LEIBNIZ RESEARCH CENTRE
FOR WORKING ENVIRONMENT
AND HUMAN FACTORS
AT THE TU DORTMUND

technische universität
dortmund

Master’s thesis (Statistics and /or Data Science)

Topic: Advancing Model Selection Methods for Extended Maximal Interaction Two-Mode Clustering

This project aims to address a critical gap in the extended maximal interaction two-mode clustering (E-ReMI) methodology: the development of robust model selection techniques to determine the optimal number of row and column clusters $(P; Q)$ in complex two-mode data. Current model selection criteria, such as the Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC), incorporate penalties for model complexity but there is no clear way to measure the complexity of clustering algorithms using classification likelihood. Specifically, the challenge lies in accounting for the unknown 0/1 cell entries of the $I\times P$ row partitioning matrix Z and the $J\times Q$ column partitioning matrix K in the penalty.

The project proposes two innovative approaches to address this challenge. The first involves extending the work of Hofmeyr (2020), who derived approximations for the effective degrees of freedom in k-means clustering, to E-ReMI model. The second approach makes use of a Bayesian framework, with P and Q being estimated parameters. This approach seeks to incorporate model selection directly into the clustering procedure by making use of mixture models with priors on the number of mixture components, as outlined by Miller and Harrison (2018).

This research will provide practical tools for determining the optimal structure in two-mode data, hence improving the applicability of maximal interaction clustering across several domains, including neuroscience, social sciences, and genomics. The project's outcomes will enhance theoretical underpinnings and provide practical insights for real-world data analysis, fostering advancements in probabilistic model-based clustering methodologies.

Key Requirements

- Master’s students in Statistics or Data Science with a strong foundation in **Mathematical Statistics**, particularly in **likelihood estimation**, and **inference**.
- Proficiency in **R and/or Python**, with experience in statistical computing and data analysis.
- Willingness to present findings at research meetings and contribute to academic publications.

References

1. Ahmed, Z., Cassese, A., van Breukelen, G., and Schepers, J. (2023). E-ReMI: Extended maximal interaction two-mode clustering, *Journal of Classification*, 40, 298-331.
2. Hofmeyr, D. P. (2020). Degrees of freedom and model selection for k-means clustering, *Computational Statistics & Data Analysis*, 149, 106974.
3. Miller, J. W., and Harrison, M. T. (2018). Mixture Models with a Prior on the Number of Components, *Journal of the American Statistical Association*, 113, 340-356.

Contact: Dr. Zaheer Ahmed (ahmed@ifado.de) or Prof. Dr. Katja Ickstadt (ickstadt@statistik.tu-dortmund.de)

Leibniz
Association