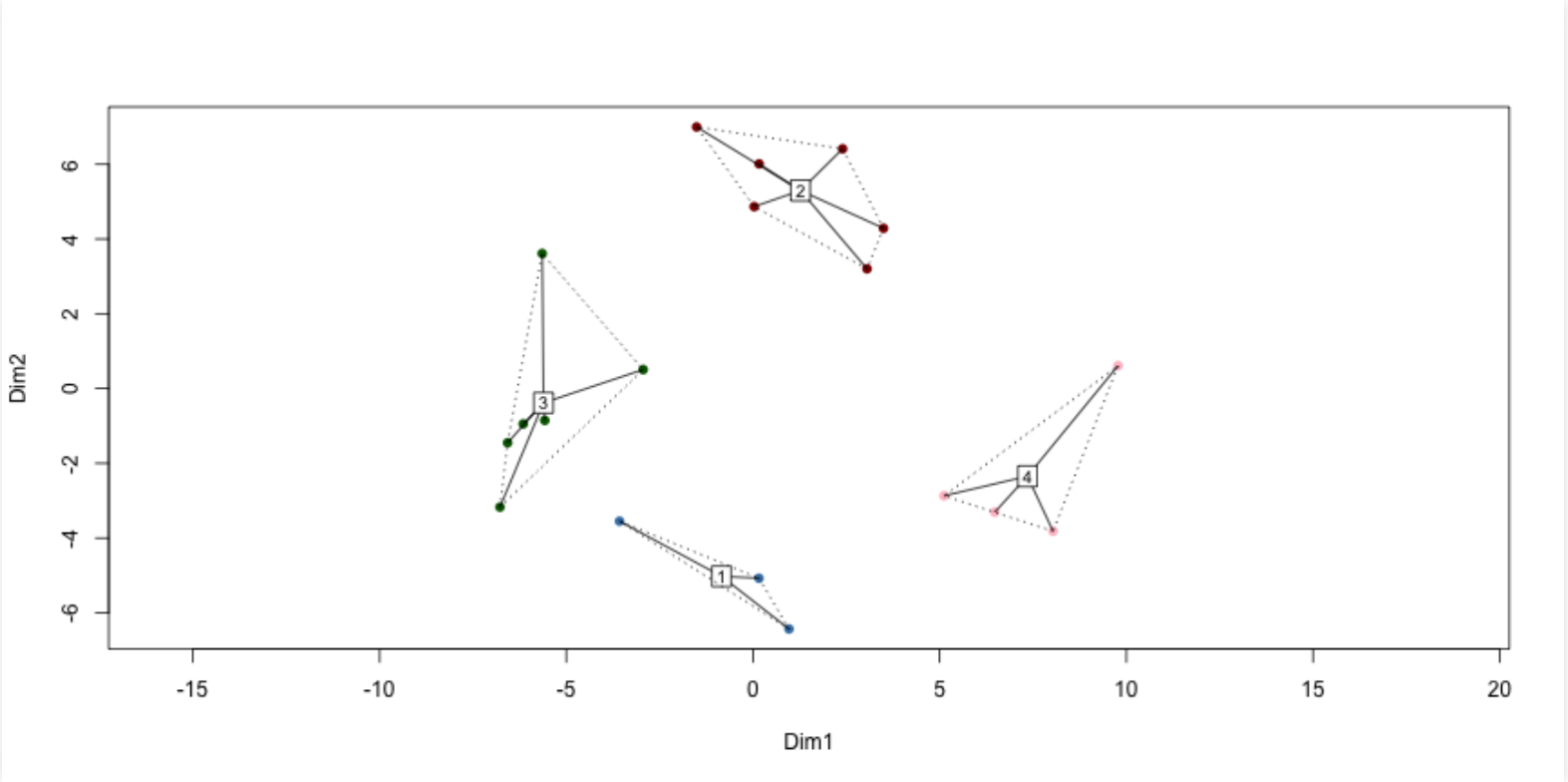
Session 9-10, Clustering and Segmentation (Technical Slides)

T. Evgeniou, A. Ovchinnikov, INSEAD

Clustering and Segmentation



What is Clustering and Segmentation?

Processes and tools to organize data in a few segments, with data being as similar as possible within each segment, and as different as possible across segments

Example Usage

- Market Segmentation
- Co-Moving Asset Classes
- Geo-demographic segmentation
- Recommender Systems
- Text Mining

A Segmentation Process

1. Confirm data is metric
2. Scale the data (Optional)
3. Select Segmentation Variables
4. Define similarity measure
5. Visualize Pair-wise Distances
6. Method and Number of Segments
7. Profile and interpret the segments
8. Robustness Analysis

Example Data: Market Research Survey

V1: Shopping is fun (scale 1-7)

V2: Shopping is bad for your budget (scale 1-7)

V3: I combine shopping with eating out (scale 1-7)

V4: I try to get the best buys while shopping (scale 1-7)

V5: I don't care about shopping (scale 1-7)

V6: You can save lot of money by comparingprices (scale 1-7)

Income: the household income of the respondent (in dollars)

Mall.Visits: how often they visit the mall (scale 1-7)

Step 1: Confirm data is metric

	Variables	V1	V2	V3	V4	V5	V6
1	1	6	4	7	3	2	3
2	2	2	3	1	4	5	4
3	3	7	2	6	4	1	3
4	4	4	6	4	5	3	6
5	5	1	3	2	2	6	4
6	6	6	4	6	3	3	4
7	7	5	3	6	3	3	4
8	8	7	3	7	4	1	4

Step 2: Scale the data (Optional)

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	V1	1	2	4	3.85	5.25	7	1.87
2	V2	2	3	4	4.1	5	7	1.39
3	V3	1	2	4	3.95	6	7	1.99
4	V4	2	3	4	4.1	5.25	7	1.5
5	V5	1	2	3.5	3.45	4.25	7	1.74
6	V6	2	3	4	4.35	5.25	7	1.48

Data Standardization: Example Code

```
ProjectData_segment_scaled=apply(ProjectData_segment,2
 function(r) {
  if (sd(r)!=0) {
    res=(r-mean(r))/sd(r)
    } else {
      res=0*r; res
      }
  })
```

Standardized Data: Summary Statistics

	Variables	min	X25.percent	median	mean	X75.percent	max	std
1	V1	-1.52	-0.99	0.08	0	0.75	1.68	1
2	V2	-1.51	-0.79	-0.07	0	0.65	2.08	1
3	V3	-1.49	-0.98	0.03	0	1.03	1.54	1
4	V4	-1.4	-0.73	-0.07	0	0.77	1.93	1
5	V5	-1.41	-0.83	0.03	0	0.46	2.04	1
6	V6	-1.59	-0.91	-0.24	0	0.61	1.79	1

Step 3. Select Segmentation Variables

The choice of the variables used for clustering is critically important

Tpically we use different variables for segmentation (the “segmentation variables”) and different ones for profiling (the “profiling variables”)

Remember: Segmentation is an iterative process

Step 4. Define similarity measure

Defining what we mean when we say “similar” or “different” observations is a key part of cluster analysis which often requires a lot of contextual knowledge and creativity

There are literally thousands of rigorous mathematical definitions of distance between observations/vectors

The user can manually define such distance metrics

Distances across our data using the Euclidean distance

Pairwise Distances between the first 5 observations using The Euclidean Distance Metric

1	2	3	4	5
0.0				
8.0	0.0			
2.8	8.2	0.0		
5.6	5.6	6.6	0.0	
8.3	2.6	9.1	6.6	0.0

Distances across our data using the Manhattan distance

Pairwise Distances between the first 5 observations using The Manhattan Distance Metric

1	2	3	4	5
0.0				
16.0	0.0			
6.0	16.0	0.0		
13.0	13.0	15.0	0.0	
17.0	5.0	19.0	16.0	0.0

Manually Defined Distances: an Example

```
My_Distance_function<-function(x,y)
{sum(abs(x-y)>2)}
```

Manually Defined Distances: an Example

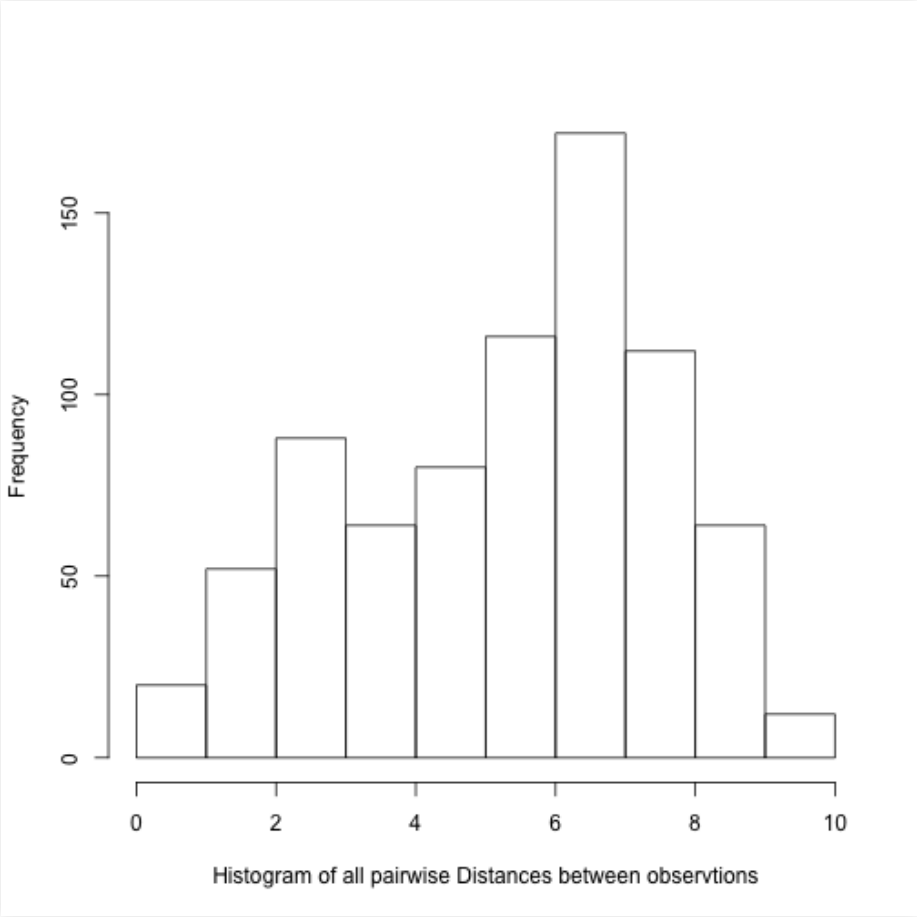
Pairwise Distances between the first 5 observations using a simple manually defined Distance Metric

1	2	3	4	5
0.0				
3.0	0.0			
0.0	3.0	0.0		
2.0	2.0	3.0	0.0	
3.0	0.0	3.0	4.0	0.0

Step 5. Visualize Pair-wise Distances

Histogram of all pairwise distances

Distance Used: euclidean



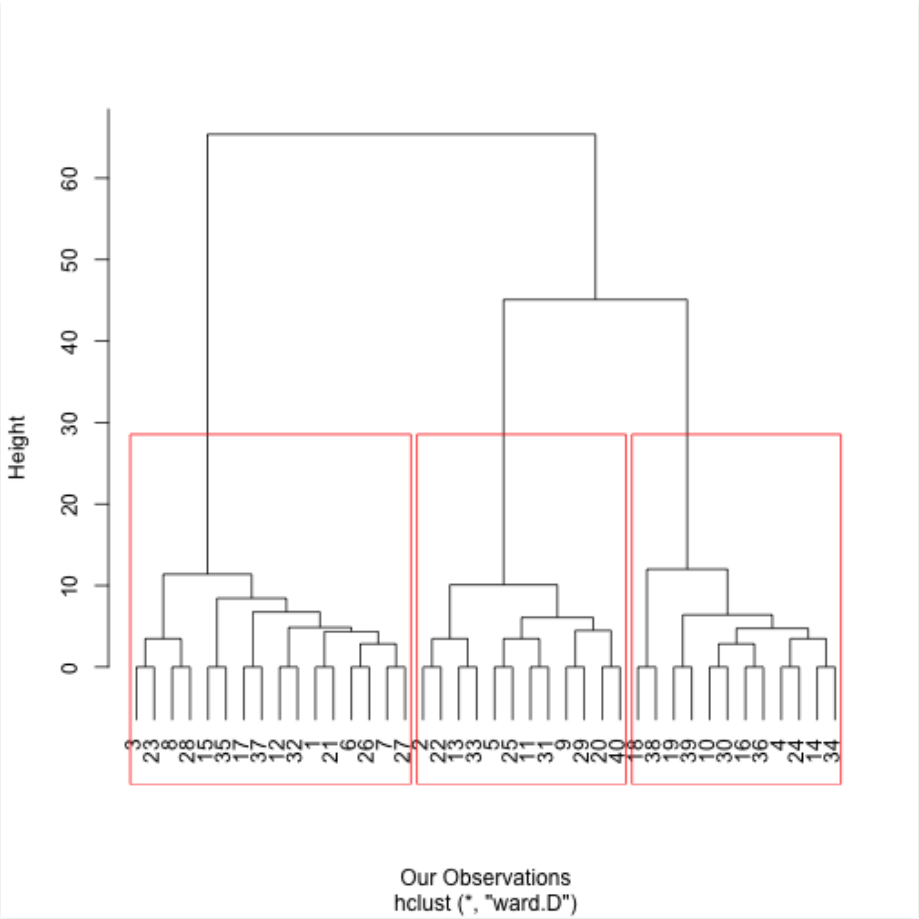
Step 6. Method and Number of Segments

There are many clustering methods. Two common ones are:

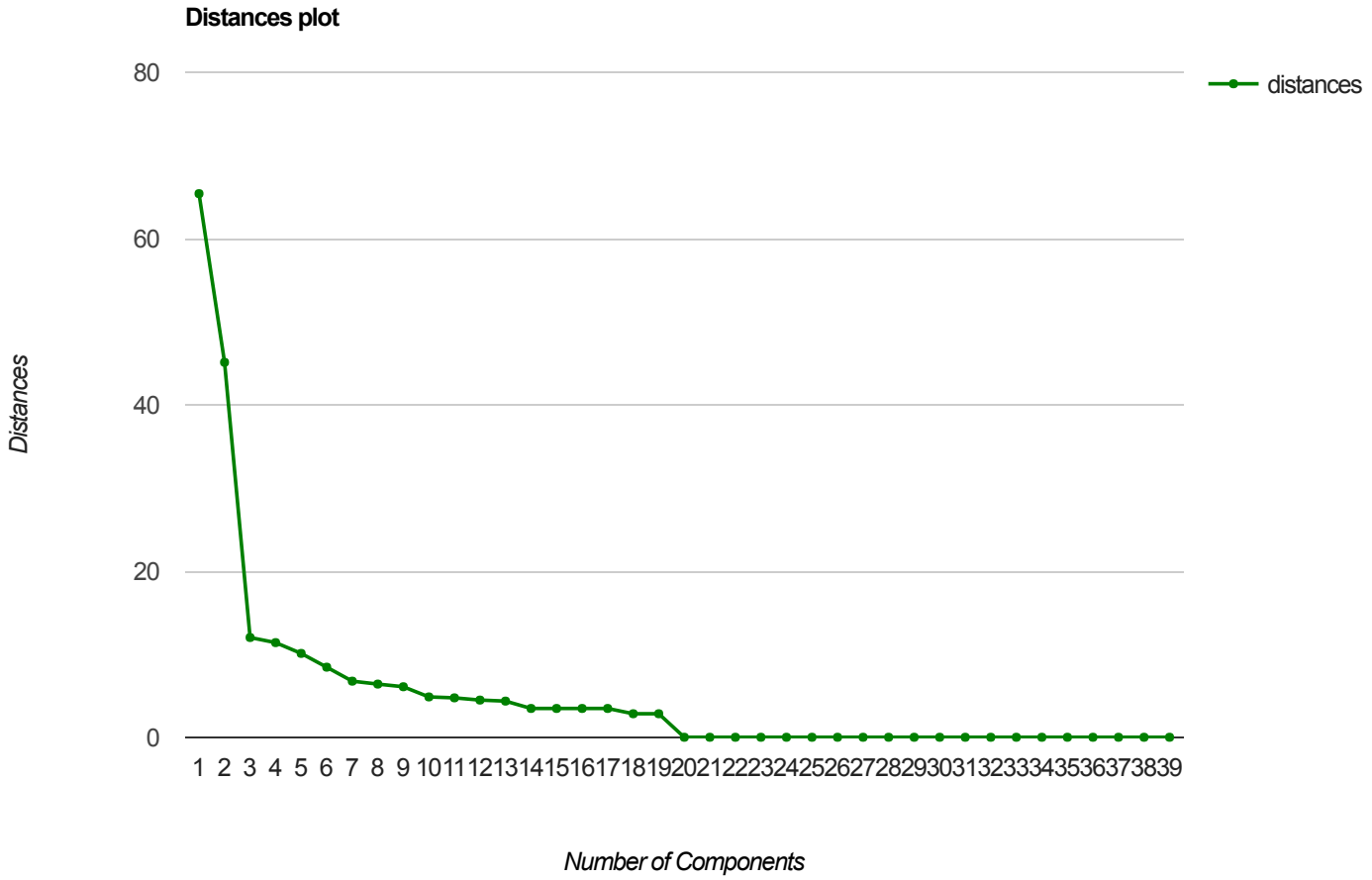
- Hierarchical Methods
- Non-Hierarchical Methods (e.g. k-means)

We can plug-and-play (with CARE) this “black box” in our analysis

Hierarchical Clustering: Dendrogram



Hierarchical Clustering Dendrogram Heights Plot



Cluster Membership: Hierarchical Clustering

	Observation	Observation.Number	Cluster_Membership
1	1	1	1
2	2	2	2
3	3	3	1
4	4	4	3
5	5	5	2
6	6	6	1
7	7	7	1
8	8	8	1

Cluster Membership: K-means Clustering

Note: K-means does not necessarily lead to the same solution every time you run it

	Observation	Observation.Number	Cluster_Membership
1	1	1	3
2	2	2	2
3	3	3	3
4	4	4	1
5	5	5	2
6	6	6	3
7	7	7	3
8	8	8	3

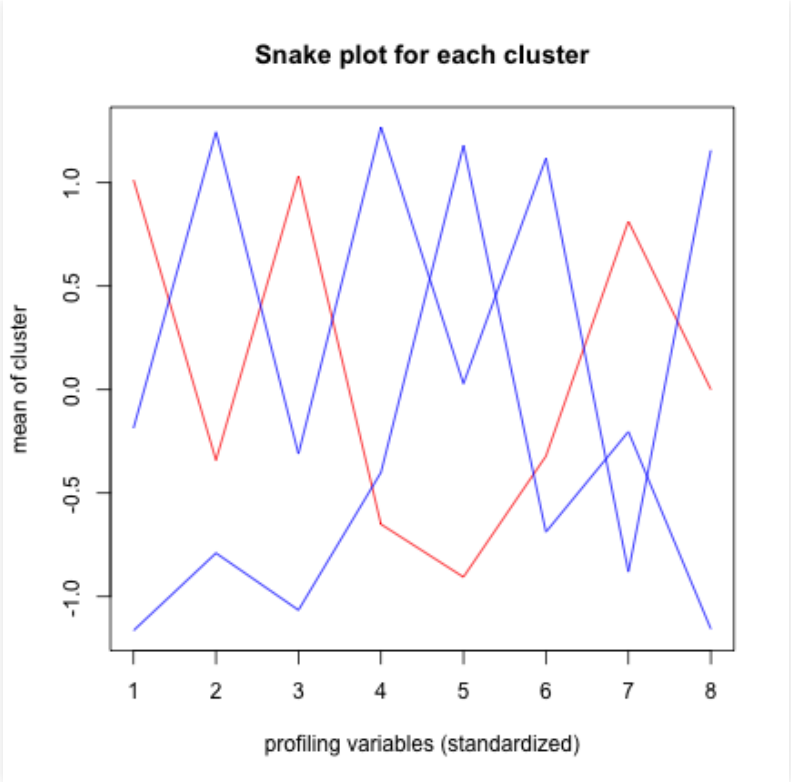
Step 7. Profile and interpret the segments

Data analytics is used to eventually make decisions, and that is feasible only when we are comfortable (enough) with our understanding of the analytics results, including our ability to clearly interpret them.

Cluster Profiling: Cluster Centers of Profling Variables

	Variables	Population	Segment.1	Segment.2	Segment.3
1	V1	3.85	5.75	1.67	3.5
2	V2	4.1	3.62	3	5.83
3	V3	3.95	6	1.83	3.33
4	V4	4.1	3.12	3.5	6
5	V5	3.45	1.88	5.5	3.5
6	V6	4.35	3.88	3.33	6
7	Income	46,000	60,000	42,500	30,833.33
8	Mall.Visits	3.25	3.25	1	5.5

Interpretation: Snake Plots



Interpretation: Ratio to Average of Total Population -1 (0 = Average)

Segment.1	Segment.2	Segment.3
0.49	-0.57	-0.09
-0.12	-0.27	0.42
0.52	-0.54	-0.16
-0.24	-0.15	0.46

The data

V1: Shopping is fun (scale 1-7)

V2: Shopping is bad for your budget (scale 1-7)

V3: I combine shopping with eating out (scale 1-7)

V4: I try to get the best buys while shopping (scale 1-7)

V5: I don't care about shopping (scale 1-7)

V6: You can save lot of money by comparingprices (scale 1-7)

Income: the household income of the respondent (in dollars)

Mall.Visits: how often they visit the mall (scale 1-7)

Step 8. Robustness Analysis

Are the clusters stable when we use:

- using different subsets of the original data
- using variations of the original segmentation attributes
- using different distance metrics
- using different segmentation methods
- using different numbers of clusters

Example Robustness Test: Different Methods

The percentage of observations belonging to the same segment is

100 %.

Segment 1	Segment 2	Segment 3
100.0	100.0	100.0

How much overlap should we expect across clustering solutions?

Key Technical Terms and Lessons

- Segmentation Variables
- Profiling Variables
- Distance Metrics
- Hierarchical Clustering
- Dendrogram
- K-means Clustering
- Robustness: Statistics, Interpretation, Decisions
- Actionability, Interpretability, Statistical Robustness

Group Work

1. How many market segments (clusters) are there? Why?

2. How would you describe the segments?

3. How would the segments inform the strategy of CreeqBoat?