Clustering Methods and Correlated Data: Simulations

Parker Knight

9/29/2021

Overview

1. Review of methods
2. Simulation design
3. Results
4. Directions

Methods of Interest

1. K-means
2. Hierarchical clustering
3. Spectral clustering

K-means

Let $X \in \mathbb{R}^{n \times p}$ be a given data matrix. Our goal is to partition the n data points into k non-overlapping clusters, denoted by the set $S = \{S_1, S_2, ..., S_k\}$.

The k-means algorithms does so by minimizing the Euclidean distance between each point and the centroid of it's assigned cluster:

$$\min_{S} \sum_{i=1}^{k} \sum_{X_j \in S_i} ||X_j - \mu_i||_2^2$$

where μ_i is centroid of the points in S_i.

K-means (cont.)

Key points:

- k must be specified up front
- Computation is NP-hard (objective function is not convex)
- By default, uses Euclidean distances
  - alternate versions exist (K-medoids)
- Assumes separation between clusters is convex

Hierarchical clustering

Spectral clustering