

Nonparametric Clustering on Mixtures of Functional Data

Haiyan WANG
Kansas State University

Abstract: This talk is based on joint work with James Neill and Forrest Miller.

In this talk, I will present a method for effectively detecting unknown clusters in high-dimensional functional data. Examples of such data include gene expression levels measured over time from microarray experiments, functional magnetic resonance imaging (fMRI), mass spectrometry data from proteomics, lipidomics, etc. We define clusters through the unknown high-dimensional multivariate distributions of all observations. Kullback-Leibler information and Mahalanobis generalized squared distance can fail to provide meaningful measures of distance between distributions in such high-dimensional settings.

We propose a new similarity measure and an agglomerative clustering algorithm, called PCLUST, to effectively differentiate among high-dimensional populations. The algorithm produces invariant results under monotone transformations of data and does not require users to specify the number of clusters. Simulations show that PCLUST significantly outperforms nine other popular algorithms in both clustering accuracy and robustness. An application in identifying biomarkers using time course gene expression data from Arabidopsis in response to environmental stresses is illustrated.