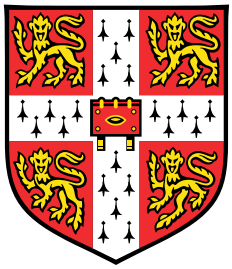
MPhil DIS Project 24

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)

This project aims to evaluate the reproducibility and robustness of the method proposed by Callaham et al. (2021) [1] for identifying dominant balances in complex physical systems. Dominant balance analysis simplifies equations involving numerous terms and complex differential equations, reducing computational costs while maintaining accuracy. Traditionally, this analysis requires significant expertise and manual effort. Callaham et al.'s novel approach leverages machine learning to automate this process. By replicating their study and applying their methods to new datasets, this project assesses both the accuracy and generalizability of their approach, contributing to the validation and potential enhancement of machine learning techniques in physical sciences.

0.2 Objectives and Motivation (150 words)

Explain the importance of reproducibility in scientific research. Highlight the problem of complex differential equations in engineering and physical sciences. Discuss the motivation to simplify these equations using dominant balance or scale analysis and the novel approach proposed by Callaham et al.

0.3 Methodology (250 words)

Summarize the three-step method proposed by Callaham et al.: Data & Equation Space Representation Gaussian Mixture Model (GMM) Clustering Sparse Principal Component Analysis (SPCA) Briefly describe how each step contributes to identifying dominant balance models. Mention any modifications or additional algorithms tested, such as Spectral Clustering, K-Means, and Weighted K-Means.

Figures to include:

Figure 4.1: Example of a Gaussian Mixture Model fit to data. Figure 4.2: Example of a 2D dataset projected onto its first 2 principal components.

0.4 Conducted Research and Results (350 words)

Discuss the portability and reproducibility of the code, including any issues encountered and resolved. Detail the turbulent boundary layer case: How the original and alternative codes were used to reproduce the results. Key findings from the reproduction of the RANS equation's terms and GMM clustering. Results from applying SPCA to identify active terms. Summarize the exploration of other algorithms: Results from Spectral Clustering, K-Means, and Weighted K-Means. Highlight the stability assessment: Impact of different numbers of clusters and alpha values on the results. Effect of training set size on the identified balance models.

Figures to include:

Figure 5.1: Plot of the 6 terms in the RANS equation using the original Callaham code. Figure 5.2: Plot of the 6 terms in the RANS equation using alternative code. Figure 5.3: Covariance matrices for each cluster found by the sklearn and custom GMM algorithms. Figure 5.4: Plot of the clusters found by the sklearn and custom GMM algorithms. Figure 5.5: Plot of unique balance models found after applying SPCA with the original and alternative code. Figure 5.6: Plot of unique balance model clusters in physical space with the original and alternative code.

0.5 Application to Elasto-Inertial Turbulence (100 Words)

Briefly describe the application of the method to a new dataset involving elasto-inertial turbulence. Highlight the potential of the method to uncover new dominant balance regimes in complex flows.

Figures to include:

Figure 6.1: Plot of the attractors found in the DNS of the FENE-P model.

0.6 Conclusion (100 Words)

Summarize the key findings and their implications. Reflect on the strengths and limitations of the method. Suggest potential future work or improvements.

Bibliography

[1] 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.