Hierarchical Modal Clustering based on the Topography of High-dimensional Mixtures

Surajit Ray
Department of Biostatistics
University of North Carolina, Chapel Hill

Wednesday, March 22, 2006, 4:00 - 5:00pm (NOTE: SPECIAL DAY!)
MCS Room B33 (Tea and Cookies at 3:30pm in MCS 153)

Abstract: With the advent of new high throughput technologies in scientific areas such as medical imaging and genomics, the need for efficient data analyses is ever increasing. This necessitates the development of statistical methodology in a non-standard setup, often referred to as the high dimensional low sample size (HDLSS) setup. In this talk I will propose new methodologies for addressing two important aspects of analyzing high-dimensional data: (i) Finding contextually relevant partitions and dimension reduction, (ii) Feature extraction and asymptotics for HDLSS data.

Multivariate mixtures provide flexible methods for both fitting and partitioning high-dimensional data. But in reality the true clusters may not arise from standard statistical distributions. We propose a new concept of modal clusters, where we start from a conventional mixture analysis or a kernel based density estimator, and cluster together those components whose contributions are actually unimodal. Ray and Lindsay (2005) show that the topography of multivariate mixtures, can be analyzed rigorously in lower dimensions by use of a ridgeline manifold that contains all critical points as well as the ridges of the density. Based on this fundamental result, we have developed a comprehensive modal clustering technique, which uses a MM algorithm (generalized version of EM) to find the modes within the ridge line manifold. Additionally, as different levels of smoothing provide different aggregations of data points, modal clustering also lays the foundation for model based hierarchical clustering.

In the last part of my talk I will demonstrate a new technique for feature extraction in HDLSS setup. This is based on the geometry and asymptotics of large dimension and small sample size. Further, based on the special geometric structure imposed by these datasets, we propose a novel technique for model-based bi-clustering. Performance of these two newly proposed methods will be demonstrated through applications on medical image segmentation, clustering of gene expression data and other simulated data sets.

For directions and maps, please see http://math.bu.edu/research/statistics/statseminar.html.