MPhil DIS Report 24

CRSiD: tmb76

University of Cambridge

June 6, 2024

Contents

Chapter 1

Executive Summary

Chapter 2

Introduction

One of the key steps of scientific method is reproducibility. Results must be reproducible by others, ensuring that the same conclusions can be drawn multiple times. If a result cannot be reproduced, it may be considered erroneous, or simply a random occurence. An important aspect of this project will be to evaluate the reproducibility of the results of Callaham et al. (2021) [2] and to test the robustness of their method.

For many problems in engineering and physical sciences, equations involve a large number of terms and complex differential equations. Simulating them can be computationally expensive or unnecessarily so, due to multiple asymptotic local behaviours where the system is dominated by a subset of the terms. In such cases, one can simplify the equations to a balance between these dominant terms, and simulate the system with sufficient accuracy and relatively lower computational cost (REFERENCE FOR THIS). This method, known as dominant balance or scale analysis, has been a powerful tool in physics.

And though extremely useful, dominant balance also requires expertise and is usually done by hand in time-consuming proofs. This report discusses and verifies a novel approach, developped by Callaham et al. (2021) [2], an applies it to new data. First, this report will focus on the paper and the research surrounding it. Delving into what the rationale behind the method is, and how it performed on a series of case studies, as well as verifying it through reproducing the results with alternative code. This will be done primarily focusing on one of the case studies, but also for

most of the others. Additionally, other algorithms than the method's chosen one are used to test the robustness of the method. Second, the method will be used on a new dataset, from simulations of elasto-inertial turbulence, a property of of polymer laden flow.

Chapter 3

Background

Dominant balance is a powerful tool in simplifying the modelling of physical processes. Importantly, it helps better understand the physics at play in a system by identifying the subset of terms that matter in an equation for a specific asymptotic case. This understanding of interactions between terms and the physics they represent can be illustrated through examples in multiple fields of physics.

Taking the example of:

There has been some research to automate the process of finding the dominant balance in a system. First is the Portwood et al. (2016) [7] paper which uses a cumulative distribution function on the local density gradient to separate each region. THis method is then highly tailored to this case, where a gradient of one of the terms is used, knowing it can help discern dynamically distinct regions, and then interprets many results from expert analysis of the identified regions. Second is the Lee & Zaki (2018) [5] which introduces an algorithm to detect different dynamical regions, but again through the use of case-specific varaibles (vorticity), which restrict the use of this algorithm to specific flows. Finally, the Sonnewald et al. (2019) [9] paper which uses a K-Means clustering algorithm to identify different regions in the ocean. And they do introduce the concept of using the terms in the governing equations as features, but identification of active terms is done through comparison of the magnitudes of each term in the equation. In other words, identifciation of dominant terms is not done algorithmically but "manually". They either focused on a specific application and used expert-knowledge on it or only clustered the data

and then relied on expert knowledge to determine which terms were active in that cluster.

"Automating the process of finding the dominant balance in a system has been the focus of several studies. For instance, Portwood et al. (2016) use a cumulative distribution function on the local density gradient to separate regions dynamically, tailored to their specific case. Lee & Zaki (2018) introduce an algorithm to detect different dynamical regions using vorticity, but this is also specific to certain flows. Sonnewald et al. (2019) use a K-Means clustering algorithm to identify different regions in the ocean, employing terms in the governing equations as features. However, the identification of dominant terms is done manually by comparing magnitudes."

A similar interesting data science and machien learning challenge has been to directly find the laws and equations that govern a system from data. Schmidt & Lipson (2009) [8] contributed to a breakthrough using symbolic regression to find linear and non-linear differential equations. Brunton et al. (2016) [1] is another example of this, improving on Schmidt & Lipson's (2009) [8] work. As symbolic regression is expensive, the problem was now approahced with sparse regression, which for high-diemensioanl problems means identifying a sparse governing equation and this makes use of the governing equations usually having only a subset of terms being important. In Lejarza & Baldea (2022) [6], the governing equations are learned from noisy data using multiple basis functions, and a non-linear moving horizon optimization. Though it can only handle ODEs, it provides an approach for thresholding hte basis functions that result in the found governing equation are of lower complexity than those obtained thorugh sparse regression based approches.

"A similar challenge in data science and machine learning is to directly find the laws and equations that govern a system from data. Schmidt & Lipson (2009) used symbolic regression to find differential equations, followed by Brunton et al. (2016) who improved on this work using sparse regression. This approach is efficient for high-dimensional problems, as it identifies a sparse governing equation where only a subset of terms are important. Lejarza & Baldea (2022) further advanced this by using multiple basis functions and a non-linear moving horizon optimization to learn governing equations from noisy data"

Much like in more recent work where neural networks are used to determine hte physics at play from data [4] [3].

The method developped by Callaham et al. (2021) [2] is a novel approach to finding the dominant balance in a system. And it could be used in conjunction with the above governing equation identifying methods. But just like Schmidt & Lipson (2009) [8] points out for their own work, this method should not be considered as a perfect tool to find all the dominant physical processes in a certin case study but rather as a guiding tool to help indicate where scientists should focus their attention. "The method developed by Callaham et al. (2021) is a novel approach to finding the dominant balance in a system. It could be used alongside other methods for identifying governing equations. However, as Schmidt & Lipson (2009) noted for their work, this method should be seen as a guiding tool to help indicate where scientists should focus their attention, rather than a perfect tool for finding all dominant physical processes in a given case study."

Chapter 4

Methodology

Get simulated data of the terms in the equation of physical variables from which terms in the equation can be derived

Group the data into feature space, with each term as a feature

Cluster the data using GMM

SPCA to identify which terms are active in each cluster

Group together clusters that have the same active terms

Chapter 5

Conducted research

Callaham has used the algorithm in a few different cases to validate how it functions.

5.1 Portability of the code

5.2 Reproducibility of the results

5.3 Exploration of other algorithms

5.3.1 Spectral clustering

5.3.2 K-Means

5.3.3 Weighted K-Means

5.4 Stability Assessment

5.4.1 Under different number of clusters set

5.4.2 Under different training set size

Chapter 6

Elasto-inertial turbulence

6.1 Background

6.2 Methodology

6.3 Results

6.4 Discussion

Chapter 7

Data analysis pipeline

Bibliography

[1] Steven L. Brunton, Joshua L. Proctor, and J. Nathan Kutz. Discovering governing equations from data: Sparse identification of nonlinear dynamical systems. *Proceedings of the National Academy of Sciences*, 113(15):3932–3937, 2016.

[2] J.L. Callaham, J.V. Koch, and B.W. et al. Brunton. Learning dominant physical processes with data-driven balance models. *Nature Communications*, 12:1016, 2021.

[3] Miles Cranmer, Sam Greydanus, Stephan Hoyer, Peter Battaglia, David Spergel, and Shirley Ho. Lagrangian neural networks. *arXiv preprint arXiv:2003.04630*, 2020.

[4] Miles Cranmer, Alvaro Sanchez-Gonzalez, Peter Battaglia, Rui Xu, Kyle Cranmer, David Spergel, and Shirley Ho. Discovering symbolic models from deep learning with inductive biases. *arXiv preprint arXiv:2006.11287*, 2020.

[5] Jin Lee and Tamer A. Zaki. Detection algorithm for turbulent interfaces and large-scale structures in intermittent flows. *Computers & Fluids*, 175:142–158, 2018.

[6] Fernando Lejarza and Michael Baldea. Data-driven discovery of the governing equations of dynamical systems via moving horizon optimization. *Scientific Reports*, 12(1):11836, 2022.

[7] G. D. Portwood, S. M. de Bruyn Kops, J. R. Taylor, H. Salehipour, and C. P. Caulfield. Robust identification of dynamically distinct regions in stratified turbulence. *Journal of Fluid Mechanics*, 807:R2, 2016.

[8] Michael Schmidt and Hod Lipson. Distilling free-form natural laws from experimental data. *Science*, pages 81–85, 2009.

[9] Matthias Sonnewald, Carl Wunsch, and Patrick Heimbach. Unsupervised learning reveals geography of global ocean dynamical regions. *Earth and Space Science*, 6:784–794, 2019.