Outlier Detection and Missing Value Imputation using Mixed Graphical Models – A Metabolomics Use Case

Jonas Leins[1], *Michael Altenbuchinger*[2], *Helena U. Zacharias*[1]

[1] Peter L. Reichertz Institute for Medical Informatics of TU Braunschweig and Hannover Medical School, Hannover Medical School, Hannover, Germany.

[2] Department of Medical Bioinformatics, University Medical Center Göttingen, Göttingen, Germany

Metabolomics data are high-dimensional and typically originate from Mass Spectrometry (MS) or Nuclear Magnetic Resonance (NMR) measurements. Given the highly complex technical nature of these measurements, the occurrence of outliers or missing values—often resulting from errors in the data acquisition process—is frequent and significantly challenges downstream data analyses. We are developing a data-driven, network-based approach for automatic outlier detection and missing value imputation using Mixed Graphical Models. This approach allows the estimation of the probability of a metabolite's concentration in an individual sample taking on a specific value by considering all other metabolite concentration values as well as additionally available clinical measurements in that particular sample. Using this approach, we aim to identify outlier values that do not fit within the sampling cohort, as well as replace missing values with more plausible values mathematically inferred from the available metabolomics data. This network-based approach not only improves the overall interpretability of downstream workflows but also enables the incorporation of clinical data and other information layers into the imputation process.