Clustering Longitudinal Categorical Data

Dr. Brianna Heggeseth, Ellen Graham, Zuofu Huang, Kieu-Giang Nguyen

Project Goals

•Explore clustering literature
•Build a tool to streamline using the methods
•Compare clustering results using two real-world datasets

Datasets

•Cancer patients moving through the health-care system: Home, Hospital, SNF, Hospice
•Sleep stages: Wake, Light, Deep, REM

Methods

•Distance-based: Measure similarity between two sequences
- Distribution Distance: The difference between the distributions of states
- Feature Distance: The length of longest common subsequence
- Edit Distance: The cost of transforming one sequence into another

•Model-based: Assume data were generated by a model with group structure
- Mixture of Markov Models: Probability of transitioning to next state solely depends on current state
- Dirichlet Multinomial Models: An extension of a mixture of Markov models allowing for within-cluster variation

Results

•Distance-based
- **Healthcare**: For edit and feature distances, clusters defined by length of life after diagnosis; for distribution distances, clusters defined by the state distribution of a sequence
- **Sleep**: No meaningful clusters produced

•Model-based
- **Healthcare**: One cluster lives a longer time after diagnosis and often transitions back to home; the other cluster lives a shorter time and stays in the same state
- **Sleep**: No meaningful clusters produced

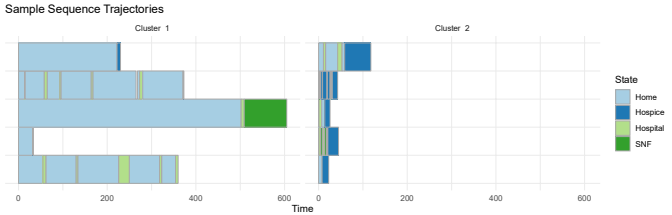
Future Work

•Develop clustering methods that make meaningful clusters of sleep data
•Explore relationship of covariates (e.g. age or time of night) and cluster assignments

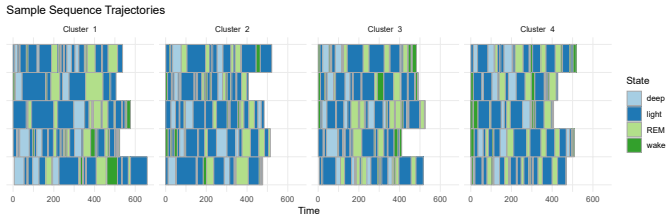
Different methods to **find common patterns** highlight some but not all features in **longitudinal categorical data.**
A good clustering method requires **context-specific knowledge.**

What is Longitudinal Categorical Data?

It consists of repeated measurements over time of a categorical variable, observed for many units or individuals.



A sample of patient's trajectories from the healthcare data set displayed by cluster



A sample of patient's trajectories from the sleep data set displayed by cluster

Clustering Comparison

Two-Way Table of Cluster Labels

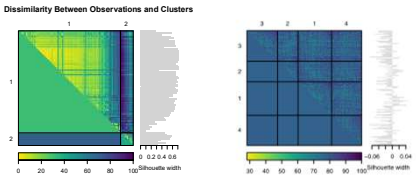
	Second Method	
	1	2
First Method		
1	89	0
2	0	11

Showing 1 to 3 of 3 entries.

Comparison Indices

Adjusted Rand Index	Jaccard Index	Normalized Mutual Information	Normalized Variation of Information
1	1	1	0

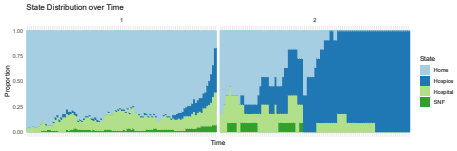
Dissimilarity Matrices



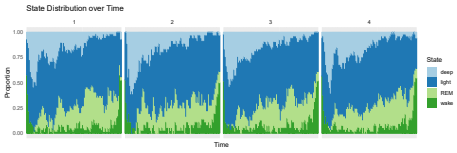
Healthcare data, 2 clusters, PAM, Gower distance

Sleep data, 4 clusters, PAM, Gower distance

State Distributions over Time



Healthcare data, 2 clusters, PAM, Gower distance



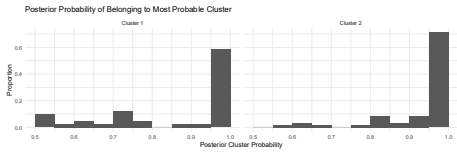
Sleep data, 4 clusters, PAM, Gower distance

Markov Chains

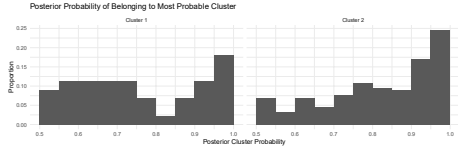
Healthcare data, 2 clusters, MMM (Bayesian Estimation)

Sleep data, 2 clusters, MMM (Bayesian Estimation)

Posterior Probabilities



Healthcare data, 2 clusters



Sleep data, 2 clusters

MACALESTER