Mathematics, Statistics, and Computer Science **@ UIC**

Statistics and Data Science Seminar

CESME: High-Dimensional Clustering via Latent Semiparametric Mixture Models

Boxiang Wang (University of Iowa)

Abstract: Cluster analysis is a fundamental task in machine learning. Several clustering algorithms have been extended to handle high-dimensional data by incorporating a sparsity constraint in the estimation of a mixture of Gaussian models. Though it makes some neat theoretical analysis possible, this type of approach is arguably restrictive for many applications. In this talk, I will introduce a novel latent variable transformation mixture model for clustering in which a mixture of Gaussians is assumed after some unknown monotone data transformation. A new clustering algorithm named CESME is developed for high-dimensional clustering under the assumption that optimal clustering admits a sparsity structure. The use of unspecified transformation makes the model far more flexible than the classical mixture of Gaussians. On the other hand, the transformation also brings quite a few technical challenges to the model estimation as well as the theoretical analysis of CESME. I will present a comprehensive analysis of CESME including identifiability, initialization, algorithmic convergence, and statistical guarantees on clustering. In addition, the convergence analysis has revealed an interesting algorithmic phase transition for CESME, which has also been noted for the EM algorithm in the literature. Leveraging such a transition, a data-adaptive procedure is developed and substantially improves the computational efficiency of CESME. Extensive numerical study and real data analysis show that CESME outperforms the existing high-dimensional clustering algorithms including CHIME, sparse spectral clustering, sparse K-means, sparse convex clustering, and IF-PCA.

Wednesday, April 12 at 4:00 PM in 636 SEO