Nonparametric Bayes tensor factorizations for big data

David B. Dunson, Duke University

In modern applications, it is routine to collect big data from a variety of modalities. For example, one may obtain video, images, text, real observations, non-Euclidean data from networks, shapes, etc. There is a need for general purpose nonparametric Bayes models and computational algorithms that can flexibly characterize dependence in huge data sets, while favoring a low-dimensional representation and computational scaling. One relatively simple, but nonetheless extremely important example, corresponds to huge sparse contingency tables defined by high-dimensional categorical variables. Initially focusing on characterizing dependence in multiway tables and classification problems, we propose novel classes of nonparametric Bayes tensor factorizations for big multivariate categorical data, while also developing efficient computational algorithms. Strong theoretical results are provided on rates of convergence including in settings in which the dimension increases with sample size exponentially. These approaches are shown to have an important applied impact in genomics applications. (Time permitting) Generalizations to accommodate mixed measurement scale data, and general nonparametric Bayesian joint modeling of multimodal data are described.