Finding Groups in Data

Course objectives: The goal of this course will be to explore methods for finding groups in data. The focus will begin with a “regular” data sets (e.g., observations by variables) and discuss both traditional clustering techniques and more recent statistical approaches. Additionally, focus will be given to both algorithmic approaches (e.g., optimization algorithms) and statistical models (e.g., finite mixture models). In addition to classic data structures, methods for finding groups in the context of network data will also be explored (e.g., blockmodeling, community detection, etc.) as well as detecting structure in longitudinal and repeated measures data.

Finally, the course will make the connection between traditional latent variable models (e.g., factor analysis) and more recent statistical approaches (e.g., finite mixture modeling). Along the way, it is shown how the gap between the two can be bridged by traditional clustering approaches – and how these different approaches can lead to insights in different parts of the data.

Computing: MATLAB or R. Students will be provided with programs and code to conduct all analyses that are discussed. Additionally, each procedure will be demonstrated when presented.
Day 1: Introduction to Concepts

- **What is a group or cluster? What constitutes “structure?” Visualization Techniques (2 hours)**

Day 2: Mixture Models (e.g., Model-based Clustering)

- **Introduction to Finite Mixture Modeling (6 hours)**

This will cover everything from model estimation, to fit statistics, to model building approaches.

- **Connection between mixture modeling and latent profile analysis (previous day) and cluster analysis (next day) (2 hours)**

Day 3: Traditional Cluster Analysis

- **Hierarchical Clustering (2 hours)**

- **Minimum Variance Clustering (e.g., K-means Clustering) (2 hours)**

- **p-median (L1-norm; k-median) clustering (1 hour)**

- **k-modes clustering (30 minutes)**
Day 4: Issues of Concern for Mixture Modeling and Cluster Analysis

- **Neural Networks (1 hour)**

- **Simulated Annealing (1 hour)**

- **Variable Selection (e.g., model building) (1 hour)**

- **Variable Standardization (1 hour)**

- **Choosing the Number of Clusters (2 hours)**

- **Cluster Validation (4 hours)**
**Additional Topics that May be Covered**

- **Network Data**

- **Adjacency Matrices and Eigenstructures (Spectral Clustering)**

- **Clusterwise Regression Models and Longitudinal Data**