November 27, 2000

MSRI Extended Abstract

David W. Scott
Dept Statistics, MS-138
Rice University
POB 1892
Houston, TX 77005-1892

scottdw@rice.edu

713-348-6037
713-348-5476 (F)

Title: Clustering by Partial Mixture Estimation

Abstract:

The identification of components of a mixture estimate as clusters in a population is a powerful model for uncovering natural groups. Many well- accepted clustering algorithms lack any probabilistic underpinnings, making inference difficult. In recent work, we (Scott, D.W. 1999, "Remarks on Fitting and Interpreting Mixture Models," Computing Science and Statistics, K. Berk and M. Pourahmadi, Eds., 31, 104–109) described an alternative to the EM algorithm for fitting mixtures based on a formula which provides an unbiased estimate of risk as measured by integrated square error, $\int (\hat{f}(x|\theta) - f(x))^2 dx$. With either approach, an incorrect specification of the mixing density leads to a single cluster being represented by two or more mixture components, making interpretation more difficult.

An interesting and novel feature of the integrated squared error approach is that $\hat{f}(x|\theta)$ need not be a complete density. In particular, $\hat{f}(x|\theta)$ can be a single component of the mixture density, for example, $\hat{f} = wN(\mu, \sigma^2)$. Random starts permit a complete exploration of the cluster structure in many cases, even when the mixture component is misspecified. The approach extends easily to the multivariate case. When clusters are not well-separated, then a subset of the parameters may be estimated to discover the cluster structure.

A graphical brushing algorithm exists to indicate which clusters are noise and which are not. Illustrations of the methodology will be given.