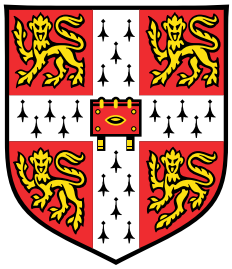
MPhil DIS Project 24

Executive Summary



CRSiD: tmb76

Department of Physics

University of Cambridge

Submitted in partial fulfilment of the requirements of the MPhil degree in Data Intensive Science

Hughes Hall June 30, 2024

Contents

0.1 Introduction (100 words) 1
0.2 Methodology (250 words) 2
0.3 Conducted Research and Results 3
0.4 EIT 7

0.1 Introduction (100 words)

In a majority of simulation problems in physics and engineering, it is usually unnecessary to consider the full system of equations that govern it. And it possible to simplify the equations by considering only a subset of the terms to actually matter. Simulating the system with only these dominant terms can provide a good approximation, whilst being computationally cheaper. This concept is the fundamental idea behind the method of dominant balance, where the subset of dominant terms is assumed to be in balance. And this method has proven extremely useful in multiple fields, including meteorology [2, 4, 6, 9].

Historically, this method has been applied manually by field experts, often over a long period of researching and the manipulation of complex mathematical equations. Over the recent years, there has been a growing interest in automating this process using machine learning techniques. Most of these however, were limited to specific systems and sometimes relied on expert interpretation rather than automatic identification of the dominant balances [**?**, 5, 8].

In this project, a novel method proposed by Callaham et al. (2021) [3] is explored, verified, and applied. This method offers a highly generalizable approach to identifying dominant balance models across various physical systems with minimal user input. One of the core aims of this project is to reproduce the results of Callaham et al.'s original paper using alternative code to ensure that the findings are robust and not dependent on specific implementations. Additionally, the project discusses the choice of clustering algorithm and explores the stability of the method

under different hyperparameters. The method is then applied to simulation data of elasto-inertial turbulence [7], a recently discovered flow, to demonstrate its potential in uncovering new dominant balance regimes in complex flows.

0.2 Methodology (250 words)

The proposed method can be summarized into 3 steps:

1. The first step involves transforming the data from the physical system into an equation-space format. This is done by computing the terms of the governing equations of the system, at each point in the physical space, and then have each point in physical space be a sample with features: the values of the equation's terms (see Fig. 1).

2. The second step then makes full use of this new format by clustering the samples in equation space, therefore grouping points with terms of similar magnitude. This is done using a Gaussian Mixture Model (GMM) clustering algorithm, which has the advantage of dealing very well with clusters of different shapes and sizes, and only requires for the number of clusters to be set (see Fig. 1).

3. The final step involves applying Sparse Principal Component Analysis (SPCA) to each. SPCA is a variant of PCA where a regularization constraint is applied to the number of non-zero coefficients in the principal components. Essentially, instead of returning a full principal component that is a linear combination of all the features, SPCA returns a sparse principal component that is a linear combination of only a few features, with some of the coefficients being zero. By applying this to the points in each cluster, and only taking the leading principal component, this gives a sparse vector describing which terms in the cluster dominate, and form together a dominant balance model (see Fig. 2).

Then, if there are any clusters that have the same dominant balance, they are combined.

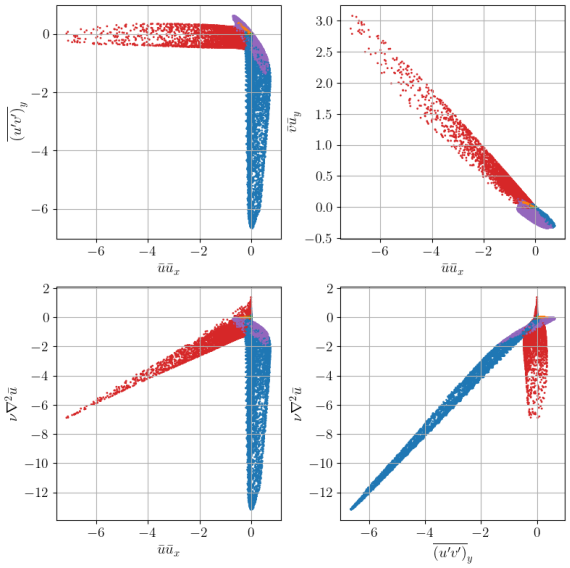
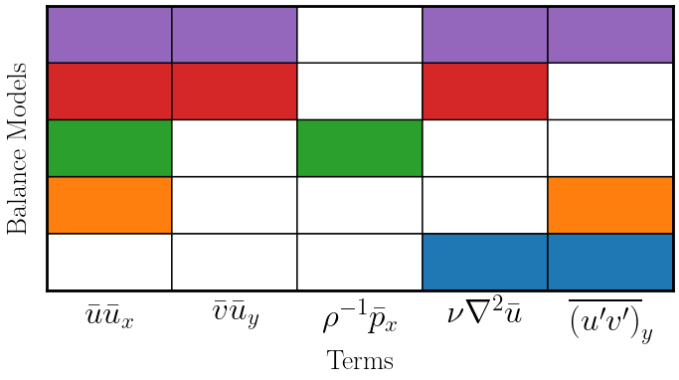


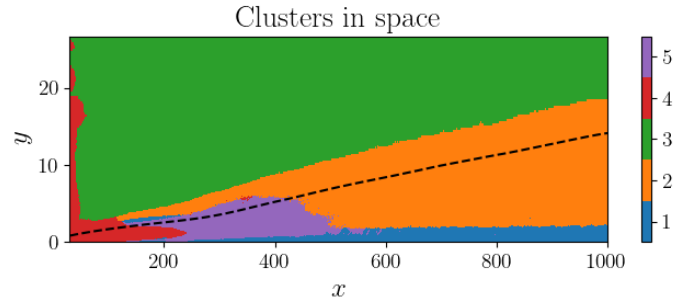
Figure 1: This plot shows the GMM clustered data in equation space for the turbulent boundary layer case. These specific plots are 2-D projections of the 6-D equation space, with the axes being the values of terms of the Reynolds-Averaged Navier-Stokes (RANS) equations. The different colours represent the different clusters identified by the GMM algorithm.

0.3 Conducted Research and Results

The Callaham et al. (2021) shared some code as runnable notebooks, and though there were some portability challenges like missing dependencies and data generation discrepancies, the code was successfully reproduced confirming the main results. Then, the turbulent boundary layer case was reproduced using alternative code primarily written in Pandas. The alternative code also made the use of a custom written GMM class. With that code, results sufficiently similar to the original were obtained, meaning the paper's results were not due to chance, and that the method, even when written differently does work well (see Fig. 3). Key steps included computing the terms of the Reynolds-Averaged Navier-Stokes (RANS) equations using numpy.gradient for derivatives and employing a Gaussian Mixture Model (GMM) for clustering. Despite minor differences in the clustering outcomes due to different implementations, the alternative code successfully replicated the identification of significant dynamics in the boundary layer.

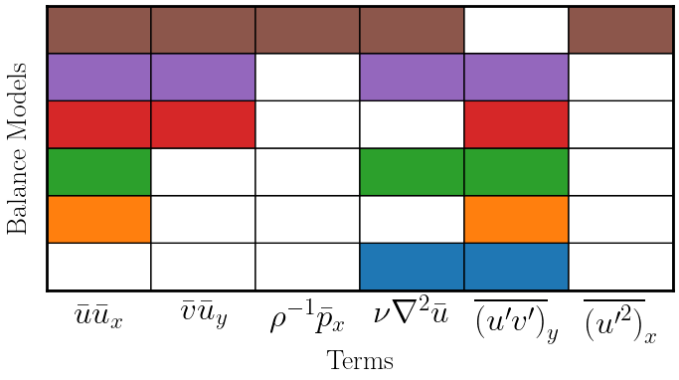


(a)

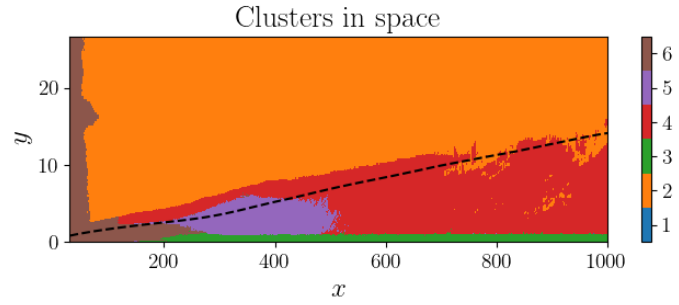


(b)

Figure 2: Plot of the unique balance models found after applying SPCA, when using the original Callaham et al. code, in grid form (a) and back in physical space (b). Color coding is consistent between the two subplots.



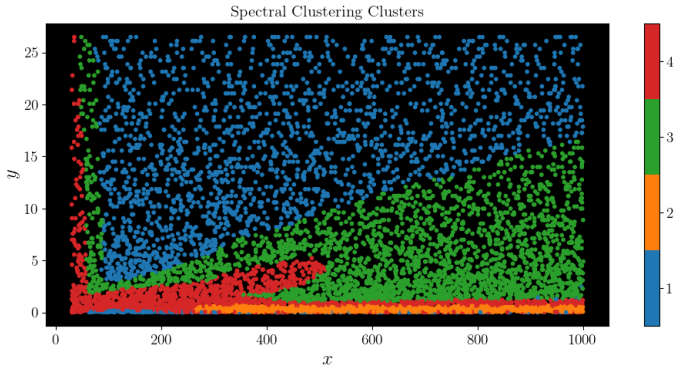
(a)



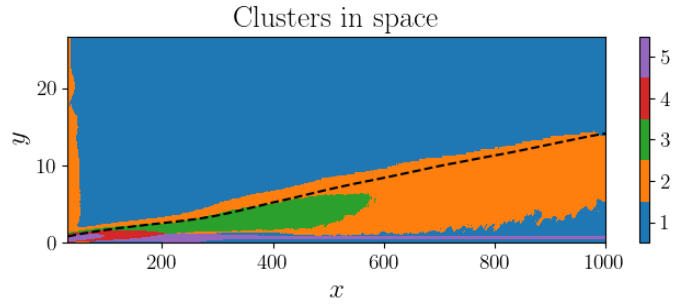
(b)

Figure 3: Plot of the unique balance models found after applying SPCA, when using the custom written alternative code, in grid form (a) and back in physical space (b). Color coding is consistent between the two subplots.

Testing the method with different clustering algorithms, the spectral clustering algorithm which was thought to be a good candidate [3, Supplementary Information] was found to have poor performance, especially in the context of practicality with the dataset sizes. K-Means and weighted K-Means algorithms were also tested, giving results that captured some of the expected dominant balances for the turbulent boundary layer case (see Fig. 4). Furthermore, the dominant balances found to be associated with the clusters were overall consistent with original results and fluid dynamics theory, with some exceptions.



(a)



(b)

Figure 4: Plot of the unique balance models found after applying SPCA, when using spectral clustering (a) and K-Means clustering (b).

Finally, the method was used under different hypeparameter values, and showed quite varying results. Putting this in the context of an unknown physical system and this poses a challenge in determining the best hyperparameters to use, as many combinations could be considered valid. This is the main limitation of the Callaham et al. (2021) method.

0.4 EIT

Finally, the method was applied to simulation data of elasto-inertial turbulence (EIT), a recently discovered flow [7]. The data used was of a simulation of one of the discovered coherent states of this flow [1], a Chaotic ARrowhead structure (CAR) (see Fig. 5).

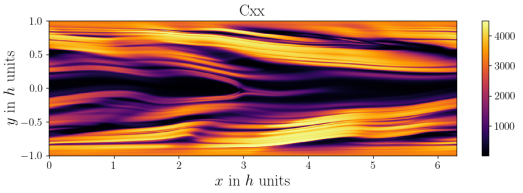


Figure 5: A Chaotic Arrowhead structure (CAR) in elasto-inertial turbulence, visualised with one of the components of the conformation tensor, C_{xx}.

In order to select the hyperparameters, the number of clusters was chosen based on when GMM identified clusters only had repeating covariance matrices (see Fig. 6). Then, results were obtained for multiple values of α. The results clearly showed there being 2 groups of dominant balances. The first only had inertial and pressure terms that were dominant, and the second had only viscous and elastic stresses terms active (see Fig. 7).

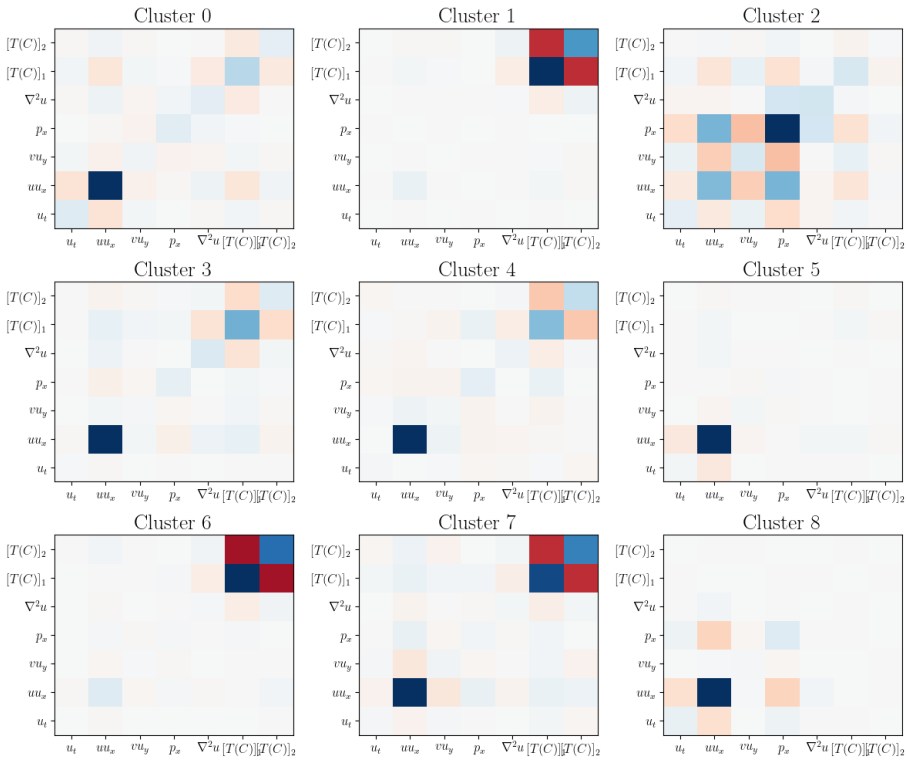
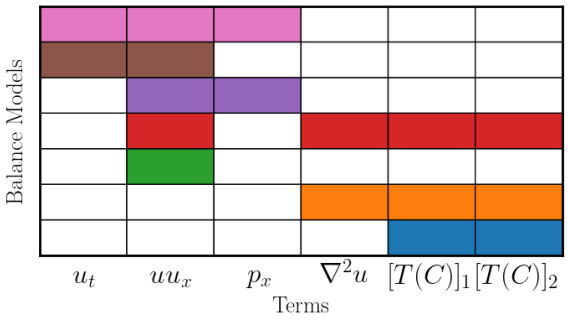
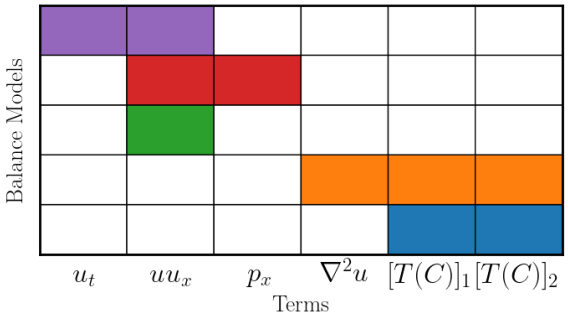


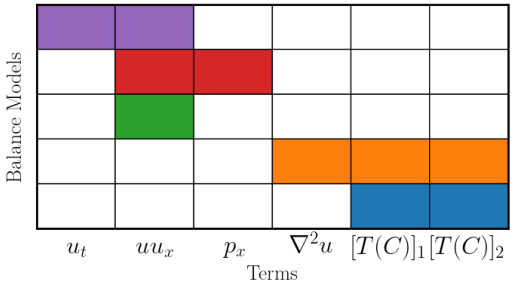
Figure 6: The covariance matrices of the 9 clusters identified by the GMM algorithm for the EIT case. As can be seen, multiple of the clusters have very similar covariance matrices, indicating that they are likely clusters with the same dominant balance, and the data has been sufficiently clustered.



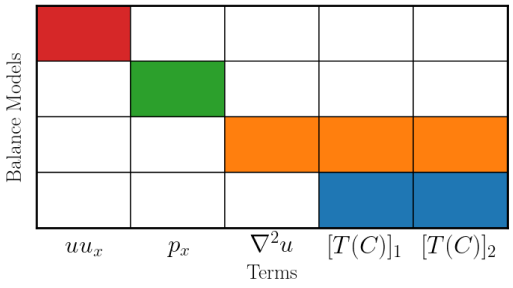
(a)



(b)



(c)



(d)

Figure 7: Plot of the unique balance models found for 9 clusters, and α values of 1.25 (a), 1.5 (b), 1.75 (c), and 2 (d).

Bibliography

[1] Miguel Beneitez, Jacob Page, Yves Dubief, and Rich R. Kerswell. Multistability of elasto-inertial two-dimensional channel flow. *J. Fluid Mech.*, 981:A30, 2024.

[2] A. P. Burger. Scale consideration of planetary motions of the atmosphere. *Tellus*, 10(2):195–205, 1958.

[3] 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.

[4] J. G. Charney. The dynamics of long waves in a baroclinic westerly current. *Journal of Atmospheric Sciences*, 4:136–162, 1947.

[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] Norman A. Phillips. Geostrophic motion. *Reviews of Geophysics*, 1:123–176, 1963.

[7] Devranjan Samanta, Yves Dubief, Markus Holzner, Christof Schäfer, Alexander Morozov, Christian Wagner, and Björn Hof. Elasto-inertial turbulence. *Proceedings of the National Academy of Sciences*, 110:10557 – 10562, 2012.

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

[9] Jun-Ichi Yano and M. Bonazzola. Scale analysis for large-scale tropical atmospheric dynamics. *Journal of Atmospheric Sciences*, 66:159–172, 2009.