Data Mining by Skewering a Data Cloud

David W. Scott (speaker) *Rice University, USA*, E-mail: scottdw@rice.edu

Key words: Clustering, eigenvector, robust covariance matrix....

Clustering p-dimensional data by fitting a mixture of K normals has enjoyed renewed interest (see, for example, the Splus function *mclust* [1]). However, the number of parameters for the model grows rapidly with dimension p. Even if all the covariance matrices are assumed to be equal, the number of parameters is $(K-1)+K\,p+p(p+1)/2$ for the weights, means and covariance matrix. Scott [2] introduced the partial mixture component algorithm which fits a single component of the mixture model at a time. This algorithm requires only $1+p+p\,(p+1)/2$ parameters for the weight, mean vector, and covariance matrix. In this talk, we introduce a new algorithm which attempts to find the "best line" through an individual cluster. This model requires only $2\,p$ parameters. In particular, the new algorithm is linear rather than quadratic in p. By repeatedly re-initializing the search algorithm, many clusters may be identified. Intuitively, the line found is approximately the eigenvector corresponding to the largest eigenvalue of the local covariance matrix. The GGobi visualization program will be used to illustrate the success of this algorithm on real and simulated data.

References

[1] Banfield, J.D. and Raftery, A.E. (1993), "Model-Based Gaussian and Non-Gaussian Clustering," *Biometrics*, 49, 803–822.
[2] Scott, D.W. (2001), "Parametric Statistical Modeling by Minimum Integrated Square Error," *Technometrics* , 43, 274–285.