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

Topic: Dealing with missing data in Extended Maximal Interaction Two-Mode Clustering (E-ReMI)

Many real-world studies, such as those exploring how people react in different situations, may have missing data. For instance, participants in a study might skip certain tasks or refuse to answer specific questions, resulting missingness in the dataset. This hinders clustering methods like E-ReMI, which require a complete data matrix. This project proposes extending E-ReMI by incorporating classical statistical methods to handle missing data under the Missing Completely at Random (MCAR), where missingness does not depend on observed or unobserved data. This extension will rely on classical methods like the Expectation-Maximization (EM) algorithm and Multiple Imputation (MI). However, a Bayesian approach incorporates prior infomation and uncertainty into the model to treat missingness probabilistically.
Using the MCAR assumption, we propose a Bayesian extension to E-ReMI to handle missing data. Bayesian approaches to handle missing data as a latent variables with prior distributions. Markov Chain Monte Carlo (MCMC) is used to iteratively update missing data and model parameter estimates. The Bayesian approaches considered in this project include Bayesian Likelihood with Missing Data Terms and Bayesian Multiple Imputation. The Bayesian extension not only addresses missing data but also improves the robustness of clustering results by incorporating prior knowledge and probabilistic reasoning. This makes it particularly well-suited for complex datasets with inherent uncertainties.

Key Requirements

- Master's students in Statistics or Data Science with a strong foundation in **Mathematical Statistics**, particularly in **likelihood estimation**, and **Bayesian 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. Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39 (1), 1-38.
3. Rubin, D. B. (1996). Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91 (434), 473-489.

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