Working Group on Statistical Learning Seminar

Title: Missing value imputation using a robust infinite factor model with application to metabolomic data

Speaker: Kate Finucane (University College Dublin)

Date: Thu 28th September 2023 at 3:00PM

Location: E0.32 (beside Pi restaurant)

Abstract: Missing data are present in metabolomic mass spectrometry data by two mechanisms: Missing at random (MAR) missingness occurs at random throughout the data due to various reasons including acquisition and processing error, while missing not at random (MNAR) missingness occurs due to data below the threshold for accurate detection. Imputation of missing values is important as many data analysis methods in metabolomic research require a full set of data. Uninformed imputation methods, such as zero-imputation or mean-imputation, can bias any further analysis and as such a data-informed imputation method is preferred. A sparse Bayesian infinite factor model (IFA) has previously been developed to handle the high-dimensionality of the spectrometry data and to impute missing data entries, with separate imputation methods for MAR and MNAR data handled elegantly by the Bayesian framework. While typically in IFA factors are assumed to be multivariate Gaussian, in cases of heavier-tailed data this assumption can result in poor model fit. Therefore, here a robust infinite factor model (rIFA) is developed where factors are assumed to be multivariate t-distributed. The ability to impute MAR and MNAR data separately is maintained. Motivation for the robust model is established as the IFA model's performance degrades when applied to simulated t-distributed data compared to normal data, leading to reduced correlation between true and imputed

data. Robust model performance is demonstrated through simulation studies where the extent and type of missingness is varied.