A Nonparametric Bayesian Model for Local Clustering with Application to Proteomics

Peter Müller

Austin, Texas

We propose a nonparametric Bayesian local clustering (NoB-LoC) approach for heterogeneous data. The NoB-LoC model defines local clusters as blocks of a two-dimensional data matrix and produces inference about these clusters as a nested bidirectional clustering. Using protein expression data as an example, the NoB-LoC model clusters proteins (columns) into protein sets and simultaneously creates multiple partitions of samples (rows), one for each protein set. In other words, the sample partitions are nested within the protein sets. Any pair of samples might belong to the same cluster for one protein set but not for another. These local features are different from features obtained by global clustering approaches such as hierarchical clustering, which create only one partition of samples that applies for all proteins in the data set. As an added and important feature, the NoB-LoC method probabilistically excludes sets of irrelevant proteins and samples that do not meaningfully co-cluster with other proteins and samples, thus improving the inference on the clustering of the remaining proteins and samples. Inference is guided by a joint probability model for all random elements. We provide extensive examples to demonstrate the unique features of the NoB-LoC model.