Python software to generate, analyze and improve PCA-derived low-dimensional manifolds

Kamila Zdybał, Elizabeth Armstrong, Alessandro Parente, James C. Sutherland

Many research disciplines exhibit growing interest in low-dimensional data representations. Parameterizing high-dimensional data sets with fewer variables allows for an easier treatment and handling of multivariate data sets. In addition, reduced-order models (ROMs) can be built in lower-dimensional spaces. ROM is particularly appealing for combustion modeling. Combustion couples the complexity of fluid flow and chemical reactions, making simulations computationally challenging.

In recent years, Principal Component Analysis (PCA) became an established dimensionality reduction technique in combustion modeling. PCAfold [1] allows to generate low-dimensional manifolds using PCA by projecting the original training data set onto a lower-dimensional basis. PCAfold exploits the idea that the parameterization obtained via dimensionality reduction is not unique. It can be altered through data preprocessing, including scaling, sampling or subsetting. Once the manifold is obtained, novel functionalities are implemented in PCAfold that allow to assess the quality of manifold topologies. Two of the most important features of a well-defined manifold include uniqueness and moderate gradients in the dependent variable space.

The starting point of the PCAfold workflow is the user-defined multivariate data set. The data set can then be passed through the workflow defined by three main modules:

`preprocess` → `reduction` → `analysis`

Below, we demonstrate the available functionalities of PCAfold on a strained laminar flamelet data set for syngas/air combustion. The data set was generated using Spitfire software [2] and the chemical mechanism by Hawkes et al. [3].

`preprocess` Data preprocessing

Data normalization

The first step towards applying a dimensionality reduction technique is data centering and scaling. Various data scaling options are implemented in PCAfold, such as Auto, Pareto, VAST, or normalizing to the -1 to 1 or 0 to 1 range, and many more. This basic preprocessing functionality can help to address the question:

- **How do different data scaling factors affect the resulting low-dimensional manifold?**

This problem has been studied in the context of combustion in [4], where the authors demonstrated strong influence of data preprocessing on the low-dimensional data representations.

Kernel density weighting

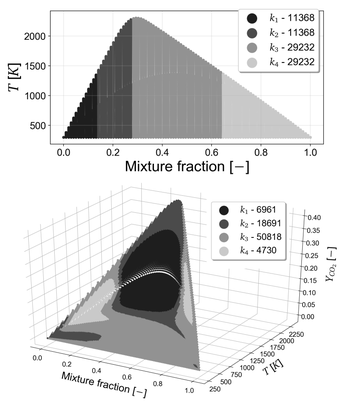
For imbalanced data sets, applying a single scaling factor to the entire thermo-chemical state variable might not be a reasonable strategy. This is especially true for combustion data sets which have abundant number of observations in the pure streams (fuel and oxidizer) and relatively few observations in the most interesting regions such as flame zone. To tackle this issue, in [5], kernel density weighting method was proposed, which allows to assign different weights to various regions of the flame. Kernel density weighting of data sets allows to find:

- **What is a more robust data scaling strategy that can address the problem of sample density bias?**

Data clustering

A few data clustering options and clustering utilities are implemented in PCAfold. For non-premixed combustion data sets, a feasible clustering strategy is binning based on mixture fraction. These can be particularly helpful in addressing the question:

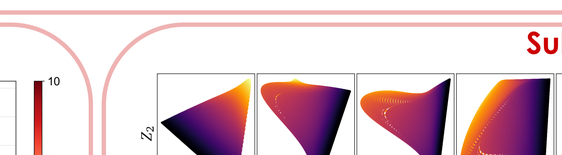
- **Can we find meaningful clusters in combustion data sets for implementing local techniques or stratified sampling?**

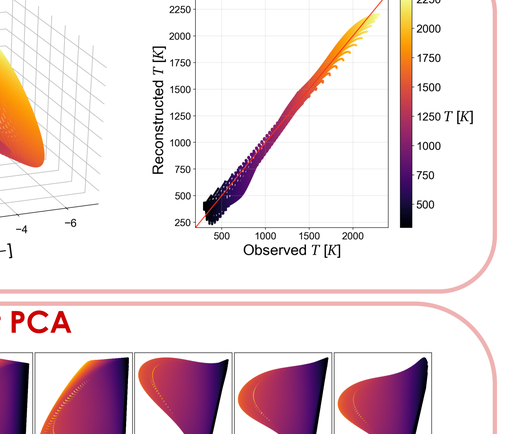


Data sampling

Another way of tackling imbalanced data sets is data sampling. Samples can be taken from local clusters of data, allowing for under-sampling large clusters (stratified sampling). As a result, a smaller data set can be formed for subsequent application of a dimensionality reduction technique.

- **How can we generate unbiased train and test data samples for training machine learning algorithms?**





Outlier detection

Removing statistical outliers from the data is possible with two popular approaches: based on the Mahalanobis distance and based on principal component classifier (PCC) [4]. The user can explore:

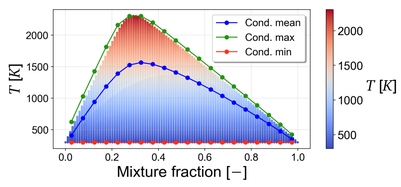
- **How does the principal component structure change after removing out-of-sample data?**

Conditional statistics

Information about conditional statistics, such as:

- **What is the conditional mean, minimum, maximum or standard deviation?**

can be easily obtained. The user can specify the conditioning variable (for instance mixture fraction) and the number of bins to create.

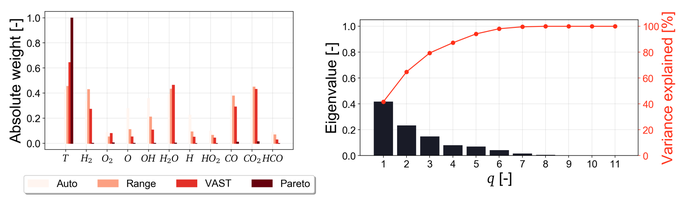


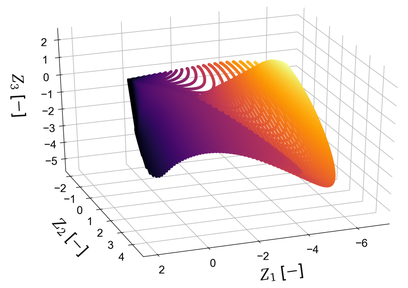
`reduction` Manifold generation & improvement

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) is the main workhorse of the PCAfold package. By reducing the dimensionality of the original data set, the user can generate low-dimensional manifolds defined by the first few principal components (PCs). Numerous plotting functionalities are available to help the user find out:

- **What is the resulting PC structure?**
- **How much information do the first few PCs convey?**
- **How does the resulting low-dimensional manifold look like (in 2D or 3D)?**
- **How well are the original variables reconstructed by the inverse PCA basis transformation?**





Local PCA

PCA can also be applied on local portions of the data, instead of the entire data set. This allows to find parameterizations tied to local regions. Applying local PCA to combustion data sets is particularly appealing, since various flame regions are characterized by different features. Local PCA functionalities can thus help in understanding:

- **What are the physical processes governing local regions of the high-dimensional data sets?**

Studying the local PC structure has recently been the subject of [6, 7].

Sample PCA

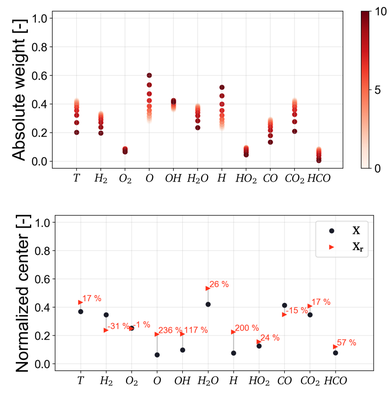
PCAfold introduces numerous interesting functionalities for performing PCA on sampled data sets. This can be helpful in addressing the question:

- **What is the effect of data sampling on the resulting PC structure?**

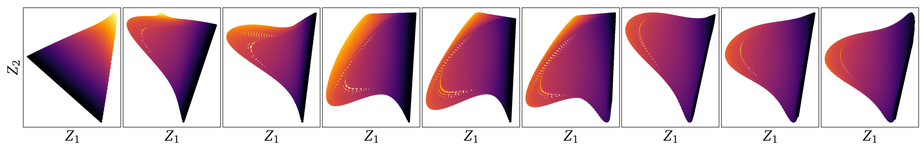
A history of iterative sub-sampling of data, heading towards equal cluster populations can be obtained and plotted. The user also has tools that allow to analyze:

- **How is sample bias removed from the data set?**

This includes visualizing the change in normalized centers of each original variable.



Subset PCA



It is a frequent practice to remove certain thermo-chemical state-space variables from combustion data sets before applying a dimensionality reduction technique. This process is more commonly known in the machine learning community as feature selection. The subset PCA functionalities allow to tackle the question:

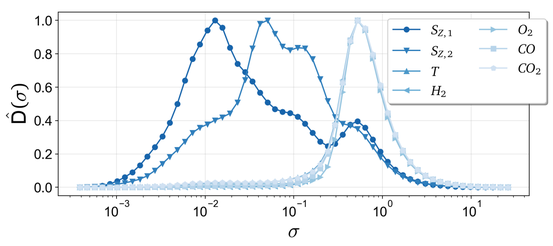
- **How does the manifold topology change when certain variables are removed from the data set?**

A novel algorithm that orders the original variables through a manifold-informed optimization is available. These tools can also help in answering the following questions:

- **What is the optimal state vector subset that results in a well-defined and regressible manifold?**
- **Which variables have the strongest influence on manifold topology?**

`analysis` Manifold analysis

Manifold assessment



Projecting the data onto a lower-dimensional basis can introduce undesired behaviors on manifolds. For instance, observations that are distant in the original space can be collapsed into a single, overlapping region. In the overlapping region, those observations are indistinguishable and the projection can become multi-valued. This is not desired from the ROM perspective. One of the questions that PCAfold can help to answer is:

- **Does the manifold possess certain desired topological characteristics?**

Ideally, we would like to search for such PCs defining the manifold, that it uniquely represents all dependent variables. A novel normalized variance derivative metric [8] can be computed for manifolds of arbitrary dimensionality. Analyzing the locations of peaks can facilitate in choosing the optimal manifold topology.

Kernel regression

Kernel regression using Nadaraya-Watson Gaussian kernel is implemented in PCAfold. Often, ROMs incorporate regression to effectively obtain dependent variable values at any location on a low-dimensional manifold [9, 10]. Manifold then becomes the regressor – an input variable for a regression technique. The application and assessment of regression on top of an identified manifold can be the first test to answer the question:

- **Is a given low-dimensional manifold suitable for building robust reduced-order models?**

Error metrics

The most straightforward assessment of regression performance is measuring the errors between the observed and the predicted variable. PCAfold allows for quick printing of many popular error metrics in the tabular format, as raw text and in the `tex` format. A color-coded summary can be printed to find out:

- **How do errors from two regression models compare?**

	R2	NRMSE	GDE
$S_{Z,1}$	0.9605	0.1987	90.7143
$S_{Z,2}$	0.8858	0.3380	90.7143
$S_{Z,3}$	0.9444	0.2357	90.7143

	R2	MAE	RMSE	NRMSE
T	0.99924	8.21372	16.00075	0.02754
Y_{H_2}	0.99986	0.00001	0.00003	0.01183
Y_{O_2}	0.99963	0.00048	0.00128	0.01922
Y_O	0.99872	0.00002	0.00004	0.03571
Y_{OH}	0.99784	0.00001	0.00003	0.04645
Y_{H_2O}	0.99901	0.00011	0.00021	0.03142
Y_H	0.99580	0.00000	0.00001	0.06479
Y_{HO_2}	0.98181	0.00000	0.00000	0.13487
Y_{CO}	0.99967	0.00272	0.00600	0.01823
Y_{CO_2}	0.99919	0.00150	0.00320	0.02837
Y_{HCO}	0.99210	0.00000	0.00000	0.08890

Regression assessment

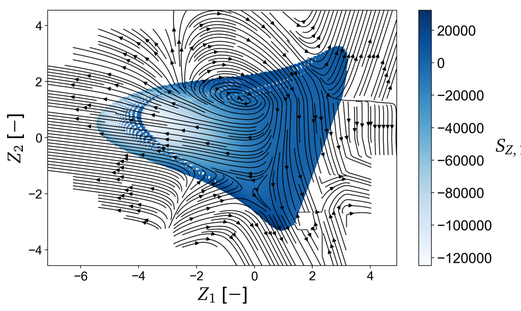
A few plotting functionalities allow to visually assess the regression performance. For instance, it can be useful to visualize stream plots of a regressed vector quantity. This can be particularly helpful in addressing the question:

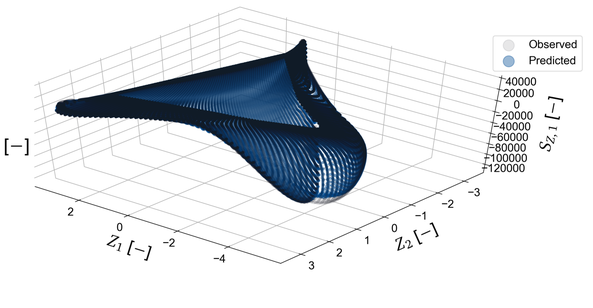
- **Are there regions where the numerical solution would be pushed off a manifold?**

2D and 3D plots of an observed versus predicted dependent variable can further help to find out:

- **Are there regions where regression struggles in particular?**

These can for instance be regions of non-uniqueness or steep gradients in the dependent variable space.





References

[1] K. Zdybał, E. Armstrong, A. Parente, J.C. Sutherland, 2020. *PCAfold: Python software to generate, analyze and improve PCA-derived low-dimensional manifolds*. SoftwareX, 12, p.100630.
[2] M.A. Hansen, 2020. Spitfire. URL: `https://github.com/sandialabs/Spitfire`.
[3] E.R. Hawkes, R. Sankaran, J.C. Sutherland, J.H. Chen, 2007. *Scalar mixing in direct numerical simulations of temporally evolving plane jet flames with skeletal CO/H2 kinetics*. Proceedings of the Combustion Institute, 31(1), pp.1633 – 1640.
[4] A. Parente, J.C. Sutherland, 2013. *Principal component analysis of turbulent combustion data: Data pre-processing and manifold sensitivity*. Combustion and Flame, 160(2), pp.340 – 350.
[5] A. Coussement, O. Gicquel, A. Parente, 2012. *Kernel density weighted principal component analysis of combustion processes*. Combustion and Flame, 159(9), pp.2844-2855.
[6] K. Zdybał, G. D'Alessio, G. Aversano, M.R. Malik, A. Coussement, J.C. Sutherland, A. Parente, *Advancing reacting flow simulations with data-driven models*, In M.A. Mendez, A. Ianiro, B.R. Noack, S.L. Brunton, editors, *Data-Driven Fluid Dynamics: Combining First Principles and Machine Learning*, Cambridge University Press, 2021
[7] Z. Li, S. Tomasch, Z.X. Chen, A. Parente, I.S. Ertesvåg, N. Swaminathan, 2021. *Study of MILD combustion using LES and advanced analysis tools*. Proceedings of the Combustion Institute, 38(4), pp.5423-5432.
[8] E. Armstrong, J.C. Sutherland, 2021. *A technique for characterising feature size and quality of manifolds*. Combustion Theory and Modelling, pp.1-23.
[9] B.J. Isaac, J.N. Thornock, J.C. Sutherland, P.J. Smith, A. Parente, 2015. *Advanced regression methods for combustion modelling using principal components*. Combustion and Flame, 162(6), pp.2592-2601.
[10] M.R. Malik, B.J. Isaac, A. Coussement, P.J. Smith, A. Parente, 2018. *Principal component analysis coupled with nonlinear regression for chemistry reduction*. Combustion and Flame, 187, pp.30-41.

Acknowledgments

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

The authors gratefully acknowledge the support of Sandia National Laboratories. Sandia National Laboratories is a multimission laboratory managed and operated by National Technology and Engineering Solutions of Sandia, LLC., a wholly owned subsidiary of Honeywell International, Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

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.

Aspects of this material are based upon work supported by the National Science Foundation under Grant No. 1953350.