



# Reduced-order modeling of turbulent reacting flows using data-driven approaches

Ph.D. dissertation, Kamila Zdybał, 2023

Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory

BRITE: BRussels Institute for Thermal-fluid systems and clean Energy

# Abstract

Turbulent multicomponent reacting flows are described by a large number of coupled partial differential equations. With such large systems of equations, the current computational capabilities are insufficient for detailed simulations. At the same time, accurate simulations are crucial to support the rapidly developing combustion technologies. Dimensionality reduction and machine learning approaches appear well-suited for building reduced-order models (ROMs) of complex systems with many degrees of freedom. Dimensionality reduction techniques project a high-dimensional system onto a lower-dimensional basis. Projections can be computed from the available training data and are referred to as low-dimensional manifolds (LDMs). Dimensionality reduction is often coupled with nonlinear regression to bypass the errors associated with the inverse basis transformation. Regression allows to reconstruct the target thermo-chemical state quantities from the LDM parameters. A data-driven ROM workflow provides substantial reduction to the number of transport equations solved in combustion simulations, but the quality of the manifold topology is one of the decisive aspects in successful modeling. Numerous manifold challenges of turbulent combustion have been reported in the literature and ought to be addressed. The present work advances the performance of ROMs of reacting flows. Our main focus is in addressing the outstanding manifold challenges. We provide novel tools and algorithms that can help further reduce the order, and improve the predictive capabilities of the model.

The significant original contribution of this work is the development of tools to quantify the quality of LDMs from the perspective of ROM. We propose a metric that reduces the LDM topology to a single number, based on two aspects that affect modeling in particular: (1) steep gradients and (2) non-uniqueness in dependent quantities of interest (QoIs). Such quantitative tool was not available in the literature thus far. The metric becomes particularly informative when building nonlinear regression models on top of a low-dimensional projection.

We demonstrate that LDM topologies can be improved using our quantitative metric as a cost function in optimization algorithms. The next contribution of this work is development of strategies to improve topologies of low-dimensional data representations. In particular, two new algorithms for variable (feature) selection are developed, that return a subset of the thermo-chemical state vector. The subset is optimized to yield an improved LDM quality once it is projected onto a lower-dimensional basis. We also use our quantitative tools to assess other means of data preprocessing, including data scaling and data sampling. We show that quantitative rankings of various data preprocessing and manifold learning strategies can be created *a priori* at the modeling stage. This allows for automating decisions which thus far had to be performed manually – either through trial and error or using heuristic guidelines. We discover that among many data preprocessing scenarios, adequate data scaling combined with optimized variable selection has the potential to affect the LDM topologies the most. We argue that further improvements in parameterization quality can be achieved in many areas of science and engineering if the low-dimensional parameter space is thoroughly explored and then assessed using the proposed quantitative metric.

While principal component analysis (PCA) has been established in the combustion literature as a dimensionality reduction technique, we develop an alternative approach to obtain LDMs from data. We propose to combine dimensionality reduction and nonlinear regression within an encoder-decoder neural network architecture. Research efforts have thus far considered dimensionality reduction and nonlinear reconstruction as two separate steps. We show significant improvements in LDM topology

when these two steps are allowed to communicate with each other through backpropagation. Data projection becomes directly optimized to represent the QoIs regressed at the output of a decoder. The significant discovery of this work is that a nonlinear reconstruction error optimality promotes finding improved LDM topologies as compared to a linear reconstruction error optimality (*e.g.*, as in PCA). Our approach can become an effective replacement of standalone dimensionality reduction techniques, such as PCA, whenever nonlinear regression is anticipated in the downstream use.

We demonstrate our predictive tools inside a full ROM of a simple system of a zero-dimensional reactor. We first generate a good quality manifold using the proposed tools. We then benchmark several nonlinear regression models: artificial neural networks (ANNs), Gaussian process regression (GPR), kernel regression, and radial basis function (RBF) regression. We show that improved manifold topologies correlate with improved manifold regressibility. We transport the LDM parameters, instead of the high-dimensional thermo-chemical state variables. We demonstrate *a posteriori* insights on the benefits that improved manifold topologies and improved nonlinear regression bring in ROMs. The challenges that remain are linked with nonlinear regression performance, especially at the boundaries of the training manifold. We propose strategies that may help improve kernel-based regression methods. Among these are local kernel rotations based on gradients in QoIs, and local anisotropic bandwidth selections based on local feature sizes in QoIs.

Finally, we provide insights into physical interpretability of low-dimensional data parameterizations obtained using data science tools. We apply local PCA to combustion datasets of varying complexity in order to find settings that support finding physically meaningful information from data. Our approach connects with the recent trends in semi-supervised learning to incorporate any existing information about the system being studied. The results indicate that physics-based knowledge of the system can be used to enhance data-driven algorithms.

Two new Python libraries are developed in this work. The first library is **PCAfold**, a Python software package that can be used to generate, analyze and improve low-dimensional data representations. This software is paramount to generating and reproducing results in this dissertation. Each tool developed in this work is available in the **PCAfold** library. **PCAfold** can be applied broadly in other disciplines of research. The second library, **multipy**, has a mostly didactic purpose. This library can accompany and support a graduate course on multicomponent mass transfer. It can become a helpful study tool for students performing research in the area of reacting flows.

# Contents

<b>1</b>	<b>Setting the stage</b>	<b>17</b>
1.1	The big picture . . . . .	17
1.2	Developed software and code . . . . .	19
<b>2</b>	<b>Introductory &amp; background information</b>	<b>21</b>
2.1	Introduction . . . . .	22
2.2	Background on multicomponent mass transfer . . . . .	23
2.2.1	Governing equations for multicomponent mixtures . . . . .	23
2.2.2	Common simplifications to the governing equations . . . . .	25
2.2.3	Premixed and nonpremixed combustion . . . . .	26
2.2.4	Turbulence-chemistry interaction . . . . .	26
2.3	Background on data science . . . . .	28
2.3.1	Obtaining training datasets for data-driven approaches . . . . .	28
2.3.1.1	The zero-dimensional reactor model . . . . .	30
2.3.1.2	The steady laminar flamelet model . . . . .	30
2.3.2	Data normalization . . . . .	31
2.3.3	Dimensionality reduction and manifold learning . . . . .	31
2.3.3.1	Principal component analysis (PCA) . . . . .	32
2.3.3.2	Local principal component analysis (LPCA) . . . . .	33
2.3.4	Data clustering . . . . .	35
2.3.4.1	K-Means clustering . . . . .	35
2.3.4.2	Vector quantization PCA clustering . . . . .	35
2.3.4.3	Mixture fraction bins . . . . .	36
2.3.5	Data subsetting (variable/feature selection) . . . . .	36
2.3.6	Data sampling . . . . .	36
2.3.7	Density estimation of point-cloud data . . . . .	37
2.3.8	Nonlinear regression . . . . .	40
2.3.8.1	Artificial neural network . . . . .	40
2.3.8.2	Gaussian process regression . . . . .	41
2.3.8.3	Kernel regression . . . . .	42
2.3.8.4	Radial basis function . . . . .	42
2.3.9	Regression assessment metrics . . . . .	43
2.3.9.1	Global metrics . . . . .	43
2.3.9.2	Stratified regression metrics . . . . .	44
2.4	Background on reduced-order modeling . . . . .	44
2.4.1	Reducing the number of governing equations . . . . .	45
2.4.2	Principal component transport . . . . .	45
2.4.3	Manifold generated from PCA . . . . .	46
2.5	Low-dimensional manifold topology . . . . .	48
2.5.1	Manifold challenges and undesired behaviors on manifolds . . . . .	48
2.5.2	Manifold assessment metrics . . . . .	49
2.5.3	Regression in the presence of non-uniqueness . . . . .	50

2.5.4	Improving the low-dimensional manifold topology . . . . .	50
2.5.4.1	The effect of data preprocessing . . . . .	51
2.5.4.2	The effect of data sampling . . . . .	51
2.5.4.3	The effect of subsetting the state vector . . . . .	52
<b>3</b>	<b>Local manifold learning and its link to domain-based physics knowledge</b>	<b>53</b>
3.1	Introduction . . . . .	54
3.2	The analyzed datasets . . . . .	54
3.3	Global and local PCA . . . . .	56
3.3.1	Global PCA . . . . .	56
3.3.2	Local PCA . . . . .	56
3.3.3	Correlation between the known and the retrieved local parameterization . . . . .	57
3.3.4	Clustering based on variable bins . . . . .	58
3.3.5	Clustering using the VQPCA algorithm . . . . .	58
3.4	Results and discussion . . . . .	61
3.4.1	The Burke-Schumann model . . . . .	61
3.4.2	The chemical equilibrium model . . . . .	62
3.4.3	The homogeneous reactor model . . . . .	65
3.4.4	The high-fidelity DNS dataset: data-aided interpretation and relation with the training manifold . . . . .	68
3.4.5	Can global PCA detect the stoichiometric mixture fraction value? . . . . .	77
3.4.5.1	Single-component fuel streams . . . . .	77
3.4.5.2	Multi-component fuel streams and the effect of fuel dilution . . . . .	77
3.4.5.3	Perspective for future study . . . . .	79
3.5	Summary . . . . .	80
<b>4</b>	<b>Cost function for low-dimensional manifold topology assessment</b>	<b>81</b>
4.1	Introduction . . . . .	82
4.2	Cost function formulation . . . . .	83
4.2.1	Mathematical formulation . . . . .	84
4.2.2	Cost function response to feature size and non-uniqueness . . . . .	86
4.2.3	A simple example of assessing functions with varying regressibility . . . . .	87
4.2.4	Effect of hyper-parameters on the cost function . . . . .	88
4.2.5	On the computational impact of evaluating the cost function . . . . .	90
4.3	Results and discussion . . . . .	90
4.3.1	Assessing data preprocessing strategies . . . . .	90
4.3.2	Detecting large gradients on manifolds . . . . .	93
4.3.3	Manifold assessment across dimensionality . . . . .	94
4.3.4	Manifold assessment across various dimensionality reduction and manifold learning techniques . . . . .	95
4.3.5	Improved manifold topologies yield more accurate regression . . . . .	95
4.3.6	Detecting overlap between classes in categorical data . . . . .	97
4.3.7	Tuning t-SNE hyper-parameters . . . . .	98
4.3.8	Sensitivity to data sampling . . . . .	101
4.3.9	Practical information for reproducing the results . . . . .	102
4.3.9.1	Reacting flow data generation . . . . .	102
4.3.9.2	Outlier removal . . . . .	102
4.3.9.3	Autoencoder . . . . .	103
4.3.9.4	Nonlinear regression using artificial neural networks (ANNs) . . . . .	103
4.3.9.5	Nonlinear regression using kernel regression . . . . .	103
4.4	Summary . . . . .	106
<b>5</b>	<b>Manifold-informed state vector subset for reduced-order modeling</b>	<b>109</b>

5.1	Introduction . . . . .	110
5.2	Data-driven approach for model reduction . . . . .	111
5.3	Manifold-informed subset of state variables . . . . .	112
5.4	Results and discussion . . . . .	113
5.4.1	Choice of the target dependent variables . . . . .	114
5.4.2	State vector subset selected by the proposed algorithm . . . . .	114
5.4.3	Effect of variable selection versus data scaling . . . . .	115
5.4.4	Choice of the manifold dimensionality . . . . .	116
5.4.5	Impact on the reduced-order model performance . . . . .	117
5.5	Forward variable addition, or backward variable elimination? . . . . .	122
5.5.1	Manifold-informed backward variable elimination . . . . .	122
5.5.2	Manifold-informed forward variable addition . . . . .	123
5.5.3	Comparison of both algorithms . . . . .	123
5.5.4	Insights from the forward variable addition . . . . .	127
5.5.4.1	Can we beat the automatic bootstrap? . . . . .	132
5.5.5	How sensitive is the forward variable addition algorithm to data sampling? . . . . .	135
5.6	Summary . . . . .	138
<b>6</b>	<b>Tackling imbalance in combustion data – a manifold perspective</b>	<b>139</b>
6.1	Introduction . . . . .	140
6.2	Training datasets . . . . .	141
6.3	PCA on sampled datasets . . . . .	142
6.3.1	Data clustering . . . . .	142
6.3.2	Data sampling . . . . .	143
6.3.3	Informing PCA with the sampled dataset . . . . .	143
6.3.4	Choice for the number of clusters . . . . .	144
6.4	Results and discussion . . . . .	146
6.4.1	Improving the state-space representation through data sampling . . . . .	146
6.4.2	Effect of data re-sampling on the LDM topology . . . . .	148
6.4.3	Nonlinear regression performance on an improved LDM . . . . .	150
6.4.4	Application to other datasets . . . . .	153
6.4.5	Practical information for reproducing the results . . . . .	159
6.4.5.1	Code used . . . . .	159
6.4.5.2	Further details on the number of clusters . . . . .	159
6.4.5.3	Further details on kernel regression . . . . .	159
6.5	Summary . . . . .	160
<b>7</b>	<b>Reduced-order modeling with a regression-aware autoencoder</b>	<b>161</b>
7.1	Introduction . . . . .	162
7.2	The proposed regression-aware autoencoder . . . . .	162
7.3	Results and discussion . . . . .	165
7.3.1	Application to a synthetic two-dimensional dataset . . . . .	165
7.3.1.1	PCA projections might not be optimal for representing target dependent variables . . . . .	165
7.3.1.2	Assessing one-dimensional subspaces found by the regression-aware autoencoder . . . . .	166
7.3.1.3	Nonlinear reconstruction error promotes improved manifold topologies . . . . .	169
7.3.1.4	Effect of multiple outputs in the decoder . . . . .	170
7.3.1.5	Does more training data promote finding optimal projections? . . . . .	171
7.3.2	Application to reacting flow datasets . . . . .	174
7.3.2.1	Convergence in the mean-squared-error (MSE) loss function . . . . .	174

7.3.2.2	Improvements in parameterizing the projected source terms . . . . .	175
7.3.2.3	Benefits of mixing the projected source terms with state variables at the output of a decoder . . . . .	177
7.3.2.4	Assessing projection qualities in the proposed regression-aware autoencoder . . . . .	179
7.3.2.5	Comparison of the regression-aware AE and PCA parameterizations . . . . .	184
7.3.3	A gallery of emerging manifold topologies . . . . .	189
7.4	Summary . . . . .	191
<b>8</b>	<b>Reduced-order model for a zero-dimensional reactor</b>	<b>193</b>
8.1	Introduction . . . . .	194
8.2	Formulating the reduced-order model . . . . .	194
8.3	Computing the low-dimensional manifold parameters . . . . .	195
8.3.1	Principal component analysis projections . . . . .	196
8.3.2	Regression-aware autoencoder projections . . . . .	198
8.4	Building nonlinear regression models . . . . .	198
8.4.1	Benchmark of nonlinear regression closure models . . . . .	198
8.4.2	Artificial neural network model for the prediction of the thermo-chemistry . . . . .	199
8.5	Results and discussion . . . . .	203
8.5.1	How does the manifold quality affect the reduced-order model? . . . . .	203
8.5.1.1	Steep gradients, manifold edge effects, and manifolds with sharp turns .	203
8.5.1.2	Manifold with a severe overlap . . . . .	207
8.5.1.3	Manifold with a subtle overlap . . . . .	209
8.5.2	Predicting the thermo-chemistry . . . . .	211
8.5.3	The outstanding challenges in reduced-order modeling . . . . .	214
8.5.3.1	Manifold edge effects . . . . .	214
8.5.3.2	Adjusting to manifold locality . . . . .	214
8.6	Summary . . . . .	216
<b>9</b>	<b>Conclusions and future work</b>	<b>217</b>
<b>Appendix A</b>	<b>Non-conservative form of the governing equations</b>	<b>219</b>
<b>Appendix B</b>	<b>Manifold-optimized reaction variables</b>	<b>223</b>
B.1	Introduction . . . . .	224
B.2	Manifold-informed Bayesian optimization . . . . .	224
B.3	Results . . . . .	224
B.3.1	Models parameterized by $(f, \mathcal{Y})$ . . . . .	225
B.3.1.1	Hydrogen/air combustion . . . . .	225
B.3.1.2	Syngas/air combustion . . . . .	225
B.3.2	Models parameterized by $(f, \mathcal{Y}, h)$ . . . . .	226
B.3.2.1	Hydrogen/air combustion . . . . .	226
B.4	Summary . . . . .	226
<b>Appendix C</b>	<b>Local feature size estimation for kernel methods</b>	<b>229</b>
C.1	Setting the goal . . . . .	230
C.2	Building the tool . . . . .	230
C.3	Testing the tool, remaining issues and future work . . . . .	231
C.3.1	Testing on synthetic datasets . . . . .	231
C.3.2	Testing on combustion datasets . . . . .	232
C.3.3	Future work . . . . .	232

<b>Appendix D Manifold edge effects</b>	<b>235</b>
D.1 Introduction . . . . .	236
D.2 Improving kernel-based predictions with an anisotropic Gaussian kernel rotation . . . . .	236
D.2.1 Informing the kernel rotation by local PCA . . . . .	238
D.2.2 Informing the kernel rotation by the direction of the projected source term . . . . .	239
D.2.3 Informing the kernel rotation by gradients in the dependent variable . . . . .	239
D.3 Combining the kernel size with the kernel rotation . . . . .	240
D.4 The effect of loss functions in ANN predictions . . . . .	241
D.4.1 Mean squared logarithmic error as a loss function . . . . .	241
D.5 Can trimming the flamelet dataset help? . . . . .	242
D.6 Can a nonlinear transformation of the manifold parameters help? . . . . .	243
D.7 Future work . . . . .	243
<b>Appendix E PCAfold</b>	<b>245</b>
E.1 Software overview . . . . .	246
E.1.1 The <code>preprocess</code> module . . . . .	246
E.1.1.1 Data centering and scaling . . . . .	246
E.1.1.2 Data clustering . . . . .	247
E.1.1.3 Data sampling . . . . .	247
E.1.1.4 Kernel density weighting . . . . .	247
E.1.1.5 Density estimation . . . . .	247
E.1.1.6 Outlier detection . . . . .	247
E.1.1.7 Conditional statistics . . . . .	247
E.1.2 The <code>reduction</code> module . . . . .	247
E.1.2.1 Principal Component Analysis (PCA) . . . . .	247
E.1.2.2 Local PCA . . . . .	248
E.1.2.3 VQPCA . . . . .	248
E.1.2.4 Subset PCA . . . . .	248
E.1.2.5 Sample PCA . . . . .	248
E.1.3 The <code>analysis</code> module . . . . .	248
E.1.3.1 Manifold topology assessment . . . . .	248
E.1.3.2 Manifold-informed variable selection . . . . .	248
E.1.3.3 Kernel regression . . . . .	248
E.1.3.4 Nonlinear regression assessment . . . . .	249
<b>Appendix F multiply</b>	<b>251</b>
F.1 Software overview . . . . .	252
F.1.1 The <code>Composition</code> class . . . . .	252
F.1.2 The <code>Velocity</code> class . . . . .	252
F.1.3 The <code>Flux</code> class . . . . .	253
F.1.4 The <code>Diffusion</code> class . . . . .	253
F.1.5 The <code>Transform</code> class . . . . .	253
F.1.6 The <code>Check</code> class . . . . .	254
F.1.7 The <code>Templates</code> class . . . . .	254
F.2 Computational example: the non-reacting Stefan tube problem . . . . .	254
F.2.1 Problem set-up . . . . .	254
F.2.2 Pause and ponder . . . . .	256
F.2.3 Compute species mole fractions . . . . .	257
F.2.3.1 Solve numerically . . . . .	257
F.2.3.2 Solve for the total molar fluxes using an optimization algorithm . . . . .	259
F.2.3.3 Solve analytically . . . . .	261