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

www.stat.rutgers.edu

Seminar

Speaker: David Tyler, Rutgers University

Title: Invariant Coordinate Selection: A new method for exploring multivariate data

Date: Wednesday September 30, 2009

Time: 3:20 PM

Place: 552 Hill Center

Abstract

When sampling from a multivariate normal distribution, the sample mean vector and sample variance-covariance matrix are a sufficient summary of the data set. To protect against nonnormality, and in particular against longer tailed distributions and outliers, one can replace the sample mean and covariance matrix with robust estimates of multivariate location and scatter. Outliers can often be detected by examining the corresponding robust Mahalonobis distances. Such an approach is appropriate if the bulk of the data arises from a multivariate normal distribution or more generally from an elliptically symmetric distribution. However, if the data arises otherwise, then different location/scatter estimates do no estimate the same population quantities, but rather are reflecting different aspects of the underlying distribution. This suggests comparing different estimates of multivariate location/scatter may reveal interesting structures in the data, ones which may not be apparent from a plot of robust Mahalonobis distances.

In this talk, new multivariate methods based upon the comparison of different estimates of multivariate scatter are introduced. These methods are based on the eigenvalue-eigenvector decomposition of one estimate of scatter relative to another. An important property of this decomposition is that the corresponding eigenvectors generate an affine invariant coordinate system (ICS) for the multivariate data. Consequently, this leads to the development of a wide class of affine equivariant coordinatewise multivariate methods. In particular, by plotting the data with respect to this new invariant coordinate system, various data structures can be revealed. Under certain independent component analysis models, which are currently popular within computer science and engineering disciplines, the invariant coordinates correspond to the independent components. When the data comes from a mixture of elliptical distributions, a subset of the invariant coordinates correspond to Fishers linear discriminant subspace, even though the class identification of the data points are unknown. Several examples are given to illustrate the utility of the proposed methods.