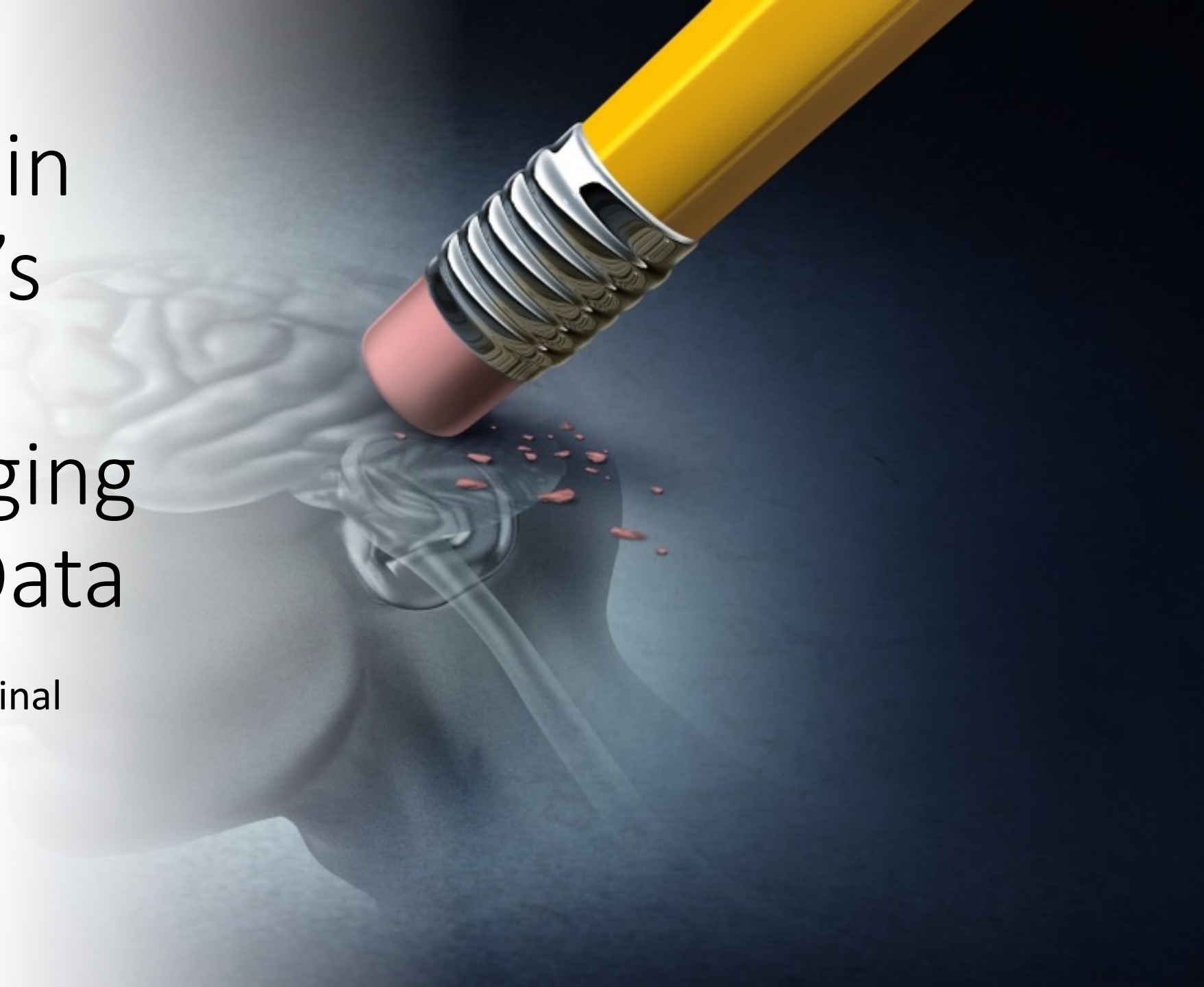
Clustering in Alzheimer's Disease Neuroimaging Initiative Data

Insights from Longitudinal Metabolomics Data

Yao Chen

Dec 7, 2023



Introduction

Alzheimer's Disease (AD) is a brain disorder that gets worse over time, and the most common cause of **dementia**.

- *Increasing Age*: The majority of people with Alzheimer's are 65 and older.
- *Genetics*: Certain genes have been linked to Alzheimer's, making it more likely in some families.
- *Brain Changes*: Alzheimer's leads to the buildup of abnormal proteins in and around brain cells, causes the brain to shrink and brain cells to eventually die.
- *Symptoms*: Characterized by memory loss and cognitive decline.

Data Source

The Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset is a rich collection of multi-dimensional data designed to study the progression of Alzheimer's. It includes neuroimaging, genetic, clinical, and, importantly, longitudinal metabolomics data, providing insights into the disease's trajectory.

Objective

Our analysis aims to uncover patterns in longitudinal metabolomics data. By exploring these datasets, we seek to identify AD progression trends and potential targets for therapeutic intervention.

This is not just an analytical challenge but a step towards unraveling the complexities of Alzheimer's Disease.

Data Overview

ADNI data for clustering:

patient ID, visit time (in months), metabolomics abundance

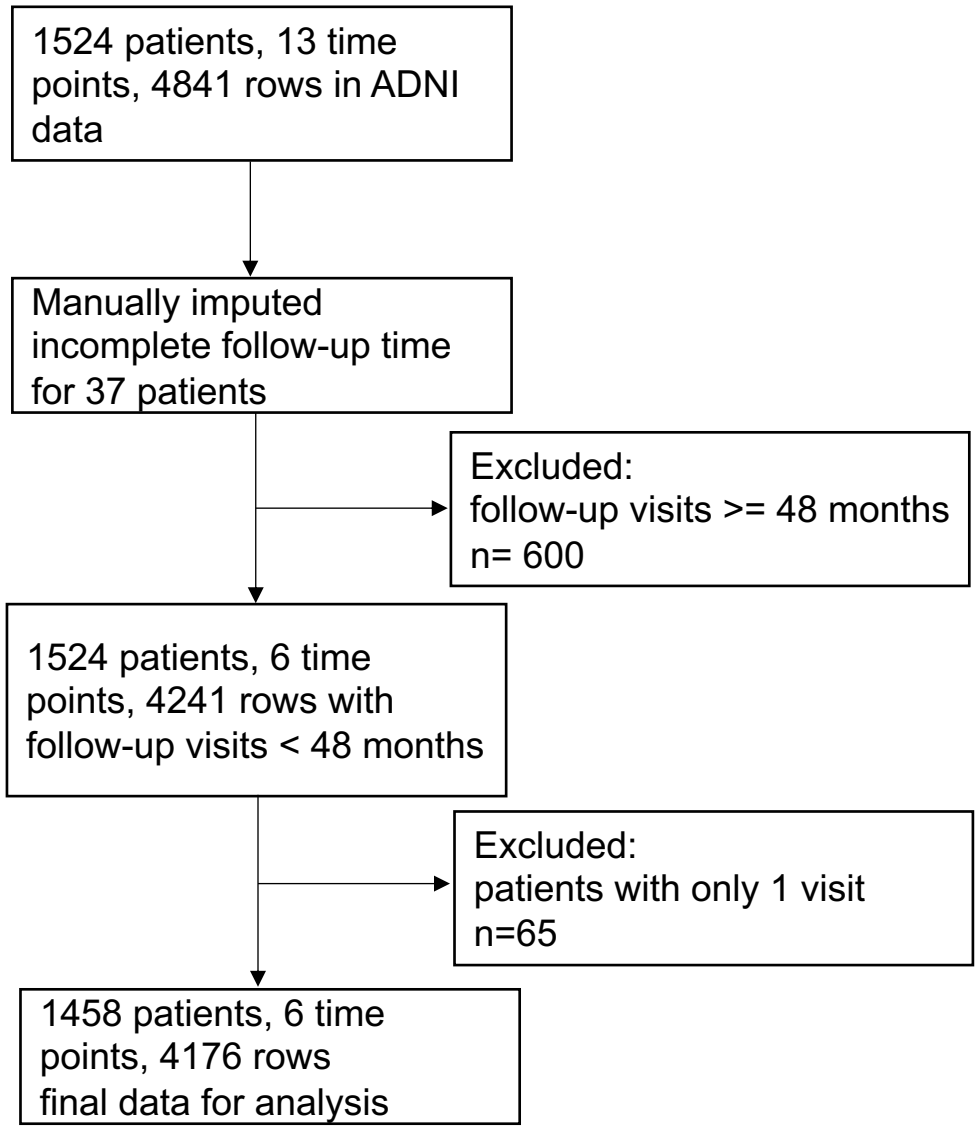
Other variables used in analysis:

age, gender, years of education,

CDRSB, ADAS11, ADAS13, MMSE, DX

Table of follow-up visit time (in months)

Months	0	6	12	18	24	36
Count	1418	158	1220	101	1118	226



Flowchart of data pre-processing

Data Overview

Baseline characteristics and cognitive performance

Characteristics	**CN 443, 30.4%**	**MCI 757, 51.9%**	**Dementia 258, 17.7%**
Age mean(SD)	74.2 (5.95)	72.8 (7.43)	75.2 (7.60)
Gender			
Female N(%)	226 (51.0)	312 (41.2)	111 (43.0)
Male N(%)	217 (49.0)	445 (58.8)	147 (57.0)
Years of education mean(SD)	16.4 (2.67)	16.0 (2.79)	15.2 (2.94)
CDRSB mean(SD)	0.04 (0.14)	1.53 (0.90)	4.30 (1.62)
ADAS11 mean(SD)	5.79 (2.87)	10.2 (4.56)	18.9 (6.45)
ADAS13 mean(SD)	9.00 (4.27)	29.1 (7.72)	29.1 (7.72)
MMSE mean(SD)	29.0 (1.14)	27.6 (1.81)	23.4 (1.98)

Generalized Estimating Equations (GEE)

- Adjusting for Age, Gender, Years of Education to control for confounding effects :

  Time-Variant Factor (Age):

    Age in this dataset does not change over time but coded to be time variant.

  Time-Invariant Factors (Gender, Years of Education):

    may significantly influence metabolomics abundance

- Metabolomics Abundance ~ Age + Gender + Years of Education
- The residuals from GEE models are used for clustering.

K-Means for Joint Longitudinal Data

Traditional k-means clustering

- Partition a dataset into 'k' clusters in which each observation belongs to the cluster with the nearest mean.
- This mean is typically a point in a multidimensional space.

Longitudinal data

- Measurements are taken over multiple time points for each subject.

Implementation with `kml3d`

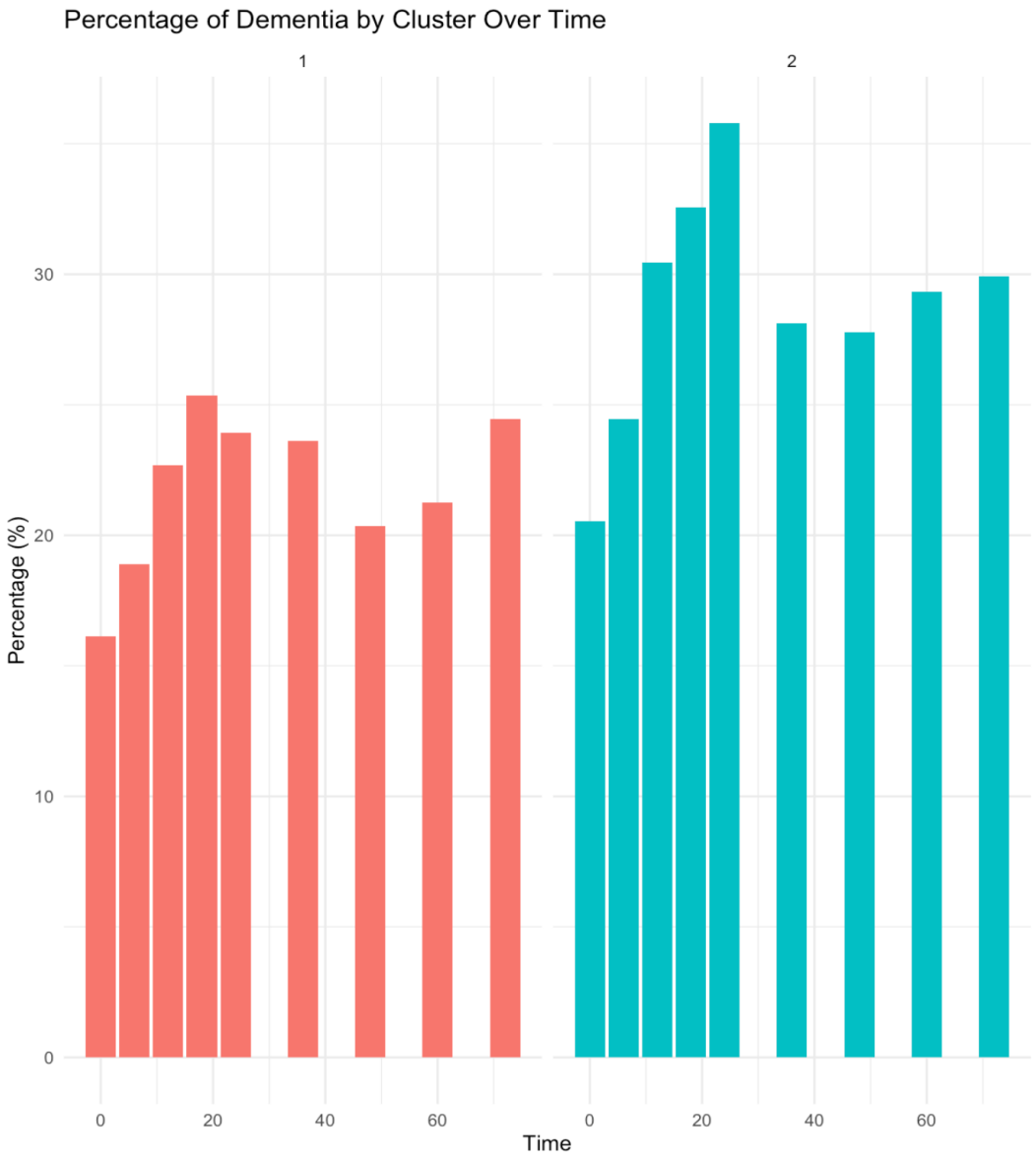
Trajectories as Features: Each individual's data is considered as a trajectory over time, rather than as a single point in multidimensional space.

Cluster Centroids: The centroid of a cluster in longitudinal k-means is also a trajectory, representing the average path taken by all trajectories in that cluster.

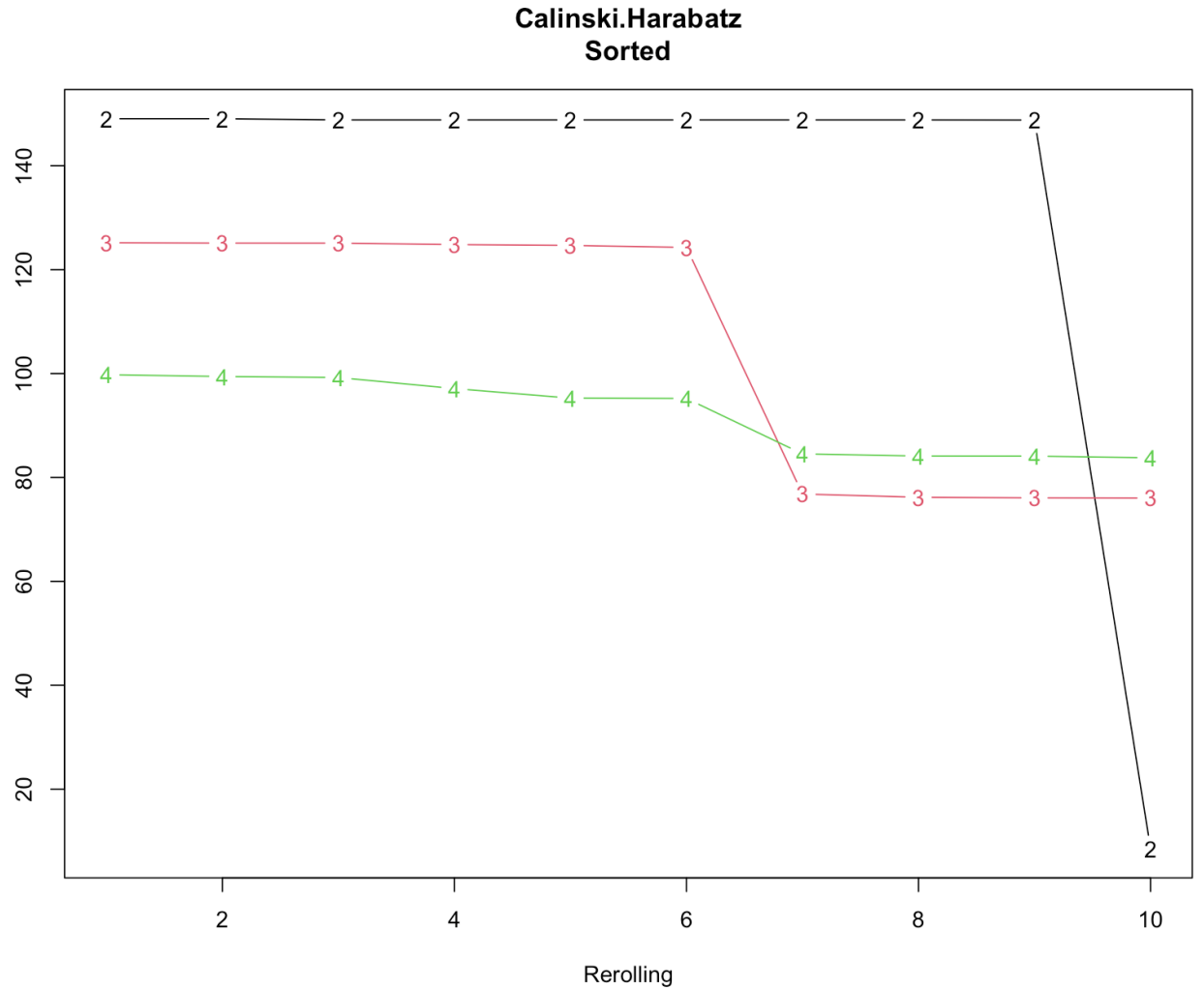
Distance Measures: The distance measure used to assign trajectories to clusters must account for the temporal nature of the data. This often involves measures that can handle time series data, like dynamic time warping.

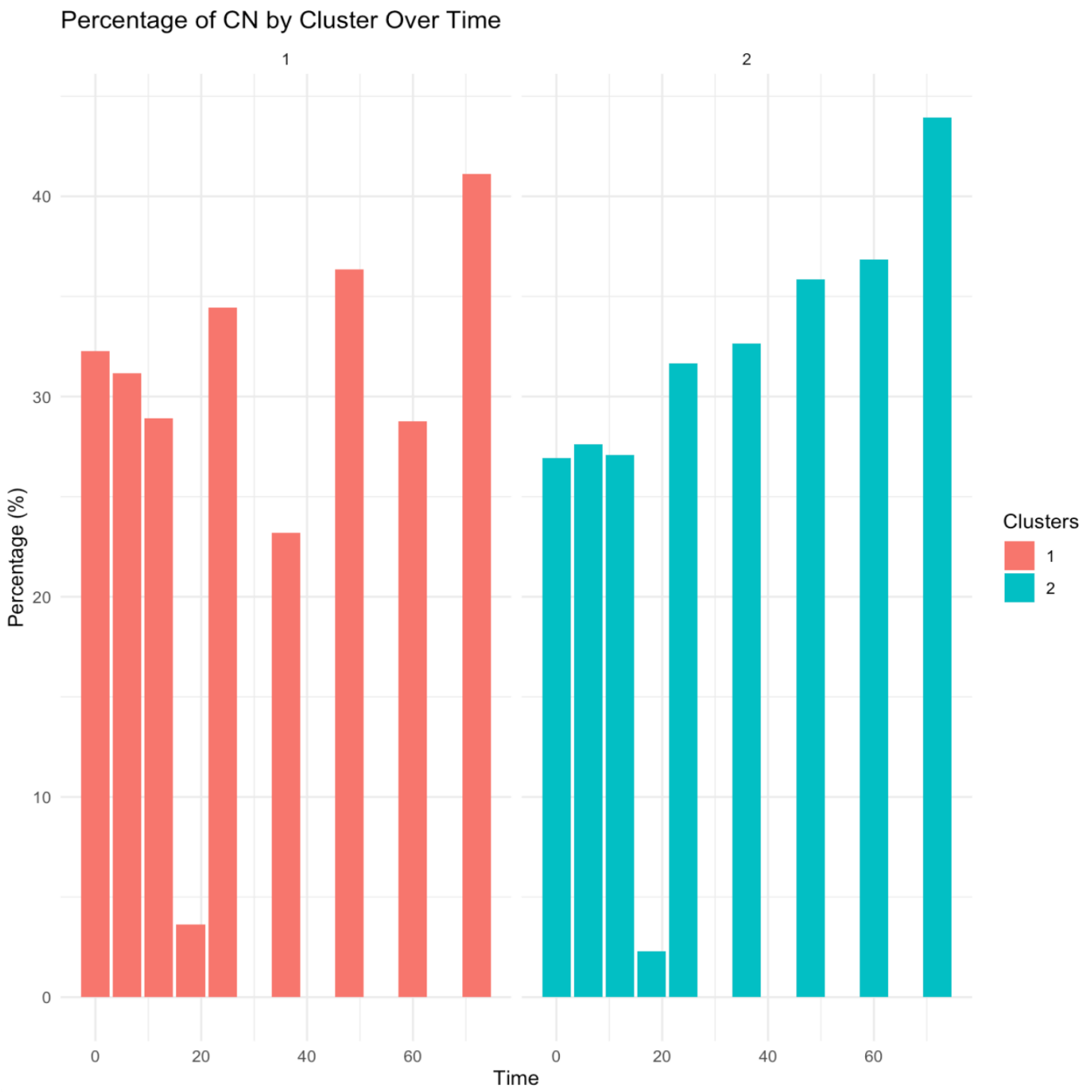
Handling Missing Data and Time Points: Longitudinal data often has missing values or differing time points across individuals. kml3d is designed to handle such irregularities, making it suitable for real-world longitudinal datasets.

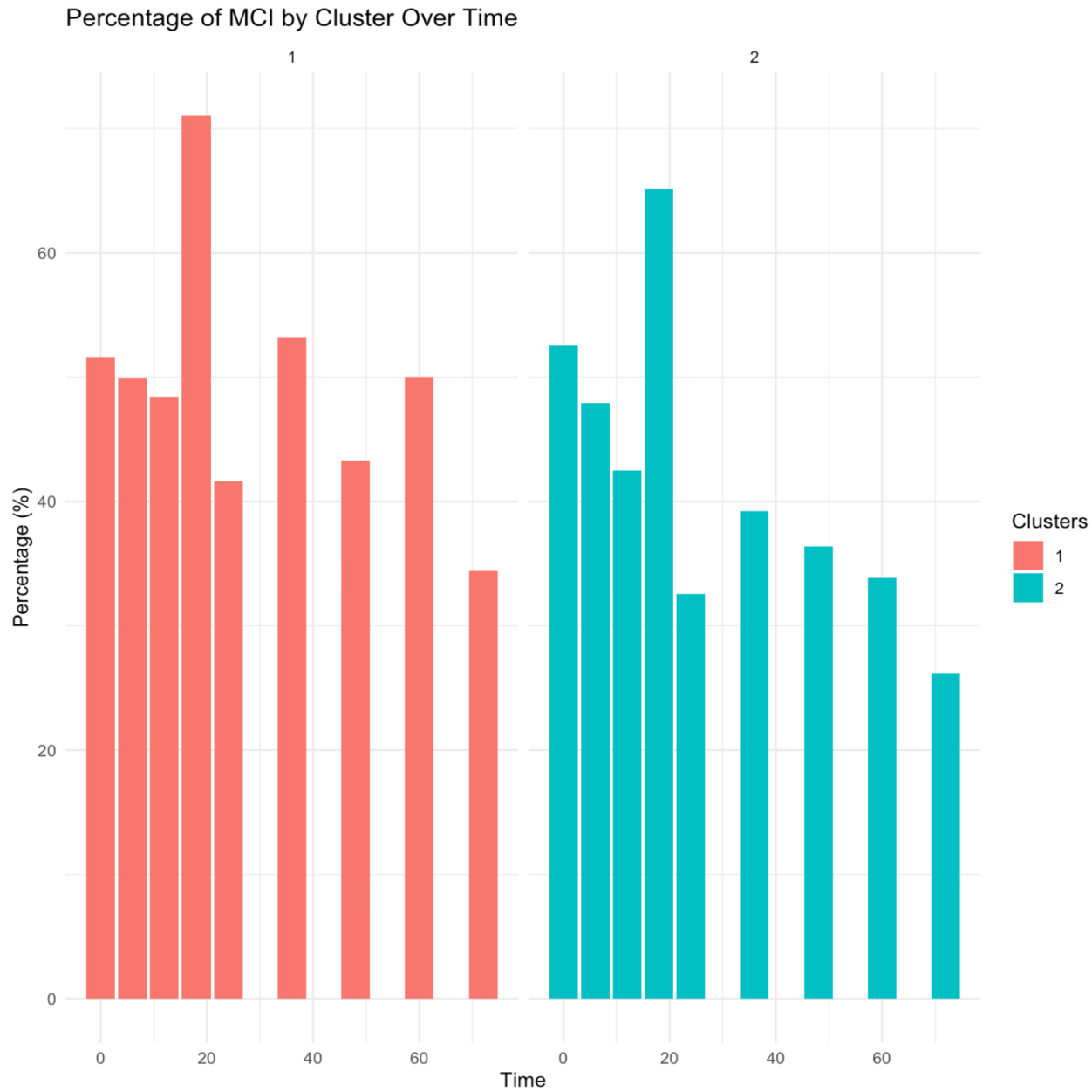
Clustering Results



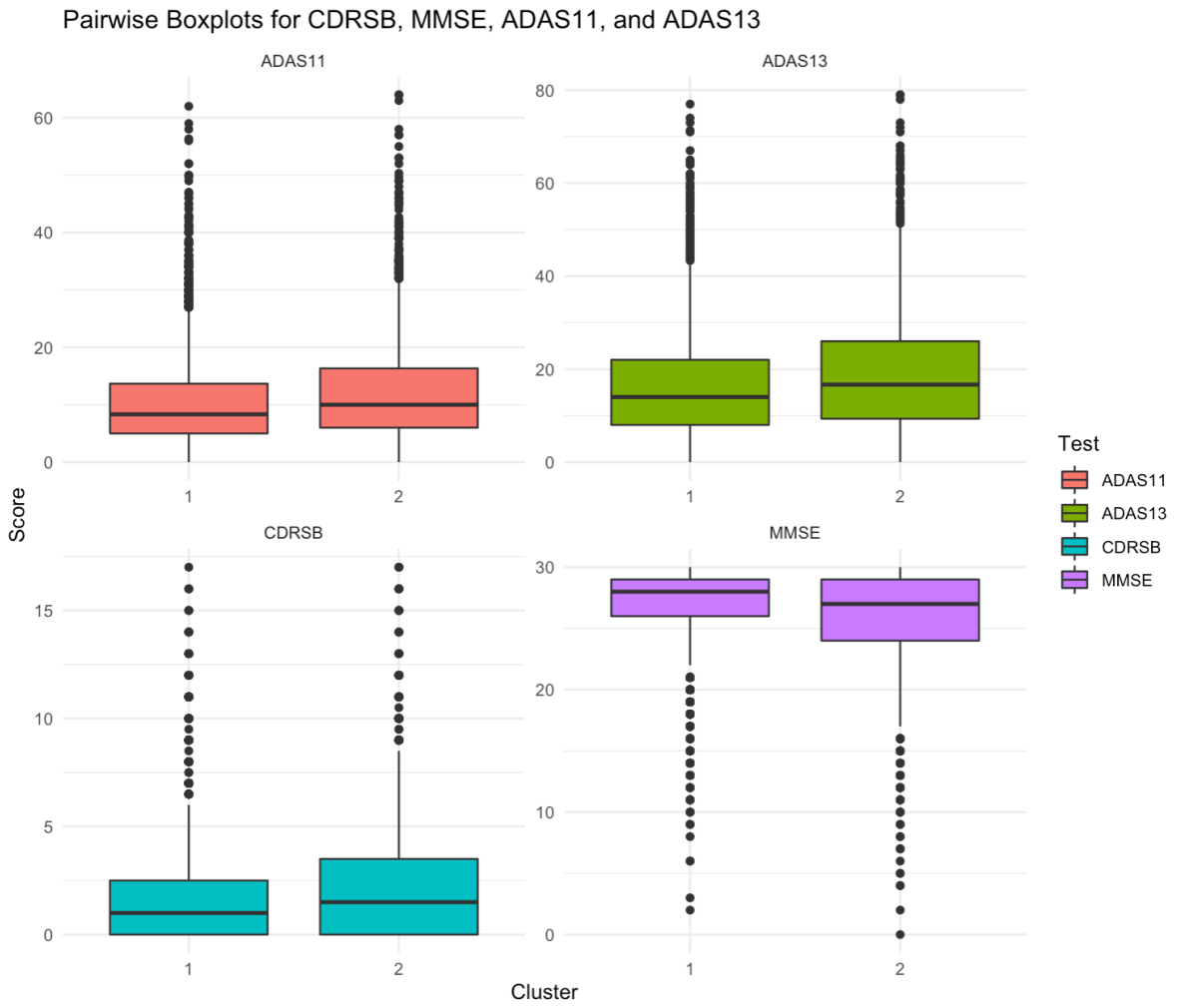
- Best clustering: 2 clusters (64%, 36%)
- Cluster 2 consistently shows a higher percentage of dementia across all time points when compared to Cluster 1




Percentage of CN by Cluster Over Time
1
2
Percentage (%)
0
10
20
30
40
0
20
40
60
Time
Clusters
1
2


Percentage of MCI by Cluster Over Time
1
2
Percentage (%)
0
20
40
60
0
20
40
60
Time
Clusters
1
2

- Cluster 2 has higher scores on ADAS11, ADAS13, CDRSB, and lower scores on MMSE
- Significant p-values in Chi-sq test for all measurements



Diagnosis Stage	ADAS11	ADAS13	CDRSB	MMSE
Cognitively Normal (CN)	Higher score is worse performance.	Higher score is worse performance.	0 – 1.0	27 - 30
Mild Cognitive Impairment (MCI)			1.0 – 2.5	20 - 26
Dementia			> 2.5	< 20

Key Findings

- **Summary of Clusters:**

When compared to Cluster 1,

- Cluster 2 consistently shows a higher percentage of dementia across all time points.
- Cluster 2 has higher scores on ADAS11 and ADAS13 and higher CDRSB scores but lower scores on MMSE compared to Cluster 1.

The significant p-values from the chi-squared tests for all measurements suggest that the differences in cognitive scores between the clusters are statistically significant.

Cluster 2 is evidently associated with a higher risk and faster progression of cognitive decline leading to dementia.

- **Impact of the Study:** These results underscore the potential of using cluster analysis in identifying subgroups within the AD spectrum, which could be crucial for targeted therapeutic interventions and personalized patient management.

Challenges and Considerations

- **Data Limitations**
  1. Loss of follow up
  2. Unbalanced data: Different participants have different numbers of follow-up visits.
- **Methodological Challenges**
  1. **Dimension Reduction**

     Tried Furry's Common PCA, Functional PCA, Longitudinal Functional PCA but are computationally intensive and time-consuming.
  2. **Longitudinal Multivariate Clustering Methods**

     Methods such as k-means for longitudinal data (kml3d) are used, but selecting the number of clusters also requires long execution time.
  3. **Interpreting Clusters**
- Individual trajectories comparison

Thank you for listening