Presentation at Merck

An Overview of Monothetic Cluster Analysis and Two Collaboration Projects

Tan Tran, PhD

September 12, 2019

Introduction

Clustering

- Unsupervised learning techniques for grouping (multivariate) responses with the goal of:
  - homogeneity — internal cohesion
  - separation — external isolation
- When to use clustering:
  - Find underlying patterns where little or no information about the data are known or to compare to known groups
  - Prediction of cluster membership based on the common characteristics of the clusters
- "A classification of a set of objects is not like a scientific theory and should perhaps be judged largely on its usefulness [...]." (Everitt, Landau, Leese, et al., 2011)

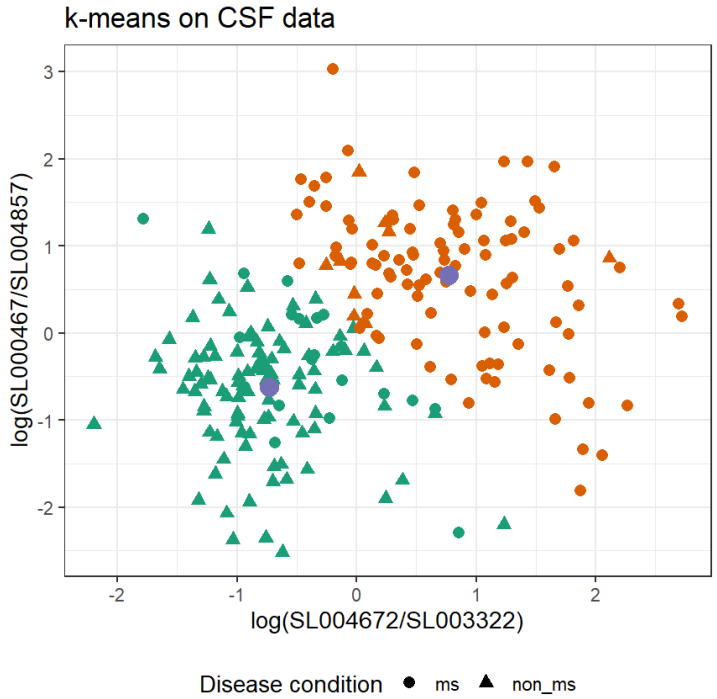
Two Clustering Techniques

- Optimization clustering techniques
  - The number of clusters, K, have to be pre-determined
  - Move the objects between clusters as long as it improves the criterion
  - k-means and partitioning around medoids (PAM, or k-medoids) are two examples of this technique
- Hierarchical clustering techniques
  - Distance measures between objects and between clusters must be defined
  - single linkage, complete linkage, Ward's method, etc.
  - Objects are fused together (agglomerative), or separated from each other (divisive) in each step based on the distance metric
  - The result is usually presented by dendrogram

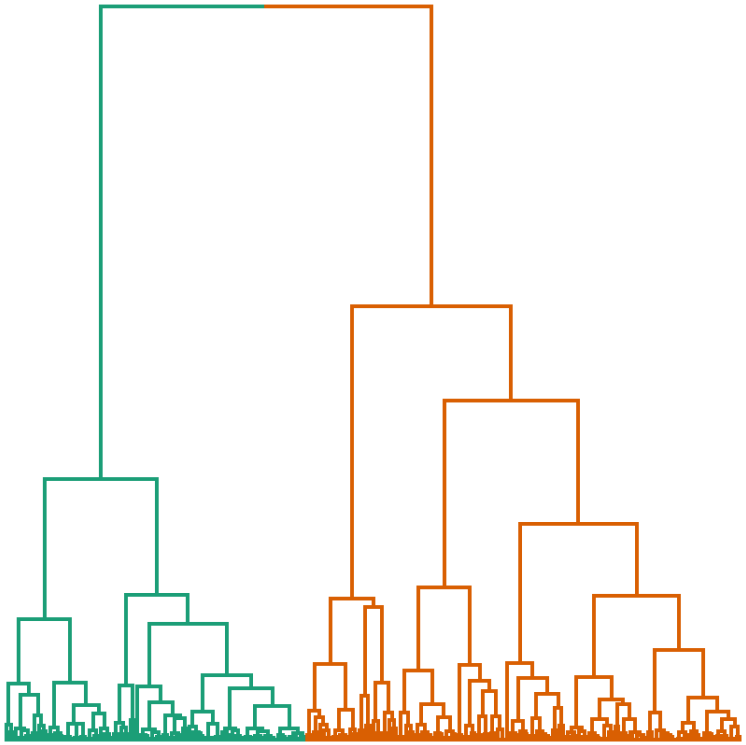
An Example: k-Means

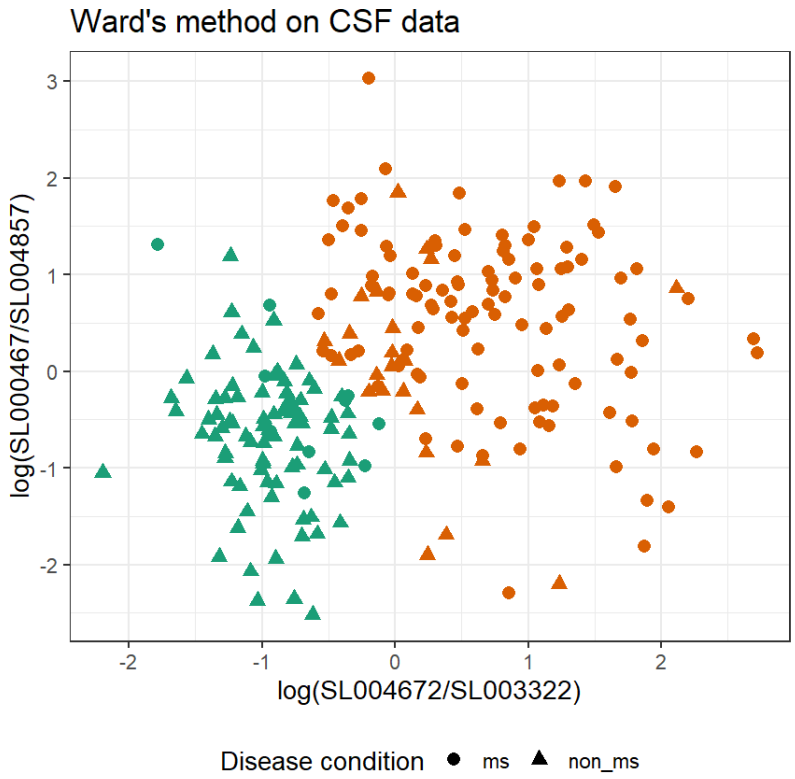
Data from Barbour, Kosa, Komori, et al. (2017) cerebro-spinal fluid (CSF) biomarker data set with $n = 225$ subjects. The variables of interest are the standardized log ratios of 22 biomarkers (proteins).

- Multiple Sclerosis (MS) and non-MS patients are known
- $Q = 2$ proteins are displayed to visually demonstrate the method



An Example: Hierarchical with Ward's Method





Polythetic vs. Monothetic Clustering

- Popular methods like k-means and Ward's are **polythetic methods**
  - Clustered using the combined information of variables
  - Observations in a cluster are similar "on average" but may share no common characteristics
- There are also **monothetic divisive methods**
  - Data are bi-partitioned based on values of one variable at a time
  - Observations share common characteristics: in the same interval or category

Monothetic Clustering Algorithm

- Introduced in Chavent (1998) and Piccarreta and Billari (2007), inspired by classification and regression trees (Breiman, Friedman, Stone, et al., 1984)
- A global criterion called **inertia** for a cluster C_k is defined as

$$I(C_k) = \frac{1}{n_k} \sum_{(i,j) \in C_k, i > j} d^2(\mathbf{y_i}, \mathbf{y_j})$$

  where $d(\mathbf{y_i}, \mathbf{y_j})$ is the distance between observations $\mathbf{y_i}$ and $\mathbf{y_j}$ and n_k is the cluster size

- Let s be a binary split dividing a cluster C_k into two clusters C_{kL} and C_{kR}. The decrease in inertia is

$$\Delta I(s, C_k) = I(C_k) - I(C_{kL}) - I(C_{kR})$$

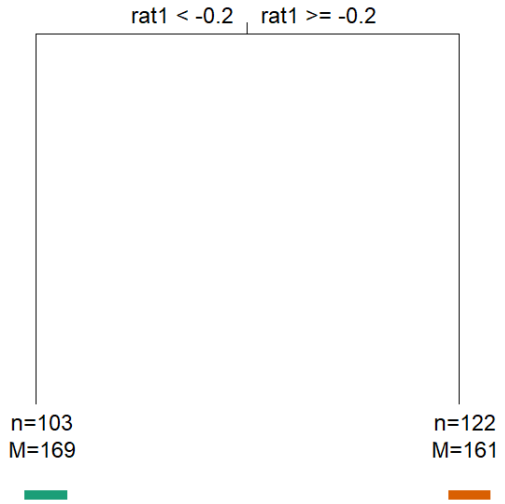
- The best split is selected as

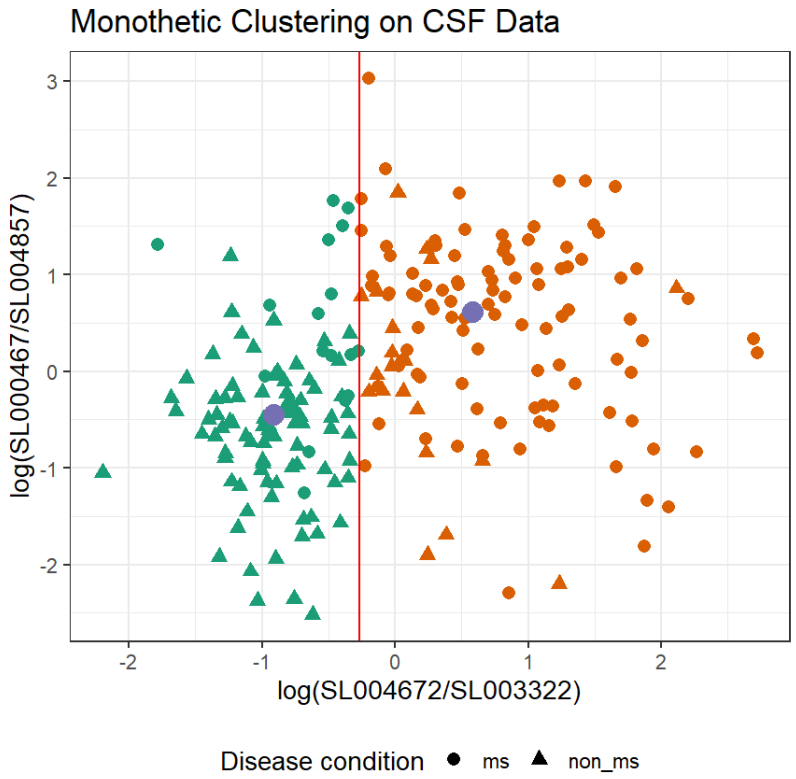
$$s^*(C_k) = \arg\max_s \Delta I(s, C_k)$$

Properties of Monothetic Clustering

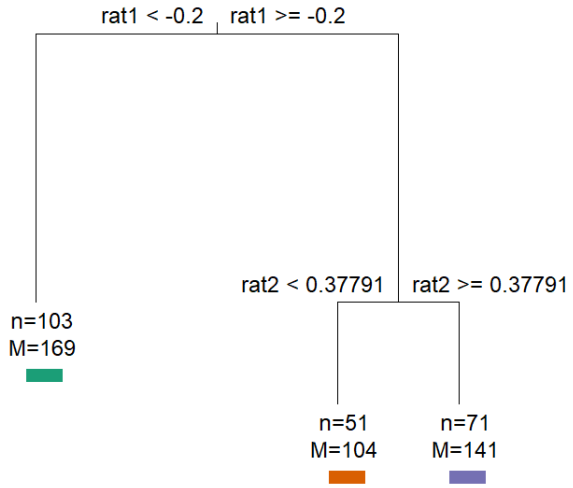
- Inertia is a global optimization criterion
- Bi-partition observations based on one variable at a time, making the method monothetic
- Defines rules for cluster membership
    - Easy classification of new members
- For the CSF data, monothetic clustering can be useful to allow classification into groups with shared characteristics that *might* relate to disease presence/absence

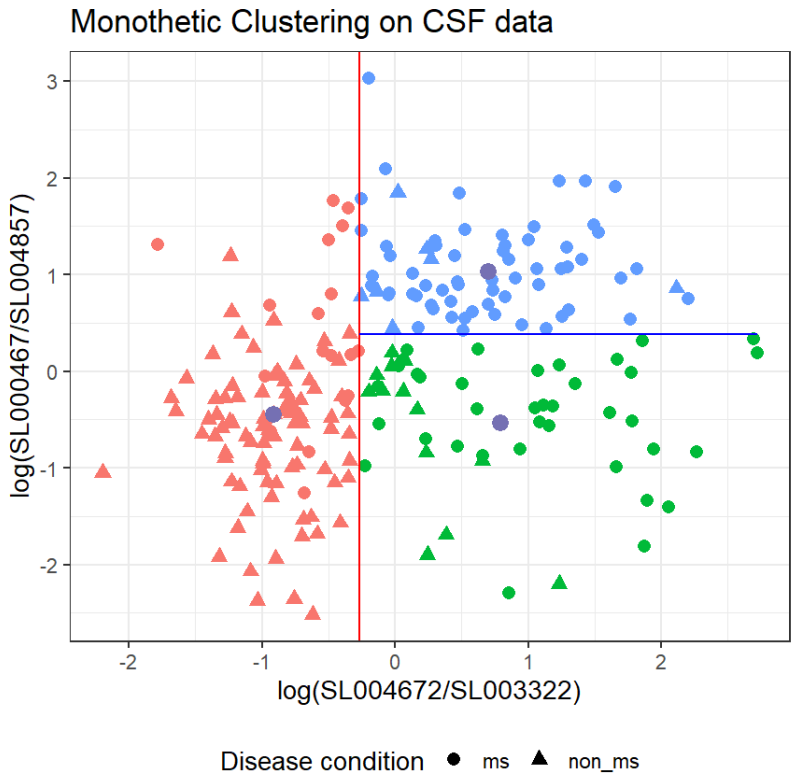
Monothetic Clustering on the CSF Data





One more split

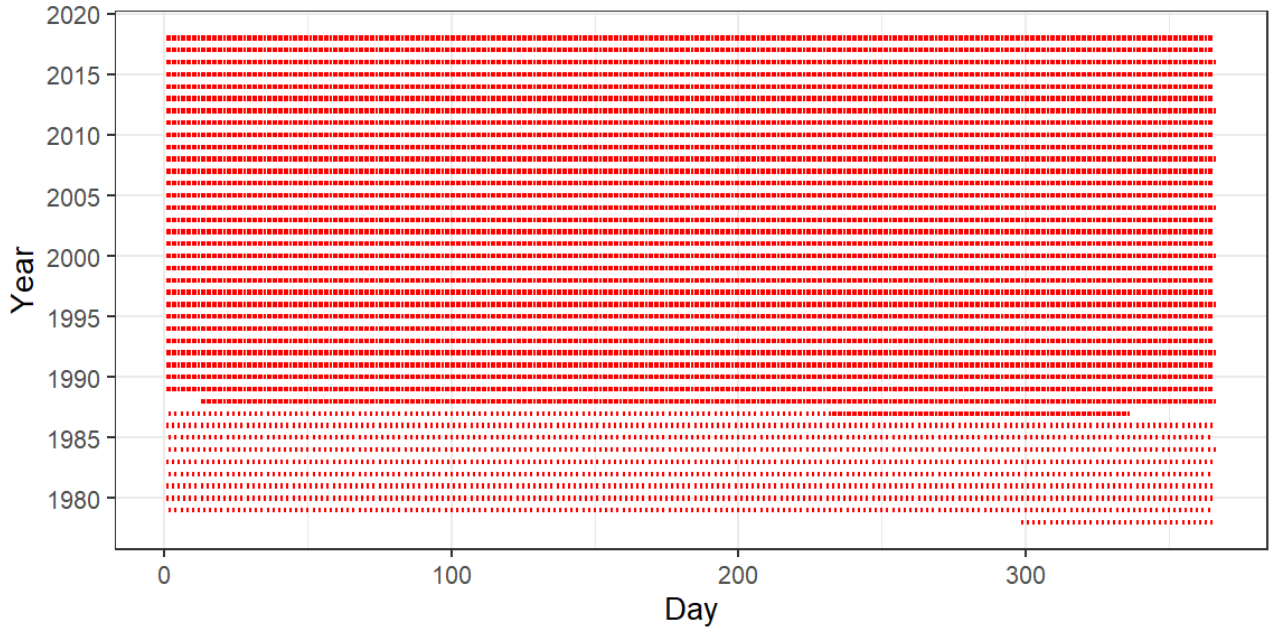




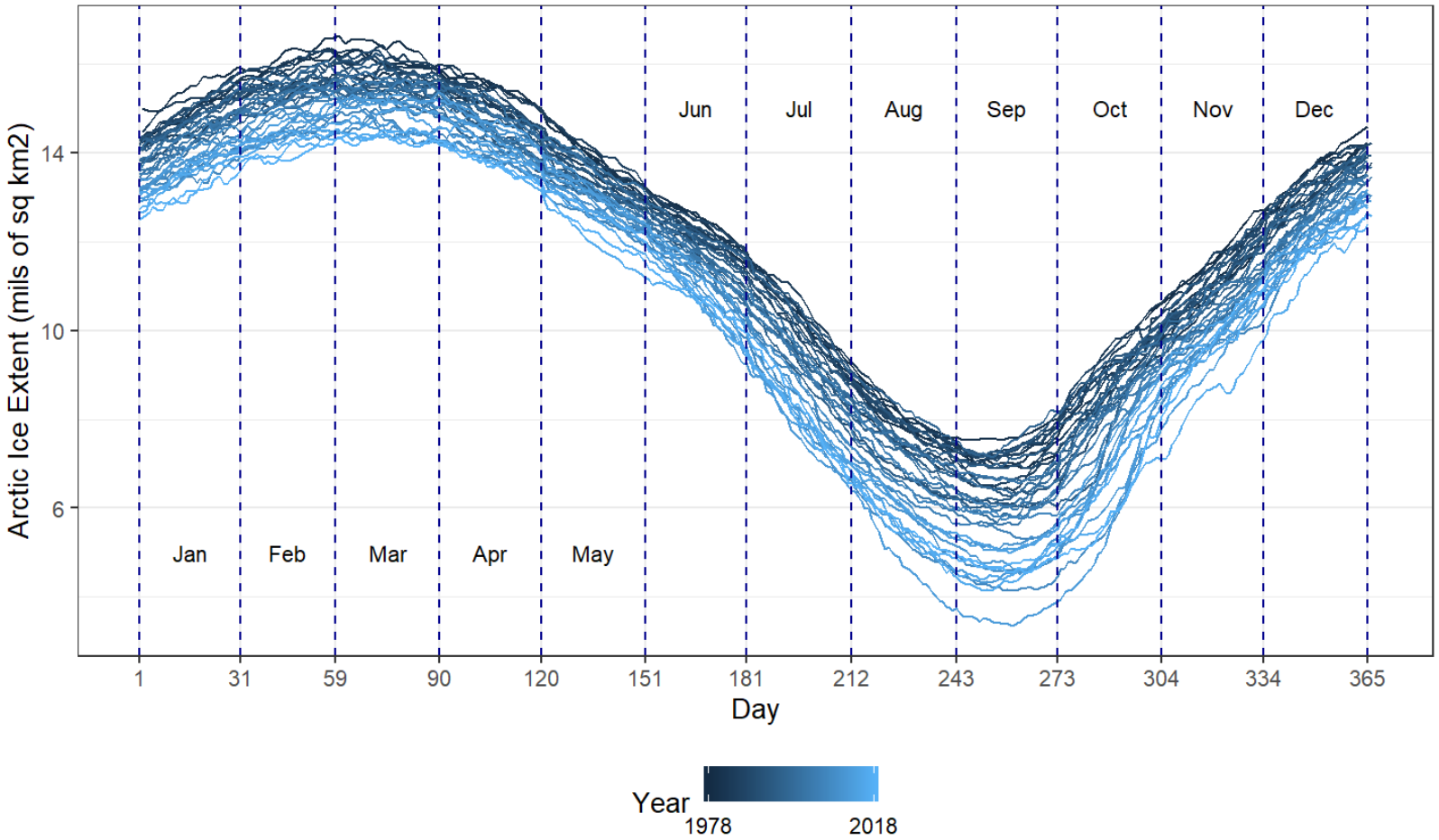
Clustering Functional Data

Arctic Sea Ice Extent Data

- Arctic Sea ice extent data set has been collected by National Snow & Ice Data Center since November 1978 (Fetterer, Knowles, Meier, et al., 2018)



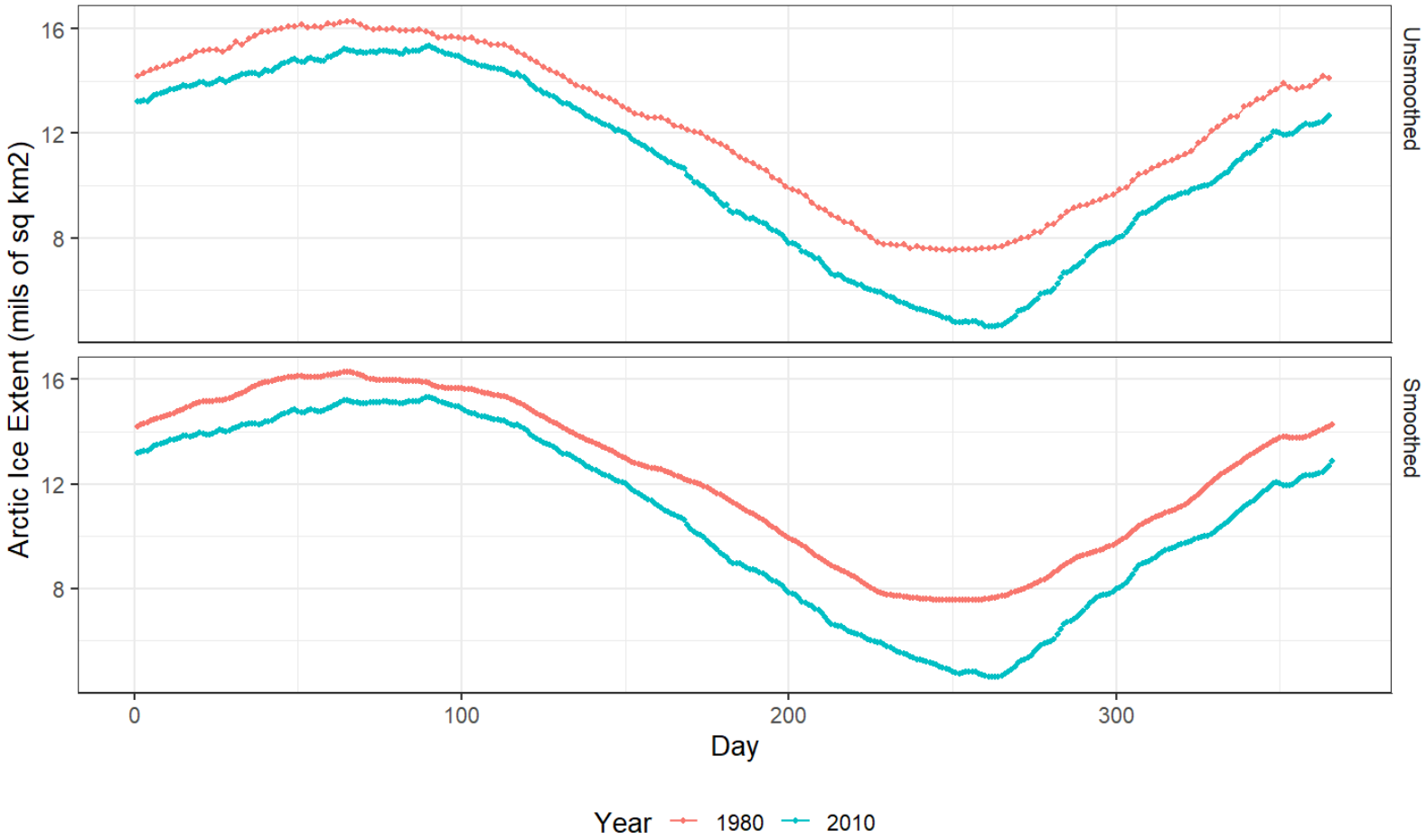
Raw Arctic Ice Extent



Functional Data

- When measurements were taken over some ordered index, such as time, frequency, or space (Ramsay and Silverman, 2005)
  - Responses are continuous as a function of the index
  - Possibly high frequency of observations and smooth underlying process or observations
- Observations are converted to functional data using basis functions:
  - **B-splines basis**, Fourier basis, Wavelets, etc.
- Penalized B-splines with knots at every day optimized with cross-validation for each curve
- Ice extent area in a year can be expressed as a function of time, $y_i(t)$
  - where i is the year, t is the day of year, and y is the ice extent at that time point

Smoothed and Interpolated Curves



Clustering Functional Data

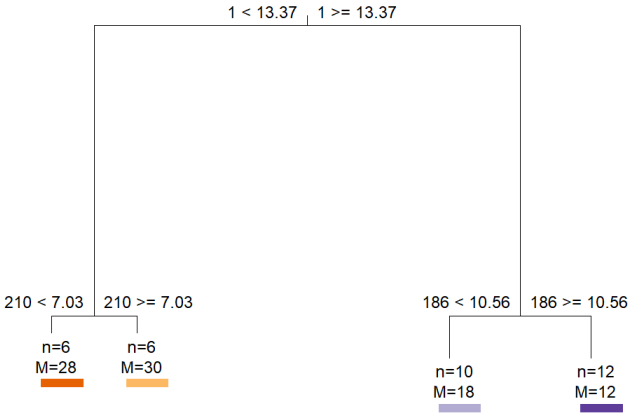
- The L_2 distance matrix can be be calculated and used for non-functional clustering algorithms

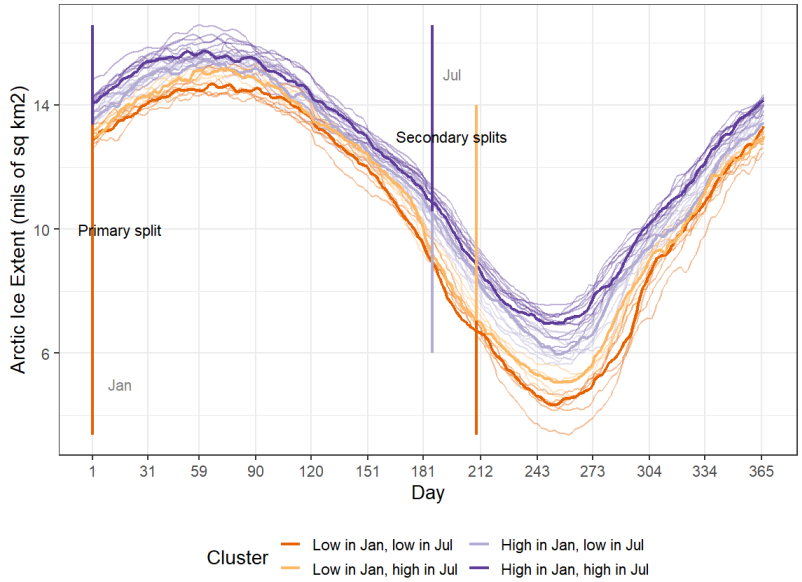
$$d(y_i, y_j) = \sqrt{\int_T [y_i(t) - y_j(t)]^2 dt}$$

- Monothetic clustering uses functional data in their discretized form
  - Transform the data into functional presentations
  - Data are then estimated to a common fine grid of t, y_{it} from $y_i(t)$
  - Missing values are imputed and the data set is balanced with $t = 1, \ldots, 366$ days for all years

Monothetic Clustering Result

- Evaluate functional data for each year, cluster years using the 366 "variables" (days) each year
- Splitting the data based on one variable (day) at a time
- Have to deal with many equivalent splits possible at "neighboring" variables





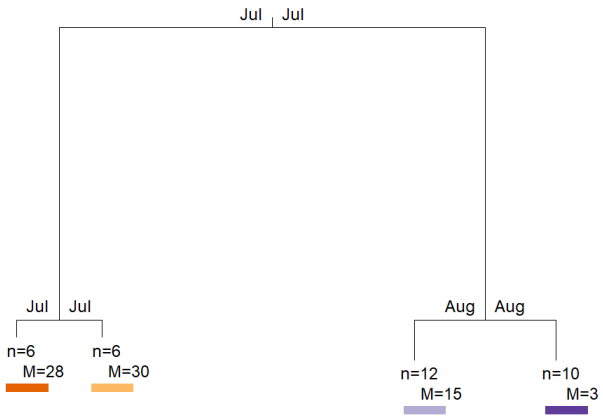
Partitioning Using Local Subregions

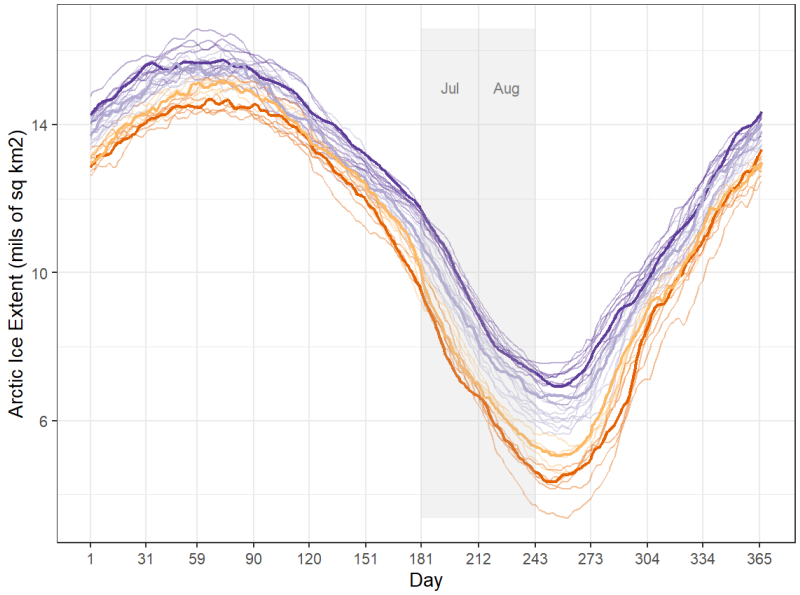
- By aggregating over regions of time, we develop a new clustering algorithm related to monothetic clustering that is more suited to functional data
- PULS recursively bi-partitions functional data using only groups from subregions.
- For each subregion $[a_1, b_1], \ldots, [a_R, b_R]$, calculate an L_2 distance matrix

$$d_R(y_i, y_j) = \sqrt{\int_{a_r}^{b_r} [y_i(t) - y_j(t)]^2 dt},$$

- Apply a clustering algorithm (PAM, Ward's method, etc.) to create 2-group cluster solutions in R subregions
- Pick the solution that maximizes the difference in the global inertia
- Recursively apply to the newly created clusters

Partitioning Using Local Subregions Result





Comparison of Results and Cluster Prediction

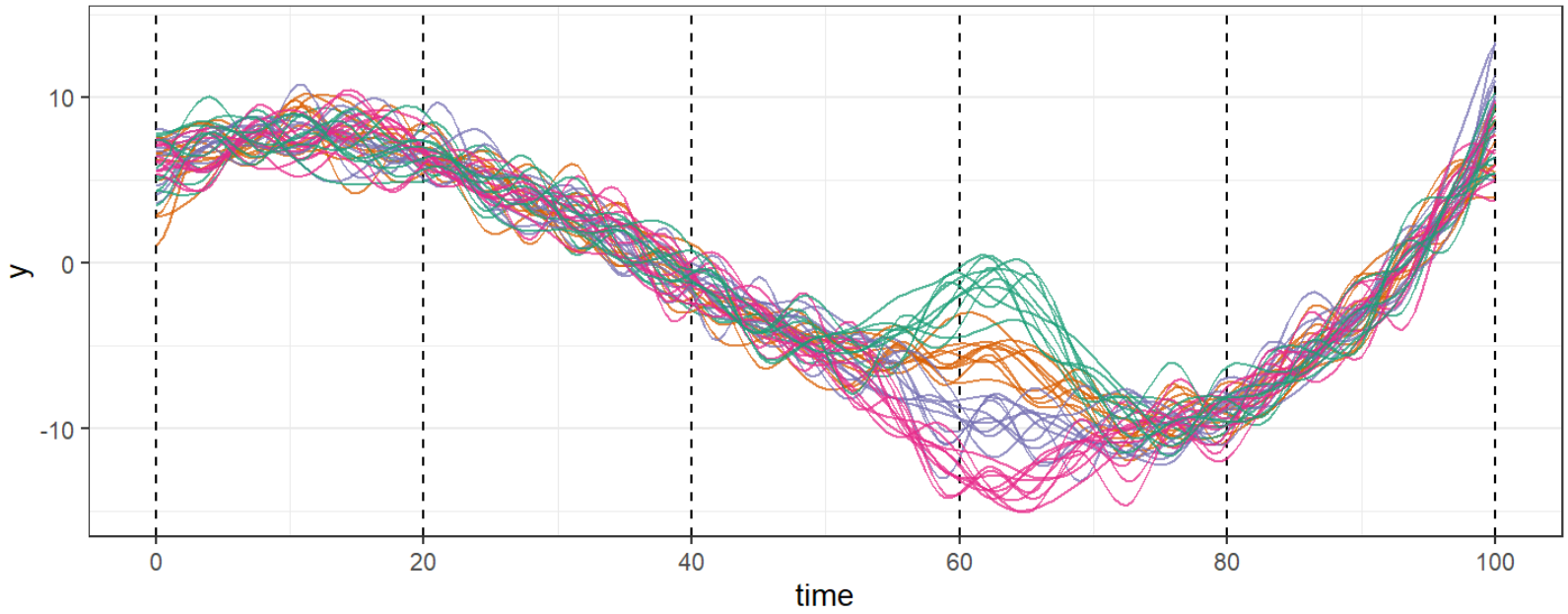
Name	PULS	MonoClust
High Jan, High Jul	1979-1981, 1983-1984,	1979-1981, 1983-1986, 1989,
	1986, 1989, 1992, 1994, 1996	1992-1994, 1996
High Jan, Low Jul	1985, 1991, 1993, 1995,	1991, 1995,
	1997-2004	1997-2000-2004
Low Jan, High Jul	2005, 2008-2010, 2013, 2014	2005, 2008-2010, 2013, 2014
Low Jan, Low Jul	2007, 2011, 2012, 2015-2017	2007, 2011, 2012, 2015-2017

- One year in each decade was randomly withheld from the test data set:
  - 1982, 1990, 2006, 2018
- Predict the cluster from monothetic clustering's splitting rule tree

High Jan, High Jul	High Jan, Low Jul	Low Jan, High Jul	Low Jan, Low Jul
1982	1990	2006	2018

Simulation Study

- Compare the performance of various clustering techniques on functional data
  - Monothetic clustering, PAM, and Ward's method
  - PULS with PAM and PULS with Ward's method on on five subregions



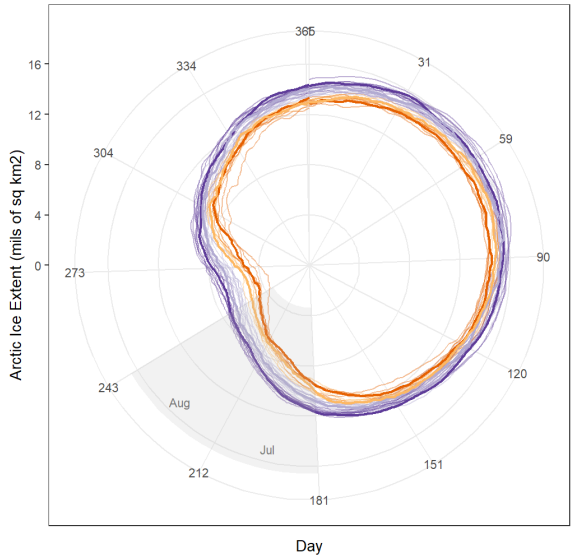
- Corrected Rand Index (Rand, 1971) and Pseudo-R2 are used to evaluate
- PULS and monothetic clustering both perform better than PAM and competitive with Ward's method
- Ward's method works well both within PULS and by itself

Other Problems (Not Mentioned Here)

- Choosing the number of clusters in monothetic clustering
  - A hybrid method of permutation test and an error-based method such as average silhouette width or pseudo-F
- Application on data with circular variables
  - Dissimilarity measures
  - Visualizations
- R packages `monoclust` and `PULS` are available on Github

Possible Extensions

- Clustering the derivative of the Arctic ice extent data
- Consider the Arctic ice extent data as circular functional data
- Search algorithms improvement



Statistical Consultation Collaborations: Power Analysis

- A research grant proposal to the USDA ARS Pulse Crop Health Initiative Research Project Plan
- Principal Investigator: Mary Miles, PhD, FACSM, of Montana State University
- My role: Statistician, involved in designing the analysis plan and statistical power analysis for the project
- Was aprroved for funding last week

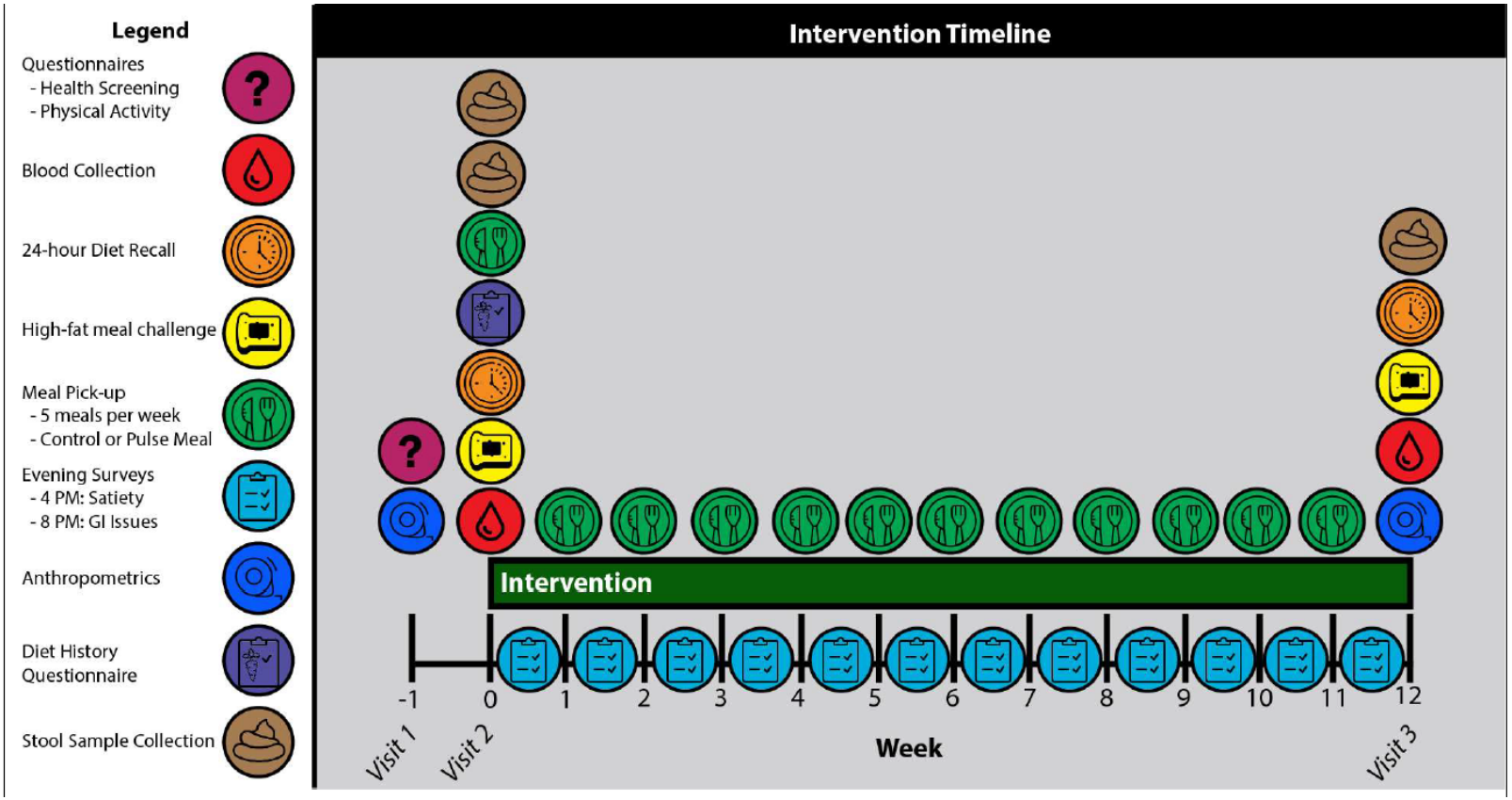
Objectives

1. **Determine the impact of pulse (green lentil and black bean) consumption on postprandial triglyceride (TG) and inflammation responses to a high-fat meal challenge.**

2. Determine the extent to which the gut microbiome and changes in the gut microbiome induced by pulse consumption influence health impacts

3. Measure metabolomic profiles to elucidate underlying mechanisms linking pulse consumption to improved health.

Study Plan

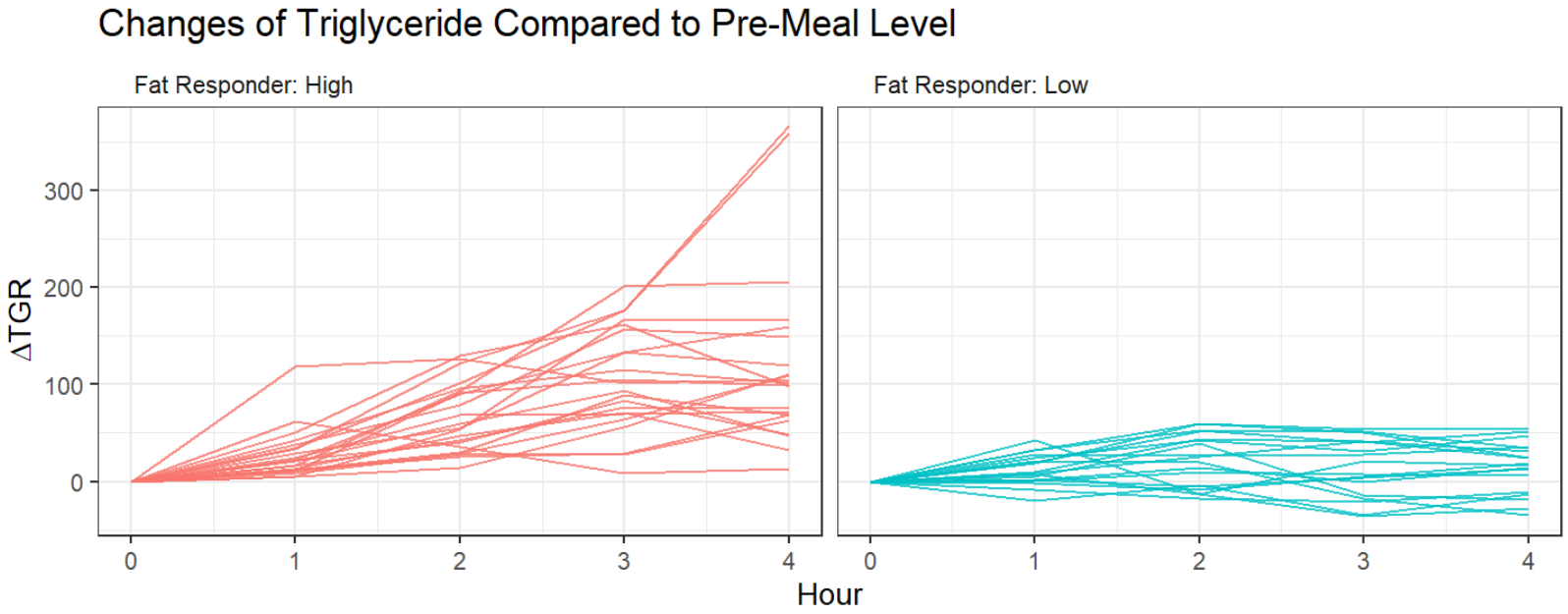

Legend
Questionnaires
- Health Screening
- Physical Activity
Blood Collection
24-hour Diet Recall
High-fat meal challenge
Meal Pick-up
- 5 meals per week
- Control or Pulse Meal
Evening Surveys
- 4 PM: Satiety
- 8 PM: GI Issues
Anthropometrics
Diet History Questionnaire
Stool Sample Collection
Intervention Timeline
Intervention
-1
0
1
2
3
4
5
6
7
8
9
10
11
12
Visit 1
Visit 2
Week
Visit 3

Collaboration

1. Learn the study
    - Biological terminologies
    - Microbiome data analysis
    - Metabolomics data analysis
2. Prioritize the objective
    - Objective 1
    - Objectives 2 and 3 are explanatory
3. Design the analysis plan
    - Obj. 1: one-sided t-test with possible multiple testings
    - Obj. 2: PLS-DA and variable importance to specify species differentiate the two treatment groups
    - Obj. 3: untargeted analysis using XCMS and mummichog
4. Power analysis based on Objective 1
    - Budget limit: can only recruit 36 to 48 people

Determine the Effect Size

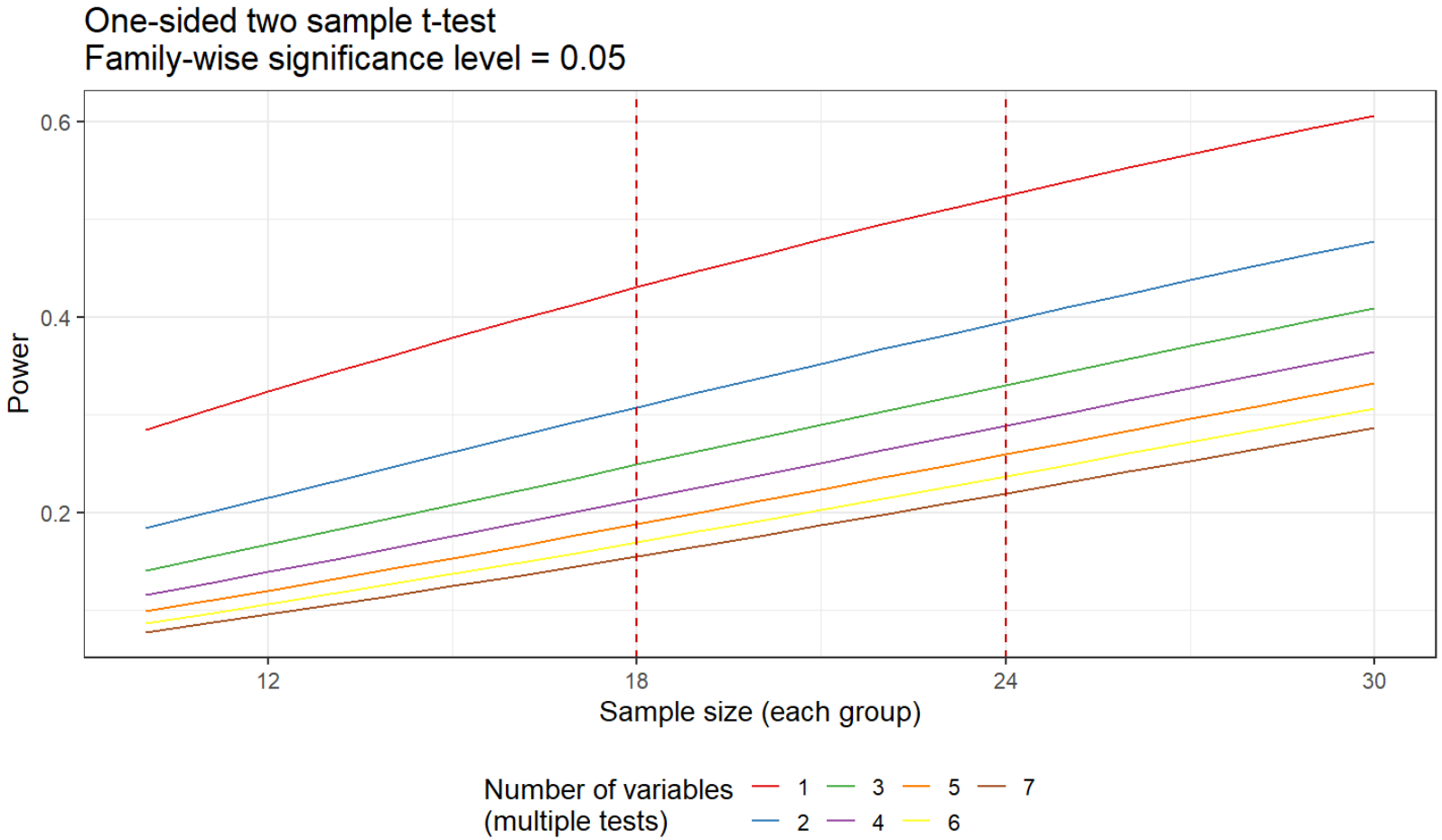
- There was a pilot study with similar design
  - Two groups of participants: high fat responder (n = 22) and low fat responder (n = 18)
  - The high fat meal challenge was done and the change in triglyceride levels were measured several times
  - There was no 12-week experiment



Max Triglyceride

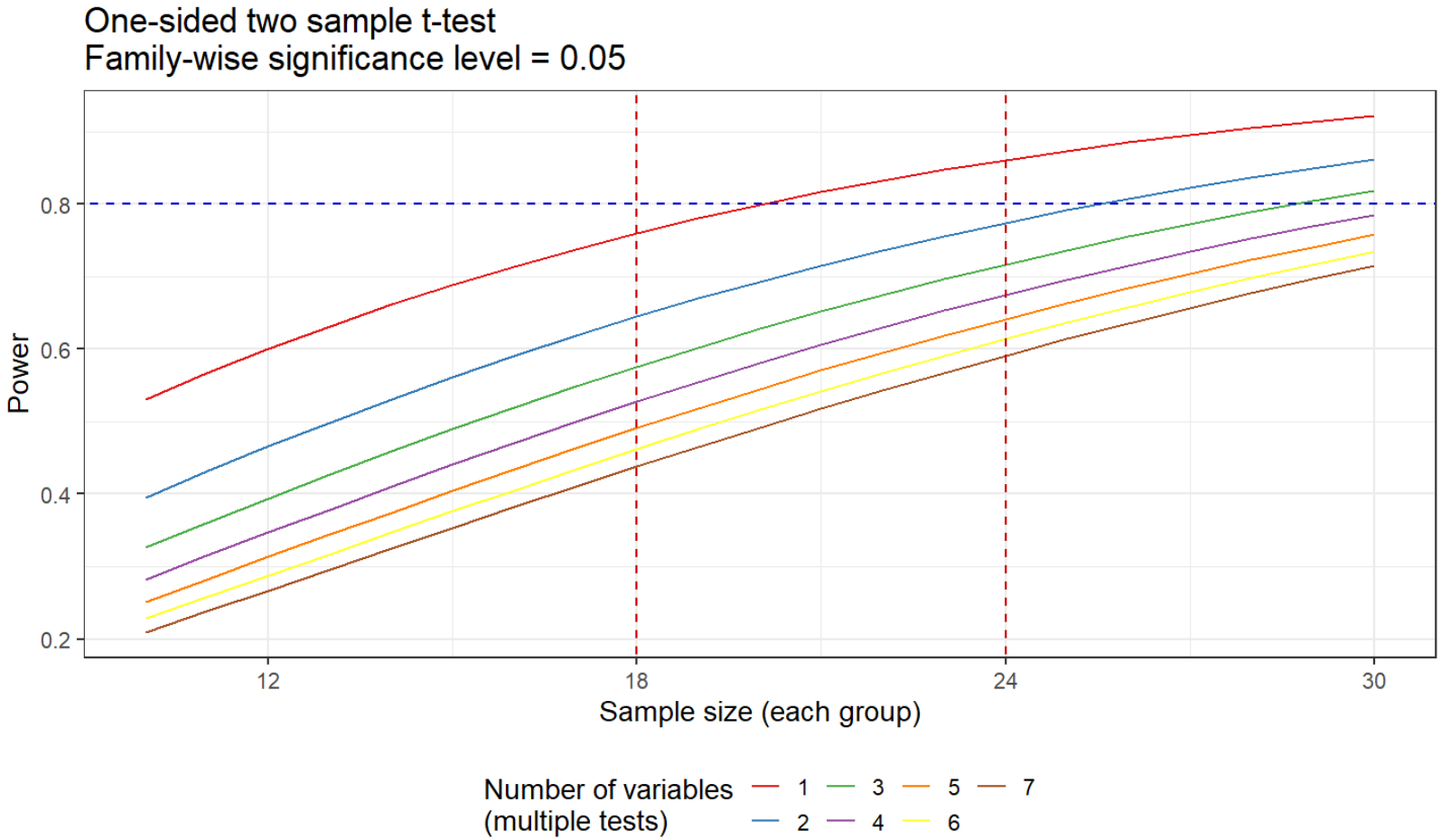
sample	Mean	Expected Mean After (50%)	SD	Effect size
High and Low Responders	87.75	43.88	81.67	0.54
High Responders Only	134.23	67.11	83.68	0.80

Power Analysis Plot with Effect Size d = 0.54



Another Power Analysis Calculation with d = 0.8

- Another meeting happened and the researcher agreed to screen the participants further to not include people with low high fat responder.



Deliverables

- An interactive report to assist the researcher to decide the sample size and the power of the test.

Sensitivity of Analysis Paths in MetaboAnalyst

- The Proteomics, Metabolomics and Mass Spectrometry Facility (Mass Spec) is a research lab in Montana State University
- They are funded by INBRE and usually do LC-MS analysis for many metabolomics-related projects on campus
- MetaboAnalyst (https://www.metaboanalyst.ca/) is the web application they use for statistical analysis of metabolomics data
- My role as a statistician with computer science background
  - Learn the analyses the webapp is doing
  - Understand how biochemists have been using the webapp
  - Suggest good sets of options that will give a statistically sound result
- This is an ongoing project

Difficulties to Overcome

- Chemical and bio-chemical terminologies
- Mass spec procedures and associated data analysis problems
- Dig into R packages used by MetaboAnalyst and reproduce the results offline
- Learn how researchers have been using MetaboAnalyst and reproduce the results
- Automate the analysis process
- Do sensitivity analysis on various options

Current Results

- Used the metabolomics data set from Dr. Hunts' pilot dietary project
    - Detect metabolites that differentiate the high and low fat responder groups
- Divided the analysis duties between domain experts and statistician
    - The biochemists to run blanks to help with filtering process
    - The statisticians to work on appropriate data transformations
- A report on sensitivity analysis with an accompanied Shiny app

Future Work

- Learn the assumptions made by Partial Least Squares Determinant Analysis (PLS-DA)
- Give formal suggestions on what analysis paths to use for finding influencing features problems
- Move on to a different set of mass spec data with a different analysis question