Hessisches Kompetenzzentrum
für Hochleistungsrechnen

Ultra-High-Dimensional Variable Selection in Genome-Wide Association Studies



Project Manager
Dr. Jasin Machkour

Researchers
Taulant Koka, Fabian Scheidt and Helena Mehler

Principal Investigator
Prof. Dr. Michael Muma

Project Term
2023 - 2024

Institute
Robust Data Science

University
Technische Universität Darmstadt

Introduction

This research project tackles the significant challenge of identifying genuine variables associated with diseases from an extensive pool of candidates. In genome-wide association studies (GWAS), researchers aim to uncover genetic variants linked to various health conditions, often sifting through millions of potential candidates. Accurately distinguishing true associations from false positives is essential, as false discoveries can waste resources and mislead scientific understanding. To effectively address this issue, high-performance computing (HPC) plays a critical role in analyzing and processing large datasets. By leveraging HPC, researchers can manage vast genomic datasets efficiently, ensuring rigorous statistical analysis while controlling the false discovery rate (FDR). This is essential for advancing precision medicine and improving our understanding of rare diseases.

Methods

Central to this project is the development of the T-Rex selector, an innovative framework designed for high-dimensional variable selection. This method enables researchers to identify relevant genetic variants without falling victim to the challenges of false discoveries. The T-Rex selector employs advanced algorithms that ensure reproducibility in large-scale, high-dimensional environments. It is specifically tailored for handling the complexities of genomic data, allowing for the analysis of datasets that can be hundreds of gigabytes in size. To address

the limitations imposed by traditional computing resources, the T-Rex selector utilizes memory mapping techniques. This approach allows the storage and processing of data on SSDs rather than relying solely on limited RAM. By processing data in an online fashion, the T-Rex selector efficiently manages memory consumption, making it possible to conduct multiple GWAS simultaneously. The analysis was conducted using the R programming language, where we developed our own software packages, TRexSelector and tlars, which were published on CRAN. These packages encapsulate the methodologies we have devised, allowing for broader accessibility and use in the scientific community.

Results

Over the past year, the project has made significant progress, achieving
key milestones:

• **Performing GWAS**: We successfully acquired UK Biobank data and established a robust pipeline for managing this extensive dataset. The T-Rex selector was optimized to handle the massive volume of genomic data, enabling GWAS for thousands of phenotypes.
• **Extending the T-Rex Framework**: The framework has been expanded to integrate additional forward selection methods, including the Elastic Net. This enhancement improves the power of variable selection while maintaining control over the false discovery rate. By accommodating a broader range of statistical approaches, the T-Rex selector enhances its utility in genomic studies.
• **Sparse Principal Component Analysis (PCA)**: The project successfully incorporated sparse PCA into the T-Rex framework, allowing for unsupervised learning tasks to be executed with FDR control. While validation through simulations has been achieved, some tasks remain pending completion due to the recent availability of necessary data.

These accomplishments establish a solid foundation for further analysis and exploration in the project’s subsequent phases.

Discussion

The results underscore the transformative potential of the T-Rex selector in genomic research. By facilitating the analysis of vast datasets while controlling for false discoveries, the framework paves the way for more accurate and reproducible findings in GWAS. The integration of advanced variable selection methods enhances the reliability of the results, providing researchers with a powerful tool for precision medicine. Looking ahead, further analysis of the UK Biobank data will deepen our understanding of genetic associations with diseases, particularly those that are less common. This ongoing research is expected to contribute significantly to the field of genomics, fostering collaborations with computational medicine teams and improving our collective ability to address health challenges.

Publications

Machkour, J.; Muma, M., Palomar, D. P.:“False Discovery Rate Control for Fast Screening of Large-Scale Genomics Biobanks,” 2023 IEEE Statistical Signal Processing Workshop (SSP), Hanoi, Vietnam, 2023, pp. 666-670. https://doi.org/10.1109/SSP53291.2023.10207957

Machkour, J.; Breloy, A.; Muma, M.; Palomar, D.P.; Pascal, F.:“Sparse PCA with False Discovery Rate Controlled Variable Selection,” ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 2024, pp. 9716-9720. https://arxiv.org/pdf/2401.08375

Machkour, J.; Muma, M.; Palomar, D.P.:“The Informed Elastic Net for Fast Grouped Variable Selection and FDR Control in Genomics Research,” 2023 IEEE 9th InternationalWorkshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Herradura, Costa Rica, 2023, pp. 466-470. https://arxiv.org/pdf/2410.05211

Scheidt, F.; Machkour, J.; Muma, M.:“Solving FDR-Controlled Sparse Regression Problems with Five Million Variables on a Laptop,” 2023 IEEE 9th InternationalWorkshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), Herradura, Costa Rica, 2023, pp. 116-120. https://arxiv.org/pdf/2409.19088

Machkour, J.; Tien, S.; Palomar, D. P.; Muma, M.:"TRexSelector: T-Rex Selector: High-Dimensional Variable Selection & FDR Control", 2024, R package version 1.0.0. https://CRAN.R-project.org/package=TRexSelector

Machkour, J.; Tien, S.; Palomar, D. P.; Muma, M.:"tlars: The T-LARS Algorithm: Early-Terminated Forward Variable Selection", 2024, R package version 1.0.1. https://CRAN.R-project.org/package=tlars

Reference

Machkour, J.; Muma, M.; Palomar, D.P.:“False discovery rate control for grouped variable selection in high-dimensional linear models using the T-Knock filter,” in 30th Eur. Signal Process. Conf. (EUSIPCO), 2022, pp. 892–896. https://doi.org/10.23919/EUSIPCO55093.2022.9909883

Last Update: 2025-01-10 16:59