

Local manifold learning and its link to domain-based physics knowledge

Kamila Zdybał^{*a,b}, Giuseppe D'Alessio^{a,b,c}, Antonio Attili^d, Axel Coussement^{a,b}, James C. Sutherland^e and Alessandro Parente^{a,b}

^aUniversité Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Brussels, Belgium

^bUniversité Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium

^cCRECK Modeling Lab, Department of Chemistry, Materials and Chemical Engineering, Politecnico di Milano, Piazza Leonardo da Vinci 32, 20133 Milano, Italy

^dUniversity of Edinburgh, School of Engineering, Institute of Multiscale Thermo-fluids, The King's Buildings, Mayfield Road, Edinburgh, EH9 3FD, United Kingdom

^eDepartment of Chemical Engineering, University of Utah, Salt Lake City, Utah, USA

ARTICLE INFO

Keywords:

local principal component analysis, low-dimensional manifolds, manifold learning, data parameterization, data clustering

ABSTRACT

In many reacting flow systems, the thermo-chemical state-space is known or assumed to evolve close to a low-dimensional manifold (LDM). Various approaches are available to obtain those manifolds and subsequently express the original high-dimensional space with fewer parameterizing variables. Principal component analysis (PCA) is one of the dimensionality reduction methods that can be used to obtain LDMs. PCA does not make prior assumptions about the parameterizing variables and retrieves them empirically from the training data. In this paper, we show that PCA applied in local clusters of data (local PCA) is capable of detecting the intrinsic parameterization of the thermo-chemical state-space. We first demonstrate that utilizing three common combustion models of varying complexity: the Burke-Schumann model, the chemical equilibrium model and the homogeneous reactor. Parameterization of these models is known *a priori* which allows for benchmarking with the local PCA approach. We further extend the application of local PCA to a more challenging case of a turbulent non-premixed *n*-heptane/air jet flame for which the parameterization is no longer obvious. Our results suggest that meaningful parameterization can be obtained also for more complex datasets. We show that local PCA finds variables that can be linked to local stoichiometry, reaction progress and soot formation processes.

* Corresponding author: Kamila.Zdybal@ulb.ac.be

1. Introduction

Parameterization methods for modeling turbulent reacting flows rely on expressing the entire thermo-chemical state-space with fewer variables [1, 2, 3, 4, 5, 6]. This allows for an easier treatment of combustion process that is in general high-dimensional due to many species involved in chemical reactions. Recently, dimensionality reduction and manifold learning techniques have been used to infer low-dimensional parameterizations directly from the training data [7, 8, 9]. In particular, principal component analysis (PCA) identifies the new parameters as directions of the largest variance in the ambient space. The new parameters, called principal components (PCs), are linear combinations of the preprocessed thermo-chemical state variables.

The question then arises if the parameters obtained in a data-driven way can be given any physical meaning. If PCA is able to retrieve physical information about the combustion process without any *a priori* knowledge, this can yield useful implications for modeling complex systems using PCs. For instance, progress variables can be identified automatically, without the need for human insight [10]. Evidence from previous research [11, 12, 13, 14, 15] suggests that physical meaning can still be attributed to PCs. A common way to interpret PCs is to assess coefficients of the linear combination

that formed them. Since those coefficients can be negative or positive, PCA can potentially differentiate between reactants and products. Results of Parente et al. [11] and Malik et al. [14] suggest that PCA can identify the mixture fraction variable, f , as defined by Bilger et al. [16]. The PC that can be given the meaning of f is characterized by high coefficients of opposite signs for fuel and oxidizer components. Such PC is thus a compact representation of mixture stoichiometry, even though f is not explicitly included as one of the variables in the training data.

While some insights are available from globally applied manifold learning, it is less clear what parameterizations can be derived from applying dimensionality reduction locally. For instance, if PCA is applied locally in identified zones (or clusters) of state-space, local parameterizations can be obtained, tied to specific regions of the flame. Cluster analysis for physically-relevant characteristics has been the concern of various researchers in the domain of reacting flows [11, 17, 18, 19, 20]. Previous research demonstrated that local PCs can vary significantly between clusters and can be linked to physical processes associated with local regions [11, 18, 20].

The goal of this work is to interpret the low-dimensional parameterization defined by the local PCs. We adopt a bottom-up approach and test local PCA on combustion datasets of gradually increasing complexity. We first show that apply-

ORCID(s):

ing PCA in local clusters identified by the vector quantization PCA (VQPCA) algorithm [21], we are able to retrieve parameters inherent to the construction of common combustion models: Burke-Schumann (BS), chemical equilibrium (EQ) and homogeneous reactor (HR). All three models assume that compositions lie close to a low-dimensional manifold (LDM) and the thermo-chemical state-space can be parameterized with one variable. We then test the local PCA approach on a more challenging case of a turbulent non-premixed *n*-heptane/air jet flame dataset coming from direct numerical simulation (DNS). The DNS dataset used here allows to link the local PCs to various phenomena beyond what the analysis of BS, EQ and HR datasets allows for, such as soot formation processes [22, 23, 18, 24]. We further show how findings carry over between the simpler strained laminar flamelet (SLF) model and the DNS dataset, which includes more physical scatter than the simple combustion models.

2. The analyzed datasets

We first apply local PCA on three combustion models with *a priori* known parameterization:

1. The Burke-Schumann (BS) model for methane/air with temperature and 4 species mass fractions (mass fraction of N_2 is removed from PCA). 10k observations are generated for 10k equally-spaced points in the f space. For $f = 0$ and $f = 1$, the pure stream temperature is 300K.
2. The chemical equilibrium (EQ) model for methane/air with temperature and 51 species mass fractions (mass fractions of N_2 and Ar are removed from PCA) using species from the GRI-3.0 chemical mechanism [25]. 10k observations are generated for 10k equally-spaced points in the f space. For $f = 0$ and $f = 1$, the pure stream temperature is 300K and pressure is 101325Pa.
3. The adiabatic, isobaric, open homogeneous reactor (HR) for hydrogen/air with temperature and 8 species mass fractions (mass fractions of N_2 , Ar and He are removed from PCA). The initial and feed temperature is 1000K and pressure is 101325Pa. This dataset has 538 observations, generated on a temporal grid.

In the supplementary material, we apply local PCA to the EQ and HR datasets with additional fuels to cover hydrogen/air, syngas/air and methane/air combustion for both datasets. Hydrogen/air and methane/air combustion is used with the same settings as mentioned for datasets 2 and 3. When syngas/air combustion is studied, we use the chemical mechanism by Hawkes et al. [26] and the state-space is formed from temperature and 10 species mass fractions (mass fraction of N_2 is removed from PCA). In the supplement, we show that the conclusions reported in the main text still hold for other fuels.

We then demonstrate our approach on a high-fidelity simulation dataset for which parameterization is no longer obvious:

4. The direct numerical simulation (DNS) of a 3D turbulent non-premixed *n*-heptane/air jet flame at a Reynolds number of around 15,000 and with no significant extinction. The dataset represents one 2D plane obtained from a 3D domain at $t = 15\text{ms}$ and is composed of temperature and 46 species mass fractions (mass fraction of N_2 is removed from PCA) based on the reduced mechanism by Bisetti et al. [27]. This data has approximately 260k observations. More information about the dataset can be found in [22].

We make a connection between the findings from the DNS dataset and the one-dimensional flame case corresponding to the DNS dataset:

5. The strained laminar flamelet (SLF) model for *n*-heptane/air with temperature and 46 species mass fractions (mass fraction of N_2 is removed from PCA) based on the reduced mechanism by Bisetti et al. [27]. We use two variations of the SLF dataset: one with unity Lewis number (Le) assumption (approximately 21k observations) and one with a mixture-averaged differential diffusion model (approximately 11k observations). More information about the dataset can be found in [23].

Each dataset 1-5 is represented by a matrix $\mathbf{X} \in \mathbb{R}^{N \times Q}$, where N is the number of observations and Q is the number of state variables. This matrix defines the thermo-chemical state-space, $\mathbf{X} = [T, Y_1, Y_2, \dots, Y_{n_s}]$, where $T \in \mathbb{R}^{N \times 1}$ is the temperature vector, $Y_i \in \mathbb{R}^{N \times 1}$ is the mass fraction of species i vector and n_s is the number of linearly independent species mass fractions. We refer to the i^{th} observation (row) in $\mathbf{X}$ as X_{i*} and to the j^{th} variable (column) in $\mathbf{X}$ as X_{*j} .

Before applying dimensionality reduction, the original dataset has to be preprocessed (centered and scaled). The preprocessing can be performed as $\tilde{\mathbf{X}} = (\mathbf{X} - \mathbf{C})\mathbf{D}^{-1}$, where $\mathbf{C} \in \mathbb{R}^{N \times Q}$ is the matrix of centers (typically mean values of each variable) and $\mathbf{D} \in \mathbb{R}^{Q \times Q}$ is the diagonal matrix of scales. Each element, d_j , on the diagonal of $\mathbf{D}$ is the scaling factor for the j^{th} variable. In this work, we test two data preprocessing scenarios:

- Auto scaling, where $d_j = \text{std}(X_{*j})$ with the state vector defined as $\mathbf{X} = [T, Y_1, Y_2, \dots, Y_{n_s}]$,
- Pareto scaling, where $d_j = \sqrt{\text{std}(X_{*j})}$ with the state vector defined as $\mathbf{X} = [Y_1, Y_2, \dots, Y_{n_s}]$.

For brevity, we refer to the dataset formulated as $\mathbf{X} = [T, Y_1, Y_2, \dots, Y_{n_s}]$ as $\mathbf{X} = [T, Y_i]$ and to the dataset formulated as $\mathbf{X} = [Y_1, Y_2, \dots, Y_{n_s}]$ as $\mathbf{X} = [Y_i]$. The choice for the two preprocessing scenarios follows the reasoning presented in [28], where Auto scaling was recommended for exploratory analysis and Pareto scaling was shown to align the first PC with temperature. The reason why we remove the temperature variable when using Pareto scaling in this work, is to allow for more robust linear combinations of the species mass fractions that have a better chance of serving as progress variables. We come back to this discussion in §4.3. Furthermore, when using Auto

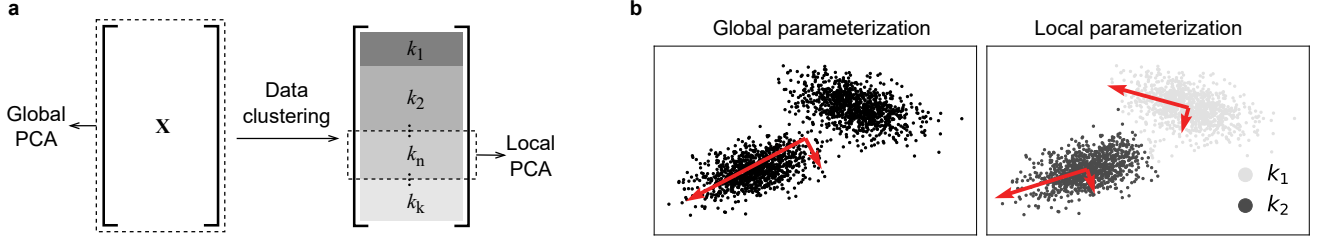


Figure 1: Schematic distinction between (a) global and local PCA and (b) global and local eigenvectors (red arrows) defining the directions of the new data parameterization. The n^{th} cluster is labeled k_n .

scaling, all scaled variables become dimensionless, while in Pareto scaling they do not.

3. Global and local PCA

In this section, we briefly present the theory for global PCA, performed on the entire data matrix, $\mathbf{X}$, and for local PCA, performed in local clusters of $\mathbf{X}$, where the n^{th} cluster is labeled k_n . The distinction between global and local PCA is schematically illustrated in Fig. 1a.

3.1. Global PCA

Global PCA is performed on the entire preprocessed data matrix, $\tilde{\mathbf{X}}$. The preprocessed dataset can then be projected onto the new basis, $\mathbf{A} \in \mathbb{R}^{Q \times Q}$, such that $\mathbf{Z} = \tilde{\mathbf{X}}\mathbf{A}$, where $\mathbf{Z} \in \mathbb{R}^{N \times Q}$ is the matrix of global PCs. The j^{th} PC (the j^{th} column of $\mathbf{Z}$) is thus given by

$$Z_{*j} = \sum_{i=1}^Q a_{ij} X_{*i}, \quad (1)$$

where the coefficients of the linear combination, a_{ij} , are the elements of the j^{th} column of $\mathbf{A}$. PCA also provides a linear route back to the original basis, $\mathbf{X} = \mathbf{Z}\mathbf{A}^T\mathbf{D} + \mathbf{C}$ since $\mathbf{A}^T = \mathbf{A}^{-1}$ for orthonormal $\mathbf{A}$. Selecting only q first PCs and eigenvectors, we can approximate the original full-rank matrix $\mathbf{X}$ by the rank- q matrix $\mathbf{X}_q$, $\mathbf{X} \approx \mathbf{X}_q = \mathbf{Z}_q\mathbf{A}_q^T\mathbf{D} + \mathbf{C}$, where the subscript q denotes truncation from Q to q components. The matrix of the truncated PCs is thus $\mathbf{Z}_q \in \mathbb{R}^{N \times q}$. For datasets exhibiting very strong low-rank structure, $q \ll Q$ is sufficient to accurately approximate $\mathbf{X}$ with $\mathbf{X}_q$.

3.2. Local PCA

Once the partitioning of the dataset is obtained using any clustering algorithm of choice, PCA can be performed in local clusters of data. Let $\mathbf{X}^{(n)} \in \mathbb{R}^{N_n \times Q}$ be the observations extracted from $\mathbf{X}$ that belong to the n^{th} cluster, k_n , with N_n observations. We also define $\tilde{\mathbf{X}}^{(n)} \in \mathbb{R}^{N_n \times Q}$ as the analogous observations extracted from the centered and scaled global dataset, $\tilde{\mathbf{X}}$. Throughout this work, whenever local PCA is performed, the extracted cluster is re-centered (but not re-scaled) through $\tilde{\mathbf{X}}^{(n)} = \tilde{\mathbf{X}}^{(n)} - \mathbf{C}^{(n)}$, where $\mathbf{C}^{(n)} \in \mathbb{R}^{N_n \times Q}$ is the matrix of centers computed from $\tilde{\mathbf{X}}^{(n)}$. The low-dimensional projection is analogous to the global PCA

case, $\mathbf{Z}^{(n)} = \tilde{\mathbf{X}}^{(n)} \mathbf{A}^{(n)}$, where $\mathbf{Z}^{(n)} \in \mathbb{R}^{N_n \times Q}$ is the matrix of local PCs, associated with the n^{th} cluster. It is worth noting, that the local eigenvectors and PCs are, in general, different from the global ones. This is illustratively demonstrated in Fig. 1b. We refer to the j^{th} local PC in the n^{th} cluster as $Z_{*j}^{(n)}$ (the j^{th} column of $\mathbf{Z}^{(n)}$). Finally, the local set of observations can be approximated with the rank- q matrix, $\mathbf{X}_q^{(n)}$, $\mathbf{X}^{(n)} \approx \mathbf{X}_q^{(n)} = \mathbf{Z}_q^{(n)} (\mathbf{A}_q^{(n)})^T \mathbf{D} + \mathbf{C} + \mathbf{C}^{(n)}$.

3.3. Correlation between the known and the retrieved local parameterization

We utilize correlation metrics to test whether parameterization with local PCs is equivalent to the *a priori* known parameterization, $\mathcal{Y}$, for datasets 1-3. Within each cluster, k_n , we measure correlation between the j^{th} local PC, $Z_{*j}^{(n)}$, and the local parameterizing variable, denoted in general as $\mathcal{Y}^{(n)}$. We adopt two correlation metrics: Pearson correlation coefficient (PCC) and distance correlation (dCor) [29, 30]. In the n^{th} cluster, the local PCC between $Z_{*j}^{(n)}$ and $\mathcal{Y}^{(n)}$ is computed as

$$r_n(Z_{*j}^{(n)}, \mathcal{Y}^{(n)}) = \left| \frac{\sum_{i=1}^{N_n} (Z_{ij}^{(n)} - \bar{Z}_{*j}^{(n)}) (\mathcal{Y}_i^{(n)} - \bar{\mathcal{Y}}^{(n)})}{\left(\sum_{i=1}^{N_n} (Z_{ij}^{(n)} - \bar{Z}_{*j}^{(n)})^2 \sum_{i=1}^{N_n} (\mathcal{Y}_i^{(n)} - \bar{\mathcal{Y}}^{(n)})^2 \right)^{1/2}} \right|, \quad (2)$$

where N_n is the total number of observations in the n^{th} cluster and the overbars denote an arithmetic average over N_n samples. We further average the PCC metric over all k clusters,

$$r(Z_{*j}, \mathcal{Y}) = \frac{1}{k} \sum_{n=1}^k r_n(Z_{*j}^{(n)}, \mathcal{Y}^{(n)}). \quad (3)$$

The above equation yields a single averaged PCC value, r , for a given clustering solution. The r values are bounded between 0 and 1. In the general case, $r = 0$ is not a sufficient condition for independence between two variables. In the n^{th} cluster, the local dCor between $Z_{*j}^{(n)}$ and $\mathcal{Y}^{(n)}$ is computed as

$$\text{dCor}_n(Z_{*j}^{(n)}, \mathcal{Y}^{(n)}) = \left(\frac{\text{dCov}_n(Z_{*j}^{(n)}, \mathcal{Y}^{(n)})}{\text{dCov}_n(Z_{*j}^{(n)}, Z_{*j}^{(n)}) \text{dCov}_n(\mathcal{Y}^{(n)}, \mathcal{Y}^{(n)})} \right)^{1/2},$$

(4)

with the local distance covariance in the n^{th} cluster between two random variables, x and y , defined as $\text{dCov}_n(x, y) = (1/N_n^2 \sum_{i,j=1}^{N_n} x_{ij}y_{ij})^{1/2}$, where x_{ij} and y_{ij} are the elements of the double-centered Euclidean distances matrices for x and y observations respectively. The dCor metric is also averaged over k clusters,

$$\text{dCor}(Z_{*j}, \mathcal{Y}) = \frac{1}{k} \sum_{n=1}^k \text{dCor}_n(Z_{*j}^{(n)}, \mathcal{Y}^{(n)}). \quad (5)$$

The dCor values are bounded between 0 and 1, where $\text{dCor} = 0$ iff two variables are independent. The PCC metric can only detect linear dependence between random variables. In contrast, the dCor metric is more suitable when nonlinear association exists between variables. Thus, values of dCor larger than PCC can suggest that nonlinear dependencies remain in local portions of data. The convergence of the two metrics can indicate that the dataset has been partitioned into clusters that represent nearly linear segments of $\mathcal{Y}$.

3.4. Clustering based on variable bins

One way to cluster non-premixed combustion datasets is to use mixture fraction, f , as the conditioning variable and assign observations to bins generated in the range of f . This method belongs to the family of supervised clustering. The first split to two clusters can be performed at the stoichiometric value of mixture fraction, f_{st} , which has to be known beforehand. This first split separates observations for which $f < f_{st}$ (fuel-lean) from observations for which $f \geq f_{st}$ (fuel-rich). If the number of clusters $k > 2$, the fuel-lean and the fuel-rich branch can be further divided into segments equal in length in the f -space, starting from the branch which has a larger range in the values of f . This technique clusters the training dataset explicitly based on the chemical composition stoichiometry. For premixed systems, other physical variable, such as one of the state variables, can be selected to create the bins. In this work, whenever variable other than f is used for clustering, we create k equally-spaced bins between the minimum and the maximum value of that variable.

3.5. Clustering using the VQPCA algorithm

VQPCA is an iterative clustering algorithm that partitions the data based on the minimum reconstruction error from local PCA [21, 31, 32]. Performing clustering through dimensionality reduction is a unique feature of VQPCA. Compared to other unsupervised techniques such as K-Means clustering, VQPCA has a potential advantage of being less sensitive to noise or outliers in the data. The convergence of the algorithm is assessed with: (1) the convergence of cluster centroids (the centroid of each cluster should not move more than a threshold between two consecutive iterations) and (2) the convergence of the PCA reconstruction error (the error should be smaller than an agreed threshold). The algorithm takes the preprocessed dataset, $\tilde{\mathbf{X}}$, and the following hyperparameters as inputs:

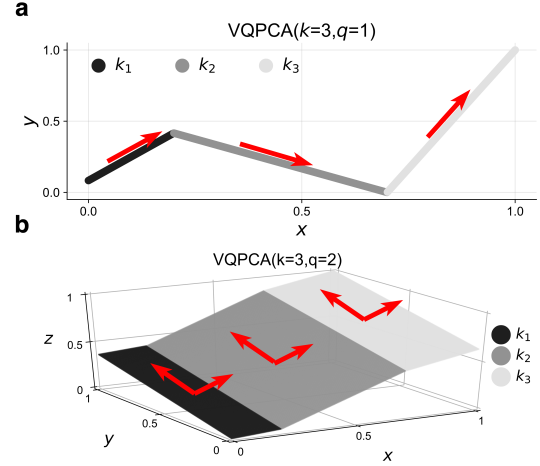


Figure 2: (a) Two-dimensional and (b) three-dimensional synthetic datasets composed of hyperplanes and the result of their VQPCA clustering. Red arrows represent the expected local eigenvectors from the converged iteration, associated with each cluster. The VQPCA algorithm classifies each hyperplane as a separate cluster based on the computed eigenvectors.

- the number of clusters, k , to create in the dataset,
- the number of local eigenvectors, q , with which each observation is approximated at each iteration, and
- the initialization of cluster centroids, $v_0 \in \mathbb{Z}_+^{N \times 1}$, defining the initial cluster classifications.

For simplicity, we refer to the VQPCA solution with k clusters and q local eigenvectors as VQPCA(k, q). The VQPCA algorithm is bootstrapped with the initial cluster centroids, v_0 , which can be obtained using any data partitioning technique of choice. The most straightforward way is to randomly assign original observations into k clusters. Another option is to uniformly stratify the data matrix into initial clusters. In this work, we use bins of mixture fraction as the cluster initializations on datasets BS, EQ, DNS and SLF and equally-spaced bins of H_2O mass fractions for the HR dataset (as described in §3.4).

At each iteration of the VQPCA algorithm, the dataset is approximated using a separate set of local eigenvectors computed from each of the k clusters. A matrix of reconstruction errors, $\epsilon \in \mathbb{R}^{N \times k}$, is populated, such that each column is linked to one of the k clusters. Each (i, j) -th element, ϵ_{ij} , of that matrix represents the reconstruction error of the i^{th} observation if it was assigned to the j^{th} cluster. The error is computed based on the global scaled reconstruction error metric as per [31],

$$\epsilon_{ij} = \|\tilde{X}_{i*} - \tilde{X}_{q,i*}^{(j)}\|, \quad (6)$$

where $\tilde{X}_{i*}$ is the preprocessed i^{th} observation of all Q variables in the dataset and $\tilde{X}_{q,i*}^{(j)}$ is the rank- q approximation of that observation using local eigenvectors from the j^{th} cluster. The minimum reconstruction error is then searched across

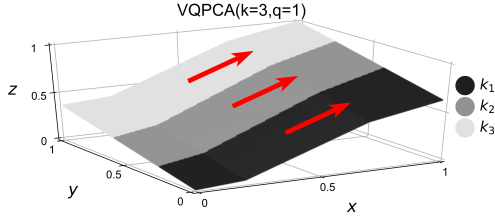


Figure 3: The result of clustering the synthetic dataset from Fig. 2b using VQPCA($k=3, q=1$). Red arrows represent the possible local eigenvectors from the converged iteration, associated with each cluster.

every row of $\mathbf{E}$. An individual observation is eventually assigned to the cluster in which that error was the smallest. This allows to form a new vector of cluster classifications and re-compute the local eigenvectors to be used at the next iteration. If the number of observations assigned to a particular cluster is less than the number of variables in a dataset, the cluster is removed. The observations from that removed cluster are redistributed at the next iteration. This prevents forming spurious clusters. Once the VQPCA algorithm converges, we obtain the final vector defining cluster classifications, $\mathbf{v} \in \mathbb{Z}_+^{N \times 1}$.

In essence, each cluster obtained with the VQPCA algorithm is associated with some q -dimensional subspace of the original Q -dimensional space. We define linearity in q dimensions as observations that lie within the span of a q -dimensional orthogonal basis. In a special case, when the dataset is composed of k q -dimensional linear manifolds, we can expect that VQPCA(k, q) will classify each manifold as a separate cluster. In a more general case, for datasets that are not composed of q -dimensional manifolds or, when we fail to guess how many underlying manifolds there are, each obtained cluster lies close to a q -dimensional manifold. The “closeness” is measured by the reconstruction errors as per Eq. (6). Fig. 2 visualizes this concept on two synthetic datasets composed of hyperplanes. The first one, in Fig. 2a, is composed of three line segments embedded in a two-dimensional space. When VQPCA($k=3, q=1$) is applied to this data, a single eigenvector is allowed to define each of the three clusters. Since the span of a single vector can define a line, the expectation is that VQPCA will partition this dataset such that each line segment becomes one cluster. The second dataset, in Fig. 2b, is composed of three planar segments embedded in a three-dimensional space. This time, when VQPCA($k=3, q=2$) is applied, we allow two orthogonal eigenvectors to be associated with each cluster. This enables detecting planar structures and, in an analogous way to Fig. 2a, we can expect VQPCA to assign each of the three planes to a separate cluster. We observe that the algorithm performed as expected; each hyperplane is classified as a separate cluster based on the computed eigenvectors. The expected eigenvectors from the converged iteration are schematically illustrated with red arrows.

A potential shortcoming of the VQPCA algorithm emerges following our discussion. VQPCA cannot guarantee a “rea-

sonable” clustering solution if one fails to guess the dimensionality, q , of the underlying manifolds. This is demonstrated in Fig. 3, where we apply VQPCA($k=3, q=1$) on the synthetic dataset from Fig. 2b. Constraining VQPCA to fit one-dimensional manifolds ($q=1$) results in a different clustering solution from the one presented in Fig. 2b. Unsurprisingly, the existence of the three planar segments is now ignored. The clustering result still satisfies optimality from the perspective of the reconstruction error metric (as per Eq. (6)). Hence, in Fig. 3, we see three elongated clusters, each being a result of assigning observations to the span of a single local eigenvector (schematically visualized with the red arrows). Comparing clustering results in Figs. 2b and 3, it becomes evident that changing the hyper-parameter q can result in very different clusters.

This shortcoming calls for measures that allow for automatic detection of the intrinsic dimensionality of the local manifolds. While several metrics exist for approximating the effective or intrinsic dimensionality [33], this task remains somewhat challenging. One way in which we can guide the choice for q in an *a posteriori* analysis of clustering solutions is by measuring the average reconstruction accuracy of each cluster using the local eigenvectors associated with that cluster. For example, the coefficient of determination, R^2 , is equal to unity when the three clusters from the VQPCA($k=3, q=2$) solution in Fig. 2b are reconstructed. For the VQPCA($k=3, q=1$) solution, the average R^2 for the reconstruction of the three clusters seen in Fig. 3 is 0.69. The supplementary material presents extended results for this synthetic dataset (see Fig. S1) for other combinations of k and q . A non-unity value of R^2 conveys the fact that observations in a cluster have some level of scatter around the span of their defining eigenvector(s). Hence, observations cannot be reconstructed perfectly and, at best, lie close to that span. For datasets such as the DNS data, this is generally the case due to large level of physical scatter in the data. Tied to this discussion, although perhaps less cumbersome, is the need to find the optimal number of clusters, k , that sufficiently represent distinct physical phenomena. This point can be addressed using clustering quality metrics that

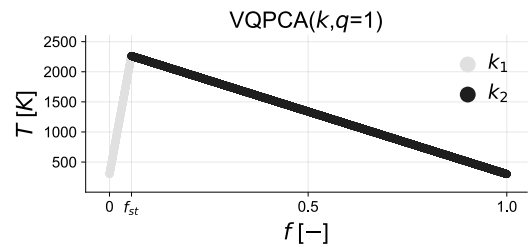


Figure 4: The result of VQPCA clustering of the BS dataset using $q=1$ (irrespective of the number of clusters, k , originally requested). f_{st} denotes the stoichiometric value of f . Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. An identical result is obtained when local PCA is performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. Clusters were initialized by binning f .

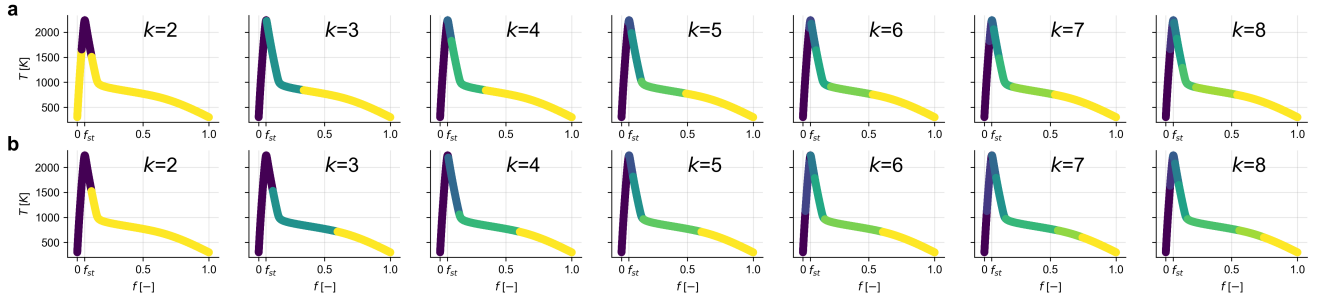


Figure 5: The result of VQPCA clustering of the EQ dataset to k clusters using $q=1$. f_{st} denotes the stoichiometric value of f . (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning f .

measure intra-cluster homogeneity and inter-cluster heterogeneity. We come back to this discussion in §4.4 and present an approach for a more informed selection of the pair (k, q) tied to a given dataset, when no *a priori* physical knowledge is available to guide this choice.

4. Results and discussion

4.1. The Burke-Schumann model

The BS dataset carries similarities with the synthetic dataset presented in Fig. 2a; by construction, this dataset is parameterized by $\mathcal{Y} = f$ only. Since the lean and rich branch are linear functions of f , clustering with VQPCA($k=2, q=1$) should separate the lean and rich branch into two clusters. The result of this clustering is presented in Fig. 4. VQPCA splits the dataset at the stoichiometric mixture fraction value, f_{st} , and the two clusters correspond to the fuel-lean and the fuel-rich side. Such partition leads to 100% correlation in terms of r and dCor between the first local PC and the local f in both clusters. Further, we have found that no matter how many clusters, k , are requested for partitioning this dataset, VQPCA ultimately finds only two, removing spurious clusters that have too few observations. This behavior is understandable, since the BS dataset is composed of only two locally one-dimensional manifolds. It is important to note here, that this finding does not prove that our local PCA approach “sees” the underlying physical meaning of the two clusters *a priori*. Rather, retrieving the two clusters is an aftermath of the mechanics of the VQPCA algorithm (see §3.5). Since the BS model is constructed using two one-dimensional linear manifolds, the one-dimensional manifolds found by VQPCA concur entirely with our physical interpretation of these manifolds.

4.2. The chemical equilibrium model

Similarly to the BS dataset, the EQ dataset is inherently parameterized by $\mathcal{Y} = f$ but the local manifolds are not linear. Fig. 5 shows the result of VQPCA clustering of the EQ dataset to k clusters using $q=1$, where k varies from 2 to 8. We compare two data preprocessing scenarios. In Fig. 5a, local PCA was performed using Auto scaling on a dataset

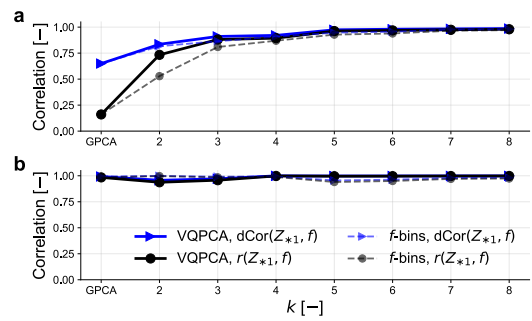


Figure 6: The r and dCor values between the first local PC and the local $\mathcal{Y} = f$ when clustering the EQ dataset with VQPCA with an increasing number of clusters, k , and using $q=1$. For comparison, dashed lines show analogous correlations but for clustering solutions obtained directly from binning the f vector. We also show correlations coming from global PCA (marked with GPCA). (a) Global and local PCA were performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Global and local PCA were performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning f . Legend applies to (a) and (b).

$\mathbf{X} = [T, Y_i]$. In Fig. 5b, local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In both cases, clusters were initialized by binning the f vector. In Fig. 5a, when creating two clusters, the first cluster captures all observations from the pure streams. In the space of f - T , this first cluster is discontinuous. It is worth noting, however, that f does not constitute the dimensions of the ambient space. The fact that VQPCA classified the observations in the oxidizer and fuel stream into one cluster means that those observations lie sufficiently close to the same local manifold, embedded in the ambient space. Starting from $k=3$ clusters, VQPCA uses the additional third eigenvector to construct separate manifolds for the oxidizer and fuel stream. It is visible in Fig. 5a that also starting from $k=3$, VQPCA splits the dataset close to f_{st} . In the space of f - T , this results in splitting the dataset at the location of the largest visible non-linearity. For the data preprocessing case from Fig. 5b, we do not observe discontinuous clusters. However, the split

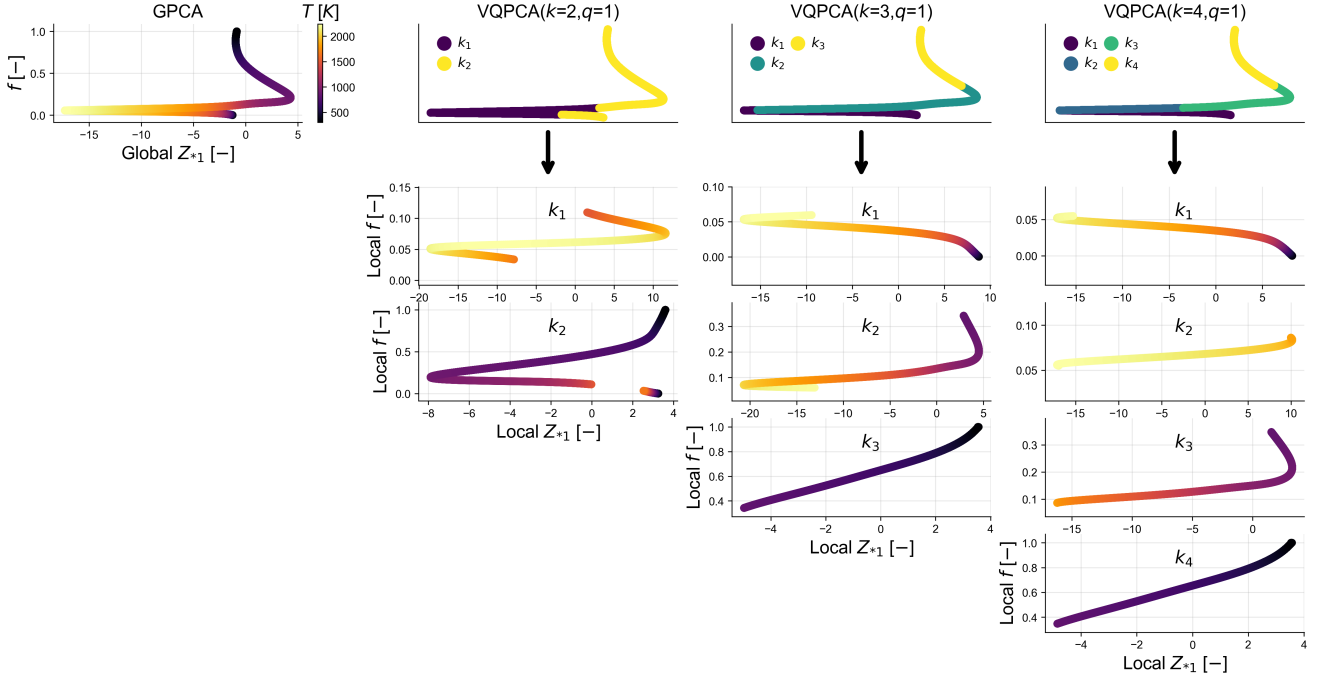


Figure 7: True shapes of the global and local one-dimensional manifolds in the EQ dataset as functions of f . We report the one-dimensional manifolds found by global PCA (GPCA) and local PCA when k was varied from 2 to 4 in $VQPCA(k, q=1)$. Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. Clusters were initialized by binning f . The same temperature colorbar applies to global and local manifolds.

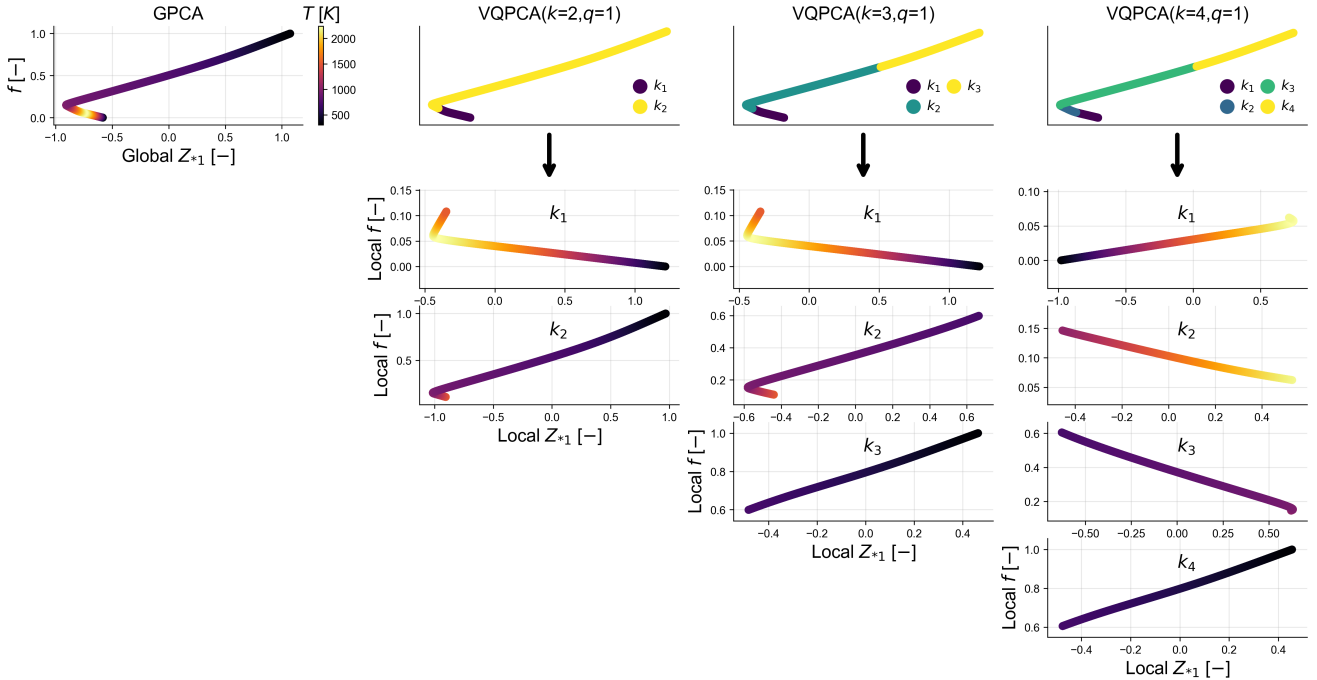


Figure 8: True shapes of the global and local one-dimensional manifolds in the EQ dataset as functions of f . We report the one-dimensional manifolds found by global PCA (GPCA) and local PCA when k was varied from 2 to 4 in $VQPCA(k, q=1)$. Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. Clusters were initialized by binning f . The same temperature colorbar applies to global and local manifolds.

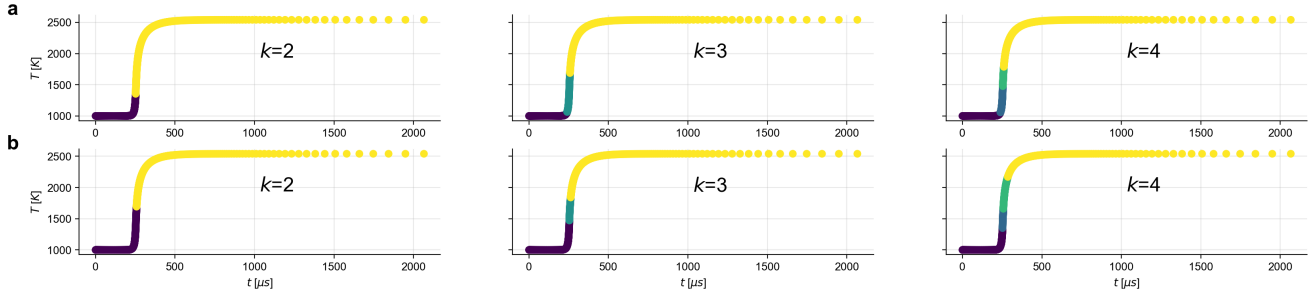


Figure 9: The result of VQPCA clustering of the HR dataset to k clusters using $q = 1$. (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning Y_{H_2O} .

close to f_{st} happens only starting from $k=4$.

In order to interpret the one-dimensional local manifolds that VQPCA finds for this dataset, we compute the r and dCor values between the first local PC and the local $\mathcal{Y} = f$, for each clustering solution from Fig. 5. These values, for increasing k , are reported in Fig. 6 with solid lines. We obtain a converging behavior in the correlation metrics as the number of clusters increases. For Fig. 6a, which corresponds to the preprocessing case with Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$, the minimum correlation in terms of r is 73% when $k=2$. For Fig. 6b, which corresponds to the preprocessing case with Pareto scaling on a dataset $\mathbf{X} = [Y_i]$, the minimum correlation in terms of r is 94% when $k=2$. Thus, there is a good agreement between the parameterization obtained with local PCA and the known intrinsic parameterization of this EQ dataset ($\mathcal{Y} = f$).

We note, that even though clusters were initialized from binning the f vector, VQPCA re-distributes observations during the iterative process with respect to the initialized solution. The re-distribution is such that correlations are generally higher for the final VQPCA clustering solution as compared to the clustering solution coming from binning f alone (cf. solid and dashed lines in Fig. 6). While we have found binning the f vector to be a good choice for initializing clusters in the VQPCA algorithm, the final VQPCA clustering solution is improved with respect to simply binning the f vector in terms of retrieving the locally one-dimensional manifolds.

There is a visible larger discrepancy between r and dCor values when $k < 5$, especially for the case of Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. This suggests that for small k , local manifolds still exhibit some level of nonlinearity. It is thus instructive to inspect the true shapes of the local one-dimensional manifolds in functions of $\mathcal{Y} = f$ for the EQ dataset. Fig. 7 shows these true shapes for global PCA and for local PCA performed in clusters found by VQPCA where k varies from 2 to 4. We note that the global manifold (denoted GPCA) and the local manifolds associated with VQPCA($k=2, q=1$), exhibit significant nonlinearity. For the VQPCA($k=3, q=1$) solution, we obtain at least one local manifold (in cluster k_3) which is nearly a linear function

of f . For the VQPCA($k=4, q=1$) solution, the linearity of local manifolds is more pronounced, with manifolds in at least three clusters (k_1 , k_2 and k_4) being nearly piecewise linear one-dimensional manifolds. Fig. 8 shows analogous true shapes of the local one-dimensional manifolds in the EQ dataset, but for the preprocessing case using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. We observe that the one-dimensional manifold from global PCA is already close to being a linear in function of f , as opposed to the global PCA manifold seen in Fig. 7. With local PCA performed in VQPCA clusters, we obtain nearly linear one-dimensional manifolds, especially when creating four clusters (the VQPCA($k=4, q=1$) solution). This qualitative comparison of local manifolds in Figs. 7-8 between the two preprocessing scenarios can help explain the generally much higher (nearing 100%) correlations seen in Fig. 6b (for Pareto scaling with $\mathbf{X} = [Y_i]$), as compared to Fig. 6a (for Auto scaling with $\mathbf{X} = [T, Y_i]$). This also demonstrates that even for qualitatively similar clustering solutions (cf. Figs. 5a-b), data preprocessing can play a significant role in local manifold parameterization and interpretation. Some means of data preprocessing (such as Pareto scaling applied on $\mathbf{X} = [Y_i]$) can help align manifold parameterizations with the physically expected parameterization of the system. The supplementary material presents analogous results for hydrogen/air and syngas/air combustion (see Figs. S2-S9).

4.3. The homogeneous reactor model

We apply the local PCA approach to a single time trajectory of the HR dataset. The HR data results from solving a system of ODEs with physical time, t , being the independent variable. For this dataset, time is the variable describing the reaction progress, from initial condition in the reactor, through mixture ignition to chemical equilibrium. Fig. 9 shows the result of VQPCA clustering of the HR dataset to k clusters, in the space of time, t , (the independent variable) and temperature, T . Here, k varies only from 2 to 4 due to small number of observations in the HR dataset. For $k > 4$, some clusters become too small to be considered meaningful for the correlation analysis. Since a single time trajectory is used, we select $q=1$ in the VQPCA algorithm. Since the

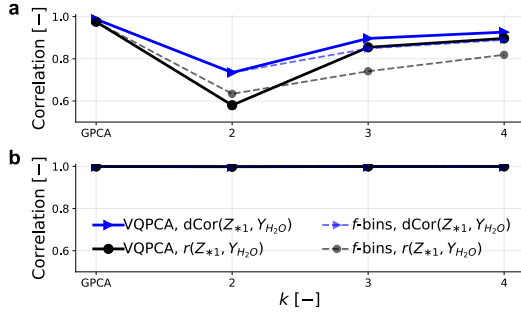


Figure 10: The r and dCor values between the first local PC and the local $\mathcal{Y} = Y_{H_2O}$ when clustering the HR dataset with VQPCA with an increasing number of clusters, k , and using $q=1$. For comparison, dashed lines show analogous correlations but for clustering solutions obtained directly from binning the Y_{H_2O} vector. We also show correlations coming from global PCA (marked with GPCA). (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning Y_{H_2O} . Legend applies to (a) and (b).

HR dataset describes a premixed system, VQPCA is now initialized with equally-spaced bins of the H_2O mass fraction, Y_{H_2O} . We observe that clusters are continuous along the time dimension. This suggests that the VQPCA algorithm clusters the HR dataset based on a variable describing progress of chemical reactions.

Mass fractions of combustion products, (in the case of hydrogen fuel, H_2O) can be used to define progress variables [34, 35, 36, 37]. In general, a progress variable, c , is defined in the literature as $c = \sum_{i=1}^{n_s} w_i Y_i$, where the weighting factors, w_i , are typically non-zero only for stable species and such that their linear combination changes monotonically across the flame. There is a similarity of this definition with Eq. (1), which suggests that under certain settings, a PC can satisfy the role of a progress variable. The dataset for PCA transformation needs to be composed of species mass fractions only (i.e. the temperature variable is not included in the training data). In Fig. 10, with solid lines we show correlations between the local PC in each cluster and the H_2O mass fraction ($\mathcal{Y} = Y_{H_2O}$). For Fig. 10a, which corresponds to the preprocessing case with Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$, the minimum correlation in terms of r is 58% when $k=2$. For Fig. 10b, which corresponds to the preprocessing case with Pareto scaling on a dataset $\mathbf{X} = [Y_i]$, the correlation in terms of r is 100% for all k considered. We thus see consistency of the local PCA results with the clusters visualization presented in Fig. 9. The VQPCA algorithm partitioned the dataset based on parameters highly correlated with a progress variable. This is especially the case when Pareto scaling was used on a dataset with the temperature variable removed, following our earlier expectation about how c is formed. In the supplement, we show analogous results for syngas/air and methane/air combustion (see Figs. S10-S13). We also show correlations between the first

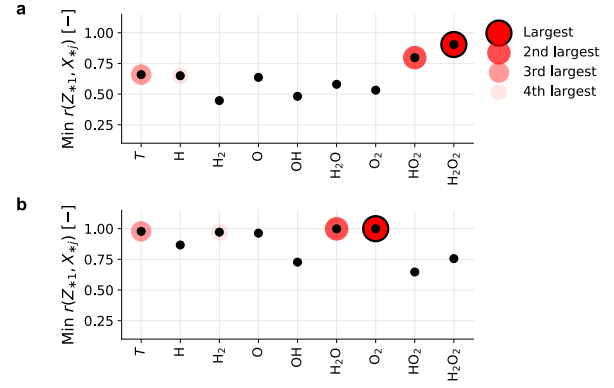


Figure 11: The minimum r values between the first local PC and each local state variable, X_{*j} , when clustering the HR dataset for **hydrogen/air** combustion with VQPCA with an increasing number of clusters, k , and using $q=1$. (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning Y_{H_2O} . Legend applies to (a) and (b).

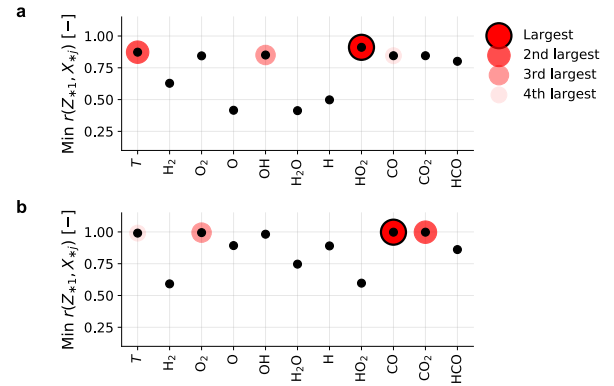


Figure 12: The minimum r values between the first local PC and each local state variable, X_{*j} , when clustering the HR dataset for **syngas/air** combustion with VQPCA with an increasing number of clusters, k , and using $q=1$. (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning Y_{H_2O} . Legend applies to (a) and (b).

local PC and $\mathcal{Y} = Y_{CO_2}$, which are of relevance for hydrocarbon fuels (see Figs. S14-S15).

Similarly as we have done for the EQ dataset, we can compute analogous correlations but for the HR data clustering obtained from binning the H_2O mass fraction vector. These results are shown in Fig. 10 with dashed lines. We again observe smaller correlations in terms of r and dCor as compared to clustering with the VQPCA algorithm, with the exception of VQPCA clustering with $k=2$ in Fig. 6a. Even though correlations associated with Pareto scaling are nearing 100%, we see smaller correlations for all k when we bin the H_2O mass fraction vector as compared to clustering using VQPCA. This suggests that the data preprocessing case

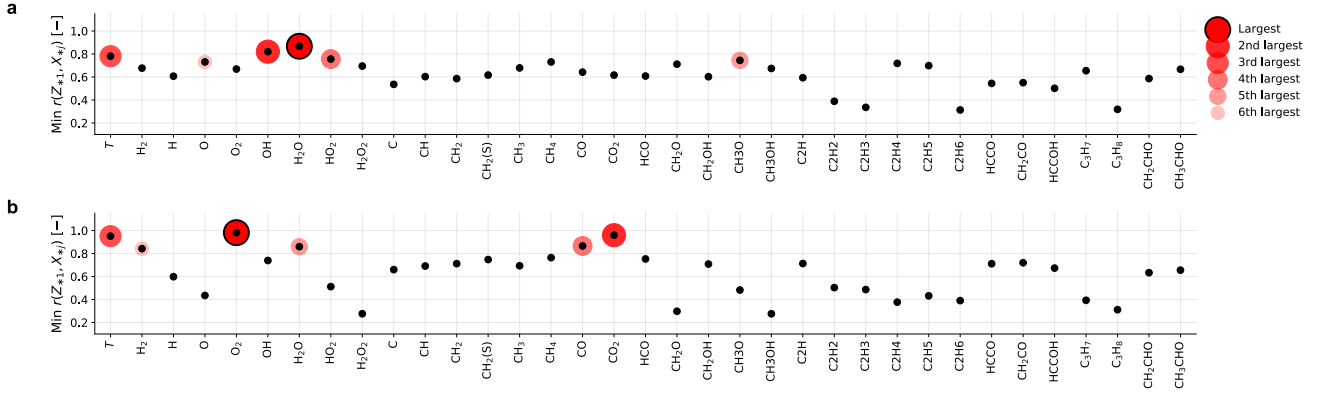


Figure 13: The minimum r values between the first local PC and each local state variable, X_{*j} , when clustering the HR dataset for **methane/air** combustion with VQPCA with an increasing number of clusters, k , and using $q=1$. (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning Y_{H_2O} . Legend applies to (a) and (b).

based on Pareto scaling of the species mass fractions is more appropriate when optimized progress variables need to be formulated.

In addition, Figs. 11-13 report the minimum r values for correlation between the first local PC and each local state variable in the HR datasets across three considered fuels: hydrogen, syngas and methane. The minimum r was searched across k changing from 2 to 4; Figs. 11-13 are thus a more compact presentation of correlations with respect to Fig. 10. With the red shaded outlines, we mark the first few largest r values. To summarize the findings from Figs. 11-13, when Auto scaling is used on a dataset $\mathbf{X} = [T, Y_i]$, the highest correlations are observed for

- H_2O_2 , HO_2 , T , H for hydrogen/air,
- HO_2 , T , OH , CO for syngas/air,
- H_2O , OH , T , HO_2 , CH_3O , O for methane/air,

and when Pareto scaling is used on a dataset $\mathbf{X} = [Y_i]$, the highest correlations are observed for

- O_2 , H_2O , T , H_2 for hydrogen/air,
- CO , CO_2 , O_2 , T for syngas/air,
- O_2 , CO_2 , T , CO , H_2O , H_2 for methane/air.

For methane/air combustion we neglect NOx species from this analysis. The species mass fractions exhibiting highest correlations are consistent with variables typically used in the literature to construct progress variables [4, 38, 37, 39, 40]. This is especially the case when Pareto scaling is used on a dataset $\mathbf{X} = [Y_i]$. This result is consistent with the work done in [28], where the authors reported that Pareto scaling emphasizes major and stable species in the definition of the first global PCs. Correlation analysis as we have performed here, can thus help in formulating optimized local c as a linear combination of state variables [10].

4.4. The high-fidelity DNS dataset: data-aided interpretation and relation with the training manifold

Thus far, we have used our physical knowledge of the system to guide the selection of q . Since the intrinsic parameterization of the DNS dataset is not known *a priori*, we first search the space of possible pairs (k, q) to apply on the DNS dataset. Our approach for obtaining a reasonable pair (k, q) is based on measuring the reconstruction quality from local PCA. We look at how closely the conditional means of five selected reconstructed state variables (T , Y_{H_2O} , Y_{O_2} , Y_{CO_2} and $Y_{n-C_7H_{16}}$) approximate their true conditional means. The choice for k and q is made once the selected variables conditioned on f are reconstructed reasonably well. We denote the true conditional mean of variable X_{*j} as $\langle X_{*j} | f \rangle_t$ and the reconstructed conditional mean as $\langle X_{*j} | f \rangle_r$. To define the conditioning variable, we generate 20 bins in the space of f . Bins have varying sizes such that they are tighter close to f_{st} and larger close to $f = 0$ and $f = 1$. For the j^{th} selected variable, X_{*j} , we measure the j^{th} coefficient of determination for the conditional mean as

$$R_j^2 = 1 - \frac{\sum_{i=1}^{20} \left(\langle X_{*j} | f \rangle_{t,i} - \langle X_{*j} | f \rangle_{r,i} \right)^2}{\sum_{i=1}^{20} \left(\langle X_{*j} | f \rangle_{t,i} - \text{mean}(\langle X_{*j} | f \rangle_{t,i}) \right)^2}. \quad (7)$$

We then find the minimum R_j^2 that happened for any of the five selected variables, which becomes our final reconstruction quality measure, R^2 . We set the threshold constraint for selecting the smallest q when $R^2 \geq 0.99$. If locally linear regions of the low-dimensional manifold are found, a multi-linear model with a correct dimensionality would reconstruct the data perfectly (recall Fig. S1). With the physical scatter present in the DNS data, we do not expect R^2 to be equal to unity with $q < Q$. Supplementary Fig. S16 shows a colormap of the R^2 values over a Cartesian product $k \times q$, where $k = [2, 3, \dots, 8]$ and $q = [1, 2, \dots, 10]$.

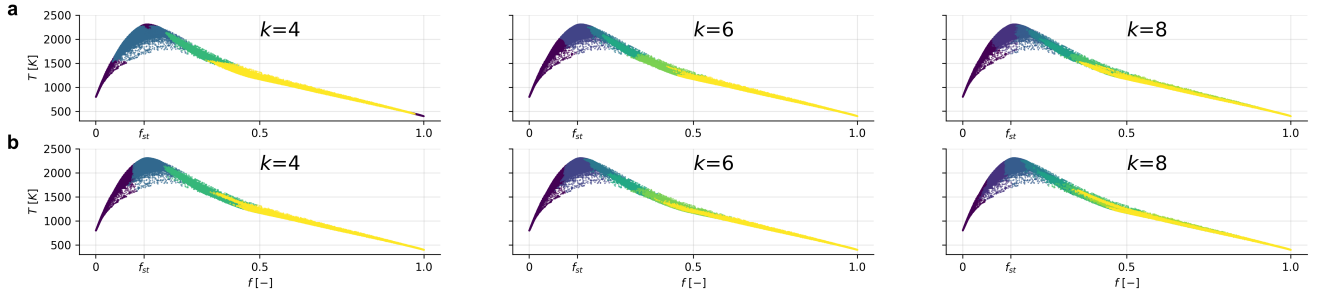


Figure 14: The result of VQPCA clustering of the DNS dataset to k clusters using $q=3$. (a) Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. (b) Local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In (a) and (b), clusters were initialized by binning f .

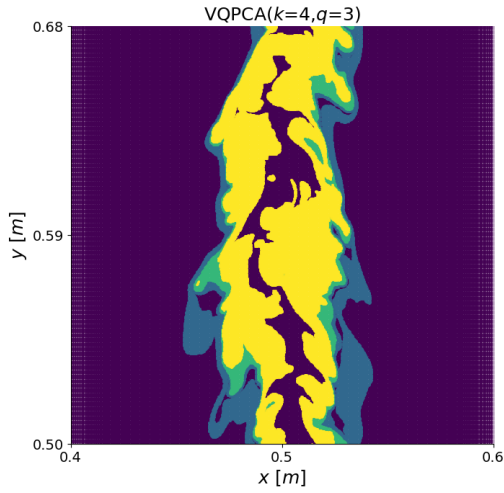


Figure 15: The result of clustering the DNS dataset using VQPCA($k=4, q=3$) plotted in the physical space, x - y . Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. Clusters were initialized by binning f .

The highest R^2 values happen for $q \geq 3$, suggesting that the DNS dataset should be approximated with at least three-dimensional manifolds. The black cross in Fig. S16 marks the maximum R^2 which happened for ($k=7, q=10$). The black circle marks the smallest dimensionality q for which R^2 is at least 0.99 which happened for ($k=5, q=3$). In addition, Fig. S17 shows a comparison of conditional means for the case marked with the black circle (Fig. S17a) and with the black cross (Fig. S17b). We observe that both cases reconstruct the conditional means reasonably well. We thus select $q=3$ for subsequent DNS data analysis.

Fig. 14 shows the result of VQPCA clustering of the DNS dataset to $k=4, k=6$ and $k=8$ clusters using $q=3$. The clustering solutions are plotted in the f - T space. Similarly as before, we compare two preprocessing scenarios. In Fig. 14a, local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. In Fig. 14b, local PCA was performed using Pareto scaling on a dataset $\mathbf{X} = [Y_i]$. In both cases, clusters were initialized by binning the f vector. In Fig. 15, we fur-

ther visualize one particular clustering solution, VQPCA($k=4, q=3$), corresponding to Auto scaling with $\mathbf{X} = [T, Y_i]$ in the physical space, x - y . We take this particular solution for subsequent detailed analysis.

We begin with the qualitative interpretation of clusters seen in Fig. 15. The first cluster, k_1 , contains the non-reacting observations from the n -heptane core jet and the air co-flow. To quantitatively support this evidence, we carry out Procrustes analysis between the shape of this cluster and the shapes of the spatial profiles of various chemical species. Procrustes analysis is a statistical tool to compare two geometrical shapes, S_1 and S_2 , after accomplishing operations of translation, rotation, reflection, and scaling. It minimizes the sum of squared differences between the two considered shapes as an objective function [41, 42]. Procrustes analysis returns a dissimilarity coefficient, $d(S_1, S_2) \in [0, 1]$, which is equal to zero when two shapes are identical. When the shape of k_1 is geometrically compared to the spatial profiles of chemical species, the minimum value obtained for d is associated with the O_2 mass fraction, $d(k_1, Y_{O_2}) = 0.34$. In contrast, the median value of d obtained when the shape of k_1 is compared with any of the remaining species in the dataset is 0.75. We also look at the isolated region where the fuel jet is injected (at $x \approx 0.5$ m). The dissimilarity coefficient between the central section of k_1 and the region where the n -heptane mass fraction is greater than 0.3 is $d(k_1, Y_{n-C_7H_{16}}) = 0.33$. From the analogous geometrical analysis, the second cluster, k_2 , appears to represent the layer where oxygen starts to react and hydroxyl and oxygen radicals are formed, since $d(k_2, Y_{OH}) = 0.4$ and $d(k_2, Y_O) = 0.46$. In the supplementary material, we provide additional figures for a qualitative visual comparison of the shapes of clusters k_1 and k_2 and the relevant species profiles (see Figs. S18-S19).

The third cluster, k_3 , is particularly interesting from the physical and chemical point of view. In Fig. 16, we examine the eigenvector weights associated with the first two local PCs in k_3 and we highlight the five highest weights (in terms of an absolute value) in red. Variables with the highest weights are the ones associated with the polycyclic aromatic hydrocarbons (PAH) growth and consumption. In particular, the high weight for propargyl (C_3H_3) can be, from a

For the very same chemical system and flame configuration, Attili et al. [22] observed that the peak in acetylene (C_2H_2) and naphthalene mass fractions occurred at $f \approx 0.3$. Concurrently, soot was absent at smaller values of f ($f < 0.3$) due to oxidation. For the DNS simulation time that we study here ($t = 15ms$), soot was also present in the entire rich side of the flame. We carry out an additional quantitative physical interpretation of cluster k_3 from the VQPCA($k=4, q=3$) solution in the f -space. In Fig. 17a, we examine probability density functions (PDFs) in f among clusters obtained with VQPCA($k=4, q=3$). The interval of f where the flame is expected to exhibit sooting behavior is contained within k_3 (marked with the red outline). In addition, VQPCA seems to be always finding a cluster dedicated

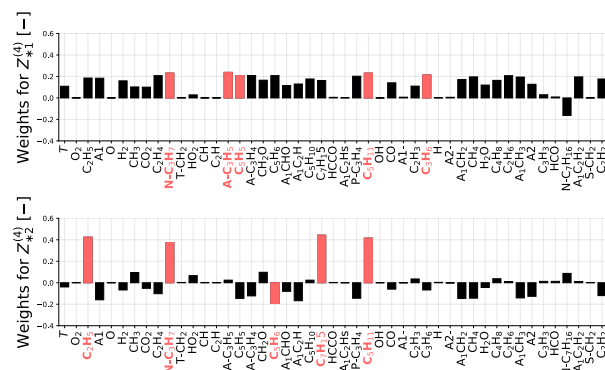
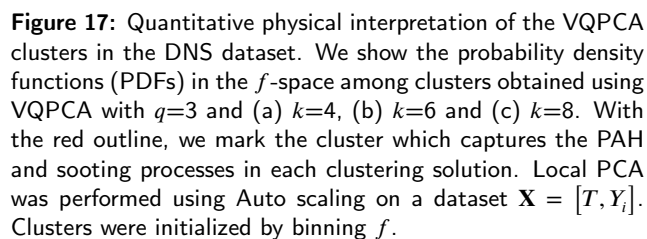


Figure 18: Quantitative physical interpretation of the first two local eigenvectors from the DNS dataset in cluster k_4 resulting from the VQPCA($k=4, q=3$) solution. Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. Clusters were initialized by binning f .

to PAHs and soot, even when the algorithm is applied with a different number of clusters. This is shown in Figs. 17b-c, where the analogous PDFs for f are reported for the VQPCA solutions using $k=6$ and $k=8$, respectively. In each case, we mark the cluster which captures the sooting processes occurring at $f \approx 0.3$ with the red outline. The supplementary Figs. S22-S23 demonstrate the eigenvector weights in the outlined clusters from the VQPCA($k=6, q=3$) and the VQPCA($k=8, q=3$) solutions. We note the similar structure to the one presented in Fig. 16. This observation suggests that our local PCA approach can be generalizable for different values of k with a fixed dimensionality q .

Finally, we look into the fourth cluster, k_4 , which encapsulates the region close to the rich side of the flame. Fol-

lowing the results of [22], we expect that this cluster can still capture the soot formation chemistry, but for conditions $f \rightarrow 1$. Fig. 18 shows the eigenvector weights for the first two PCs in k_4 . We note a qualitatively different PC structure to the one coming from k_3 . In particular, the two PCs do not exhibit high weights for A2; this is in contrast to what we observe in Fig. 16. Since k_4 spans the range in f from $f \approx 0.35$ to 1.0 (see Fig. 17a), we hypothesize, that cluster k_4 represents regions where A2 does not necessarily form due to high dissipation rates. Both acetylene and naphthalene have been shown to be sensitive to scalar dissipation rate in laminar [44] and turbulent flames [43, 22]. The presence of A2 within cluster k_4 can be mostly due to transport of soot towards rich side of the flame [22].

We now extend the application of our approach to the SLF dataset for n -heptane/air combustion which is essentially a one-dimensional case corresponding to the analyzed DNS dataset. Attili et al. [23] studied the effect of differential diffusion on accurate representation of the non-premixed sooting flames. The authors made an interesting observation that unity Le number assumption should be used in flamelet models to better represent turbulent combustion. In this section, we apply our local PCA approach to two flamelet models as introduced in §2: one that assumes unity Le and one that assumes mixture-averaged diffusion. Our goal is to show that we arrive at the same observation as Attili et al. [23] but in a purely data-driven way. We use the DNS VQPCA clustering solutions to inform the SLF dataset clustering. Our procedure is to use $q=3$ and use the local eigenvectors from the DNS dataset to partition the SLF data [45]. This gives us a chance to compare local PCA performed on the DNS and the SLF data on equal terms.

In an analogous way as for the DNS dataset, we compute the PDFs in the f -space. Fig. 19 shows the PDFs for VQPCA clustering solutions obtained for $k=4$, $k=6$ and $k=8$ clusters when the unity Le assumption is used. We note that also for this SLF dataset, we can identify a cluster associated with the soot formation processes occurring at $f \approx 0.3$. In each case, we mark that cluster with the red outline. We note that in contrast to what we observed for the DNS dataset, the soot-related cluster spans a wider range in f . In each case in Fig. 19, we observe that the outlined cluster contains region from f slightly lower than 0.3, all the way to $f = 1.0$. We can take a closer look at the VQPCA($k=4, q=3$) solution, in an analogous way to what we have done for the DNS dataset. In Fig. 20, we examine the eigenvector weights associated with the first two local PCs in k_3 from the VQPCA($k=4, q=3$) solution. For both PCs, we mark the five highest weights (in terms of an absolute value) in red. The first PC exhibits highest weights for variables that did not show up in the PC structure for the DNS dataset. However, weights on the second PC are highest for A1C2Hs, A1-, A2-, H2O and C3H3. This PC thus resembles the first PC from k_3 in the DNS dataset (cf. with Fig. 16). The supplementary Figs. S25-S26 demonstrate the eigenvector weights in the outlined clusters from the VQPCA($k=6, q=3$) and the VQPCA($k=8, q=3$) solutions. We note that in each case, we can identify a PC

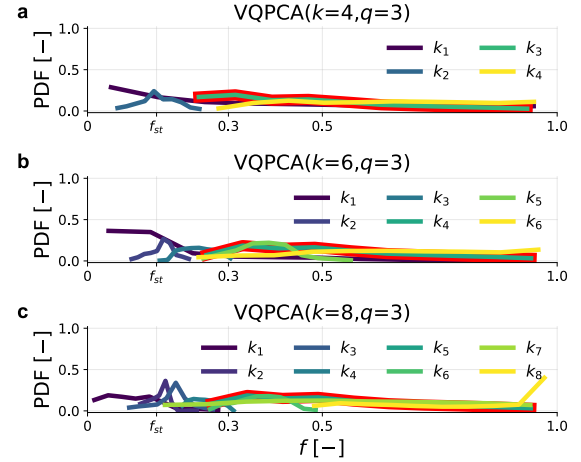
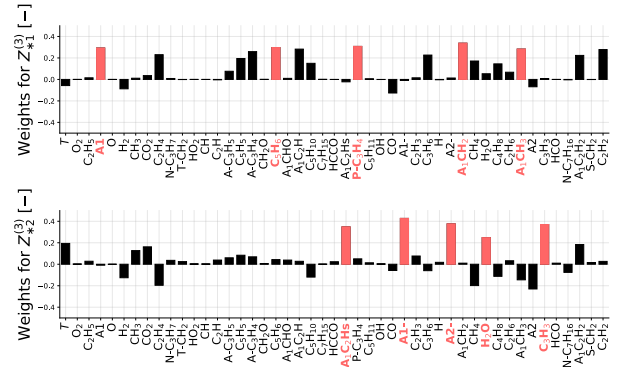


Figure 19: Quantitative physical interpretation of the VQPCA clusters in the SLF dataset in the f -space. **We use the unity Le assumption.** We show the probability density functions (PDFs) in f among clusters obtained using VQPCA with $q=2$ and (a) $k=4$, (b) $k=6$ and (c) $k=8$. With the red outline, we mark the cluster which captures the PAH and sooting processes in each clustering solution. Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. Clusters were initialized by binning f .



processes, which might have been the case if global PCA was applied instead, due to much oversampled region close to $f \approx 0$ in the DNS data.

Fig. 21 shows the analogous PDFs but for the SLF with the mixture-averaged differential diffusion model. We note different probability distributions when non-unity Le assumption is used as compared to the flamelet with unity Le assumption. Primarily, clustering this SLF dataset based on local eigenvectors coming from the DNS data caused some too small clusters to disappear in the VQPCA($k=6, q=3$) and VQPCA($k=8, q=3$) cases. Thus, in Fig. 21b the actual number of clusters found is five, and in Fig. 21c it is six. This behavior suggests that the clustering solution obtained from the DNS data is less adequate to partition the SLF dataset with the non-unity Le assumption. The cluster associated with soot formation for each VQPCA clustering solution is highlighted in red. We observe that in each case, the soot cluster does not encapsulate the region $f \approx 0.3$. The main difference between the unity and non-unity Le assumption from the perspective of the soot cluster is that in the unity Le, this cluster spans the range from $f \approx 0.3$ to $f=1.0$. In the non-unity Le assumption, the soot cluster spans the range from $f \approx 0.4$ to $f=1.0$. This suggests, that with the more accurate differential diffusion model, PAHs show up in the PC structure due to their transport towards the rich side of the flame. Conversely, with the unity Le assumption, PAHs might have showed up in the PC structure both due to their formation at $f \approx 0.3$, and due to transport. In Fig. 22, we present the eigenvector weights associated with the first two local PCs in k_3 from the current VQPCA($k=4, q=3$) solution. In the second PC, we can still identify parameterization based on the PAH species, with the highest weights for $A_1C_2H_5$, $A1-$, $A2-$, C_3H_3 and $A_1C_2H_2$. The supplementary Figs. S25-S26 demonstrate the eigenvector weights in the outline clusters from the VQPCA($k=6, q=3$) and the VQPCA($k=8, q=3$) solutions. We note that in each case, we can identify a PC with high weights for $A_1C_2H_5$, $A1-$, $A2-$, C_3H_3 and $A_1C_2H_2$, and similarly as in the unity Le SLF, this behavior can be seen in the second PC.

5. Conclusions

Dimensionality reduction techniques can be used to obtain effective parameterizations of complex systems with many degrees of freedom. Among many available linear and non-linear reduction techniques, PCA is widely used in various research fields. PCA allows to compute parameterizations directly from the training data. Areas of research strongly populated with large amounts of data can particularly benefit from this data-driven approach. In this work, we present an application of local PCA to simplify the description of high-dimensional reacting flows datasets. We perform a bottom-up exercise in interpreting local PCs. We demonstrate that applying PCA in local clusters (a locally linear approach) we can retrieve the inherent parameters used to generate data from simple combustion models: BS, EQ and HR. The local parameters can be thought of as a more compact representation of the original thermo-chemical variables. For

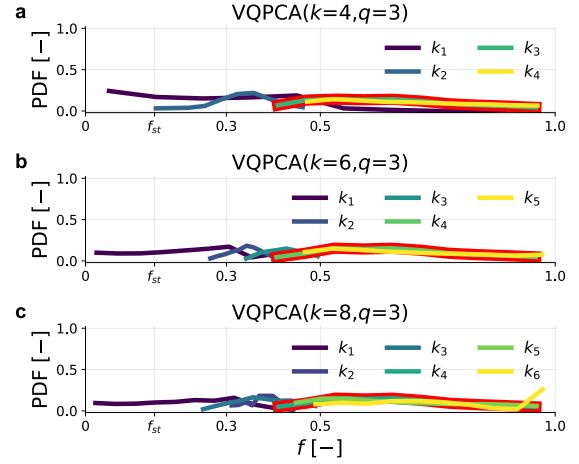
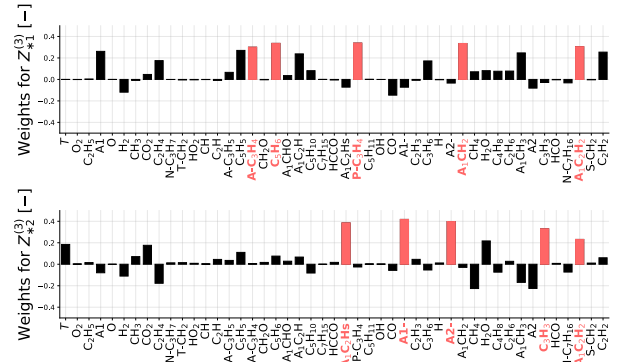


Figure 21: Quantitative physical interpretation of the VQPCA clusters in the SLF dataset in the f -space. **We use the non-unity Le assumption.** We show the probability density functions (PDFs) in f among clusters obtained using VQPCA with $q=2$ and (a) $k=6$ and (b) $k=8$. With the red outline, we mark the cluster which captures the PAH and sooting processes in each clustering solution. Local PCA was performed using Auto scaling on a dataset $\mathbf{X} = [T, Y_i]$. Clusters were initialized by binning f .



that when locally linear manifolds are anticipated in the data, clustering based on linear reconstruction error minimization can be a good choice.

Acknowledgements

The research of the first author is supported by the F.R.S.-FNRS Aspirant Research Fellow grant. GDA has received funding from the Fonds National de la Recherche Scientifique (F.R.S.-FNRS) through a FRIA fellowship. Aspects of this material are based upon work supported by the National Science Foundation under Grant No. 1953350. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program under grant agreement no. 714605.

Conflict of interest

There are no known conflicts of interest associated with this publication.

Author contributions

Conceptualization: K.Z.; G.D.; A.A.; A.C.; J.C.S.; A.P. Data curation: K.Z.; G.D.; A.A. Formal analysis: K.Z.; G.D.; Methodology: A.A.; A.C.; J.C.S.; A.P. Supervision: A.A.; A.C.; J.C.S.; A.P. Visualization: K.Z.; G.D. Writing - original draft: K.Z.; G.D. Writing - review & editing: K.Z.; G.D.; A.A.; A.C.; J.C.S.; A.P. All authors approved the final submitted draft.

Available code

All code used to produce the results can be found in the form of Jupyter notebooks in the public GitHub repository: <https://github.com/kamilazdybal/local-manifold-learning>. The training datasets were generated using **Spitfire** software [46], available at:

<https://github.com/sandialabs/Spitfire>

and **Cantera** software [47]. Local PCA results were produced using **PCAfold** software [48], more information can be found at:

<https://pcafold.readthedocs.io>.

VQPCA implementation available at:

<https://github.com/burn-research/reduced-order-modelling> was used.

Distance correlation was computed using **dcor** Python package available at:

<https://github.com/vnmabus/dcor>.

References

- [1] U. Maas, S. Pope, Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space, *Combustion and Flame* 88 (3) (1992) 239–264. doi:[https://doi.org/10.1016/0010-2180\(92\)90034-M](https://doi.org/10.1016/0010-2180(92)90034-M). URL <https://www.sciencedirect.com/science/article/pii/S001021809290034M>
- [2] J. Van Oijen, L. De Goeij, Modelling of premixed laminar flames using flamelet-generated manifolds, *Combustion Science and Technology* 161 (1) (2000) 113–137. arXiv:<https://doi.org/10.1080/00102200008935814>, doi:10.1080/00102200008935814. URL <https://doi.org/10.1080/00102200008935814>
- [3] O. Gicquel, N. Darabiha, D. Thévenin, Laminar premixed hydrogen/air counterflow flame simulations using flame prolongation of ILDM with differential diffusion, *Proceedings of the Combustion Institute* 28 (2) (2000) 1901–1908. doi:[https://doi.org/10.1016/S0082-0784\(00\)80594-9](https://doi.org/10.1016/S0082-0784(00)80594-9). URL <https://www.sciencedirect.com/science/article/pii/S0082078400805949>
- [4] J. Van Oijen, L. De Goeij, Modelling of premixed counterflow flames using the flamelet-generated manifold method, *Combustion Theory and Modelling* 6 (3) (2002) 463–478. arXiv:<https://doi.org/10.1088/1364-7830/6/3/305>, doi:10.1088/1364-7830/6/3/305. URL <https://doi.org/10.1088/1364-7830/6/3/305>
- [5] J. C. Sutherland, P. J. Smith, J. H. Chen, A quantitative method for a priori evaluation of combustion reaction models, *Combustion Theory and Modelling* 11 (2) (2007) 287–303.
- [6] P. K. Jha, C. P. Groth, Tabulated chemistry approaches for laminar flames: Evaluation of flame-prolongation of ILDM and flamelet methods, *Combustion Theory and Modelling* 16 (1) (2012) 31–57. doi:10.1080/13647830.2011.608856. URL <https://doi.org/10.1080/13647830.2011.608856>
- [7] J. C. Sutherland, A. Parente, Combustion modeling using principal component analysis, *Proceedings of the Combustion Institute* 32 (1) (2009) 1563 – 1570. doi:<https://doi.org/10.1016/j.proci.2008.06.147>. URL <http://www.sciencedirect.com/science/article/pii/S1540748908002630>
- [8] Y. Yang, S. B. Pope, J. H. Chen, Empirical low-dimensional manifolds in composition space, *Combustion and Flame* 160 (10) (2013) 1967 – 1980. doi:<https://doi.org/10.1016/j.combustflame.2013.04.006>. URL <http://www.sciencedirect.com/science/article/pii/S0010218013001508>
- [9] A. Coussement, O. Gicquel, A. Parente, MG-local-PCA method for reduced order combustion modeling, *Proceedings of the Combustion Institute* 34 (1) (2013) 1117–1123. doi:<https://doi.org/10.1016/j.proci.2012.05.073>. URL <https://www.sciencedirect.com/science/article/pii/S1540748912000740>
- [10] A. Najafi-Yazdi, B. Cuenot, L. Mongeau, Systematic definition of progress variables and intrinsically low-dimensional, flamelet generated manifolds for chemistry tabulation, *Combustion and Flame* 159 (3) (2012) 1197–1204. doi:<https://doi.org/10.1016/j.combustflame.2011.10.003>. URL <https://www.sciencedirect.com/science/article/pii/S0010218011003087>
- [11] A. Parente, J. Sutherland, B. Dally, L. Tognotti, P. Smith, Investigation of the MILD combustion regime via principal component analysis, *Proceedings of the Combustion Institute* 33 (2) (2011) 3333 – 3341. doi:<https://doi.org/10.1016/j.proci.2010.05.108>. URL <http://www.sciencedirect.com/science/article/pii/S1540748910002725>
- [12] B. J. Isaac, J. N. Thornock, J. Sutherland, P. J. Smith, A. Parente, Advanced regression methods for combustion modelling using principal components, *Combustion and Flame* 162 (6) (2015) 2592–2601. doi:<https://doi.org/10.1016/j.combustflame.2015.03.008>. URL <https://www.sciencedirect.com/science/article/pii/S0010218015000838>
- [13] R. Ranade, T. Echekki, A framework for data-based turbulent combustion closure: A priori validation, *Combustion and Flame* 206 (2019) 490 – 505. doi:<https://doi.org/10.1016/j.combustflame.2019.05.028>. URL <http://www.sciencedirect.com/science/article/pii/S001021801930241X>
- [14] M. R. Malik, P. Obando Vega, A. Coussement, A. Parente, Combustion

- tion modeling using principal component analysis: A posteriori validation on Sandia flames D, E and F, *Proceedings of the Combustion Institute* (2020). doi:<https://doi.org/10.1016/j.proci.2020.07.014>.
URL <http://www.sciencedirect.com/science/article/pii/S1540748920301966>
- [15] G. D'Alessio, A. Cuoci, A. Parente, Feature extraction and artificial neural networks for the on-the-fly classification of high-dimensional thermochemical spaces in adaptive-chemistry simulations, *Data-Centric Engineering* 2 (2021).
- [16] R. Bilger, S. Starner, R. Kee, On reduced mechanisms for methane-air combustion in nonpremixed flames, *Combustion and Flame* 80 (1990) 135–149.
- [17] E. Fooladgar, C. Duwig, A new post-processing technique for analyzing high-dimensional combustion data, *Combustion and Flame* 191 (2018) 226 – 238. doi:<https://doi.org/10.1016/j.combustflame.2018.01.014>.
URL <http://www.sciencedirect.com/science/article/pii/S001021801830018X>
- [18] G. D'Alessio, A. Attili, A. Cuoci, H. Pitsch, A. Parente, Analysis of turbulent reacting jets via principal component analysis, in: *Data Analysis for Direct Numerical Simulations of Turbulent Combustion*, Springer, 2020, pp. 233–251.
- [19] G. D'Alessio, A. Cuoci, G. Aversano, M. Bracconi, A. Stagni, A. Parente, Impact of the partitioning method on multidimensional adaptive-chemistry simulations, *Energies* 13 (10) (2020) 2567.
- [20] Z. Li, S. Tomasch, Z. X. Chen, A. Parente, I. S. Ertesvåg, N. Swaminathan, Study of MILD combustion using LES and advanced analysis tools, *Proceedings of the Combustion Institute* 38 (4) (2021) 5423–5432. doi:<https://doi.org/10.1016/j.proci.2020.06.298>.
URL <https://www.sciencedirect.com/science/article/pii/S1540748920303904>
- [21] N. Kambhatla, T. K. Leen, Dimension reduction by local principal component analysis, *Neural Computation* 9 (7) (1997) 1493–1516. arXiv:<https://doi.org/10.1162/neco.1997.9.7.1493>, doi:10.1162/neco.1997.9.7.1493.
URL <https://doi.org/10.1162/neco.1997.9.7.1493>
- [22] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, Formation, growth, and transport of soot in a three-dimensional turbulent non-premixed jet flame, *Combustion and Flame* 161 (7) (2014) 1849–1865. doi:<https://doi.org/10.1016/j.combustflame.2014.01.008>.
URL <https://www.sciencedirect.com/science/article/pii/S0010218014000133>
- [23] A. Attili, F. Bisetti, M. E. Mueller, H. Pitsch, Effects of non-unitary lewis number of gas-phase species in turbulent nonpremixed sooting flames, *Combustion and Flame* 166 (2016) 192–202. doi:<https://doi.org/10.1016/j.combustflame.2016.01.018>.
URL <https://www.sciencedirect.com/science/article/pii/S0010218016000316>
- [24] G. D'Alessio, A. Attili, A. Cuoci, H. Pitsch, A. Parente, Unsupervised data analysis of direct numerical simulation of a turbulent flame via local principal component analysis and procustes analysis, in: *International Workshop on Soft Computing Models in Industrial and Environmental Applications*, Springer, 2020, pp. 460–469.
- [25] G. P. Smith, D. M. Golden, M. Frenklach, N. W. Moriarty, B. Eiteeneer, M. Goldenberg, C. T. Bowman, R. K. Hanson, S. Song, W. Gardiner Jr, et al., *Gri-mech 3.0*, 1999 (2011).
URL http://www.me.berkeley.edu/gri_mech
- [26] E. R. Hawkes, R. Sankaran, J. C. Sutherland, J. H. Chen, Scalar mixing in direct numerical simulations of temporally evolving plane jet flames with skeletal CO/H₂ kinetics, *Proc. Combust. Inst.* 31 (1) (2007) 1633–1640.
- [27] F. Bisetti, G. Blanquart, M. E. Mueller, H. Pitsch, On the formation and early evolution of soot in turbulent nonpremixed flames, *Combustion and Flame* 159 (1) (2012) 317–335. doi:<https://doi.org/10.1016/j.combustflame.2011.05.021>.
URL <https://www.sciencedirect.com/science/article/pii/S0010218011001672>
- [28] A. Parente, J. C. Sutherland, Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity, *Combustion and Flame* 160 (2) (2013) 340 – 350. doi:<https://doi.org/10.1016/j.combustflame.2012.09.016>.
URL <http://www.sciencedirect.com/science/article/pii/S0010218012002775>
- [29] G. J. Székely, M. L. Rizzo, N. K. Bakirov, et al., Measuring and testing dependence by correlation of distances, *The annals of statistics* 35 (6) (2007) 2769–2794.
- [30] D. Edelmann, T. F. Móri, G. J. Székely, On relationships between the Pearson and the distance correlation coefficients, *Statistics & Probability Letters* 169 (2021) 108960. doi:<https://doi.org/10.1016/j.spl.2020.108960>.
URL <https://www.sciencedirect.com/science/article/pii/S0167715220302637>
- [31] A. Parente, J. Sutherland, L. Tognotti, P. Smith, Identification of low-dimensional manifolds in turbulent flames, *Proceedings of the Combustion Institute* 32 (1) (2009) 1579 – 1586. doi:<https://doi.org/10.1016/j.proci.2008.06.177>.
URL <http://www.sciencedirect.com/science/article/pii/S154074890800299X>
- [32] K. Zdybał, G. D'Alessio, G. Aversano, M. R. Malik, A. Coussement, J. C. Sutherland, A. Parente, Advancing reactive flow simulations with data-driven models, in: M. A. Mendez, A. Ianiro, B. R. Noack, S. L. Brunton (Eds.), *Data-Driven Fluid Mechanics: Combining First Principles and Machine Learning*, Cambridge University Press, 2022, Ch. 15.
- [33] M. Del Giudice, Effective dimensionality: A tutorial, *Multivariate behavioral research* 56 (3) (2021) 527–542.
- [34] Y.-C. Chen, R. W. Bilger, Experimental investigation of three-dimensional flame-front structure in premixed turbulent combustion: Ii. lean hydrogen/air bunsen flames, *Combustion and Flame* 138 (1) (2004) 155–174. doi:<https://doi.org/10.1016/j.combustflame.2004.04.009>.
URL <https://www.sciencedirect.com/science/article/pii/S0010218004000987>
- [35] A. Fiolitakis, P. Gerlinger, B. Noll, M. Aigner, W. Krebs, B. Wegner, A novel progress variable approach for predicting no in laminar hydrogen flames, in: *48th AIAA Aerospace Sciences Meeting Including the New Horizons Forum and Aerospace Exposition*, 2010, p. 608.
- [36] F. Ladeinde, W. Li, Differential turbulent supersonic combustion of hydrogen, methane, and ethylene, without assisted ignition, *AIAA Journal* 56 (12) (2018) 4870–4883.
- [37] A. Vasavan, P. de Goey, J. van Oijen, A novel method to automate fgm progress variable with application to igniting combustion systems, *Combustion Theory and Modelling* 24 (2) (2020) 221–244. arXiv:<https://doi.org/10.1080/13647830.2019.1673902>, doi:10.1080/13647830.2019.1673902.
URL <https://doi.org/10.1080/13647830.2019.1673902>
- [38] R. S. Barlow, G. Magnotti, H. C. Cutcher, A. R. Masri, On defining progress variable for raman/rayleigh experiments in partially-premixed methane flames, *Combustion and Flame* 179 (2017) 117–129.
- [39] W. T. Chung, A. A. Mishra, N. Perakis, M. Ihme, Data-assisted combustion simulations with dynamic submodel assignment using random forests, *Combustion and Flame* 227 (2021) 172–185.
- [40] D. K. Dalakoti, A. Wehrfritz, B. Savard, M. S. Day, J. B. Bell, E. R. Hawkes, An a priori evaluation of a principal component and artificial neural network based combustion model in diesel engine conditions, *Proceedings of the Combustion Institute* 38 (2) (2021) 2701–2709.
- [41] J. C. Gower, Generalized procustes analysis, *Psychometrika* 40 (1) (1975) 33–51.
- [42] D. G. Kendall, A survey of the statistical theory of shape, *Statistical Science* 4 (2) (1989) 87–99.
- [43] F. Bisetti, G. Blanquart, M. E. Mueller, H. Pitsch, On the formation and early evolution of soot in turbulent nonpremixed flames, *Combustion and Flame* 159 (1) (2012) 317–335.
- [44] H. Böhm, K. Kohse-Höinghaus, F. Lacas, C. Rolon, N. Darabiha, S. Candel, On pah formation in strained counterflow diffusion flames,

- Combust. Flame 124 (1-2) (2001) 127–136.
- [45] G. D'Alessio, A. Parente, A. Stagni, A. Cuoci, Adaptive chemistry via pre-partitioning of composition space and mechanism reduction, *Combustion and Flame* 211 (2020) 68–82.
 - [46] M. A. Hansen, E. Armstrong, J. C. Sutherland, J. McConnell, J. C. Hewson, R. Knaus, *Spitfire* (2022).
URL <https://github.com/sandialabs/Spitfire>
 - [47] D. G. Goodwin, R. L. Speth, H. K. Moffat, B. W. Weber, Cantera: An object-oriented software toolkit for chemical kinetics, thermodynamics, and transport processes, <https://www.cantera.org>, version 2.5.1 (2021). doi:10.5281/zenodo.4527812.
 - [48] K. Zdybał, E. Armstrong, A. Parente, J. C. Sutherland, PCAfold: Python software to generate, analyze and improve PCA-derived low-dimensional manifolds, *SoftwareX* 12 (2020) 100630. doi:<https://doi.org/10.1016/j.softx.2020.100630>.
URL <http://www.sciencedirect.com/science/article/pii/S2352711020303435>