A general framework for the association analysis of heterogeneous data

Multivariate association analysis is of primary interest in many applications. Despite the prevalence of high-dimensional and non-Gaussian data (such as count-valued or binary), most existing methods only apply to low-dimensional datasets with continuous measurements. Motivated by the Computer Audition Lab 500-song (CAL500) music annotation study, we develop a new framework for the association analysis of two sets of high-dimensional and heterogeneous (continuous/binary/count) data. We model heterogeneous random variables using exponential family distributions, and exploit a structured decomposition of the underlying natural parameter matrices to identify shared and individual patterns for two datasets. We also introduce a new measure of the strength of association, and a permutation-based procedure to test its significance. An alternating iteratively reweighted least squares algorithm is devised for model fitting, and several variants are developed to expedite computation and achieve variable selection. The application to the CAL500 data sheds light on the relationship between acoustic features and semantic annotations, and provides an effective means for automatic annotation and music retrieval.

Wednesday, March 22, 2017
Time: 3:00 – 4:00 PM
Location: CAB 657