SoftwareX 23 (2023) 101447

Contents lists available at ScienceDirect

SoftwareX

journal homepage: www.elsevier.com/locate/softx

Software update

PCAfold 2.0—Novel tools and algorithms for low-dimensional manifold assessment and optimization

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

[a] Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Brussels, Belgium
[b] BRITE: BRussels Institute for Thermal-fluid systems and clean Energy, Brussels, Belgium
[c] Department of Chemical Engineering, University of Utah, Salt Lake City, UT, USA

ARTICLE INFO

Article history:
Received 12 June 2023
Received in revised form 13 June 2023
Accepted 16 June 2023

Keywords:
Dimensionality reduction
Low-dimensional manifold
Reduced-order modeling
Artificial neural networks
Nonlinear regression
Python

ABSTRACT

We describe an update to our open-source Python package, PCAfold, designed to help researchers generate, analyze and improve low-dimensional data manifolds. In the current version, PCAfold 2.0, we introduce novel tools and algorithms for assessing and optimizing low-dimensional manifolds. This includes a method that generates a "map" of local feature sizes that can help pinpoint researchers to problematic regions on a manifold. We introduce a novel cost function that characterizes the quality of a manifold topology with a single number. We develop two algorithms for feature selection based on principal component analysis (PCA) that use the cost function as an objective function to minimize. We introduce a quantity of interest (QoI)-aware dimensionality reduction strategy where data projections are computed using an artificial neural network and are directly optimized towards representing various projection-independent and projection-dependent QoIs. We also introduce an implementation of partition of unity networks (POUnets) for efficient reconstruction of QoIs from low-dimensional manifolds based on combining neural network classification with localized polynomial regression. Our software can be broadly applicable in all domains of science and engineering that aim to reduce data dimensionality, as well as in the fundamental research on representation learning.

Code metadata

Current code version	2.0.0
Permanent link to code/repository used for this code version	https://github.com/ElsevierSoftwareX/SOFTX-D-23-00371
Code Ocean compute capsule	N/A
Legal Code License	MIT License
Code versioning system used	git
Software code languages, tools, and services used	Python 3.7
Compilation requirements, operating environments & dependencies	Cython, matplotlib, numpy, scipy, termcolor, pandas, scikit-learn, tensorflow, keras, tqdm
If available Link to developer documentation/manual	https://pcafold.readthedocs.io/
Support email for questions	kamilazdybal@gmail.com

1. Introduction

Multivariate datasets are collected at speed through experiments or numerical simulations in various disciplines of science and engineering. Researchers who process such high-dimensional datasets frequently rely on dimensionality reduction for surrogate modeling, efficient visualization, or data compression. Data analysis in the reduced space is often influenced by the quality of reduced data representations. Thus, quantitative tools are increasingly needed for automating assessments and improvements

DOI of original article: https://doi.org/10.1016/j.softx.2020.100630.
* Corresponding author at: Université Libre de Bruxelles, École polytechnique de Bruxelles, Aero-Thermo-Mechanics Laboratory, Brussels, Belgium.
E-mail addresses: kamilazdybal@gmail.com (Kamila Zdybał), Elizabeth.Armstrong@chemeng.utah.edu (Elizabeth Armstrong), Alessandro.Parente@ulb.be (Alessandro Parente), James.Sutherland@utah.edu (James C. Sutherland).

https://doi.org/10.1016/j.softx.2023.101447
2352-7110/

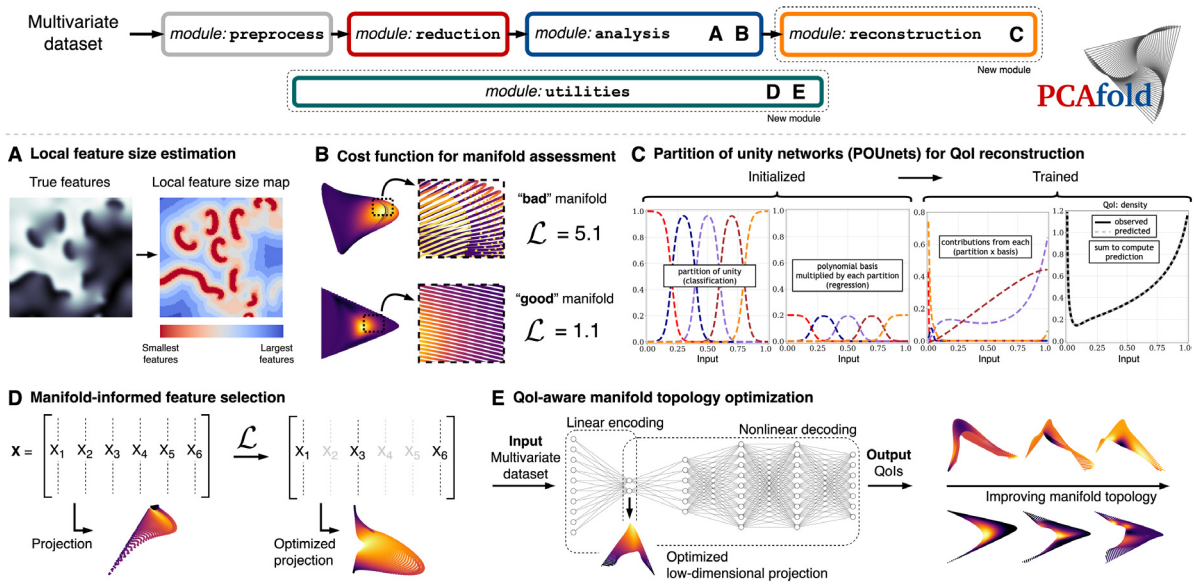

Fig. 1. Overview of the five main novel functionalities A–E introduced in PCAfold 2.0. These are described in further detail in the text.

of low-dimensional manifolds. Of interest, especially in reduced-order modeling, is also the ability to reconstruct quantities of interest (QoIs) from reduced data representations.

In [1], we have described PCAfold – an open-source Python library for generating, analyzing and improving low-dimensional manifolds. Here, we introduce PCAfold 2.0 that provides the research community with novel functionalities that significantly broaden the scope of low-dimensional manifold assessment and optimization possible as compared to the previous software version. Our software can be of relevance to numerous disciplines of science and engineering that explore low-dimensional data projections. In particular, it can serve as a useful toolbox at all stages of reduced-order modeling.

The PCAfold 2.0 workflow passes the user-provided multivariate dataset through the four main modules: `preprocess`, `reduction`, `analysis`, and `reconstruction` and an optional fifth module, `utilities`. In PCAfold 2.0, we introduce two new modules: `reconstruction`, that gathers regression tools, and `utilities`, which contains novel algorithms for optimizing the low-dimensional manifold topology. The latter draw various preprocessing, data reduction, analysis, and reconstruction tools from the four main modules and thus can be considered as utilities that span various manifold assessment and optimization tasks.

2. Overview of novel functionalities

Fig. 1 shows an illustrative overview of the five major novel functionalities introduced in PCAfold 2.0. These are powerful new tools for assessing and optimizing low-dimensional manifolds, as well as tools for reconstructing QoIs from the reduced data representations. Below, we describe these five items in more detail. For each new functionality, we point the reader to illustrative tutorials that can be found in the PCAfold documentation.

A **Local feature size estimation.** We develop an extension to the manifold assessment tools proposed thus far to obtain a "map" of local feature sizes on a manifold. This tool builds on the normalized variance and the normalized variance derivative metrics [2] available in the previous software version. The obtained map can have two main applications. First, it allows for detailed assessments of manifold topologies in the presence of a QoI. The local feature size map can inform us exactly where on a manifold are regions that exhibit non-uniqueness or steep gradients in QoIs. Second, it can inform the local bandwidth selection for kernel-based methods such as kernel regression. An illustrative tutorial can be found at: pcafold.readthedocs.io/en/latest/tutorials/demo-feature-size-map.html

B **Cost function for manifold assessment.** We present the novel quantitative measure for assessing the quality of a low-dimensional manifold from the perspective of non-uniqueness and steep gradients in QoIs [3]. This measure is built atop the normalized variance derivative metric [2], already present in the previous version of PCAfold. The new measure turns a low-dimensional data projection into a single number that quantifies how "costly" is the parameterization of a particular QoI on top of that projection. A larger number indicates a more problematic manifold topology that can include overlaps, twists, large gradients, or large curvatures. The cost function, $\mathcal{L}$, can further be used in various manifold optimization tasks. It can help automate researcher's decisions on data preprocessing, the choice of a reduction technique, or the choice of manifold dimensionality. An illustrative tutorial can be found at: pcafold.readthedocs.io/en/latest/tutorials/demo-cost-function.html

C **Partition of unity networks (POUnets) for QoI reconstruction.** POUnets are a recently introduced family of artificial neural networks (ANNs) that combine classification with regression to achieve interpolant levels of accuracy with small memory footprints [4,5]. The POUnets are implemented in `PartitionOfUnityNetwork` with a single-layer radial basis function (RBF) classification network whose neurons are each assigned a weight and a polynomial basis. Each neuron has trainable parameters controlling the location and shape of the RBFs and each monomial term in the basis functions has a trainable coefficient. These components together achieve the localized polynomial regression through learned domain partitioning and basis coefficients. Training follows the

least-squares gradient descent (LSGD) block coordinate descent strategy [4] with a sum of squared errors loss function.
For initialization, the user must specify the partition locations and shapes along with the polynomial `basis_type` with current support for constant, linear, or quadratic polynomials. The `init_uniform_partitions` helper function computes the partition initialization parameters from a specified grid and training dataset, only keeping partitions where training data exist [5]. The POUnet class also provides nonlinear power transformations of the dependent variable for training, which can help constrain positivity or squeeze variation occurring over orders of magnitude. The transformation for variable f can be written as

$$(|f + s_1|)^{\alpha}\text{sign}(f + s_1) + s_2\text{sign}(f + s_1) \quad (1)$$

where α is the specified `transform_power`, s_1 is the `transform_shift`, and s_2 is the `transform_sign_shift`. Finally, available training hyperparameters include the learning rate for gradient descent, regularization for least squares, and the loss function error type being absolute or relative. An illustrative tutorial can be found at: pcafold.readthedocs.io/en/latest/tutorials/demo-pounets.html

D **Manifold-informed feature selection.** We introduce a manifold-informed feature selection algorithm where a subset of the multivariate dataset is created based on minimizing the cost function, $\mathcal{L}$ [6]. Such subset results in an improved manifold topology once it is projected to a lower-dimensional basis. Two variants of the algorithm are introduced: backward variable elimination and forward variable addition. An illustrative tutorial can be found at: pcafold.readthedocs.io/en/latest/tutorials/demo-cost-function.html

E **QoI-aware manifold topology optimization.** We introduce an autoencoder-like dimensionality reduction strategy that optimizes the projection directly towards representing custom QoIs [7]. The QoIs can be selected as the original variables in the data (as is done in the classic autoencoder), or can be any projection-dependent quantities whose definition changes during network training. The latter can be selected as any variables required by the reduced-order model at model runtime. Connecting projection with QoI reconstruction in a single neural network naturally promotes improved projection topologies, since non-uniqueness and large gradients in QoIs immediately propagate large errors through the network. While a similar premise has been introduced in [8,9], we find that optimizing projections directly for projection-dependent QoIs, in addition to any projection-independent QoIs, can bring further benefits in reduced-order modeling. With the current implementation of PCAfold, we also allow the user to connect an ordinary neural network encoder with a POUnet decoder. Two illustrative tutorials can be found at: pcafold.readthedocs.io/en/latest/tutorials/demo-qoi-aware-encoder-decoder.html
pcafold.readthedocs.io/en/latest/tutorials/demo-qoi-aware-encoder-pounet.html

3. Summary

In addition to the five main functionalities A–E described in Fig. 1, we have added numerous smaller tools and user-interface improvements. We introduce new methods for data scaling, density estimation of point-cloud data, and computing conditional statistics from data. In addition to POUnets, we implement a class for nonlinear regression of QoIs with a standard ANN. We introduce various regression assessment metrics with their "stratified" variants that allow to assess regression performance in local regions of data.

Many of the novel functionalities of PCAfold 2.0 have been used in [3] for assessments of low-dimensional manifolds from reacting flows, plasma flows and atmospheric physics datasets, in [6] for determining optimized thermo-chemical state vector subsets, in [5] for demonstrating the capability of POUnets for reconstructing QoIs from tabulated chemistry models, and in [10] for developing a reduced-order model of a scramjet engine. Two graduate dissertations by the primary authors [11,12] have also made extensive use of PCAfold 2.0.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

Acknowledgments

K.Z. acknowledges the support of F.R.S.-FNRS, Belgium. E.A. acknowledges the support of Sandia National Laboratories, USA. 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, USA under Grant No. 1953350.

References

[1] Zdybał K, Armstrong E, Parente A, Sutherland JC. PCAfold: Python software to generate, analyze and improve PCA-derived low-dimensional manifolds. SoftwareX 2020;12:100630. http://dx.doi.org/10.1016/j.softx.2020.100630.
[2] Armstrong E, Sutherland JC. A technique for characterising feature size and quality of manifolds. Combust Theory Model 2021;25(4):646–68. http://dx.doi.org/10.1080/13647830.2021.1931715.
[3] Zdybał K, Armstrong E, Sutherland JC, Parente A. Cost function for low-dimensional manifold topology assessment. Sci Rep 2022;12:14496. http://dx.doi.org/10.1038/s41598-022-18655-1.
[4] Lee K, Trask NA, Patel RG, Gulian MA, Cyr EC. Partition of unity networks: Deep hp-approximation. 2021, URL https://arxiv.org/abs/2101.11256v1.
[5] Armstrong E, Hansen MA, Knaus RC, Trask NA, Hewson JC, Sutherland JC. Accurate compression of tabulated chemistry models with partition of unity networks. Combust Sci Technol 2022;1–18. http://dx.doi.org/10.1080/00102202.2022.2102908.
[6] Zdybał K, Sutherland JC, Parente A. Manifold-informed state vector subset for reduced-order modeling. Proc Combust Inst 2023;39(4):5145–54. http://dx.doi.org/10.1016/j.proci.2022.06.019.
[7] Zdybał K, Parente A, Sutherland JC. Improving reduced-order models through nonlinear decoding of projection-dependent model outputs. 2023, Article under review in Patterns.
[8] Perry BA, Henry de Frahan MT, Yellapantula S. Co-optimized machine-learned manifold models for large eddy simulation of turbulent combustion. Combust Flame 2022;244:112286. http://dx.doi.org/10.1016/j.combustflame.2022.112286.

[9] Scherding C, Rigas G, Sipp D, Schmid PJ, Sayadi T. Data-driven framework for input/output lookup tables reduction: Application to hypersonic flows in chemical nonequilibrium. Phys Rev Fluids 2023;8(2):023201. http://dx.doi.org/10.1103/PhysRevFluids.8.023201.

[10] Ispir AC, Zdybał K, Saracoglu BH, Magin T, Parente A, Coussement A. Reduced-order modeling of supersonic fuel-air mixing in a multi-strut injection scramjet engine using machine learning techniques. Acta Astronaut 2023;202:564–84. http://dx.doi.org/10.1016/j.actaastro.2022.11.013.

[11] Armstrong E. Development of improved parameterizations and nonlinear regression for reduced-order modeling in combustion [Ph.D. thesis], Department of Chemical Engineering, The University of Utah; 2023.

[12] Zdybał K. Reduced-order modeling of turbulent reacting flows using data-driven approaches [Ph.D. thesis], Université libre de Bruxelles; 2023, http://dx.doi.org/10.13140/RG.2.2.18843.95521, URL http://hdl.handle.net/2013/ULB-DIPOT:oai:dipot.ulb.ac.be:2013/357444.