Advances in Clustering of Incomplete Multivariate Data with Skewed and Heavy-Tailed Clusters

Andriëtte Bekker
University of Pretoria, South Africa

This presentation focuses on the significant challenge of clustering incomplete multivariate data, particularly in the presence of skewed, heavy-tailed clusters and mild outliers. Our approach is the first to accommodate missing data under a missing at random (MAR) mechanism for the full finite mixture of scale mixtures of multivariate skew-normal (FMSMSN) family, extending previous work that focused on more restrictive special cases. To mitigate the effect of possible atypical observations within an incomplete dataset, a heavy-tailed extension, the contaminated multivariate skew-normal distribution, will receive attention. We derive the distributional properties of the missing components and develop an augmented EM-type algorithm with closed-form conditional expectations for the E-step. Simulation studies demonstrate strong clustering performance and accurate parameter recovery across varying levels of missingness, sample sizes, and cluster overlap. We also present a real-data application to global CO_2 emissions, illustrating the practical advantages of our method.