HT-BAKED PCA (C).
Table 3: $R^2$ statistics for the reconstruction of 2-D flame-vortex using 4PCs using HT-BAKED PCA and T-BAKED PCA.

<table>
<thead>
<tr>
<th>Variable</th>
<th>$R^2$ Classical</th>
<th>$R^2$ T-BAKED</th>
<th>$R^2$ HT-BAKED</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Y_H$</td>
<td>0.9862</td>
<td>0.9729</td>
<td>0.9972</td>
</tr>
<tr>
<td>$Y_{H_2}$</td>
<td>0.9962</td>
<td>0.7749</td>
<td>0.9892</td>
</tr>
<tr>
<td>$Y_O$</td>
<td>0.9451</td>
<td>0.7153</td>
<td>0.9696</td>
</tr>
<tr>
<td>$Y_{O_2}$</td>
<td>0.9924</td>
<td>0.9654</td>
<td>0.9992</td>
</tr>
<tr>
<td>$Y_{OH}$</td>
<td>0.9874</td>
<td>0.9885</td>
<td>0.9938</td>
</tr>
<tr>
<td>$Y_{H_2O}$</td>
<td>0.9847</td>
<td>0.9157</td>
<td>0.9947</td>
</tr>
<tr>
<td>$Y_{HO_2}$</td>
<td>0.9999</td>
<td>0.9998</td>
<td>0.9999</td>
</tr>
<tr>
<td>$Y_{H_2O_2}$</td>
<td>0.9999</td>
<td>0.9976</td>
<td>0.9999</td>
</tr>
</tbody>
</table>

As for the 1D-premixed flame 4 PCs were used for recovering the original data-set, which allows to recover more than 98% of the original information. The T-BAKED and HT-BAKED PCA were performed using equation 26 for the bandwidth $h$. Note that numerical study also indicates that the most accurate reconstruction results are obtained using 2 PCs of each type for the HT-BAKED PCA.

As for the 1D-premixed flame data-set, results for the reconstruction of $Y_{H_2O}$, $Y_{H_2O_2}$, $Y_{OH}$ are plotted on Figure 13 and the $R^2$ statistics for the three methods are shown in Table 3. For $Y_{H_2O}$ and $Y_{OH}$ results are similar to those of the 1-D flame: in the flame front the T-BAKED PCA yields more accurate results, while from a global point of view the HT-BAKED PCA is more accurate than the T-BAKED PCA.

An comparison of the plotted reconstruction of $Y_{H_2O}$ and $Y_{OH}$, shows that the HT-BAKED PCA is superior to the classical PCA. As for the 1-D flame, HT-BAKED PCA appears then superior to classical or T-BAKED PCA.

The two proposed approaches are thus efficient to improve the performance of the premixed flame PCA. However, a large part of the combustion processes used relies on diffusion flame. Therefore the proposed methods will be applied on a DNS data-set of a turbulent CO/H$_2$/N$_2$-air diffusion flame in the next section.

### 7.2. Diffusion flame

To fully validate the proposed T-BAKED and HT-BAKED method, a turbulent diffusion flame data-set is now considered. This data-set comes from a 3-D DNS of CO/H$_2$/N$_2$-air jet diffusion flame using detailed chemistry, details about the computation set-up can be found in [23]. Twelve variables are included in this data-set: $Y_{CO}$, $Y_{CO_2}$, $Y_H$, $Y_{H_2O}$, $Y_{H_2O_2}$, $Y_{HCO}$, $Y_{HO_2}$, $Y_{H_2O_4}$, $Y_O$, $Y_{O_2}$, $Y_{N_2}$ and $Y_{OH}$.

As for the other data-sets 4 PCs where retained for the three methods, allowing to recover at least 91% of the information, and the bandwidth was computed using equation 26. Figure 14 gives the reconstruction results for $Y_{CO_2}$, $Y_O$, $Y_{O_2}$ and $Y_{OH}$. Unlike the premixed cases, the T-BAKED and HT-BAKED PCA gives similar results, except for $Y_{O_2}$ where an improvement in the reconstruction...
Figure 13: Recovered $Y_{\text{H}_2\text{O}}$, $Y_{\text{H}_2\text{O}_2}$, $Y_{\text{OH}}$ as a function of their observed counterparts with 4 PCs, using a classical PCA (A), a T-BAKED PCA (B) and a HT-BAKED PCA (C).
8. Conclusions

In this work the kernel density method was introduced into the PCA process to address the issue of uneven sampling in highly non-linear multi-dimensional data sets, as in combustion systems. The method introduces a density metric within the PCA, which allows a better reconstruction of the flame region, usually under-represented in high-fidelity numerical (and experimental) data sets. The approach was first demonstrated for a two variables data-set, and then tested on a 1-D laminar flame, on a 2-D flame vortex interaction data-set and on a 3-D turbulent diffusion flame data-set.

The 1-D laminar flame was employed to perform a benchmark of the mono-variable vs. multi-variable kernel density method, showing that temperature can be used for computing the kernel density. The resulting T-BAKED PCA allows a better preconditioning of the data-set for the PCA at lower computational cost than the multi-variable kernel density method. Moreover, an optimal value of the bandwidth was also found (equation 26).

Results on the 1-D laminar flame showed that T-BAKED PCA allows to significantly improve the description of the flame. Due to the very nature of the method, however, the information contained in the hot and cold regions disappear from the first PCs, leading to higher reconstruction errors in those zones.

To solve the problem, the HT-BAKED PCA method was introduced: it combines classic and T-BAKED PCA, to optimize the PCA reconstruction in all regions providing the best reconstruction of the original data-set for a fixed number of PCs. This conclusion was drawn both from the analysis of the 1-D flame data-set and more representative 2-D flame vortex interaction and 3-D turbulent diffusion flame data-sets. Finally, the best strategy for choosing the number of PCs of each type appears to be an equi-repartition.

In conclusion the T-BAKED PCA and HT-BAKED are powerful and robust tools to study combustion processes from high-fidelity numerical data-set or experimental data-set: if the main interest is in the flame region, T-BAKED PCA can be employed due to its remarkable performances; on the other hand, when a globally optimized reconstruction is of interest, HT-BAKED PCA provides the most accurate results for a fixed number of PCs.

9. Acknowledgments

The first author was supported by a fellowship from the Fonds National de la Recherche Scientifique, FRFNRS (Communauté Française de Belgique). Computing resources were provided by the IDRIS under the allocation 2009-1209020164 made by GENCI (Grand Equipement National de Calcul Intensif). The authors wish to thank Prof. James C. Sutherland from the University of Utah for gently providing the diffusion flame data-set.

References

Figure 14: Recovered $Y_{\text{CO}_2}$, $Y_O$, $Y_{\text{O}_2}$ and $Y_{\text{OH}}$ as a function of their observed counterparts with 4 PCs, using a classical PCA (A), a T-BAKED PCA (B) and a HT-BAKED PCA (C).


