technische universität dortmund

fakultät statistik

Evaluating novel methods for polygenic risk score estimation using individual-level SNP data from the UK Biobank

Prof. Dr. Katja Ickstadt and JProf. Dr. Christian Staerk

March 12, 2025

Problem

Polygenic risk scores (PRS) play a crucial role in predicting genetic susceptibility to complex diseases by leveraging single nucleotide polymorphism (SNP) data. The aim of this Master thesis project is to compare and evaluate novel methodologies for PRS estimation using large individual-level data from the UK Biobank. A specific focus will be on `snpboost` based on statistical boosting (Klinkhammer et al., 2023) and `snpnet` based on ℓ_1-regularization via the Lasso (Li et al., 2020), in comparison with Cross-Leverage Score methods based on sliding window and sketching approaches (Teschke et al., 2024). These approaches aim to estimate PRS while utilizing only a small fraction of the genetic variants compared to traditional methods. This research assesses these methods based on predictive accuracy, computational efficiency, and robustness to different ancestries. The models are trained on UK Biobank data, and their performance is analyzed across different phenotypes. Examples include standing height, LDL-cholesterol, blood glucose level, lipoprotein A, and BMI (Tanigawa et al., 2022). By systematically comparing these approaches, this Master thesis aims to provide valuable insights into the strengths and limitations of different PRS estimation techniques, guiding future genomic studies and applications in precision medicine.

Keywords: Polygenic risk scores, SNP data, UK Biobank, snpboost, snpnet, Cross-Leverage Score, Genomics

References

Klinkhammer, H., Staerk, C., Maj, C., Krawitz, P. M., and Mayr, A. (2023). A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Frontiers in Genetics*, 13.

Li, R., Chang, C., Justesen, J. M., Tanigawa, Y., Qian, J., Hastie, T., Rivas, M. A., and Tibshirani, R. (2020). Fast lasso method for large-scale and ultrahigh-dimensional cox model with applications to uk biobank. *Biostatistics*, 23(2):522–540.

Tanigawa, Y., Qian, J., Venkataraman, G., Justesen, J. M., Li, R., Tibshirani, R., Hastie, T., and Rivas, M. A. (2022). Significant sparse polygenic risk scores across 813 traits in uk biobank. *PLOS Genetics*, 18(3):1–21.

Teschke, S., Ickstadt, K., and Munteanu, A. (2024). Detecting interactions in high-dimensional data using cross leverage scores. *Biometrical Journal*, 66(8):e70014.