MONTANA STATE UNIVERSITY

PhD Defense

Monothetic Cluster Analysis with Extensions to Circular and Functional Data

Tan Tran
Advisor: Dr. Mark Greenwood

April 29, 2019

Table of Contents

Chapter 1: 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)

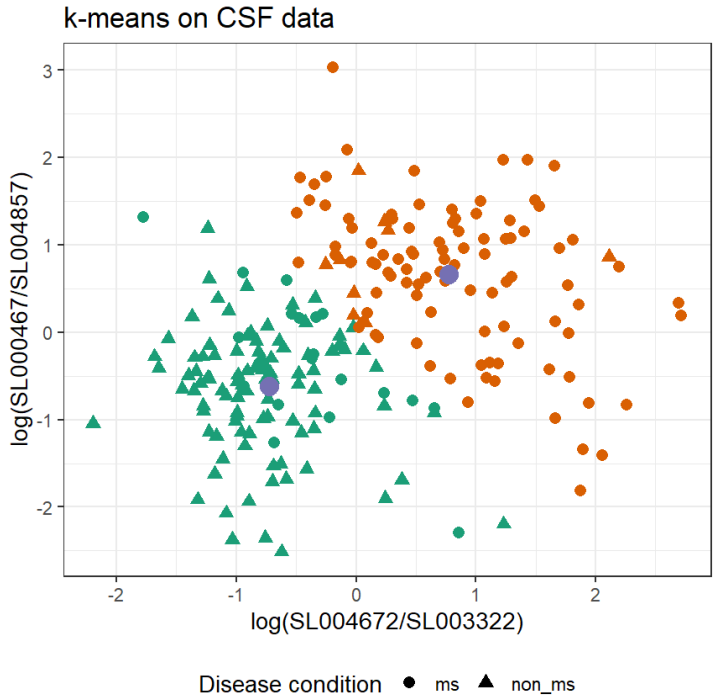
Two Clustering Techniques

- 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

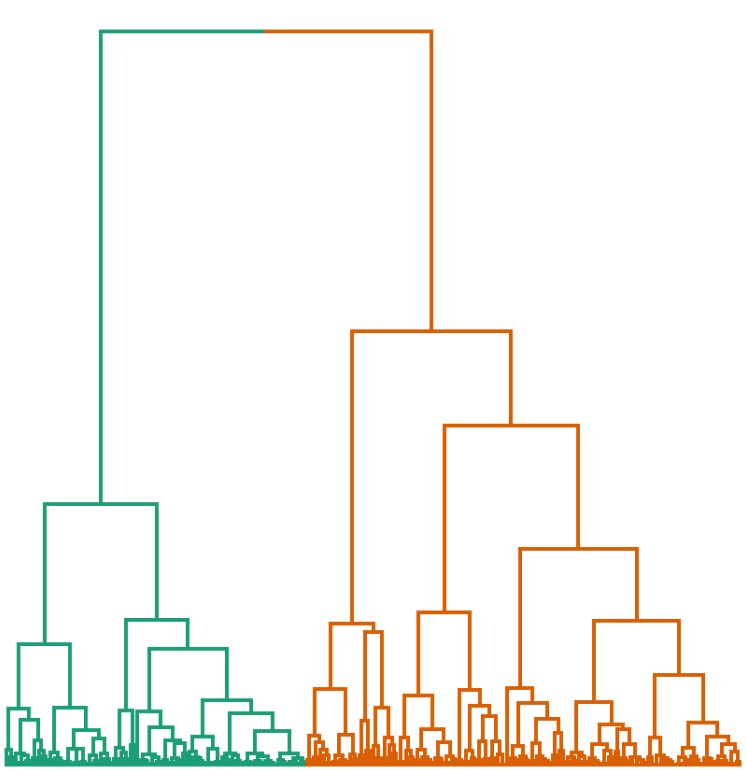
An Example: k-Means

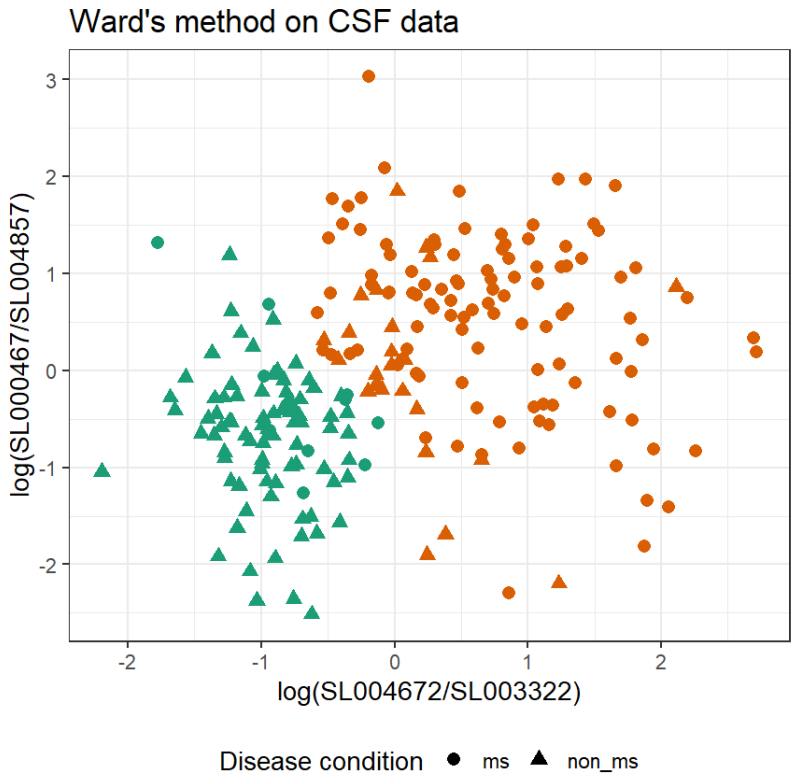
Data from Barbour, Kosa, Komori, et al. (2017) cerebro-spinal fluid (CSF) biomarker data set with $n = 225$ subjects. The variables of interest are the standardized log ratios of some proteins.

- Multiple Sclerosis (MS) and non-MS patients are known
- $Q = 2$ log ratios are used to 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

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

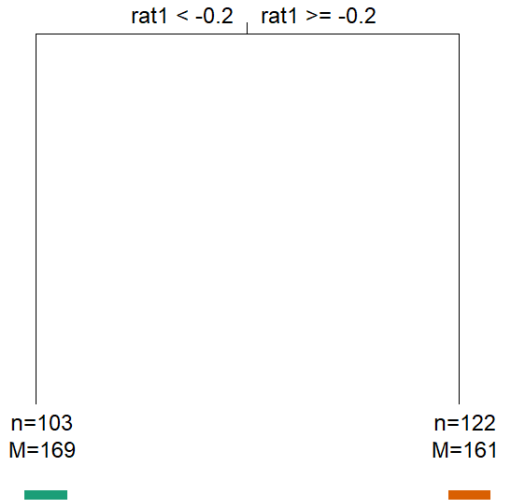
- The best split is selected as

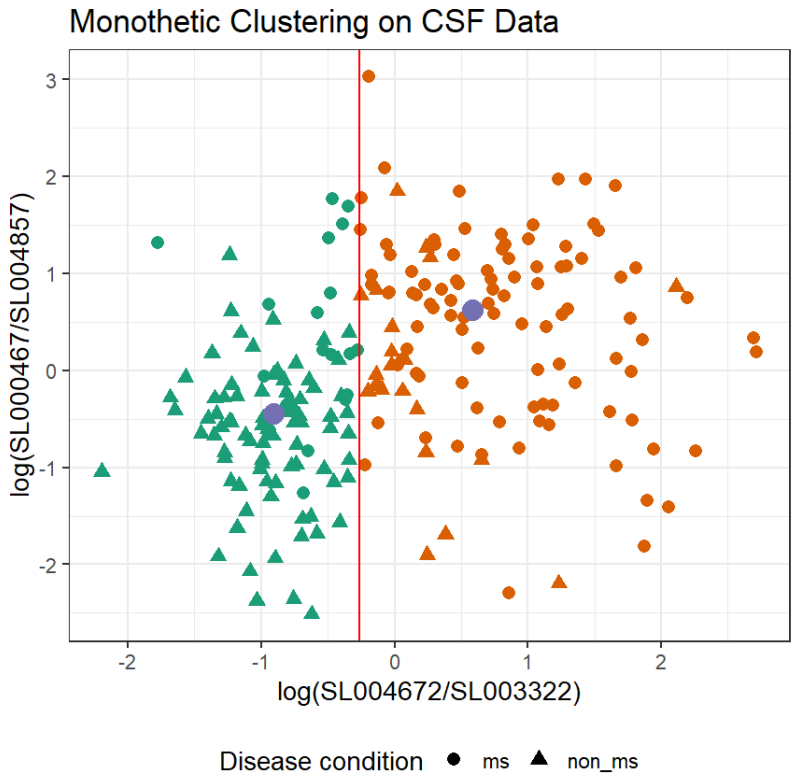
$$s^*(C_k) = \arg\max_s \Delta I(s, C_k)$$

Properties of Monothetic Clustering

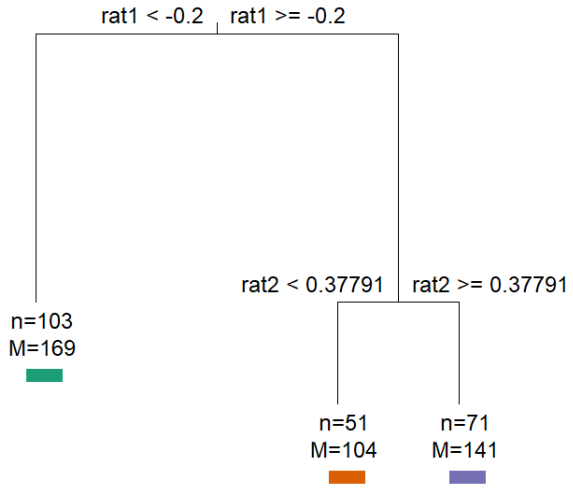
- Inertia is a global optimization criterion
- Bi-partition observations based on one variable at a time, making the method monothetic
- Defines rules for cluster membership
    - Easy classification of new members
- For the CSF data, monothetic clustering can be useful to allow classification into groups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

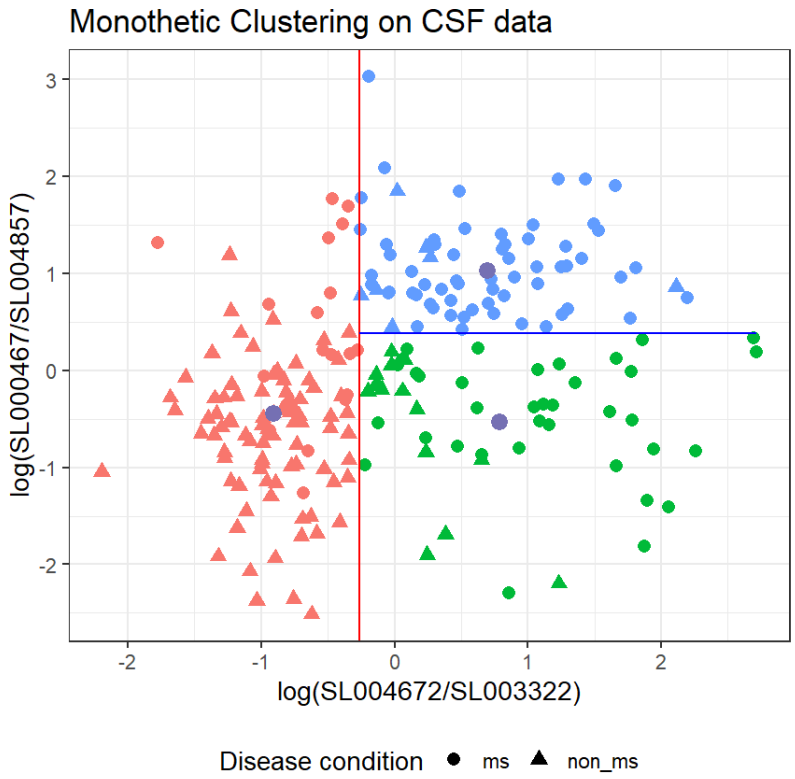
Monothetic Clustering on the CSF Data





One more split





Chapter 2: Choosing the Number of Clusters

- When the data do not have a true cluster structure, the results of a clustering algorithm can be arbitrary and misleading.
- If a cluster structure is present, the number of clusters to report has to be "estimated"
- This can be done informally by subject matter or plotting, which are very subjective, or based on application
- Many formal techniques have been suggested to overcome the subjectivity (Milligan and Cooper, 1985; Tibshirani, Walther, and Hastie, 2001)
- There is no generally "best" technique but investigations are needed to assess performance of the technique in different cluster structures and with different clustering algorithms

Average Silhouette Width (AW)

- The silhouette width $s(i)$ (Rousseeuw, 1987) is a measure of how "comfortable" an observation i is in the cluster it resides

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))},$$

  - $a(i)$: the average distance between i and other observations in the same cluster
  - $b(i)$: the minimum average distance from i to other observations in any other cluster that i is not a member
- The value obtained is from -1 to 1
- The average silhouette width of a K cluster solution

$$\overline{s_K} = \frac{\sum_{i=1}^{n} s(i)}{n}$$

- Select cluster solution size K that has maximum average silhouette width
- Not defined for $K = 1$

Caliński and Harabasz (CH)'s Pseudo-F

- Caliński and Harabasz (1974)
- Choose K to maximize variation between clusters relative to variation within clusters.
- Use

$$\text{pseudo}-F = \frac{B(K)/(K-1)}{W(K)/(n-K)}$$

  with $B(K)$ the between cluster sums of squares (possibly from dissimilarities) and $W(K)$ the within cluster sums of squares.
- It is not defined for $K = 1$ so cannot select a single cluster solution.

M-fold Cross-Validation

- Based on ideas for pruning regression trees (Breiman, Friedman, Stone, et al., 1984)
- Randomly divide data set into M equal-sized subsets, withhold a subset, and use the rest for training
- Compute a measure of prediction error for the m^{th} set of withheld observations (in Euclidean distance cases)

$$MSE_m = \frac{1}{n_m} \sum_{j=1}^{p} \sum_{i \in m} (y_{ij} - \hat{y}_{ij})^2$$

  where $\hat{y}_{ij}$ are the predicted responses, which is the centroid of the predicted cluster
- Cross-validation based estimate of the error for the tree of size K

$$CV_K = \frac{1}{M} \sum_{m=1}^{M} MSE_m$$

M-Fold Cross-Validation

- CV-based selection rules:

  1. Choose K^* that provides the smallest CV_K (*minCV* rule)

  2. Choose the smallest K satisfying

$$CV_K \leq CV_{K^*} + \gamma SE_{K^*}$$

where SE_{K^*} is the standard error estimate of CV_{K^*} and γ = 1 or 2 (*CV1SE* and *CV2SE* rules)

Permutation Tests: Cluster Shuffling

- H_0: The two new clusters are identical to each other
- Based on ideas from conditional inference trees (Hothorn, Hornik, and Zeileis, 2006)
- Using permutations of observations across split and pseudo-F (F*) test statistic (Anderson, 2001)

$$p-\text{value} = \frac{\text{count}(F^* \geq F_{obs})}{B}$$

- Apply test at each node (proposed split)
- Adjust each p-value using Bonferroni corrections based on the number of tests required to get to tested node
- Grow tree until a proposed split has an adjusted p-value over a pre-determined threshold (say, α of 0.01 or 0.05)

Permutation Tests: Variable Shuffling

- Problems with Cluster Shuffling approach:
    - Monothetic clustering guaranteed that the chosen splitting variable created the best split in terms of the change in sum of squared distances with that variable
    - F-statistic is also based on sum of squared distances
    - Type I error rate is inflated so tends to end up with too many splits
- Remedy 1: Modify test statistic to exclude splitting variables from test statistic calculation (use pseudo-F)
- Remedy 2: Permutation Tests: Variable Shuffling
    - Permute the values of the selected splitting variable
    - Re-optimize split
    - Test statistic is a "measure of clustering": average silhouette width or CH's pseudo-F

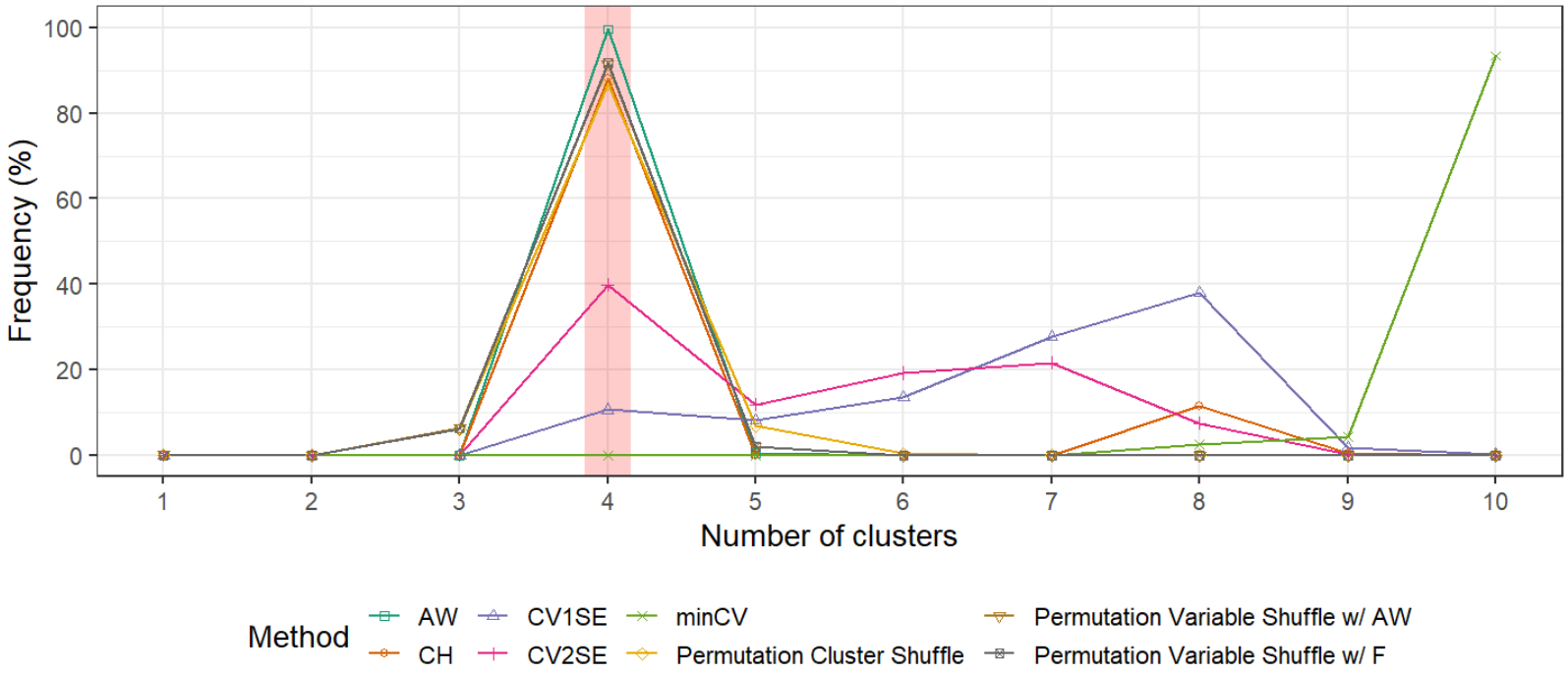
Simulation Study 1: Unclusterable Data

- Assess whether no or some cluster structure is present
- Simulated data
  - Sampled from a multivariate uniform distribution
  - Three different sizes 200 x 4, 200 x 8, and 300 x 4
  - 1,000 data sets each size
- Significance level of 0.05 is used for the hypothesis tests

	Rate of not choosing one cluster result		
Clustering method	**200 x 4**	**200 x 8**	**300 x 4**
Cluster shuffling	0.150	0.251	0.142
Variable Shuffling w/ AW	0.074	0.112	0.064
Variable Shuffling w/ F	0.076	0.111	0.065
minCV	1.000	1.000	1.000
CV1SE	0.994	0.596	1.000
CV2SE	0.419	0.029	0.804

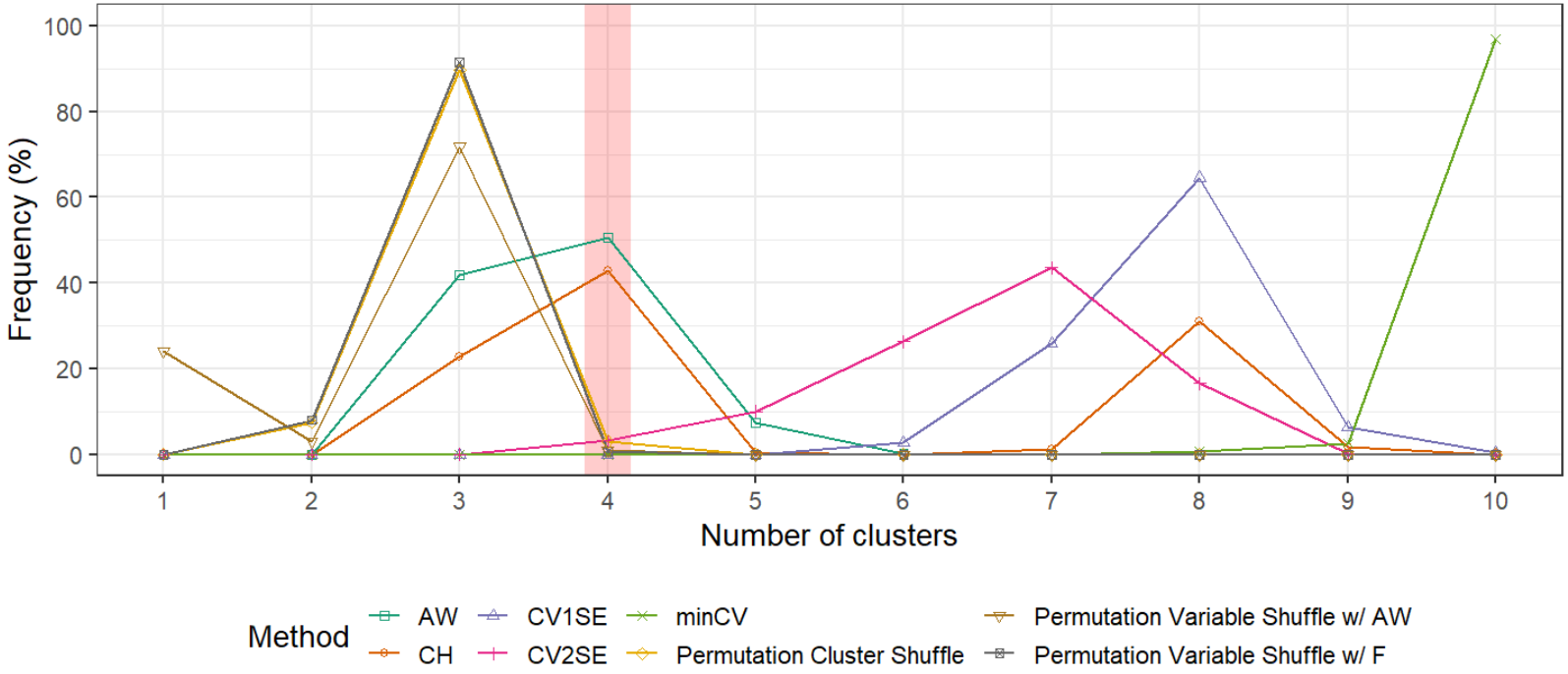
Simulation Study 2a: Accuracy in Choosing the Number of Clusters

- Assess rate of selection of correct number of clusters
- Simulated data with four true clusters:
  - Data size 200 x 4, generated from multivariate normal distributions $N(\mu_i, I_4)$ where $\mu_i \sim N(0, 5^2)$, $i = 1 \ldots 4$, the distance between two closest observations in two clusters is at least 2 units



Simulation Study 2b: Accuracy in Choosing the Number of Clusters

- Simulated data:
  - Data size 200 x 5
  - The previous scenario with an extra noise variable generated from $N(0, 1)$
  - The data sets were standardized before applying the clustering algorithms



A Hybrid Approach

- AW and CH's F performed well in choosing the "correct" number of clusters
  - But they can't pick the one cluster solution
- Permutation-based hypothesis tests has good (not great) Type I error rates in unclusterable data
  - But they cannot compete when > 1 cluster structure is present
- We suggest a hybrid approach for choosing the number of clusters in monothetic
  - Variable shuffling hypothesis test is used first to decide if there is a cluster structure
  - A classic measure of clustering is applied if evidence of rejecting the null hypothesis is strong
  - Should use the same statistic in both stages

Simulation Study 3: The Hybrid Approach

- One true cluster: 500 data sets with the size of 200 x 4, generated from a multivariate uniform distribution (similar to Simulation Study 1: Unclusterable Data)
- Two true clusters: 500 data sets with the size of 200 x 4, generated from a multivariate normal distributions $N(\mu_i, I_4)$ where $\mu_i \sim N(0, 5^2)$, $i = 1, 2$ (similar to Simulation Study 2a)

	One true cluster			Two true clusters			
Clustering method	1	> 1	Rate	1	2	> 2	Rate
CH's F	0	500	0.000	0	500	0	1.00
PVS-F*	462	38	0.924	25	455	20	0.91
Hybrid	462	38	0.924	25	475	0	0.95

* Permutation Variable Shuffling w/ F

Summary

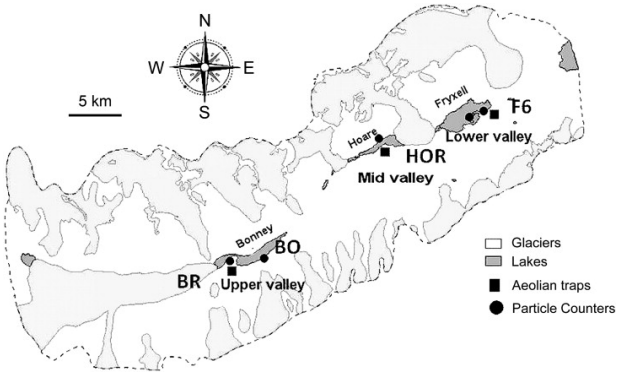
- Choosing a reasonable number of clusters assists greatly in understanding and interpreting the characteristics of a data set
- Cross-validation C_K keeps decreasing when K increases and did not work consistently in the simulation studies
- Clustering algorithms always generate groups. This should not be done if there is no cluster structure.
- The hybrid method has potential in improving the ability to choose the correct of number of clusters even when the data are on the edge of having one or two clusters
    - Larger sample sizes increase the accuracy of this approach

Chapter 3: Data with Circular Variables

Particle Counts in the Antarctica

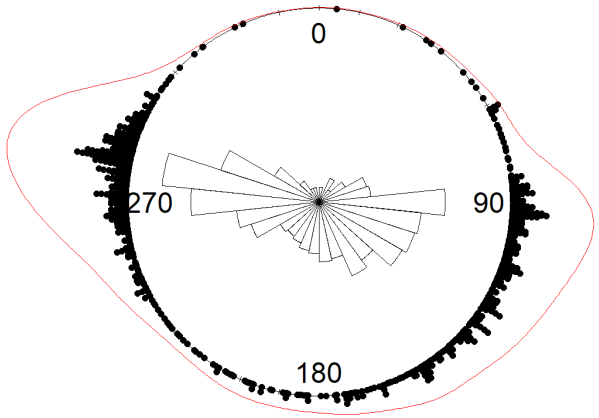
- Let's take a trip!
- The data were introduced in Šabacká, Priscu, Basagic, et al. (2012)
- Variables include:
  - Particle counts: measured over 1 minute every 15 minutes
  - Wind speed (m/s): measured every 4 seconds and averaged at 15-minute intervals
  - Wind direction (in degrees): measured every 30 seconds and averaged at 15-minute intervals

- Modifications:
  - Subset of July 7-14, 2008 (11% of the recorded winds had particles compared to overall 2.39% in the whole time recorded), $n = 673$
  - Particle counts were transformed into particle existence (`has.sensit`)



Circular Variables

- Circular variables are measured in forms of angles or two-dimensional orientations
  - Times of day of occurrences
  - Aspect of slope (directional orientation)
  - Wind and ocean current directions



- Properties and challenges (Antarctica data set)
  - The beginning coincides with the end
  - Zero value position: north
  - Clockwise or counter-clockwise: clockwise
  - Interpretation: zero is from the north to the south or from the south to the north

Dissimilarity Measure

- For two angles y_{iq} and y_{jq} (in degrees) of a circular variable q, we suggest using

$$d(y_{iq}, y_{jq}) = \frac{180 - \left|180 - |y_{iq} - y_{jq}|\right|}{180}$$

- This dissimilarity measure (Jammalamadaka and SenGupta, 2001) provides values between 0 and 1
- It fits well with Gower's dissimilarity for a data set with Q variables

$$d_{gow}(\mathbf{y_i}, \mathbf{y_j}) = \frac{1}{Q} \sum_{q=1}^{Q} d_{gow}(y_{iq}, y_{jq}),$$

  - If q is a linear quantitative variable

$$d_{gow}(y_{iq}, y_{jq}) = \frac{|y_{iq} - y_{jq}|}{\max_{i,j} |y_{iq} - y_{jq}|}$$

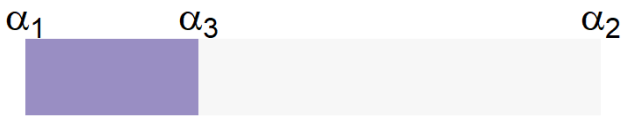
  - If q is a categorical variable
    $d_{gow}(y_{iq}, y_{jq}) = 0$ if y_{iq} and y_{jq} are in the same category, and 1 otherwise

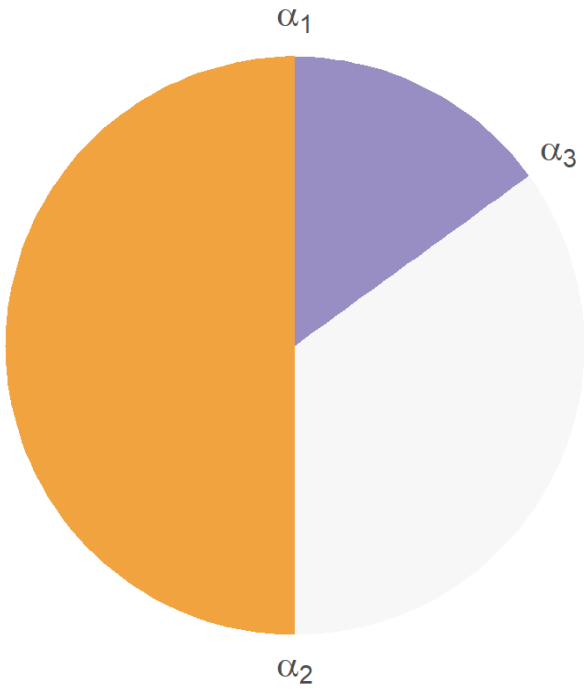
Splitting on a Circular Variable in Monothetic Clustering

- Circular variable needs two splits to create the first two clusters
  - An exhaustive search needs to be made
  - "Clocking"

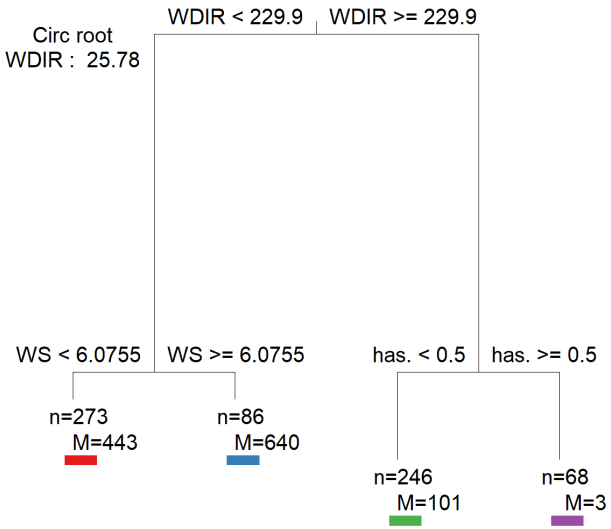
Splitting on a Circular Variable

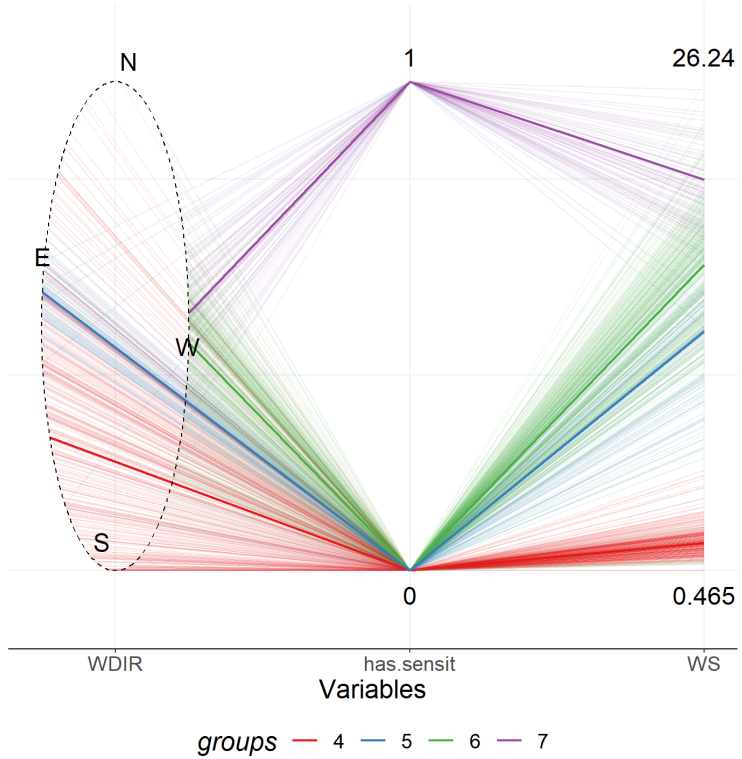
- After that, every split creates a new cluster, as in a linear variable
  - Having the first two clusters
  - Shift the zero-direction to the "hour" hand, and consider the variable as linear
- Until the circular variable is split, the clocking keeps occurring





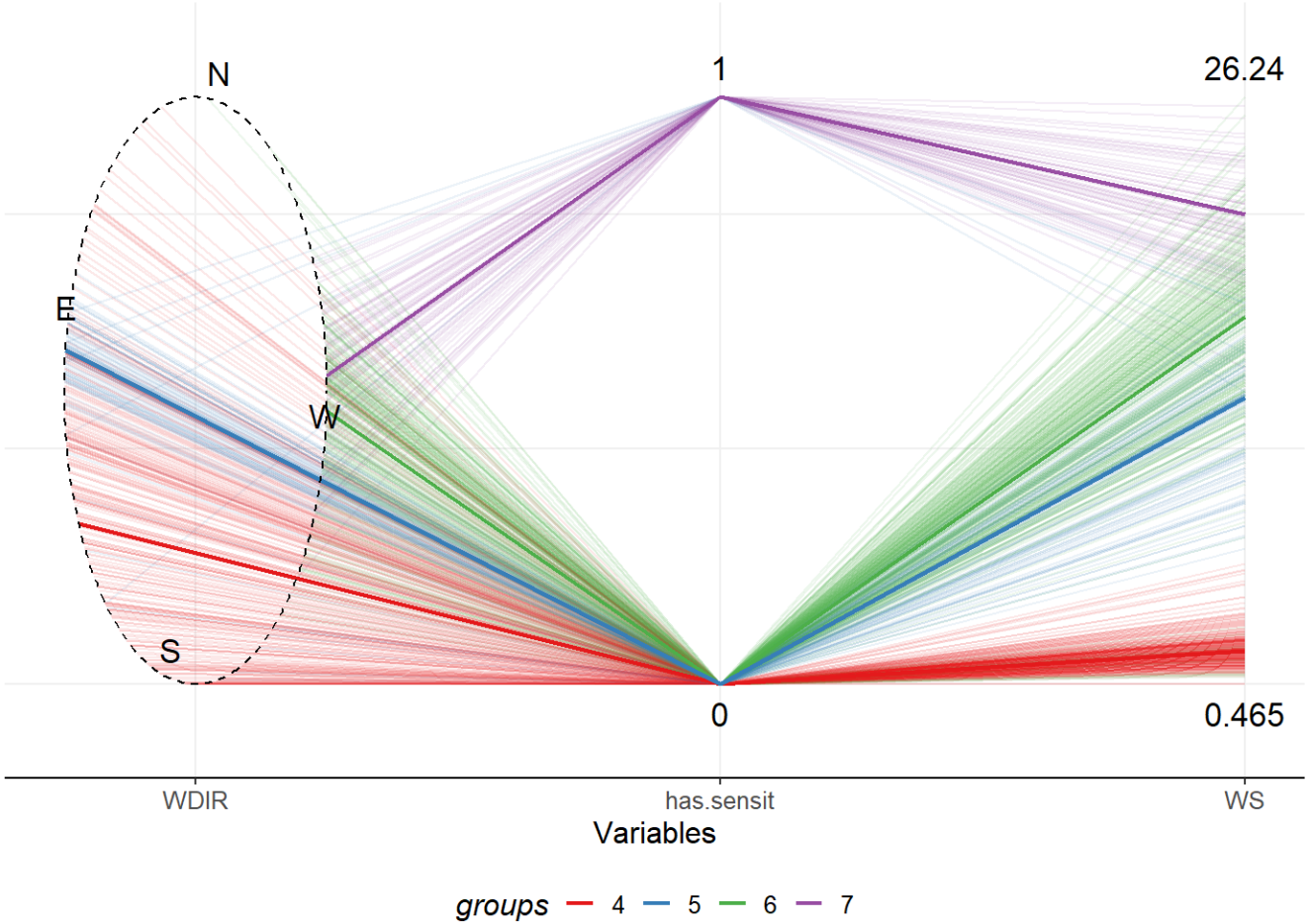
Monothetic Clustering on the Antarctica Data Set





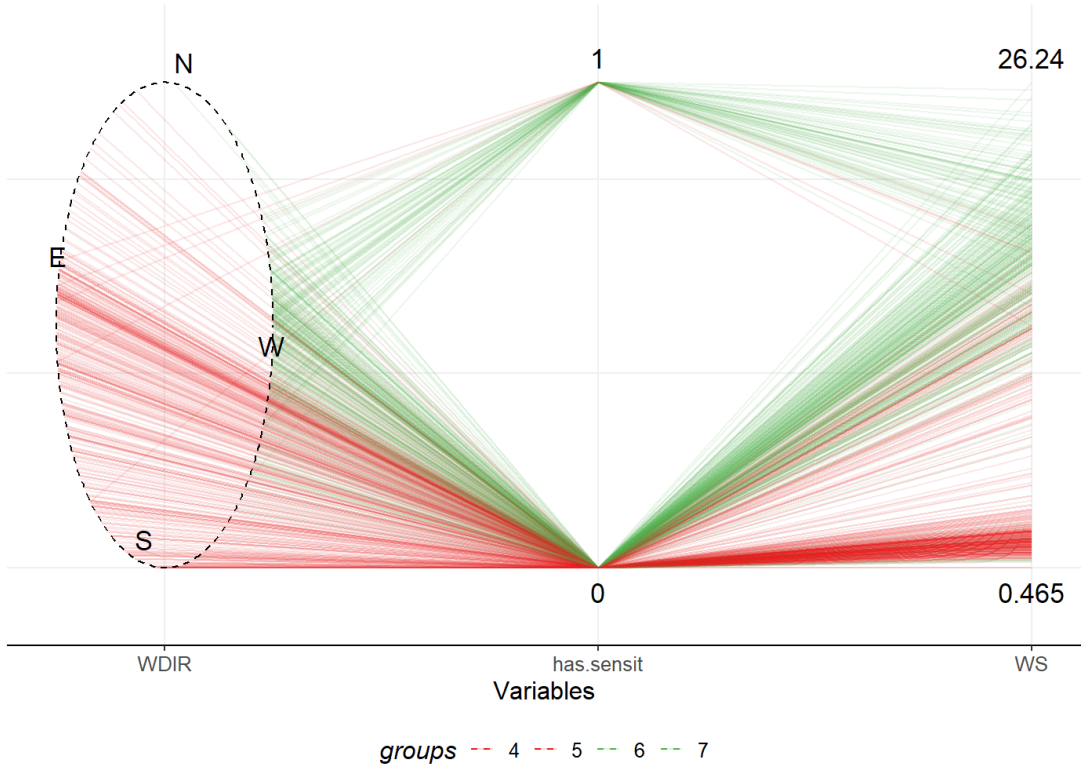
Parallel Coordinates Plot for Circular Variable

- Inspired by Will (2016)'s writing project



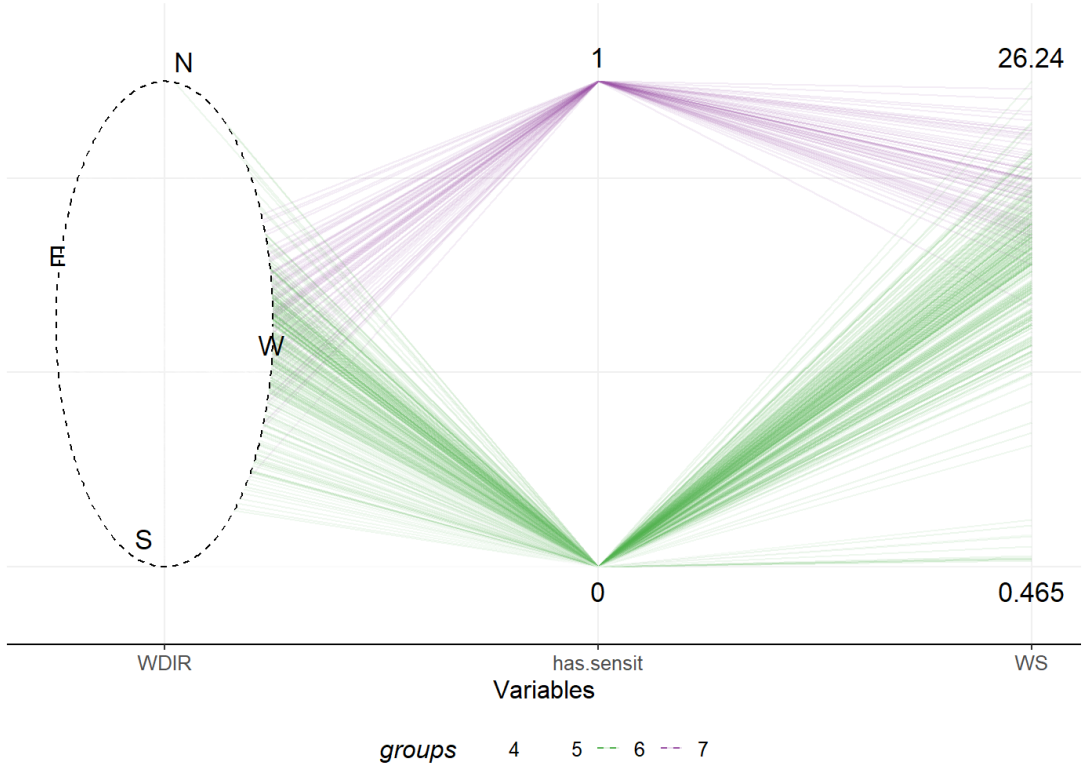
A Closer Look

- 25.78 < WDIR < 229.9 and WDIR >= 229.9 WDIR < 25.78
- Winds came from the East/South-East (up-valley sea breezes) and winds came from the West (down-valley föhn winds)



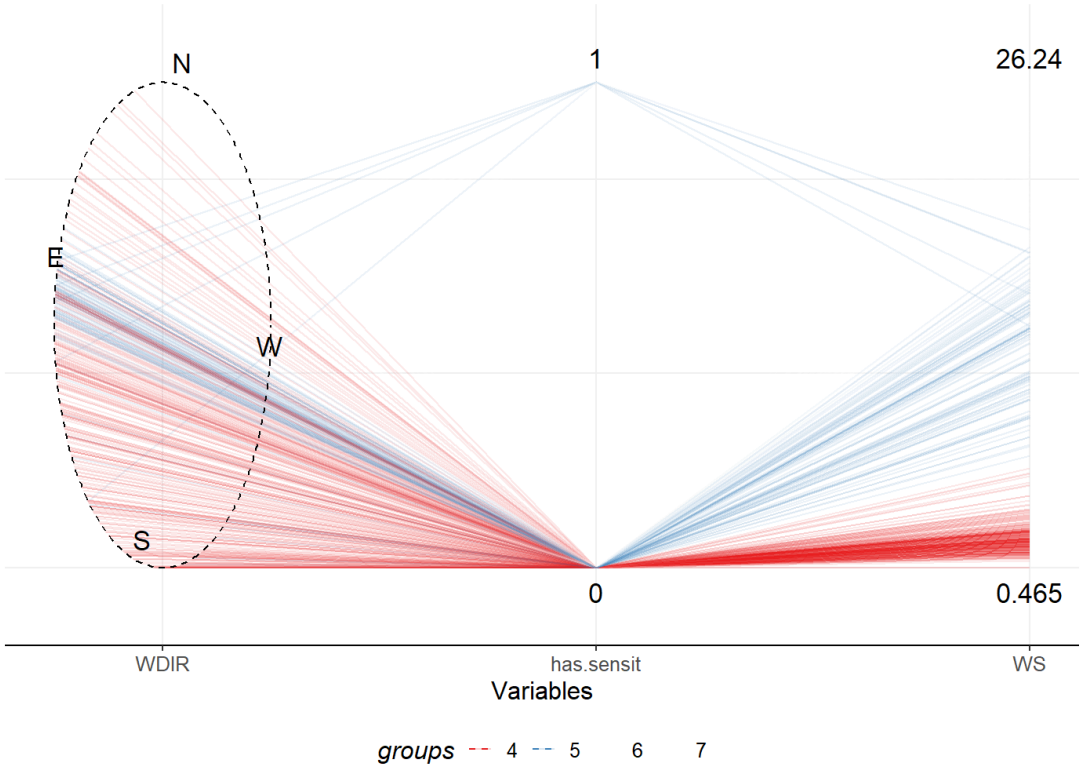
A Closer Look

- Winds from the West, has particles or does not have particles



A Closer Look

- Winds from the east, split at wind speed
- Stronger winds come directly from the east (and some had particles), winds coming from the south-east and south were very weak.



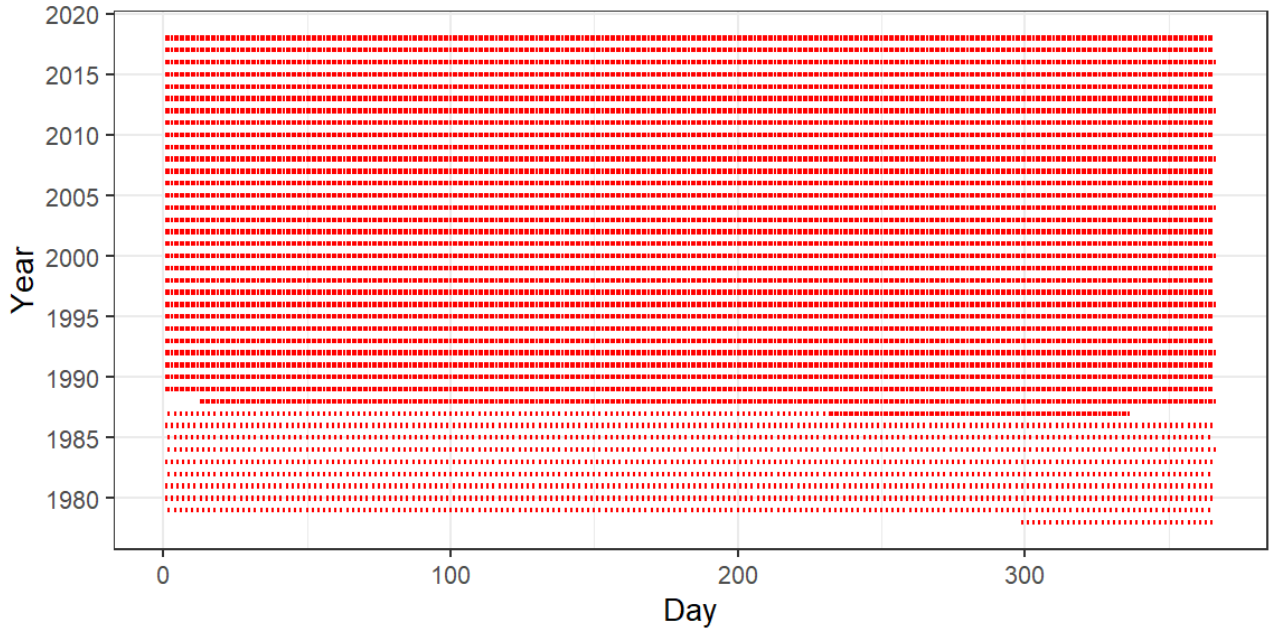
Summary of Chapter 3

- Data sets with circular variables have interesting directional information that are useful in many applications
- Monothetic clustering works with mixed variables including circular variables, using Gower's dissimilarity, without losing their natural characteristics
- Visualizations are important to assist in interpretation
- Parallel coordinates plot depicting circular variables as ellipses retain most features of the variables
- Monothetic clustering results provided interesting results of the Antarctic particle count data
    - Explored the multivariate relationships between three variables

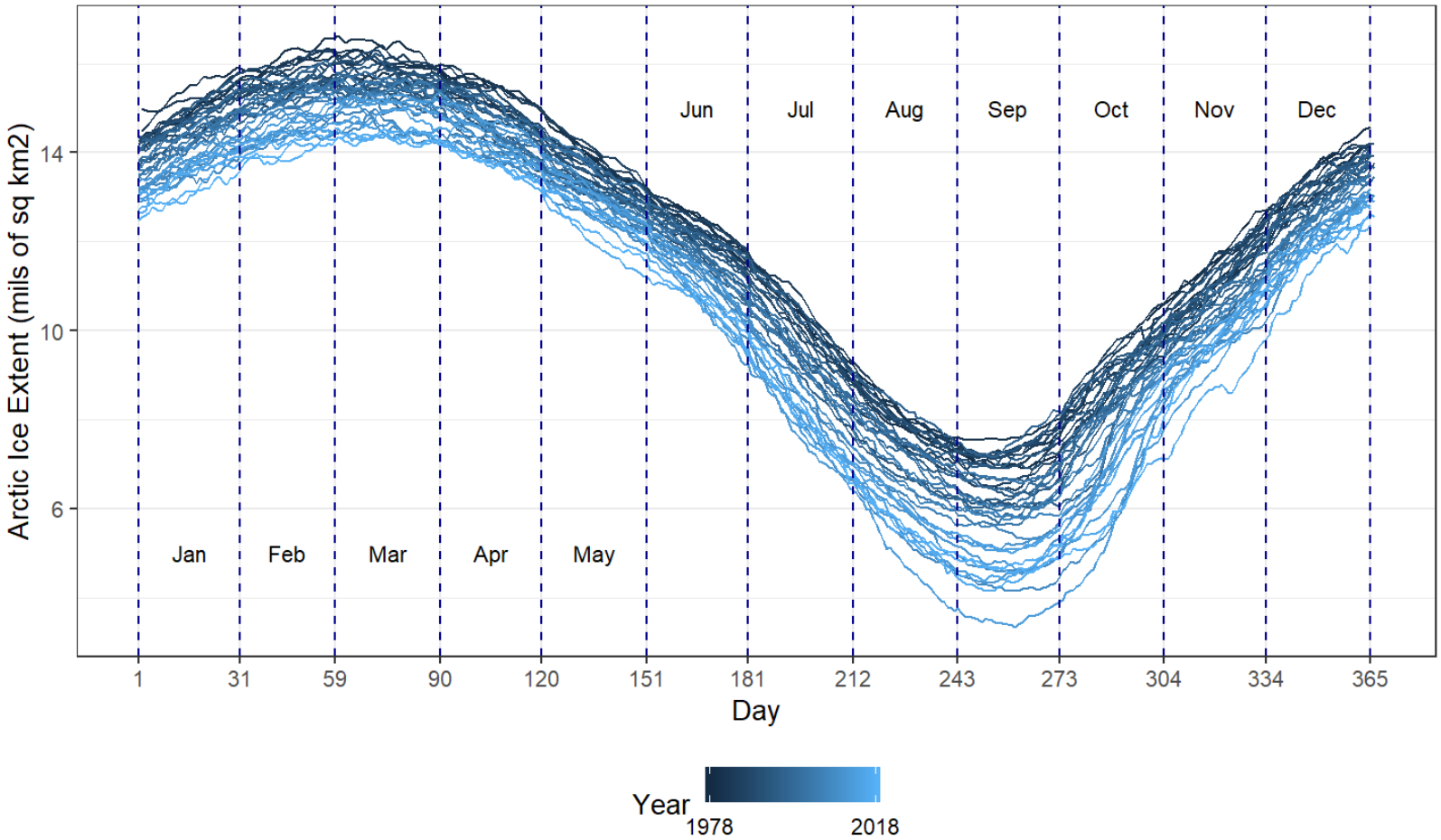
Chapter 4: Clustering Functional Data

Arctic Sea Ice Extent Data

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



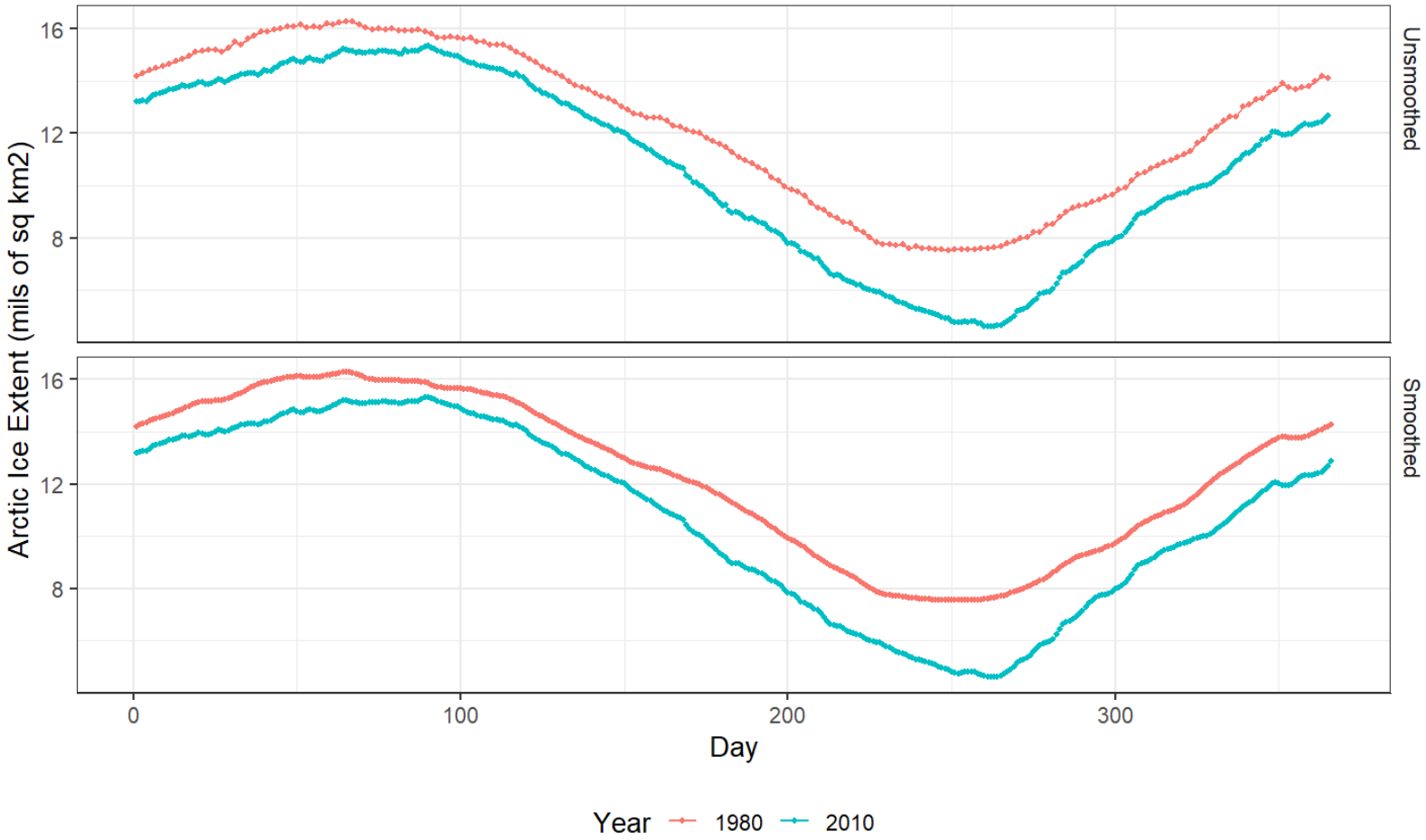
Raw Arctic Ice Extent



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

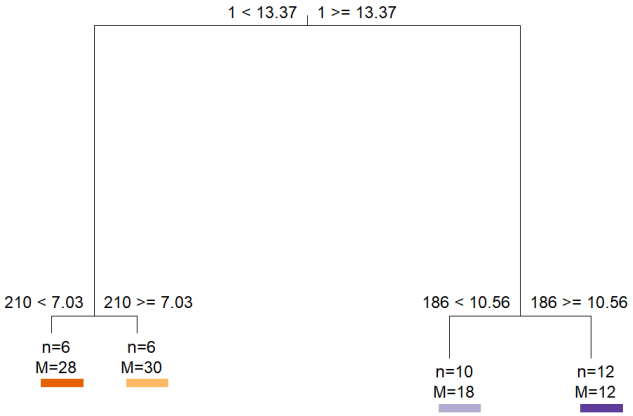
- The L_2 distance matrix can be be calculated and used for non-functional clustering algorithms

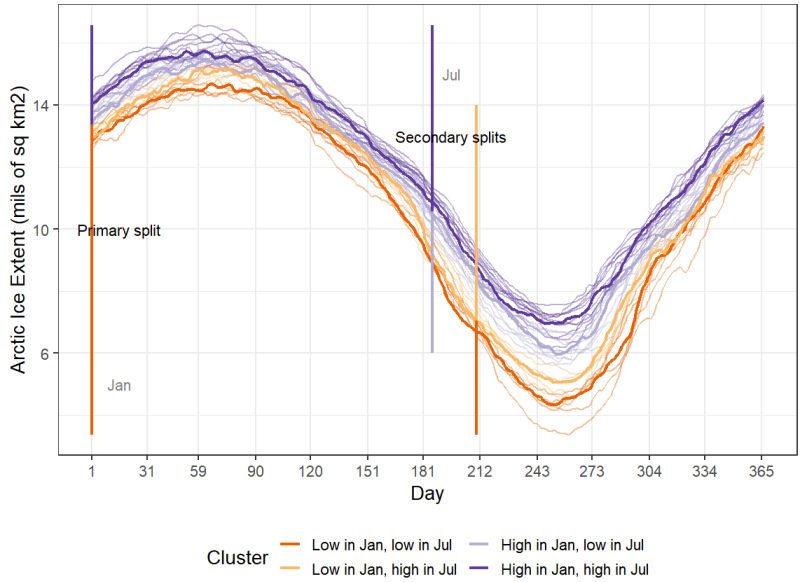
$$d(y_i, y_j) = \sqrt{\int_T [y_i(t) - y_j(t)]^2 dt}$$

- 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





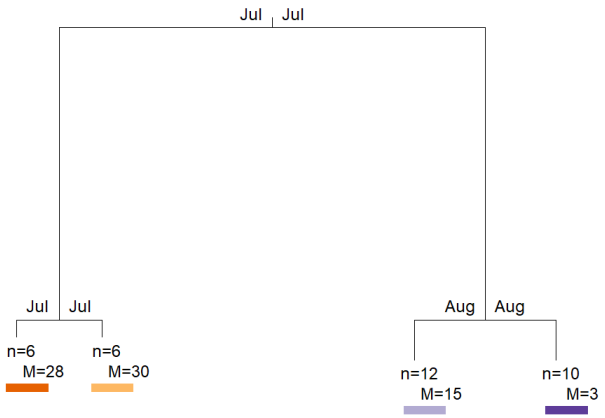
Partitioning Using Local Subregions

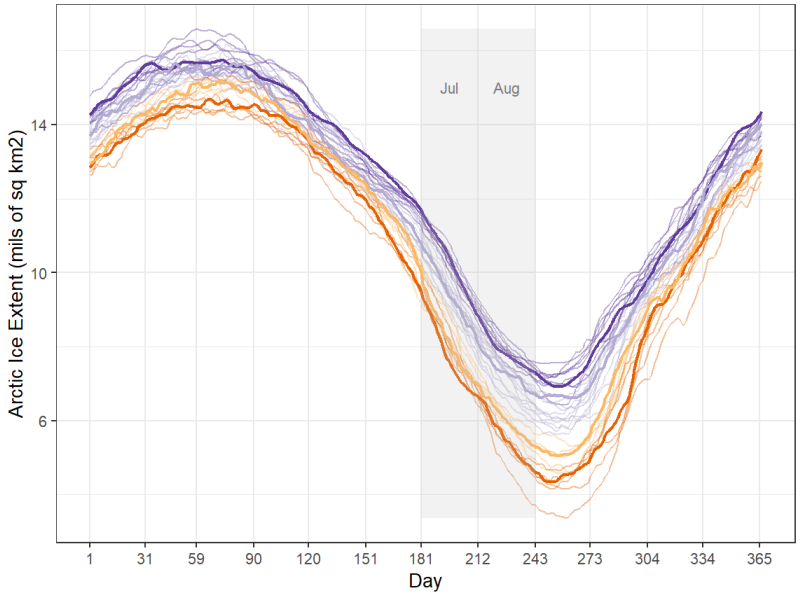
- By aggregating over regions of time, we develop a new clustering algorithm related to monothetic clustering that is more suited to functional data
- PULS recursively bi-partitions functional data using only groups from subregions.
- For each subregion $[a_1, b_1], \ldots, [a_R, b_R]$, calculate an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

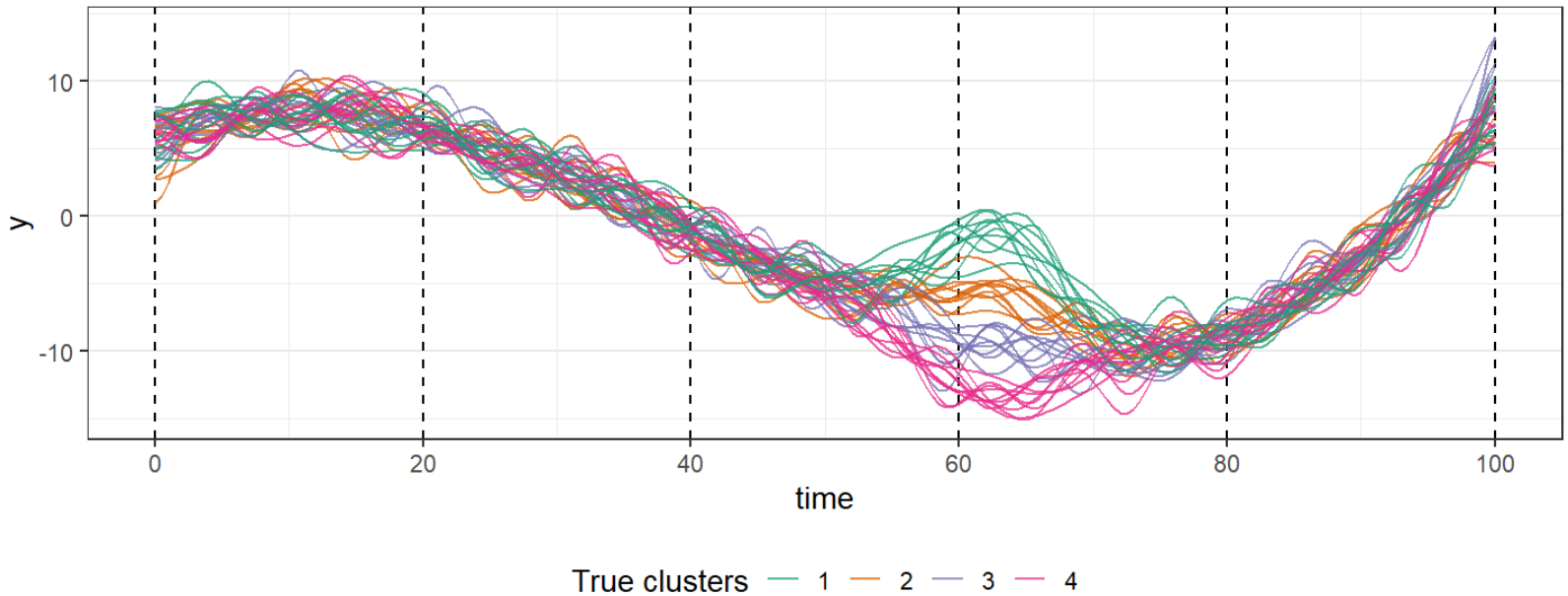
Name	PULS	MonoClust
High Jan, High Jul	1979-1981, 1983-1984,	1979-1981, 1983-1986, 1989,
	1986, 1989, 1992, 1994, 1996	1992-1994, 1996
High Jan, Low Jul	1985, 1991, 1993, 1995,	1991, 1995,
	1997-2004	1997-2000-2004
Low Jan, High Jul	2005, 2008-2010, 2013, 2014	2005, 2008-2010, 2013, 2014
Low Jan, Low 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 Jan
1982	1990	2006	2018

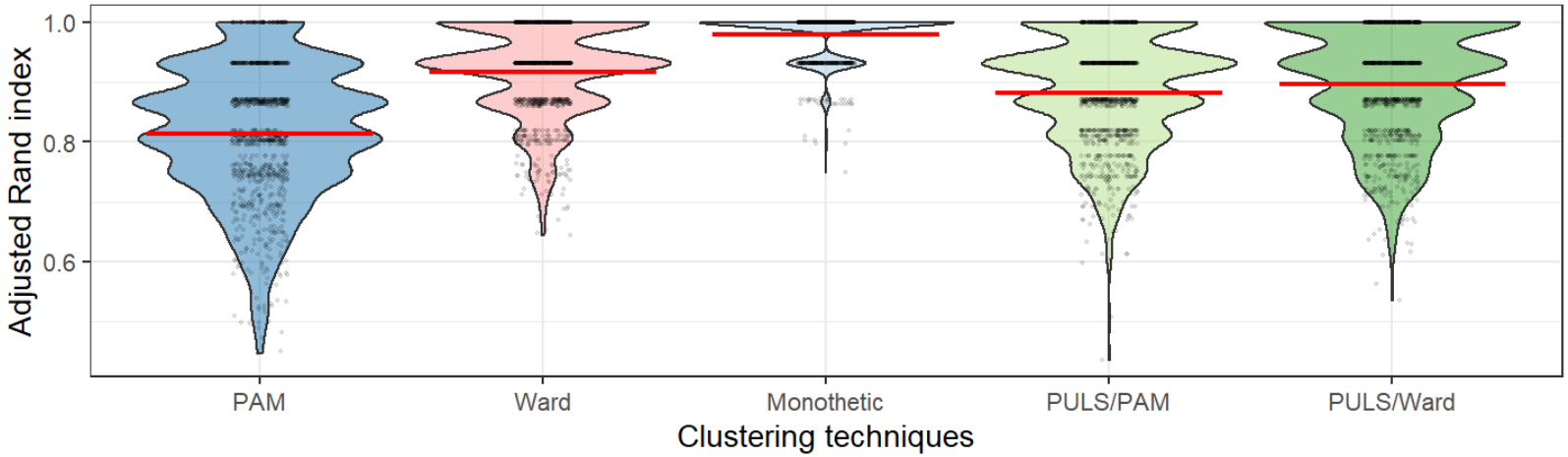
Simulation Study

- Compare the performance of various clustering techniques on functional data
  - True functional curves with index (time) ranges from 0 to 100
  - Identical except for 50-70 interval
  - White noise are added
- Monothetic clustering, PAM, and Ward's method on a fine grid at each 0.5 unit, $Q = 200$
- PULS with PAM and PULS with Ward's method on distance matrix calculated by L_2 on functions on five subregions



Simulation Study Results: Corrected Rand Index

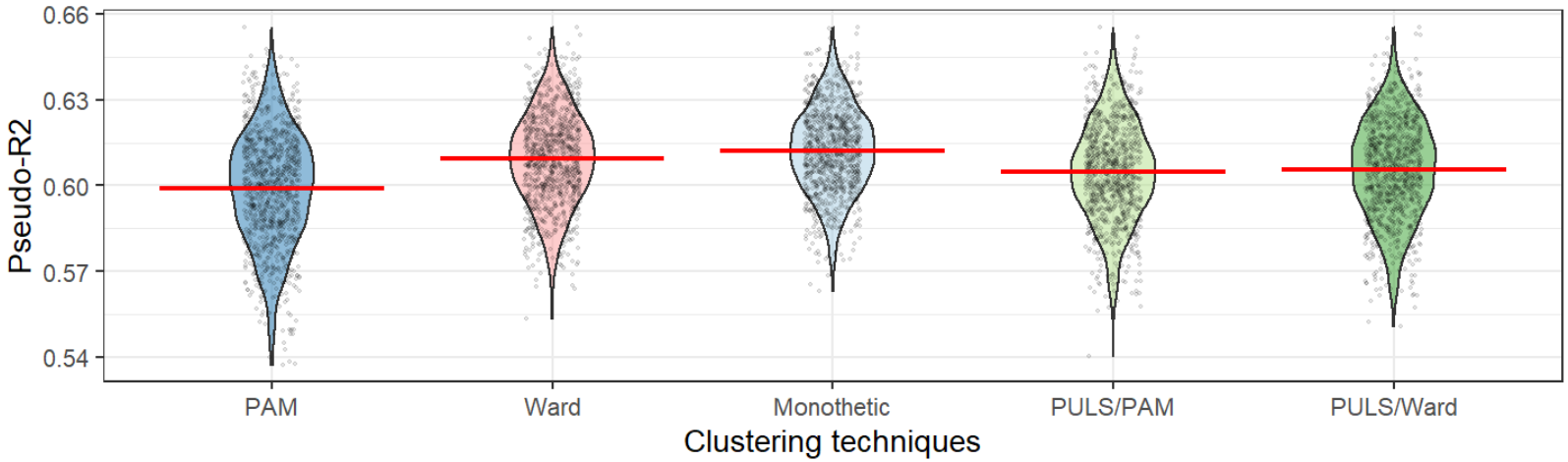
- Rand Index (Rand, 1971) is a measure of agreement between two cluster solutions
- Corrected Rand index (Hubert and Arabic, 1985) is corrected for agreement by chance, bounded between 0 and 1



- Corrected Rand index:
  - 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

Simulation Study Results: Pseudo-R2

- Pseudo-R2 measures the explained variation of the data set



- Pseudo-R2:
  - There is limited evidence of differences across the methods
  - On average, monothetic clustering explained the most variance in the data
  - Two methods in PULS have similar performance with Ward's slightly better

Chapter 5: R Packages and Vignette

- Monothetic clustering and PULS are implemented in the **MonoClust** and **PULS** R packages published on Github
- Packages' documentation and vignette are available
- Most of the calculations and plots shown are done with the two packages

MonoClust

- Inherit the object structure and borrow the tree output from the **rpart** package for classification and regression trees (Breiman, Friedman, Stone, et al., 1984; Therneau and Atkinson, 2018)
- Search for the best split is done by exhaustive search: the global optimum is guaranteed but slow when sample size and/or number of variables are large

PULS

- Take a functional data object (from **fda**, Ramsay, Wickham, Graves, et al., 2018) as input
- Many functions and contents are borrowed from **MonoClust**

Chapter 6: Conclusions and Future Extensions

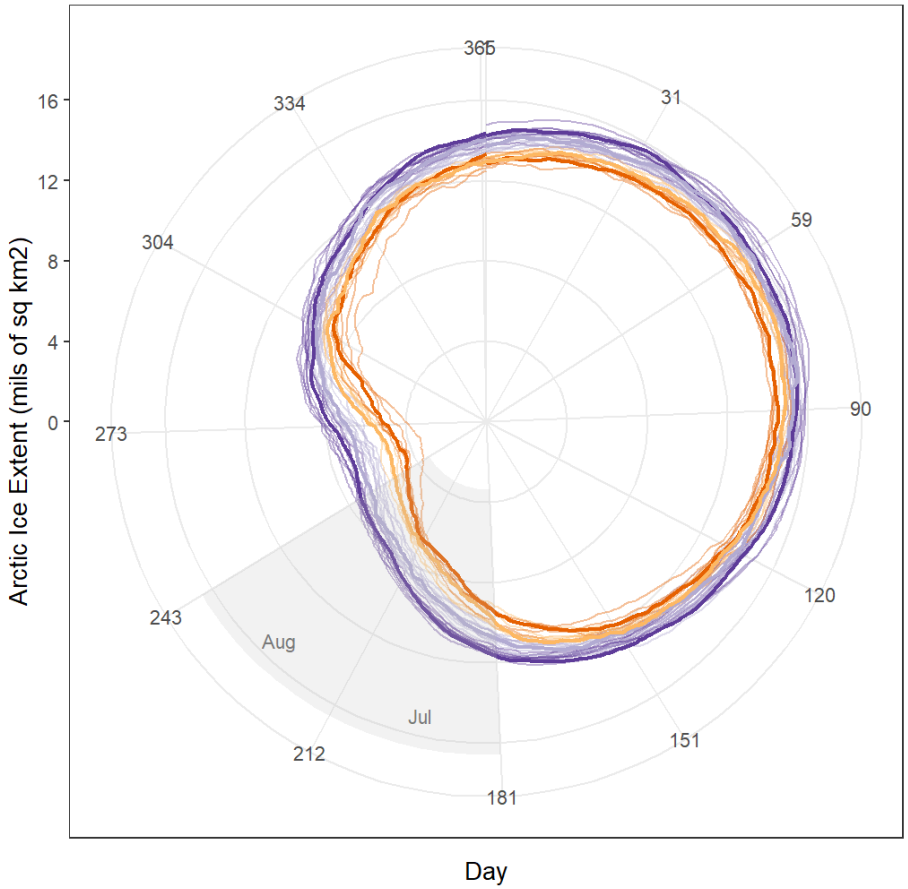
- Monothetic clustering bi-partitions data at the values of one variable at a time
  - Clusters share the same characteristics from split variables
  - Can interpret the resulting clusters
  - Can predict the cluster a new observation would fall into
- The hybrid methods for choosing the number of clusters improve the ability to choose a reasonable number of clusters, including one cluster solution
  - Worked in the simulation study of monothetic clustering, and could be used in other tree-based clustering techniques
- Monothetic clustering algorithm has been shown to work with data with mixed types of variables, including circular variables
- PULS was based on monothetic clustering and designed to work directly with functional data

Future Extensions

- Hybrid method using hypothesis test could be extended to hierarchical clustering
- A dissimilarity measure of circular variables based on trigonometric transformation should be examined and compared
- Exhaustive search is slow, heuristic search algorithms may help with some trade-offs
    - Results of best split for each variable could be stored so they won't unnecessarily repeat
- Sparse clustering (Witten and Tibshirani, 2010) can be applied to functional data to "weight" the contributions of variables (days) to the clustering process
- PULS could be applied to non-functional data with known "grouped" variables to detect the underlying structure of data.
    - Potential: ratio of proteins in MS patients (mentioned in Chapter 1)
- Clustering the derivative of the Arctic ice extent data

Future Extensions

- The Arctic ice extent data: treat the functional data as periodic (circular) data



Acknowledgement

- Dr. Mark Greenwood for the guidance and patience
- Brian McGuire, Garland Will
- Dr. Megan Higgs and Dr. Lillian Lin and SCRS (funded by MT INBRE, CAIRHE, AI-AN CTRP, MW CTR-IN)
- Department of Mathematical Sciences's Gary Sackett Fellowship and Travel Grant
- The Vietnam Education Foundation
- My PhD Committee Members:
    - Dr. Mark Greenwood
    - Dr. John Borkowski
    - Dr. Laura Hildreth
    - Dr. Megan Higgs, and
    - Dr. Nicole Carnegie
- ...And my parents, my loving wife and son, who just overcame jet lag several days ago

References I

Šabacká, M., J. C. Priscu, H. J. Basagic, A. G. Fountain, D. H. Wall, R. A. Virginia, and M. C. Greenwood (2012). "Aeolian flux of biotic and abiotic material in Taylor Valley, Antarctica". In: *Geomorphology* 155-156, pp. 102-111. ISSN: 0169555X. DOI: 10.1016/j.geomorph.2011.12.009.

Anderson, M. J. (2001). "A new method for non-parametric multivariate analysis of variance". In: *Austral Ecology* 26.1, pp. 32-46. ISSN: 14429985. DOI: 10.1111/j.1442-9993.2001.01070.pp.x.

Barbour, C., P. Kosa, M. Komori, M. Tanigawa, R. Masvekar, T. Wu, K. Johnson, P. Douvaras, V. Fossati, R. Herbst, Y. Wang, K. Tan, M. Greenwood, and B. Bielekova (2017). "Molecular-based diagnosis of multiple sclerosis and its progressive stage". In: *Annals of Neurology* 82.5, pp. 795-812. ISSN: 03645134. DOI: 10.1002/ana.25083.

Breiman, L., J. Friedman, C. J. Stone, and R. Olshen (1984). *Classification and Regression Trees*. 1st ed. Chapman and Hall/CRC. ISBN: 0412048418.

Caliński, T. and J. Harabasz (1974). "A dendrite method for cluster analysis". En. In: *Communications in Statistics* 3.1, pp. 1-27.

References II

Chavent, M. (1998). "A monothetic clustering method". In: *Pattern Recognition Letters* 19.11, pp. 989-996. ISSN: 01678655. DOI: 10.1016/S0167-8655(98)00087-7.

Everitt, B. S., S. Landau, M. Leese, and D. Stahl (2011). *Cluster Analysis*. 5th ed. Wiley, p. 346. ISBN: 0470749911.

Fetterer, F., F. Knowles, W. Meier, M. Savoie, and A. K. Windnagel (2018). *Sea Ice Index, Version 3*. DOI: 10.7265/N5K072F8. URL: https://nsidc.org/data/g02135 (visited on 2018).

Hothorn, T., K. Hornik, and A. Zeileis (2006). "Unbiased Recursive Partitioning: A Conditional Inference Framework". En. In: *Journal of Computational and Graphical Statistics* 15.3, pp. 651-674. ISSN: 1061-8600. DOI: 10.1198/106186006X133933.

Hubert, L. and P. Arabic (1985). "Comparing Partitions". In: *Journal of Classification* 2, pp. 193-218.

Jammalamadaka, S. R. and A. SenGupta (2001). *Topics in Circular Statistics*. Vol. 5. ISBN: 9789812779267. DOI: 10.1142/9789812779267.

References III

Milligan, G. W. and M. C. Cooper (1985). "An examination of procedures for determining the number of clusters in a data set". In: *Psychometrika* 50.2, pp. 159-179. ISSN: 0033-3123. DOI: 10.1007/BF02294245.

Piccarreta, R. and F. C. Billari (2007). "Clustering work and family trajectories by using a divisive algorithm". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 170.4, pp. 1061-1078. ISSN: 0964-1998. DOI: 10.1111/j.1467-985X.2007.00495.x.

Ramsay, J. O. and B. W. Silverman (2005). *Functional data analysis*. Springer, p. 426. ISBN: 9780387400808.

Ramsay, J. O., H. Wickham, S. Graves, and G. Hooker (2018). *fda: Functional Data Analysis*. R package version 2.4.8.

Rand, W. M. (1971). "Objective Criteria for the Evaluation of Clustering Methods". In: *Journal of the American Statistical Association* 66, pp. 846-850. ISSN: 0162-1459. DOI: 10.1080/01621459.1971.10482356.

References IV

Rousseeuw, P. J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53-65. ISSN: 0377-0427. DOI: 10.1016/0377-0427(87)90125-7.

Therneau, T. and B. Atkinson (2018). *rpart: Recursive Partitioning and Regression Trees*. R package version 4.1-13.

Tibshirani, R., G. Walther, and T. Hastie (2001). "Estimating the number of clusters in a data set via the gap statistic". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63.2, pp. 411-423. ISSN: 1369-7412. DOI: 10.1111/1467-9868.00293.

Will, G. (2016). "Visualizing and Clustering Data that Includes Circular Variables". Writing Project.

Witten, D. M. and R. Tibshirani (2010). "A framework for feature selection". In: *American Statistician* 105.490, pp. 713-726. ISSN: 0162-1459. DOI: 10.1198/jasa.2010.tm09415.A.

Appendix

- Table of Simulation Study 2a

method	1	2	3	4	5	6	7	8	9	10
AW	0.000	0	0.000	0.996	0.004	0.000	0.000	0.000	0.000	0.000
CH	0.000	0	0.000	0.882	0.000	0.000	0.000	0.114	0.004	0.000
CV1SE	0.000	0	0.000	0.106	0.082	0.136	0.276	0.380	0.018	0.002
CV2SE	0.000	0	0.000	0.398	0.118	0.192	0.216	0.074	0.002	0.000
minCV	0.000	0	0.000	0.000	0.000	0.000	0.000	0.024	0.042	0.934
Permutation Cluster Shuffle	0.000	0	0.060	0.868	0.068	0.004	0.000	0.000	0.000	0.000
Permutation Variable Shuffle w/ AW	0.002	0	0.064	0.914	0.020	0.000	0.000	0.000	0.000	0.000
Permutation Variable Shuffle w/ F	0.000	0	0.062	0.918	0.020	0.000	0.000	0.000	0.000	0.000

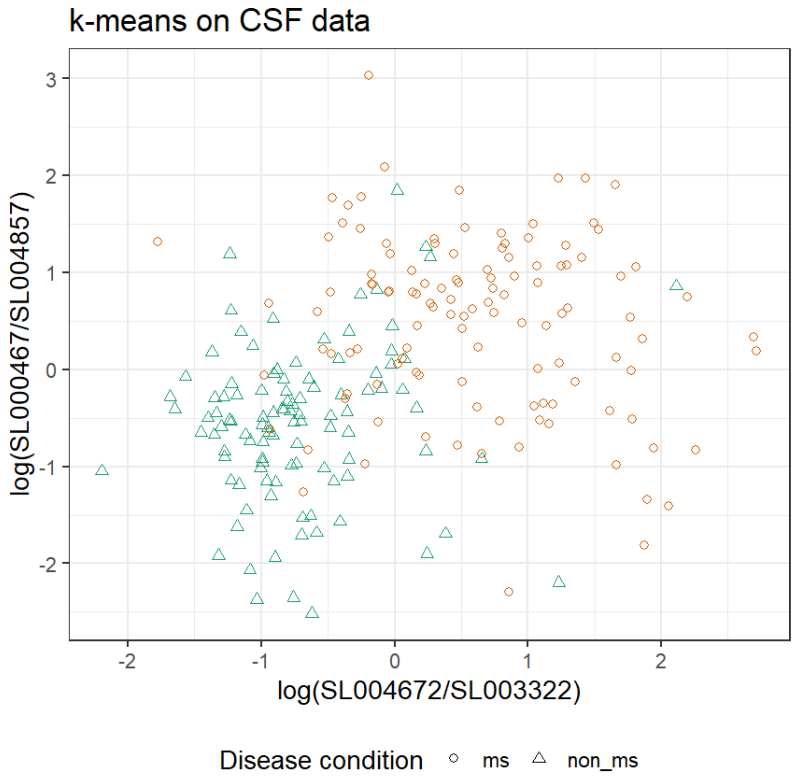
Appendix

- Table of Simulation Study 2b

method	1	2	3	4	5	6	7	8	9	10
AW	0.000	0.000	0.418	0.506	0.074	0.002	0.000	0.000	0.000	0.000
CH	0.000	0.000	0.228	0.428	0.004	0.000	0.012	0.310	0.018	0.000
CV1SE	0.000	0.000	0.000	0.000	0.000	0.028	0.258	0.646	0.064	0.004
CV2SE	0.000	0.000	0.000	0.032	0.100	0.264	0.436	0.166	0.002	0.000
minCV	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.006	0.026	0.968
Permutation Cluster Shuffle	0.000	0.074	0.896	0.030	0.000	0.000	0.000	0.000	0.000	0.000
Permutation Variable Shuffle w/ AW	0.242	0.030	0.718	0.010	0.000	0.000	0.000	0.000	0.000	0.000
Permutation Variable Shuffle w/ F	0.000	0.080	0.914	0.006	0.000	0.000	0.000	0.000	0.000	0.000

Appendix

- Ever wonder how many clusters the hybrid method suggested?



Appendix

- The permutation hypothesis test by shuffling variable using F statistic with $B = 100)$ shows $p = 0.001$
  - It suggests that this data set needs at least 2 clusters
- CH's pseudo-F shows this table

2	3	4	5	6	7	8	9	10
167.5055	161.0543	154.7617	151.6537	153.9777	151.9183	152.344	154.3431	155.4007

- This method chooses $K = 2$ cluster solution
- Ever wonder: which clustering method has the best prediction?

name	rand
k-means	0.5488889
Ward's	0.4971429
mono	0.4360317