A Summary of "Combining Mixture Components for Clustering" by Baudry et al.

Branden Olson

Multivariate Gaussian mixtures can be used to cluster data. Baudry et al. introduce a routine that allows for the number of clusters and mixtures to differ by minimizing entropy. The method is shown to successfully identify particular cell subpopulations in patients.

Clustering is an important problem with many applications. Model-based clustering often assumes a mixture of multivariate Gaussian distributions. While previous methods impose that the number of mixtures equals the number of clusters, this is not always the case.

Baudry et al introduce a routine that combines mixture components based on minimizing an entropy criterion. The procedure yields a list of solutions that can be chosen subjectively, although the authors give a way to select automatically based on entropy.

Beyond a swath of successful simulation studies, the authors apply the method to biological data to identify particular cell subpopulations, which yields results consistent with biological expectations, unlike the others.