Clustering High-Dimensional Data Using Evidence of Multimodality

Peter Hall
University of Melbourne
Department of Mathematics and Statistics
Melbourne, VIC 3010 Australia
halpstat@ms.unimelb.edu.au

Abstract

We suggest a nonparametric approach to clustering very high-dimensional data, designed particularly for problems where the mixture nature of a population is expressed through multimodality of its density. In such cases a technique based implicitly on mode-testing can be particularly effective. In principle, several alternative approaches could be used to assess the extent of multimodality, but in the present problem the excess mass method has important advantages. The resulting methodology for determining clusters is particularly effective in cases where the data are relatively heavy tailed or show a moderate to high degree of correlation, or when the number of important components is relatively small.