Manifold-informed state vector subset for reduced-order modeling

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

[a] *Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Brussels, Belgium*
[b] *Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*
[c] *Department of Chemical Engineering, University of Utah, Salt Lake City, Utah, USA*

Abstract

Reduced-order models (ROMs) for turbulent combustion rely on identifying a small number of parameters that can effectively describe the complexity of reacting flows. With the advent of data-driven approaches, ROMs can be trained on datasets representing the thermo-chemical state-space in simple reacting systems. For low-Mach flows, the full state vector that serves as a training dataset is typically composed of temperature and chemical composition. The dataset is projected onto a lower-dimensional basis and the evolution of the complex system is tracked on a lower-dimensional manifold. This approach allows for substantial reduction of the number of transport equations to solve in combustion simulations, but the quality of the manifold topology is a decisive aspect in successful modeling. To mitigate manifold challenges, several authors advocate reducing the state vector to only a subset of major variables when training ROMs. However, this reduction is often done *ad hoc* and without giving detailed insights into the effect of removing certain variables on the resulting low-dimensional data projection. In this work, we present a quantitative manifold-informed method for selecting the subset of state variables that minimizes unwanted behaviors in manifold topologies. While many authors in the past have focused on selecting major species, we show that a mixture of major and minor species can be beneficial to improving the quality of low-dimensional data representations. The desired effects include reducing non-uniqueness and spatial gradients in the dependent variable space. Finally, we demonstrate improvements in regressibility of manifolds built from the optimal state vector subset as opposed to the full state vector.

Keywords: reduced-order modeling; low-dimensional manifold; dimensionality reduction; principal component analysis; non-linear regression

1. Introduction

Parameterization approaches can be used to compress description of complex combustion systems with many degrees of freedom. Numerous physics-based parameterization techniques can be found in the literature [1–3]. An alternative to the physics-motivated parameterizations is a data-driven approach, where low-dimensional manifolds (LDMs) are constructed from the training data representing a sufficiently wide range of the thermo-chemical state-space [4, 5]. Linear [4, 6–9] and nonlinear [10, 11] dimensionality reduction techniques have been used in the past to find lower-dimensional basis to represent combustion datasets with fewer parameters.

The success of a given parameterization technique then depends on the quality of LDM topology. Characteristics of a good parameterization include soft gradients, as well as uniqueness in dependent variables [12, 13]. The question of the parameterization quality is of particular importance in data-driven model reduction. Notably, non-uniqueness can be introduced during low-dimensional data projection, resulting in ambiguity in dependent variable values. If the new manifold parameters are later used as regressors, regression can struggle in regions of non-uniqueness.

Problems with ill-behaved manifolds can be alleviated through appropriate training data preprocessing. The most straightforward strategy is data scaling [8, 9, 14–16]. Other authors have tackled manifold challenges by training combustion models on a subset of the original thermo-chemical state-space variables [9, 10, 16–21]. A closer look at the variables typically selected in the literature suggests that authors create subsets in qualitative ways, taking fuel and oxidizer components and complete combustion products, with [10, 17–19, 21] or without [9, 16, 17, 20] temperature, and rarely including minor species [19, 20]. Such variable selections are often done *ad hoc*, without detailed justification for selecting some variables and discarding another. A notable exception is the work by Hiremath et al. [22] where an algorithm for species selection was developed based on minimizing the reconstruction error from dimensionality reduction. However, the greedy algorithm proposed in [22] does not take into account LDM topology. To the authors' knowledge, no comprehensive assessment concerning the selection and quality of manifolds has been given, which can have significance when nonlinear regression on manifolds is incorporated into the reduced-order modeling (ROM) workflow.

Here, we propose a manifold-informed variable selection strategy to define a meaningful subset of the original state variables. Using the recently proposed technique to characterize manifold quality [23], we build an iterative variable selection algorithm that pays attention to two topological aspects of LDMs: feature sizes and uniqueness. The state vector subset is thus optimized to result in an improved LDM topology once the dataset is projected onto a lower-dimensional basis. While in this work we use principal component analysis (PCA) for generating LDMs, our approach is not limited to any particular dimensionality reduction technique. The variable selection algorithm proposed can also work across different target manifold dimensionalities. Moreover, the manifold can be optimized towards efficient modeling of an arbitrary set of target dependent variables. With the appropriately selected state vector subset, we demonstrate an improved regressibility of manifolds using kernel regression.

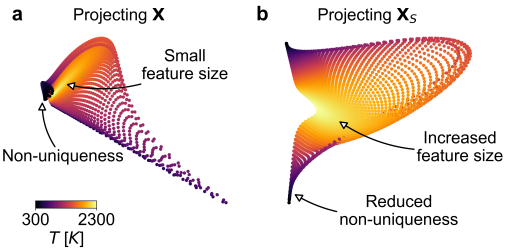

Fig. 1: Example two-dimensional PCA projections of (a) the full state vector, $\mathbf{X}$, and (b) the optimal subset of that state vector, $\mathbf{X}_S$, from a syngas/air flamelet dataset with Level scaling. Projections are colored by the temperature.

2. Data-driven approach for model reduction

A data-driven approach to obtaining LDM parameterizations relies on the availability of training data. In particular, simulating simple canonical systems allows to obtain the training dataset, which we also refer to as the thermo-chemical state vector, $\mathbf{X} \in \mathbb{R}^{N \times Q}$, where N is the number of observations and Q is the number of state variables. In the present work, we generate training data from the steady laminar flamelet model for various fuels: hydrogen [24], syngas [25] and ethylene [26]. We focus on the full state vector defined as $\mathbf{X} = \left[T, Y_1, Y_2, \ldots, Y_{n_s-1}\right]$, where T is the temperature, Y_i is the mass fraction of species i and n_s is the number of species in the chemical mechanism. For clarity, we will refer to the mass fraction of a given species by its chemical formula. Other definitions for the state vector can be adopted which follow directly from the governing equations that describe the simulated system [27].

The goal of model reduction is to decrease the number of governing equations needed to solve in a simulation. The evolution of the state vector can be described by the general transport equation written in the matrix form:

$$\frac{D\mathbf{X}^\top}{Dt} = -\nabla \cdot \mathbf{J}^\top + \mathbf{S}^\top, \tag{1}$$

where $\mathbf{J}$ is the matrix of diffusive fluxes and $\mathbf{S}$ is the source terms matrix. The ROM approach transforms the state vector, $\mathbf{X}$, to a lower-dimensional basis defined by the matrix of q modes, $\mathbf{A} \in \mathbb{R}^{Q \times q}$. In PCA, the projection onto the new basis is performed

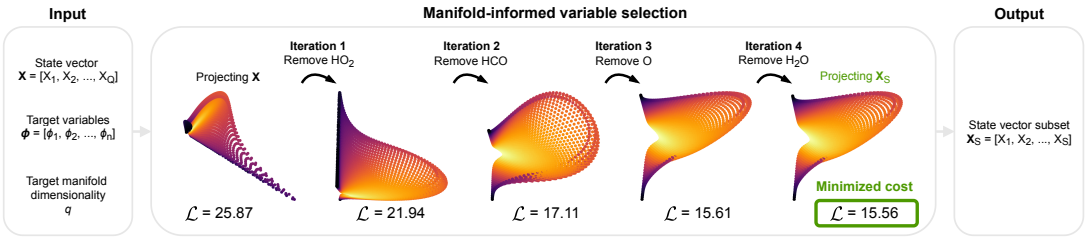

Fig. 2: Schematic illustration of the manifold-informed variable selection algorithm proposed in this work. The algorithm selects the subset of the thermo-chemical state vector, optimized with respect to an accurate parameterization of the user-defined target dependent variables, $\boldsymbol{\phi}$, using q manifold parameters. At each iteration, cumulative cost over all n target variables is computed as $\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_{\phi_i}$ and the variable that decreases this cost the most is removed. The output is the optimal state vector subset corresponding to the minimized cost from all iterations (marked in green). The changing LDM topologies in the middle frame are shown for a syngas/air flamelet dataset with Level scaling.

as $\mathbf{Z} = \mathbf{X}\mathbf{A}$. The principal components (PCs), $\mathbf{Z}$, define the parameterization of a q-dimensional manifold. As long as $\mathbf{A}$ is constant in time and space, Eq. (1) can be replaced by a reduced set of transport equations for the PCs,

$$\frac{D\mathbf{Z}^\top}{Dt} = -\nabla \cdot \mathbf{J_Z}^\top + \mathbf{S_Z}^\top, \tag{2}$$

where $\mathbf{J_Z}$ is the matrix of the projected diffusive fluxes and $\mathbf{S_Z}$ is the matrix of the projected source terms (PC source terms). We will refer to the i^{th} PC source term as $S_{Z,i}$. There are two more ingredients of the reduced model. First, the PC source terms need to be parameterized by the new manifold parameters [4, 28]. Second, it has been a frequent approach in the literature to obtain nonlinear mappings between the manifold parameters and the thermo-chemical state variables that are the desired output of a simulation [8, 9]. To tackle both aspects, we can build nonlinear regression models to effectively approximate any dependent variable, ϕ, as $\phi \approx \mathscr{F}(\mathbf{Z})$, where $\mathscr{F}$ is a regression function. To date, nonlinear regression techniques such as artificial neural networks (ANNs) [10, 29], multivariate adaptive regression splines (MARS) [7] or Gaussian process regression (GPR) [8, 9, 16] were used to obtain $\mathscr{F}$ in the context of reacting flow simulations.

ROMs can also be trained on a meaningful subset of the original state variables. Evidence from previous research suggests that variable selection can have beneficial effects on the LDM topology [9, 10, 16–21]. For the original state vector $\mathbf{X}$ with Q state variables, we define $\mathbf{X}_S \in \mathbb{R}^{N \times S}$ as the state vector subset, where $S < Q$. $\mathbf{X}_S$ is generated through discarding certain variables (columns) from $\mathbf{X}$. We can then compute the PCs by projecting $\mathbf{X}_S$ instead of $\mathbf{X}$ onto the PCA basis. Figure 1 demonstrates how significantly LDM topologies can change through data preprocessing such as variable selection. Figure 1a shows an example two-dimensional PCA projection of the full state vector, $\mathbf{X}$, from a syngas/air combustion dataset. Figure 1b shows a projection when the optimal subset of state variables, $\mathbf{X}_S$, was used for PCA. Both manifolds are colored by the temperature. In addition, Level scaling is applied on $\mathbf{X}$ and $\mathbf{X}_S$ (various scaling criteria explored in this work are summarized in the supplementary Table S1). Significant non-uniqueness is introduced on the manifold in Fig. 1a through curling low-temperature observations into a small region and overlapping them with higher-temperature observations. Different feature sizes are observed on both manifolds as well. For instance, the size of the high-temperature region is increased on the manifold in Fig. 1b as compared to the manifold in Fig. 1a. These topological considerations indicate that modeling a dependent variable such as temperature on the manifold in Fig. 1a can be challenging due to (1) non-uniqueness and (2) steep gradients of high-temperature regions over small manifold length scales. The example from Fig. 1 demonstrates conceptually that adequate choice of state variables for dimensionality reduction can be beneficial to improving the LDM topology.

3. Manifold-informed subset of state variables

Preprocessing the training dataset such as scaling or variable selection, prior to applying dimensionality reduction, can influence the LDM topology significantly. In addition, changing the target manifold dimensionality, q, can alter the verdict of what the best preprocessing strategy is. Thus, there is a need to automate the preprocessing selection process to produce optimal LDM topology. Furthermore, when regression of a set of dependent variables is employed on a manifold, the reasonable strategy is to "tune" the manifold towards well representing those variables specifically.

Instead of relying on knowledge-based selection of state variables, we propose an approach of a manifold-informed variable selection which can be performed *a priori* at the modeling stage. We develop a backward variable elimination algorithm in which we iteratively remove variables that decrease some cost function, $\mathcal{L}$, the most. Figure 2 illustrates schematically the proposed algorithm. The input to the algorithm is the full state vector, $\mathbf{X}$,

the user-selected set of n target dependent variables that should be assessed on the manifold, $\boldsymbol{\phi} = [\phi_1, \phi_2, \ldots, \phi_n]$, and the target manifold dimensionality, q. The output of the algorithm is the state vector subset, $\mathbf{X}_S$. The middle frame in Fig. 2 illustrates changing LDM topologies when iteratively eliminating variables from the state vector using the proposed algorithm. The results are shown for a syngas/air flamelet dataset with Level scaling, where the full state vector is $\mathbf{X} = [T$, H_2, O_2, O, OH, H_2O, H, HO_2, CO, CO_2, HCO$]$. The cost function used here is based on the normalized variance derivative, $\hat{\mathcal{D}}(\sigma)$, metric [23] and considers the LDM topology as well as the relevant set of target dependent variables that should be well-defined on the resulting LDM. For the i^{th} dependent variable, $\hat{\mathcal{D}}_i(\sigma)$ quantifies the information content at various manifold length scales, $\sigma = \langle \sigma_{min}; \sigma_{max} \rangle$. The cost is defined for the i^{th} dependent variable, ϕ_i, as

$$\mathcal{L}_{\phi_i} = \int_{\widetilde{\sigma}_{min}}^{\widetilde{\sigma}_{max}} P_i(\sigma, \sigma_{p,i}) \cdot \hat{\mathcal{D}}_i(\sigma) d\widetilde{\sigma}, \qquad (3)$$

where $\sigma = \sigma_{p,i}$ represents the largest feature size in ϕ_i on a manifold and $P_i(\sigma, \sigma_{p,i})$ is the penalty function defined as

$$P_i(\sigma, \sigma_{p,i}) = \left|\widetilde{\sigma} - \widetilde{\sigma}_{p,i}\right| + \frac{1}{||\widetilde{\sigma}_{p,i}||}, \qquad (4)$$

where tilde denotes a $\log_{10}$-transformed quantity. The quantity $||\widetilde{\sigma}_{p,i}|| \in \langle 0, 1 \rangle$ is a normalized $\sigma_{p,i}$ value, where the normalization is such that $\min(||\widetilde{\sigma}||) = 0$ and $\max(||\widetilde{\sigma}||) = 1$. This normalization introduces a gentle penalty for the size of the largest feature. A larger value for $\sigma_{p,i}$ is desired as it can indicate that features on a manifold will be well resolved over regressible length scales. Large values of $\hat{\mathcal{D}}_i(\sigma)$ at small σ can indicate non-uniqueness in variable ϕ_i and we penalize those in particular with the $\left|\widetilde{\sigma} - \widetilde{\sigma}_{p,i}\right|$ term in the penalty function. For a given LDM topology, we compute the overall cost from a set of n target dependent variables, $\boldsymbol{\phi}$, as a cumulative sum $\mathcal{L} = \sum_{i=1}^{n} \mathcal{L}_{\phi_i}$. This yields a single cost value characterizing the LDM topology. LDM topology can be optimized with respect to the most important variables such as temperature, fuel and oxidizer components, main products and important radicals. Dependent variables that are functions of the original state variables, such as the PC source terms, $\mathbf{S_Z}$, can be selected as well. Table 1 summarizes the sets $\boldsymbol{\phi}$ used in this work for each fuel. $\widetilde{\mathbf{S}}_\mathbf{Z}$ are the symlog-transformed PC source terms included in the set of target variables to force the optimization to also represent small values of $\mathbf{S_Z}$ well on a manifold. In Fig. 2, the cumulative costs are reported for each iteration. The optimal subset corresponds to the minimum $\mathcal{L}$ from all iterations (marked in green). Notably, in this example, the cost associated with projecting the full state vector is significantly higher ($\mathcal{L} = 25.9$) than the cost associated with projecting the optimal subset ($\mathcal{L} = 15.6$). Minimizing $\mathcal{L}$ minimizes unwanted behaviors on manifolds such as non-uniqueness and small feature sizes (compare with Fig. 1). The proposed algorithm is available in the `PCAfold` Python library [30]. The supplementary material includes an example code snippet for running the algorithm on a combustion dataset.

Table 1: The set of target dependent variables, $\boldsymbol{\phi}$, selected in this work for each fuel.

Fuel	$\boldsymbol{\phi}$
Hydrogen	$\mathbf{S_Z}$, $\widetilde{\mathbf{S}}_\mathbf{Z}$, T, O_2, OH, H_2O, H_2
Syngas	$\mathbf{S_Z}$, $\widetilde{\mathbf{S}}_\mathbf{Z}$, T, O_2, OH, H_2O, CO, CO_2, H_2
Ethylene	$\mathbf{S_Z}$, $\widetilde{\mathbf{S}}_\mathbf{Z}$, T, O_2, OH, H_2O, CO, CO_2, C_2H_4

4. Results and discussion

A few important questions can be posed to motivate subsetting the state vector. Is there a subset of the state variables that optimally represents the LDM topology? How is the optimal topology affected by the choice of target variables that we wish to represent on the LDM? How do we identify the best data scaling to represent the target variables on the LDM? The metric proposed in Eq. (3) provides a quantitative measure of LDM topology quality for a given set of target variables to allow us to answer these questions.

4.1. Choice of the target dependent variables, $\boldsymbol{\phi}$

Different dependent variables are affected differently by the LDM topology. For instance, overlapping observations on a manifold are only problematic in modeling when there is a variation in a dependent variable's values across these observations. Figure 3 demonstrates this visually on an example two-dimensional PCA projection which is severely folded over itself. The projection is computed from a syngas/air flamelet dataset with $\langle -1, 1 \rangle$ scaling and is colored by the temperature in Fig. 3a and by the O_2 mass fraction in Fig. 3b. The overlap introduced on the manifold does not impact the temperature variable greatly, as there is a similar temperature value for observations that are directly one on top of the other. The cost associated with the temperature variable for this projection is $\mathcal{L}_T = 1.3$. At the same time, the O_2 mass fraction exhibits large variation in values across the overlapping observations. In Fig. 3b we see observations corresponding to nearly zero mass fractions being projected directly below observations with high O_2 mass fractions. As a result, the cost computed for the O_2 variable is higher, $\mathcal{L}_{O_2} = 1.9$. This example indicates the impact of target variables selection. First, we note that some variables (like the temperature variable in the example shown in Fig. 3a) might not be effective at detecting non-uniqueness. Second, the final selection of the state vector subset will depend on which target variables should be well represented on an optimized manifold.

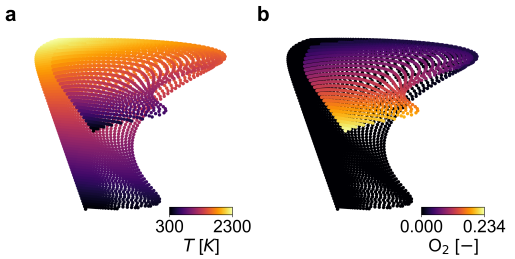

Fig. 3: Two-dimensional PCA projection of a syngas/air flamelet dataset with $\langle -1, 1\rangle$ scaling colored by (a) the temperature, and (b) the O_2 mass fraction.

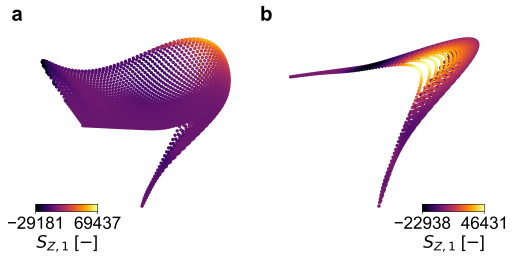

Fig. 4: Two-dimensional PCA projections of a syngas/air flamelet dataset with Auto scaling resulting from taking (a) the optimal state vector subset, $\mathbf{X}_S = [T$, O_2, OH, H, CO, HCO], and (b) the same subset but with the minor species, H and HCO, mass fractions removed. Projections are colored by the corresponding first PC source term, $S_{Z,1}$.

4.2. State vector subset selected by the proposed algorithm

Table 2 collects results of running the proposed variable selection algorithm on datasets with various fuels with the target manifold dimensionality $q = 3$. The target variables are selected as per Table 1. We report cumulative costs associated with generating the LDM from the full state vector, $\mathcal{L}^{(F)}$, and costs associated with selecting an appropriate state vector subset, $\mathcal{L}^{(S)}$. The best scaling option (resulting in the lowest cost) is reported for each data preprocessing case. The last column shows the state variables selected. In all cases, selecting an appropriate state vector subset decreased the cumulative cost with respect to taking the full state vector. The results in Table 2 show that in certain cases, it can be beneficial to remove the temperature variable from the state vector. This is consistent with the results reported in the existing literature [9, 16, 17, 20]. We further observe that the chemical species selected by the proposed algorithm are a mixture of major and minor species. This result differs from the selections typically made in the literature [9, 10, 16–18, 21]. To justify why our algorithm might select minor species, in Fig. 4, we show two-dimensional PCA projections of a subset syngas/air flamelet dataset with Auto scaling. Both projections are colored by the corresponding first PC source term, $S_{Z,1}$. The manifold seen in Fig. 4a results from taking the optimal subset of the state vector, in this case $\mathbf{X}_S = [T$, O_2, OH, H, CO, HCO]. The cumulative cost over target dependent variables (as per Table 1) is $\mathcal{L} = 14.1$. In Fig. 4b, we show a two-dimensional projection resulting from the same subset but with the minor species, H and HCO, mass fractions removed. We note that the cumulative cost now increased to $\mathcal{L} = 19.8$ and the LDM topology changed significantly. The more severe visible overlap on a manifold in Fig. 4b affects the PC source terms in particular. The individual costs increased between manifolds in Figs. 4a-b from $\mathcal{L}_{S_{Z,1}} = 1.5$ to 4.2 for the first PC source term and from $\mathcal{L}_{S_{Z,2}} = 1.2$ to 4.2 for the second PC source term.

4.3. Effect of variable selection versus data scaling

While the impact of combustion data scaling on the manifold sensitivity has been studied in [14], we examine how variable selection in combination with scaling affects LDM topologies. In Fig. 5, we compare cumulative costs, $\mathcal{L}$, resulting from only scaling the full state vector (black circles) with scaling and optimal variable selection (red triangles) for a syngas/air flamelet dataset. The results are reported for various scaling options (see Table S1). The resulting three-dimensional LDM topologies colored by the temperature are visualized at the top of the figure. The best topologies in both cases (corresponding to the lowest $\mathcal{L}$) are highlighted with thick axes. The highest cost of only scaling the full state vector (black circles) happens for $\langle 0, 1\rangle$ scaling. This high cost can be understood by looking at the visualized projection. The $\langle 0, 1\rangle$ manifold exhibits sharp changes in topology and crossing observations which introduce overlap. On the other hand, a good performance corresponding to the $\langle -1; 1\rangle$ scaling can be due to the effect of spreading the low-temperature observations over wider regions, thus increasing fea-

Table 2: Thermo-chemical state-space variables selected by the manifold-informed variable selection algorithm for the target manifold dimensionality $q = 3$, using the target variables, $\boldsymbol{\phi}$, as per Table 1. We report the cumulative cost corresponding to the manifold obtained from the full state vector, $\mathcal{L}^{(F)}$, and the optimal state vector subset, $\mathcal{L}^{(S)}$. The results are presented for flamelet datasets with various fuels and for the best scaling option (reported) for each data preprocessing case.

Fuel	$\mathcal{L}^{(F)}$, scaling	$\mathcal{L}^{(S)}$, scaling	Selected state variables
Hydrogen	12.5, Max	11.0, Auto	T, H, H_2, O, OH, O_2
Syngas	14.6, $\langle -1, 1\rangle$	13.1, Range	O, OH, H, CO, CO_2
Ethylene	16.2, Range	13.6, Auto	T, H_2, H, O_2, OH, H_2O, CH_3, CO_2, HCO, C_2H_2, C_2H_4, CH_2CO

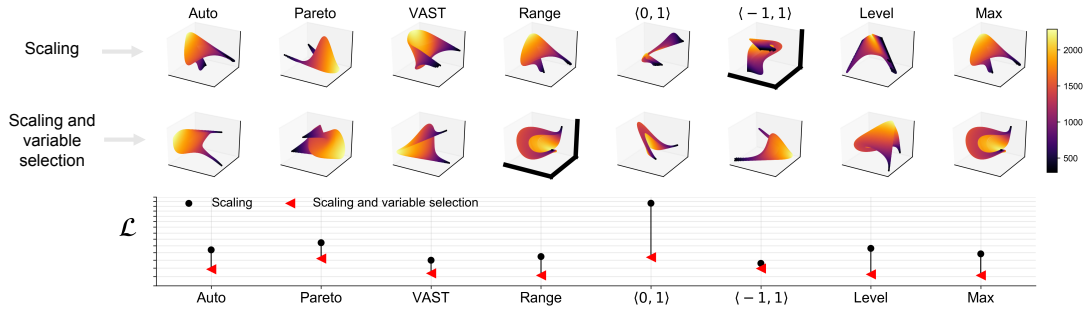

Fig. 5: Cumulative costs, $\mathcal{L}$, for three-dimensional PCA projections of a syngas/air flamelet dataset using different scaling options. Projections are generated from the full state vector (black circles) and from an optimal subset of the state vector (red triangles). The visualized projections colored by the temperature correspond to scaling only (top row) and to scaling with variable selection (bottom row). Projections corresponding to the lowest $\mathcal{L}$ in both cases are highlighted with thick axes.

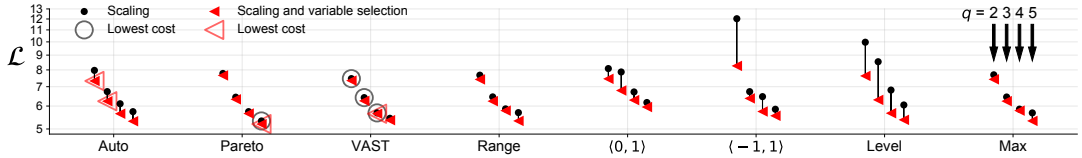

Fig. 6: Cumulative costs, $\mathcal{L}$, for PCA projections of a syngas/air flamelet dataset with $q = 2, 3, 4, 5$ using different scaling options. Projections are generated from the full state vector (black circles) and from an optimal subset of the state vector (red triangles). Markers with an outline highlight the lowest $\mathcal{L}$ for both preprocessing strategies and for each dimensionality q.

ture sizes and reducing non-uniqueness on the resulting projection. We observe that costs measured using the cumulative sum generally drop as we select an appropriate state vector subset for all scaling options explored. The same observation holds for hydrogen/air and ethylene/air flamelets (see supplementary Figs. S1-S2). This result indicates that it can be beneficial to combine scaling with variable selection as a data preprocessing strategy. Moreover, larger differences in cost values between different scaling options are observed for scaling without (black circles) than for scaling with (red triangles) variable selection. This suggests that the optimal state vector subsets yield more homogenized LDM topologies across changing data scaling option.

4.4. Choice of the manifold dimensionality, q

The cost function can further help guide the choice of the target manifold dimensionality, q. Figure 6 presents cumulative costs, $\mathcal{L}$, for scaling and scaling with variable selection analogous to Fig. 5 but across different values of $q = 2, 3, 4, 5$. In PCA-based ROM, the value for q also dictates how many PC source terms need to be computed, e.g. with $q = 3$ we have three PC source terms (three columns in the matrix $\mathbf{S_Z}$). Thus, for the cumulative costs to be compared on equal basis in Fig. 6 (for the same number n of target dependent variables), $\mathcal{L}$ is computed without including $\mathbf{S_Z}$ or $\widetilde{\mathbf{S}}_\mathbf{Z}$ in the set $\boldsymbol{\phi}$. We observe that the optimal preprocessing settings (markers with an outline) can change when requesting different manifold dimensionality. The minimal costs for scaling only happened for VAST scaling when $q = 2, 3, 4$ and for Pareto scaling when $q = 5$; the minimal costs for scaling with variable selection happened for Auto scaling when $q = 2, 3$, for VAST scaling when $q = 4$ and for Pareto scaling when $q = 5$. The optimal subsets can change significantly across different scalings and manifold dimensionalities (see supplementary Table S2). We also observe that costs generally drop with an increasing q. This is an expected outcome, since by increasing q in PCA we add another orthogonal manifold parameter and, in principle, improve the parameterization quality. Furthermore, a larger difference in costs is usually observed when changing q from 2 to 3, than when changing q from 4 to 5. The decreasing behavior of $\mathcal{L}$ with increasing q can be due to the number of manifold parameters becoming sufficient to define a quality data parameterization. While the traditional eigenvalue convergence analysis in PCA can help select q tied to the variance explained, the cost function proposed here provides different information and can guide an optimal choice for q from the LDM topology perspective.

4.5. Impact on the reduced-order model performance

To assess the impact of the state vector subset on the generation of ROMs, we can compare the performance of the optimal state vector subset with randomly selected subsets. We generate various three-dimensional PCA projections of a syngas/air flamelet dataset preprocessed with Range scaling (the best option for scaling with variable selection, see Fig. 5) by randomly selecting subsets of the state variables.

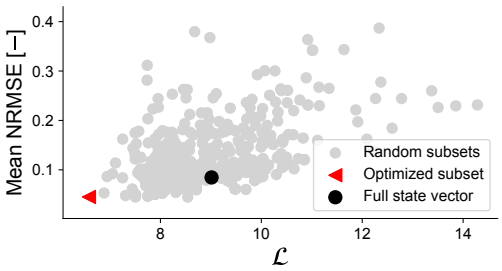

Fig. 7: Cumulative cost, $\mathcal{L}$, from three PC source terms versus the average NRMS error from kernel regression of the same PC source terms for a syngas/air flamelet dataset with Range scaling. Grey points correspond to three-dimensional PCA projections of 500 randomly selected state vector subsets. The red triangle represents projection of the optimal state vector subset and the black circle of the full state vector.

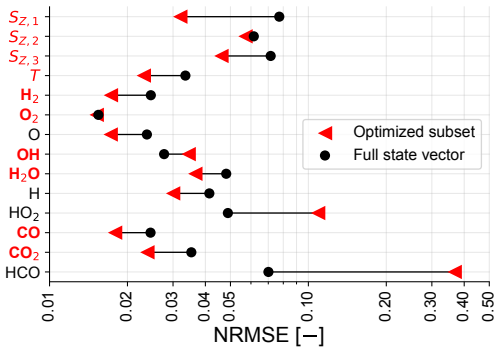

Fig. 8: NRMS errors from kernel regression of all state variables and three PC source terms from a syngas/air flamelet dataset. The full state vector is scaled with $\langle -1, 1 \rangle$ scaling and the optimal subset with Range scaling. The variables highlighted in red are the target dependent variables.

Figure 7 presents cumulative costs from the three PC source terms in the linear and symlog space, $\mathcal{L} = \mathcal{L}_{\mathbf{S_Z}} + \mathcal{L}_{\tilde{\mathbf{S}}_\mathbf{Z}}$, versus the average NRMS error from kernel regression prediction of the same three PC source terms. Throughout this work, we measure NRMS errors on test data only (20% of the data not seen by the kernel regression model). The figure shows grey points corresponding to 500 randomly selected state vector subsets, the red triangle corresponding to the optimal state vector subset and the black circle corresponding to the full state vector. Generally, lower costs yield better regression results and higher costs yield higher regression errors (Spearman correlation between the two axes is 39%). Further, the cost associated with the optimal subset is lower than costs for randomly selected subsets and significantly lower than the cost associated with using the full state vector. Supplementary Figs. S3-S4 show analogous results for other fuels with Spearman correlation equal to 61% for hydrogen fuel and 51% for ethylene fuel.

The optimization of the LDM topology should facilitate finding a better definition for the regression function $\mathscr{F}$ and, what follows, a better model for $\boldsymbol{\phi}$. We assess the regression performance of all state variables and the PC source terms using the optimized manifold parameters as regressors. Figure 8 shows the NRMS errors from kernel regression of variables for a syngas/air flamelet dataset. Analogous results for other fuels can be found in the supplementary material (Figs. S5-S6). We report errors associated with three-dimensional PCA projection of the full state vector (black circles) with the best $\langle -1, 1 \rangle$ scaling option and of the optimal state vector subset (red triangles) with the best Range scaling option. Variables highlighted in red are the target variables for this fuel. We observe that all target variables except OH mass fraction have lower regression errors when the optimal state vector subset is used. Most importantly, regression of the first three PC source terms is improved. Worse regression of two state variables, HO_2 and HCO, can be observed on an optimized manifold. Since mass fractions of HO_2 and HCO were not included in the target variables, the algorithm had no means to assess the representation of these variables on a manifold. Figure 8 shows that regression performance can improve when selecting an appropriate state vector subset as compared to only scaling the data. Finally, we note that regression techniques alone require tuning, which was not the main focus of the present work (e.g. kernel regression is dependent on the type of kernel and the kernel bandwidth). It is possible that combining our manifold improvement strategies with appropriate regression settings may yield further improvements in regression results beyond what we report in this paper.

5. Conclusions

Many factors affect the quality of low-dimensional data projections. Acting on the original dataset through data preprocessing, such as scaling or variable selection, can largely impact the quality of the LDMs. In this work, we propose a manifold-informed methodology to select an optimal subset of the full thermo-chemical state vector for training ROMs. The subset is selected through an optimization algorithm that pays attention to LDM topology aspects such as feature size and non-uniqueness. Two main strengths of the proposed algorithm are: (1) LDM topology can be optimized for any target dimensionality, and (2) LDM topology can be optimized with respect to an arbitrary set of target dependent variables. The latter is particularly appealing, since only the most important dependent variables can be included in the optimization. These can for instance be temperature and major species, as well as functions of the original variables required by the ROM. We shed more light on how the subset selection affects the low-dimensional data projections. We demonstrate that minor species (often discarded in the literature), can play an important role in achieving the desired LDM quality. While

many ROM efforts have recently focused on PCA as the dimensionality reduction technique, the methodology reported in this work can be easily extended to other manifold identification techniques.

Finally, we note that sampling observations in state-space can also affect the LDM quality. Although not explored in this work, tackling data imbalance can be another viable data preprocessing strategy. This is especially true for experimental or DNS datasets, which can be biased by uneven sample distributions in various flame regions. By balancing observations in a dataset, we can help PCA "see" variances in regions that would otherwise be overlooked due to high sample densities in other regions. Two such data balancing strategies can be helpful. First, an approach that has been introduced in the past is kernel density weighting of datasets [31, 32], prior to applying a reduction technique. This approach can be in fact viewed as data scaling, but in the observation space, instead of in the variable space. It allows to give more/less importance to individual observations. The second approach is data sampling, e.g. through undersampling observations from abundant or over-resolved regions. Techniques such as DBSCAN or Gaussian mixtures can be employed to guide the sampling process based on local data density. With the help of our cost function, manifolds that result from datasets that have been density-weighted or re-sampled can be assessed in an analogous way as we have assessed various scaling criteria and variable selection in this work. It remains to be seen in future studies which observations play crucial role for achieving desired quality of manifold topologies.

Acknowledgments

The research of the first author is supported by the F.R.S.-FNRS Aspirant Research Fellow grant. 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.

Supplementary material

Supplementary material is submitted along with this manuscript.

References

[1] U. Maas, S. Pope, Simplifying chemical kinetics: Intrinsic low-dimensional manifolds in composition space, Combust. Flame 88 (3) (1992) 239–264.

[2] J. Van Oijen, L. De Goey, Modelling of premixed counterflow flames using the flamelet-generated manifold method, Combust. Theory Model. 6 (3) (2002) 463–478.

[3] P. K. Jha, C. P. Groth, Tabulated chemistry approaches for laminar flames: Evaluation of flame-prolongation of ILDM and flamelet methods, Combust. Theory Model. 16 (1) (2012) 31–57.

[4] J. C. Sutherland, A. Parente, Combustion modeling using principal component analysis, Proc. Combust. Inst. 32 (1) (2009) 1563–1570.

[5] Y. Yang, S. B. Pope, J. H. Chen, Empirical low-dimensional manifolds in composition space, Combust. Flame 160 (10) (2013) 1967 – 1980.

[6] A. Biglari, J. C. Sutherland, A filter-independent model identification technique for turbulent combustion modeling, Combust. Flame 159 (5) (2012) 1960 – 1970.

[7] A. Biglari, J. C. Sutherland, An a-posteriori evaluation of principal component analysis-based models for turbulent combustion simulations, Combust. Flame 162 (10) (2015) 4025 – 4035.

[8] M. R. Malik, B. J. Isaac, A. Coussement, P. J. Smith, A. Parente, Principal component analysis coupled with nonlinear regression for chemistry reduction, Combust. Flame 187 (2018) 30–41.

[9] M. R. Malik, P. Obando Vega, A. Coussement, A. Parente, Combustion modeling using principal component analysis: A posteriori validation on sandia flames D, E and F, Proc. Combust. Inst. 38 (2) (2021) 2635–2643.

[10] H. Mirgolbabaei, T. Echekki, Nonlinear reduction of combustion composition space with kernel principal component analysis, Combust. Flame 161 (2014) 118–126.

[11] H. Mirgolbabaei, T. Echekki, N. Smaoui, A nonlinear principal component analysis approach for turbulent combustion composition space, Int. J. Hydrog. 39 (9) (2014) 4622–4633.

[12] M. Ihme, L. Shunn, J. Zhang, Regularization of reaction progress variable for application to flamelet-based combustion models, J. Comput. Phys. 231 (23) (2012) 7715–7721.

[13] S. B. Pope, Small scales, many species and the manifold challenges of turbulent combustion, Proc. Combust. Inst. 34 (1) (2013) 1 – 31.

[14] A. Parente, J. C. Sutherland, Principal component analysis of turbulent combustion data: Data preprocessing and manifold sensitivity, Combust. Flame 160 (2) (2013) 340 – 350.

[15] A. Biglari, J. C. Sutherland, An a-posteriori evaluation of principal component analysis-based models for turbulent combustion simulations, Combust. Flame 162 (10) (2015) 4025–4035.

[16] B. J. Isaac, J. N. Thornock, J. Sutherland, P. J. Smith, A. Parente, Advanced regression methods for combustion modelling using principal components, Combust. Flame 162 (6) (2015) 2592–2601.

[17] H. Mirgolbabaei, T. Echekki, A novel principal component analysis-based acceleration scheme for LES–ODT: An a priori study, Combust. Flame 160 (5) (2013) 898–908.

[18] T. Echekki, H. Mirgolbabaei, Principal component transport in turbulent combustion: A posteriori analysis, Combust. Flame 162 (5) (2015) 1919–1933.

[19] O. Owoyele, T. Echekki, Toward computationally efficient combustion DNS with complex fuels via principal component transport, Combust. Theory Model. 21 (4) (2017) 770–798.

[20] H.-T. Nguyen, P. Domingo, L. Vervisch, P.-D. Nguyen, Machine learning for integrating combustion chemistry in numerical simulations, Energy and AI 5 (2021) 100082.

[21] K. M. Gitushi, R. Ranade, T. Echekki, Investigation of deep learning methods for efficient high-fidelity simulations in turbulent combustion, Combust. Flame 236

(2022) 111814.
[22] V. Hiremath, Z. Ren, S. B. Pope, A greedy algorithm for species selection in dimension reduction of combustion chemistry, Combust. Theory Model. 14 (5) (2010) 619–652.
[23] E. Armstrong, J. C. Sutherland, A technique for characterising feature size and quality of manifolds, Combust. Theory Model. 0 (0) (2021) 1–23.
[24] M. P. Burke, M. Chaos, Y. Ju, F. L. Dryer, S. J. Klippenstein, Comprehensive H2/O2 kinetic model for high-pressure combustion, Int. J. Chem. Kinet. 44 (7) (2012) 444–474.
[25] 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.
[26] Z. Luo, C. Yoo, E. Richardson, J. Chen, C. Law, T. Lu, Chemical explosive mode analysis for a turbulent lifted ethylene jet flame in highly-heated coflow, Combust. Flame 159 (1) (2012) 265–274.
[27] M. A. Hansen, J. C. Sutherland, On the consistency of state vectors and Jacobian matrices, Combust. Flame 193 (2018) 257–271.
[28] J. C. Sutherland, P. J. Smith, J. H. Chen, A quantitative method for a priori evaluation of combustion reaction models, Combust. Theory Model. 11 (2) (2007) 287–303.
[29] 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, Proc. Combust. Inst. (2020).
[30] 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.
[31] A. Coussement, O. Gicquel, A. Parente, Kernel density weighted principal component analysis of combustion processes, Combust. Flame 159 (9) (2012) 2844–2855.
[32] M. R. Malik, A. Coussement, T. Echekki, A. Parente, Principal component analysis based combustion model in the context of a lifted methane/air flame: Sensitivity to the manifold parameters and subgrid closure, Combust. Flame 244 (2022) 112134.