

RUTGERS UNIVERSITY
DEPARTMENT OF STATISTICS AND BIostatISTICS
HILL CENTER #501, BUSCH CAMPUS, PISCATAWAY

www.stat.rutgers.edu

Seminar

Speaker: Tao Shi, Department of Statistics, Ohio State University

Title: Data Spectroscopy: Eigenspace of Convolution Operators and Clustering

Date: Wednesday October 22, 2008

Time: 3:20 PM

Place: 552 Hill Center

Abstract

In this talk, we focus on obtaining clustering information in a distribution when iid data are given. First, we develop theoretical results for understanding and using clustering information contained in the eigenvectors of data adjacency matrices based on a radial kernel function (with a sufficiently fast tail decay). We study which eigenvectors should be used and when the clustering information for the distribution can be recovered from the data. Second, we use heuristics from these analyses to design the Data Spectroscopic clustering (DaSpec) algorithm. Our findings not only extend and go beyond the intuitions underlying existing spectral techniques (e.g. spectral clustering and Kernel Principal Components Analysis), but also provide insights about their usability and modes of failure. Simulation studies and experiments on real world data are conducted to show the promise of our proposed data spectroscopy clustering algorithm relative to k-means and one spectral method. In particular, DaSpec seems to be able to handle unbalanced groups and recover clusters of different shapes better than competing methods.

This is joint work with Prof. Mikhail Belkin (Ohio State University) and Prof. Bin Yu (University of California, Berkeley).