Manifold-informed state vector subset for reduced-order modeling

Kamila Zdybał[1, 2], James C. Sutherland[3], and Alessandro Parente[1, 2]

[1]*Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Brussels, Belgium*

[2]*Université Libre de Bruxelles and Vrije Universiteit Brussel, Combustion and Robust Optimization Group (BURN), Brussels, Belgium*

[3]*Department of Chemical Engineering, University of Utah, Salt Lake City, UT, USA*

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 data sets representing the evolution of the thermo-chemical state-space in simple systems. For low-Mach flows, the full state vector that serves as a training data set is typically composed of temperature and chemical composition. The data set is projected onto a lower-dimensional basis and the evolution of the complex system is tracked on the low-dimensional manifold. This approach allows for substantial dimensionality reduction, 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 subsetting 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 best subset of state variables that minimizes unwanted behaviors in manifold topologies.

1 Introduction

Parameterization approaches can be used to compress descriptions of complex combustion systems with many degrees of freedom. In data-driven approaches, low-dimensional manifolds (LDMs) are constructed directly from the training data [1, 2]. Linear and nonlinear dimensionality reduction techniques have been used in the past to find the lower-dimensional basis to represent the full data set. The success of a given reduction technique then depends on the quality of low-dimensional data projection. Characteristics of a good parameterization include soft gradients, as well as uniqueness and realizability in the independent variable space [3]. The question of the parameterization quality persists in data-driven reduced-order modeling (ROM). Notably, non-uniqueness can be introduced during low-dimensional data projection, resulting in ambiguity in dependent variable values.

Problems with ill-behaved manifolds can be alleviated through appropriate data preprocessing. The most straightforward strategy is data scaling. Other authors have tackled manifold challenges by training combustion models on only a subset of the original thermo-chemical state-space variables [4, 5, 6, 7, 8, 9]. A closer look at the variables typically selected in the literature suggests that authors create the state vector subset in qualitative ways, taking fuel and oxidizer components and complete combustion products, with [4, 5, 6, 8] or without [4, 7, 9] temperature, and rarely including minor species [8]. However, such variable selections are often done *ad hoc*, without detailing justification for selecting some variables and discarding another. In particular, to the authors' knowledge, no satisfying explanation has been given as to which chemical species produce good basis for training ROMs.

The goal of this work is to define a meaningful subset of the original variables, optimized to result in an improved LDM topology. Recently developed technique to characterize manifold quality [10] can be used to assess data parameterizations. Using these LDM assessment tools, we developed a variable selection algorithm that allows to find the best set of state variables that minimize regions of undesired manifolds topologies. The desired effects include reducing non-uniqueness and spatial gradients in the dependent variable space.

2 Results and discussion

In this work, we define a cost function, $\mathcal{L}$, based on the recently proposed normalized variance derivative metric [10]. The cost function essentially distills information about variance in dependent variable's values occurring at different spatial scales on a manifold. It is defined for the i^{th} dependent variable as:

$$\mathcal{L}_i = \int_{\widetilde{\sigma}_{min,i}}^{\widetilde{\sigma}_{max,i}} P_i(\sigma, \sigma_{peak,i}) \cdot \hat{\mathcal{D}}_i(\sigma) d\widetilde{\sigma}, \tag{1}$$

where $\hat{\mathcal{D}}_i(\sigma)$ is the normalized variance derivative as per [10] and $P_i(\sigma, \sigma_{peak,i})$ is the penalty function defined as:

$$P_i(\sigma, \sigma_{peak,i}) = \left|\widetilde{\sigma} - \widetilde{\sigma}_{peak,i}\right| + \frac{\widetilde{\sigma}_{max,i} - \widetilde{\sigma}_{min,i}}{\widetilde{\sigma}_{peak,i} - \widetilde{\sigma}_{min,i}}. \tag{2}$$

The range of spacial scales at which the manifold is scanned is defined by $\sigma \in \langle \sigma_{min}, \sigma_{max} \rangle$ and σ_{peak} denotes the size of the largest feature on a manifold. Tilde denotes a $\log_{10}$-transformed quantity. Being defined as an area under a penalized $\hat{\mathcal{D}}_i(\sigma)$ curve, the cost function thus reduces manifold topology to a single number and can be implemented within optimization algorithms.

Here, we show a benchmark test performed on a data set obtained from a steady laminar flamelet model for

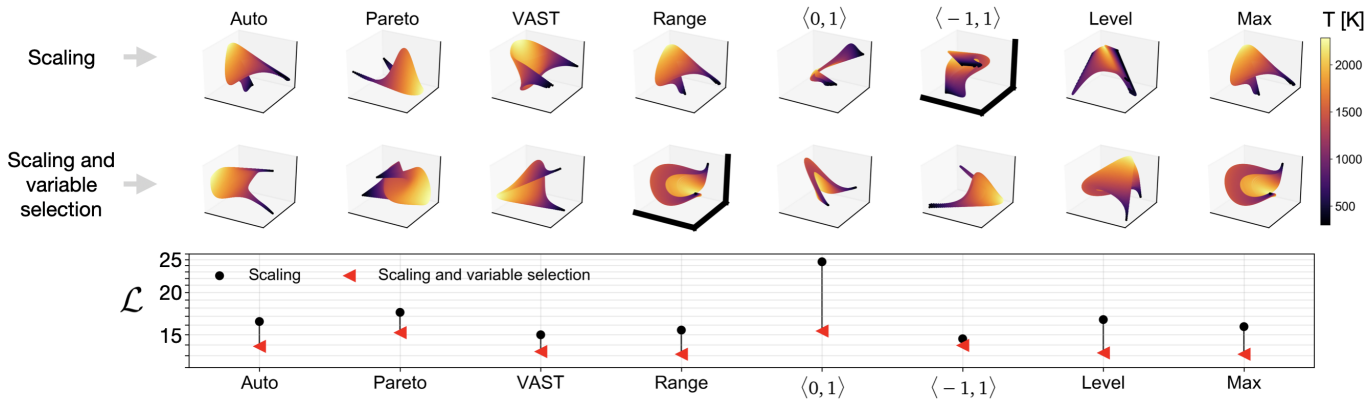

Figure 1: Ranking of different training data scaling strategies and scaling combined with subsetting strategies in their capacity to produce quality three-dimensional manifolds using PCA. Manifolds are generated from a data set describing combustion of syngas in air. Cumulative cost for the optimizing variables: S_{PC_1}, S_{PC_2}, S_{PC_3}, T, Y_{H_2}, Y_{O_2}, Y_{OH}, Y_{H_2O}, Y_{CO} and Y_{CO_2} was taken. The best manifold topology (with the lowest cost) for each strategy has been highlighted with thicker axes. All manifolds are colored by the temperature, T.

combustion of syngas in air. The data set was generated using Spitfire software [11]. We developed a variable selection algorithm, where at each iteration the variable that minimizes the cost the most is removed. The state vector subset is generated by including only the variables that minimize the overall cost from all iterations. Fig. 1 demonstrates how the cost function ranks different data scaling options in their capacity to generate quality three-dimensional manifolds using principal component analysis (PCA). For comparison, we show how costs can be further reduced if an appropriate state vector subset is selected. Subsetting the state vector allows to mitigate undesired manifold behaviors, such as compressing data observations to small regions on a manifold as is for instance happening on a manifold generated from $\langle 0, 1 \rangle$ scaling of the full training data set. Such regions can be particularly difficult to regress and should be avoided from the ROM perspective.

A powerful characteristic of the proposed approach is that manifolds can be optimized for a selection of important dependent variables. In the example shown in Fig. 1, we used the three PCA-transformed source terms of the original state variables (S_{PC_1}, S_{PC_2}, S_{PC_3}) and a few selected state variables that can be most important in accurate modeling, such as temperature and mass fractions of selected species (T, Y_{H_2}, Y_{O_2}, Y_{OH}, Y_{H_2O}, Y_{CO}, Y_{CO_2}). Moreover, manifold topology can be optimized for any target dimensionality.

While the example in Fig. 1 is shown for syngas/air combustion, we tested the approach on flamelets for other fuels as well. We observed a similar behavior of decreased costs when the optimal state vector subset is selected on hydrogen/air and ethylene/air combustion. This suggests benefits from creating a meaningful subset of the original state vector. While evidence in the literature already exists to support this conclusion, in this work, we created tools that can quantify the effect of subsetting. Many authors in the past have focused on selecting major species, but we show that a mixture of major and minor species can be beneficial. Oftentimes, the variable selection algorithms includes minor species such as O, OH or H.

Acknowledgements

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.

References

[1] J. C. Sutherland and A. Parente. Combustion modeling using principal component analysis. *Proceedings of the Combustion Institute*, 32(1):1563–1570, 2009.

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

[3] S. B. Pope. Small scales, many species and the manifold challenges of turbulent combustion. *Proceedings of the Combustion Institute*, 34(1):1 – 31, 2013.

[4] H. Mirgolbabaei and T. Echekki. A novel principal component analysis-based acceleration scheme for les–odt: An a priori study. *Combustion and Flame*, 160(5):898–908, 2013.

[5] H. Mirgolbabaei and T. Echekki. Nonlinear reduction of combustion composition space with kernel principal component analysis. *Combustion and Flame*, 161:118–126, 2014.

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

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

[8] O. Owoyele and T. Echekki. Toward computationally efficient combustion dns with complex fuels via principal component transport. *Combustion Theory and Modelling*, 21(4):770–798, 2017.

[9] M. R. Malik, P. Obando Vega, A. Coussement, and A. Parente. Combustion modeling using principal component analysis: A posteriori validation on sandia flames d, e and f. *Proceedings of the Combustion Institute*, 38(2):2635–2643, 2021.

[10] E. Armstrong and J. C. Sutherland. A technique for characterising feature size and quality of manifolds. *Combustion Theory and Modelling*, 0(0):1–23, 2021.

[11] M. A. Hansen. Spitfire, 2020.