

Statistics and Data Science Seminar

Truncated latent gaussian copula model for zero-inflated data

Irina Gaynanova (Texas A&M University)

Abstract: A great number of multivariate statistical methods, such as principal component analysis, discriminant analysis, canonical correlation analysis and graphical lasso to name a few, require the estimate of covariance or correlation matrix of variables as one of the inputs. It is typical to use Pearson sample correlation matrix, which works well at capturing dependencies between normally distributed variables. In this work we consider the problem of estimating dependencies between zero-inflated measurements, which arise in miRNA data, microbiome data, physical activity data, etc. We propose truncated latent Gaussian copula to model the data with excess zeroes, which allows us to derive a rank-based estimator of latent correlation matrix without the estimation of marginal transformation functions. The new methodology is applied for the analysis of associations between gene expression and microRNA data of breast cancer patients, and for inferring the conditional independence graph in quantitate gut microbiome data.

Wednesday, October 28 at 4:00 PM in Zoom