

STATISTICS COLLOQUIUM

MONDAY, APRIL 30, 2007

TALK: 4:00 PM — SCIENCE CENTER RM. B-10

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

“Modal Inference and its application to High-Dimensional Clustering”

Surajit Ray

Department of Mathematics and Statistics

Boston University

ABSTRACT

Multivariate mixtures provide flexible methods for both fitting and partitioning high-dimensional data. Ray and Lindsay (2005) show that the topography of multivariate mixtures, in the sense of their key features as a density, 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. To use this rich feature for data analysis we first construct an extension of EM algorithm that can be used to find the modes of a mixture density. Even in very high dimensions the computational complexity of our EM algorithm is extremely low. In addition, the method of steepest ascent can be used to assign the individual data points to modes, providing a clustering of data points through their modal association. These tools can be used in various ways. For one, we can take a conventional mixture analysis and cluster together those components whose contribution is actually unimodal. This cluster could then represent a single true component with a more complex distribution. We can also turn kernel density estimation into clustering tool in which the data points become identified with each other by their association with a common mode of the density estimator. If in addition we let the bandwidth parameter go from 0 to infinity, we can construct a hierarchical clustering of the data points. In addition to providing satisfying clustering results that lie somewhere between clustering algorithms and a formal mixture analysis, the estimation method raises interesting inferential questions that lie somewhere between the two points of view. Application of modal clustering will be discussed in the context of image segmentation.