Clustering of Mixed and Continuous Data

Ray Chen and Jung-a Kim

Methods used

- K-prototypes
- Agglomerative Clustering
- Partitioning around medoids
- DBSCAN
- Mixture of skew-t distributions
- Gaussian mixture model
- Mixture of t distributions
- Mixture of skew-Gaussian distribution
- K-means

K- Prototypes

Cost function is:

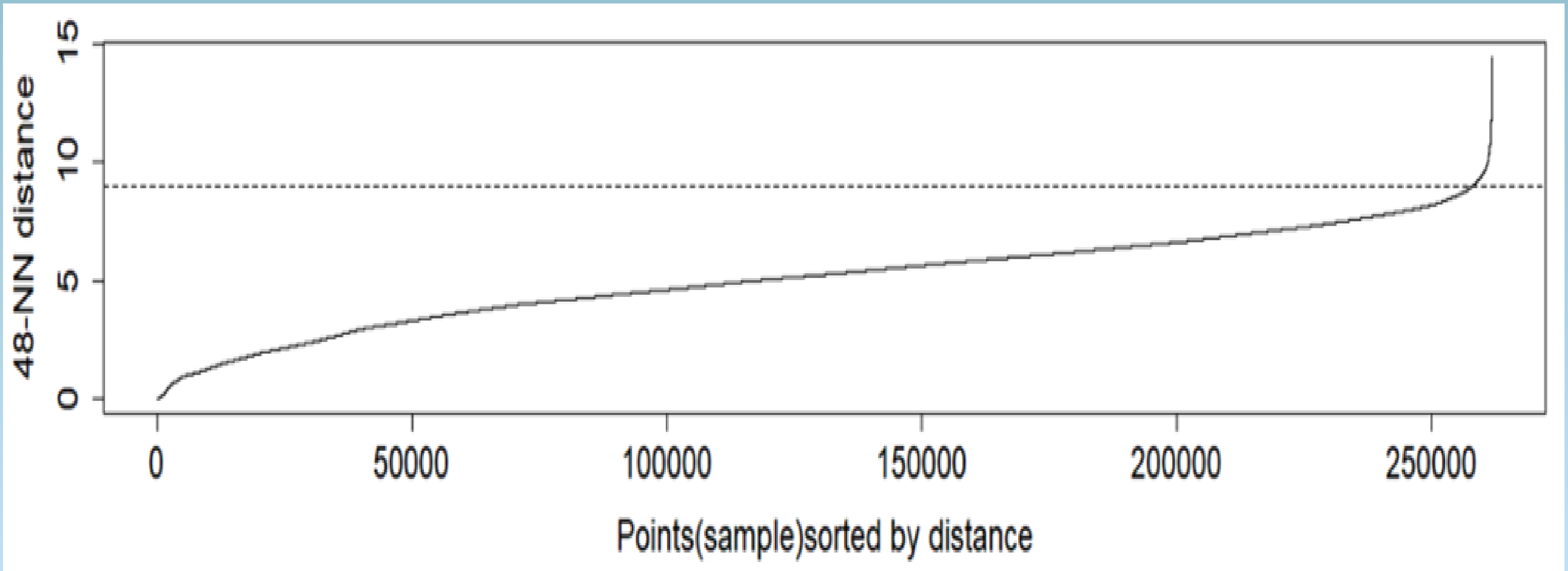
$$E_l = \sum_{i=1}^{n} y_{il} \sum_{j=1}^{m_r} \lambda_j \left(x_{ij}^r - q_{lj}^r\right)^2 + \sum_{j=m_r+1}^{m} \lambda_j n_l \left(1 - p\left(q_{lj}^c \in C_j \middle| l\right)\right) = E_l^r + E_l^c$$

- m_r is the number of numeric variables, m is the total number of variables, so $m - m_r$ is number of categorical variables
- r represents the numerical data and c represents the categorical data
- n is the number of observations
- y_{il} is the membership of X_i to the cluster l: $y_{il} = 1$, if X_i belongs to the cluster l and $y_{il} = 0$ otherwise
- n_l is the number of members in cluster l
- λ_j is the weight parameter which determines the degree of the variable's influence on the cost
  - We will use the lambda vector as it takes the inverse variability of each variable

DBSCAN

- Detects dense and sparse regions of data
- Two parameters required are:
  - ε, is the degree of density
  - Minimum sample size
- Core points are observations with number of neighbors within ε greater than the minimum sample size
- Border points are observations within ε of a core point
- Outliers are neither core nor border points

Selecting the DBSCAN epsilon



- Find epsilon at the point where the points sharply increase

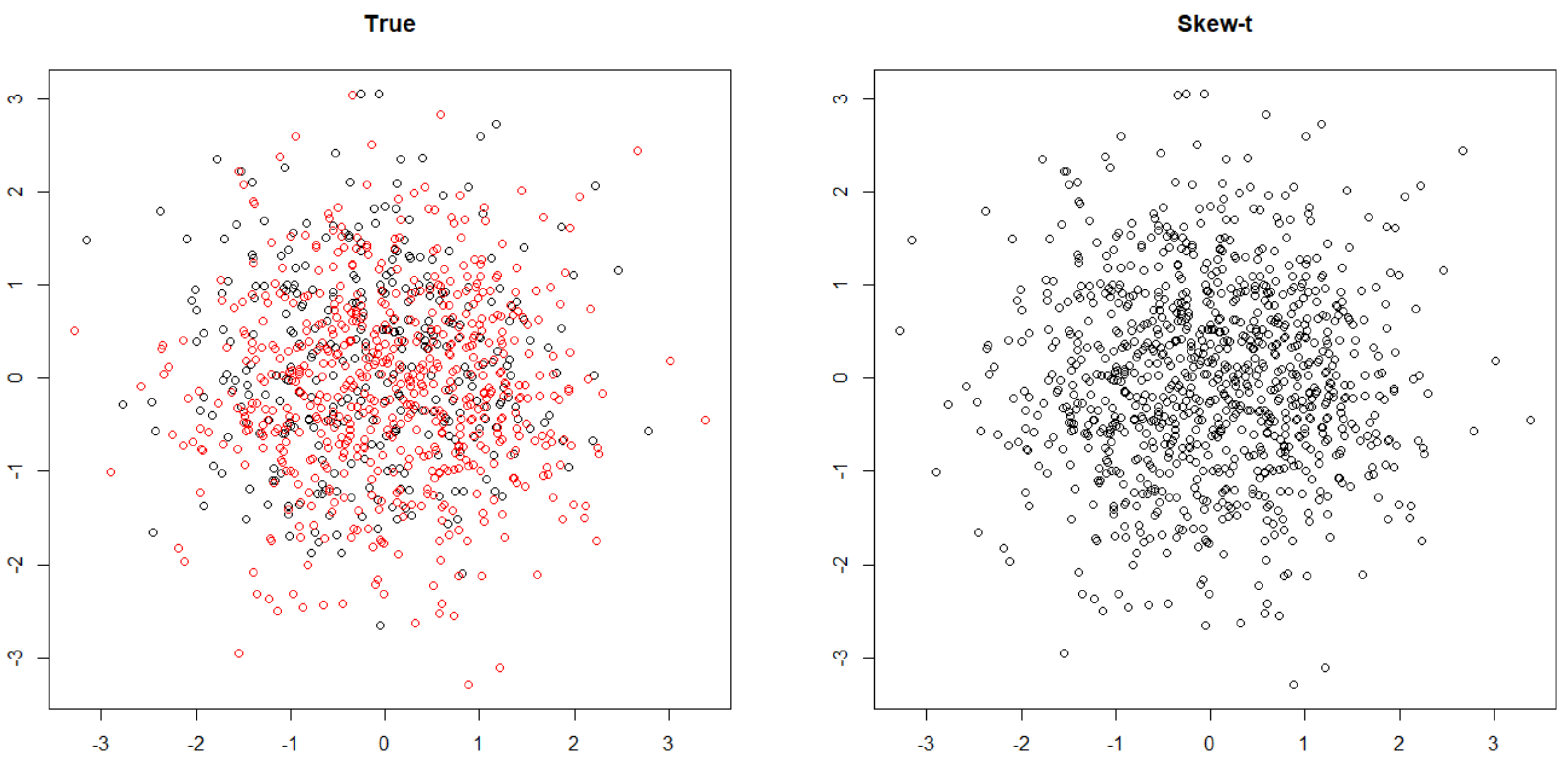
Simulation study

K-means, PAM, DBSCAN, and Skew-t mixture model

Considered

- Overlappedness
- 2 vs. 4 clusters
- No correlation vs. moderate correlation
- Types of distributions(Gaussian, skew-Gaussian, t, skew-t)

	Separation	Distribution	Correlation	k-means	PAM	DBSCAN	Skew-t
1	Clear	(G, G)	(X, X)	1.00	1.00	0.98	1.00
2	Clear	(G, G, G, t)	(X, X, X, X)	0.98	0.99	0.95	0.99
3	Clear	(G, G)	(X, mod-pos)	1.00	1.00	0.98	1.00
4	Clear	(G, G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.99	0.99	0.96	0.99
5	Clear	(G, skew-G)	(X, mod-pos)	0.63	0.64	0.93	1.00
6	Clear	(G, skew-G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.85	0.84	0.95	0.99
7	Over	(G, G)	(X, X)	0.01	0.01	0.00	0.00
8	Over	(G, G, G, skew-t)	(X, X, X, X)	0.02	0.12	0.01	0.33
9	Over	(G, G)	(X, mod-pos)	0.01	0.01	0.00	0.61
10	Over	(G, G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.02	0.12	0.00	0.86
11	Over	(G, skew-G)	(X, mod-pos)	0.08	0.17	-0.02	0.79
12	Over	(G, skew-G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.08	0.18	0.00	0.91


True
Skew-t
-3
-2
-1
0
1
2
3

	Separation	Distribution	Correlation	k-means	PAM	DBSCAN	Skew-t
1	Clear	(G, G)	(X, X)	1.00	1.00	0.98	1.00
2	Clear	(G, G, G, t)	(X, X, X, X)	0.98	0.99	0.95	0.99
3	Clear	(G, G)	(X, mod-pos)	1.00	1.00	0.98	1.00
4	Clear	(G, G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.99	0.99	0.96	0.99
5	Clear	(G, **skew**-G)	(X, mod-pos)	0.63	0.64	0.93	1.00
6	Clear	(G, skew-G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.85	0.84	0.95	0.99
7	Over	(G, G)	(X, X)	0.01	**0.01**	0.00	0.00
8	Over	(G, G, G, skew-t)	(X, X, X, X)	0.02	**0.12**	0.01	0.33
9	Over	(G, G)	(X, mod-pos)	0.01	**0.01**	0.00	0.61
10	Over	(G, G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.02	**0.12**	0.00	0.86
11	Over	(G, skew-G)	(X, mod-pos)	0.08	**0.17**	-0.02	0.79
12	Over	(G, skew-G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.08	**0.18**	0.00	0.91

	Separation	Distribution	Correlation	k-means	PAM	DBSCAN	Skew-t
1	Clear	(G, G)	(X, X)	1.00	1.00	0.98	1.00
2	Clear	(G, G, G, t)	(X, X, X, X)	0.98	0.99	0.95	0.99
3	Clear	(G, G)	(X, mod-pos)	1.00	1.00	0.98	1.00
4	Clear	(G, G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.99	0.99	0.96	0.99
5	Clear	(G, skew-G)	(X, mod-pos)	0.63	0.64	0.93	1.00
6	Clear	(G, skew-G, G, t)	(X, mod-pos, mod-pos, mod-pos)	0.85	0.84	0.95	0.99
7	Over	(G, G)	(X, X)	0.01	0.01	0.00	0.00
8	Over	(G, G, G, skew-t)	(X, X, X, X)	0.02	0.12	0.01	0.33
9	Over	(G, G)	(X, mod-pos)	0.01	0.01	0.00	0.61
10	Over	(G, G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.02	0.12	0.00	0.86
11	Over	(G, skew-G)	(X, mod-pos)	0.08	0.17	-0.02	0.79
12	Over	(G, skew-G, G, skew-t)	(X, mod-pos, mod-pos, mod-pos)	0.08	0.18	0.00	0.91

Tetragonula Bee Species

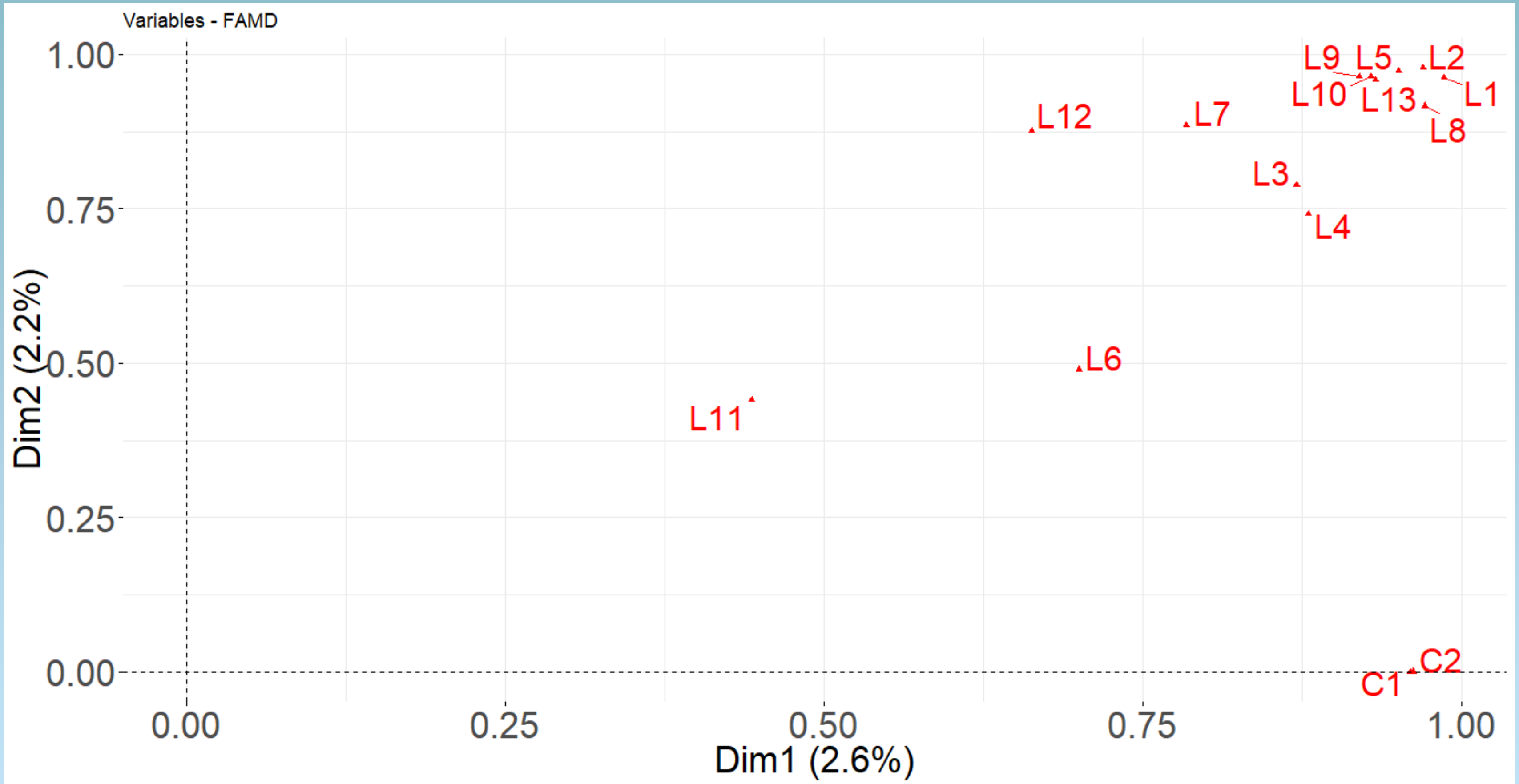
- Genetic data for 236 Tetragonula bees from Australia and Southeast Asia

- 13 categorical variables: L1 – L13 are strings of six digits which enc ode a pair of alleles with no numeric information.

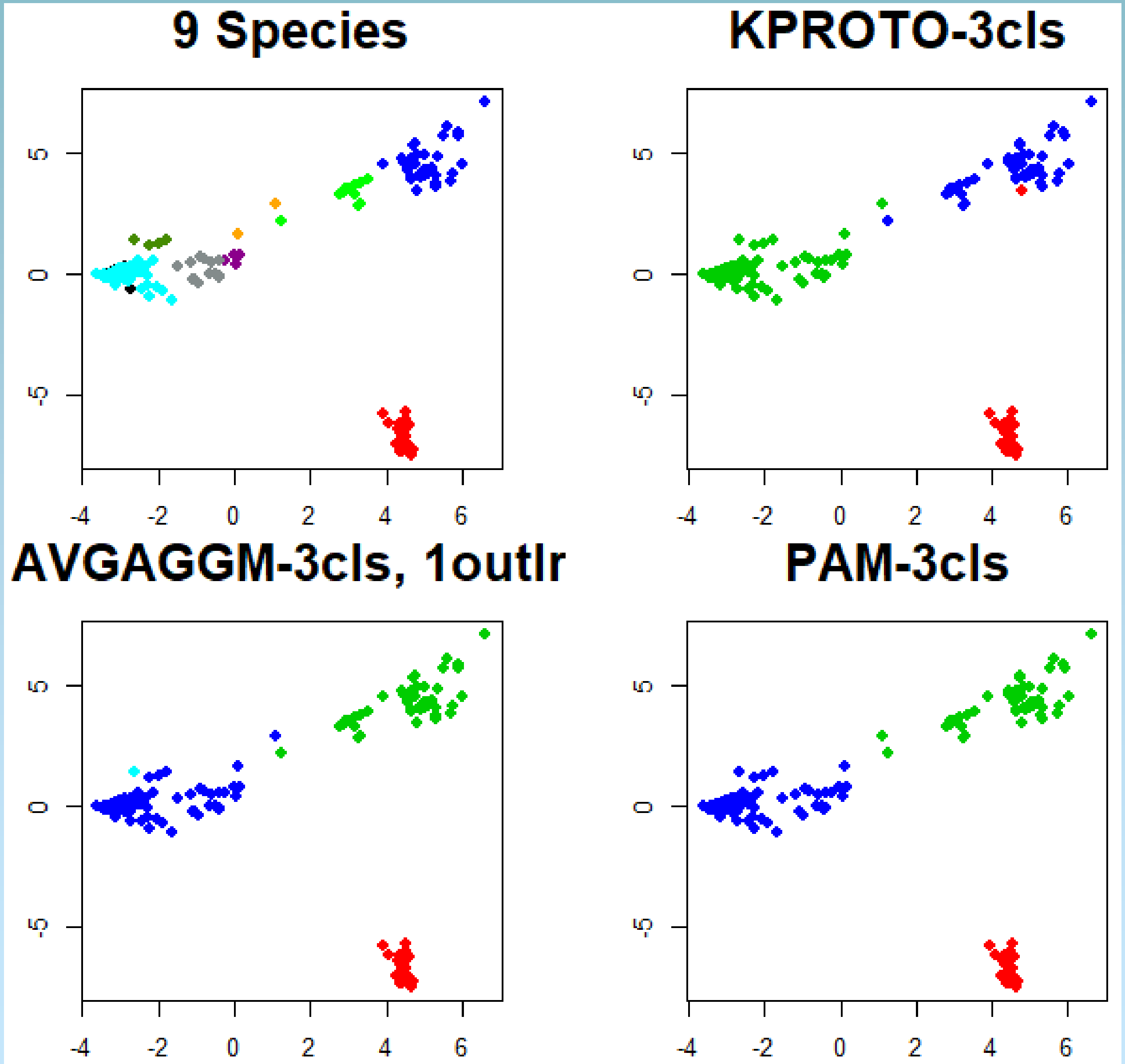
- 2 numerical variables: C1 and C2 are coordinates of locations of ind ividual bees. C1 is is latitude (negative values are South). C2 is longit ude (negative values are West).

- Species represent the species out of 9 categories labeled from 1 to 9.

Bee Species: 2 numeric & 13 categorical



Bee Species: 2 numeric & 13 categorical



ARI: 43%
AvgSil: 0.35

ARI: 48%
AvgSil: 0.32

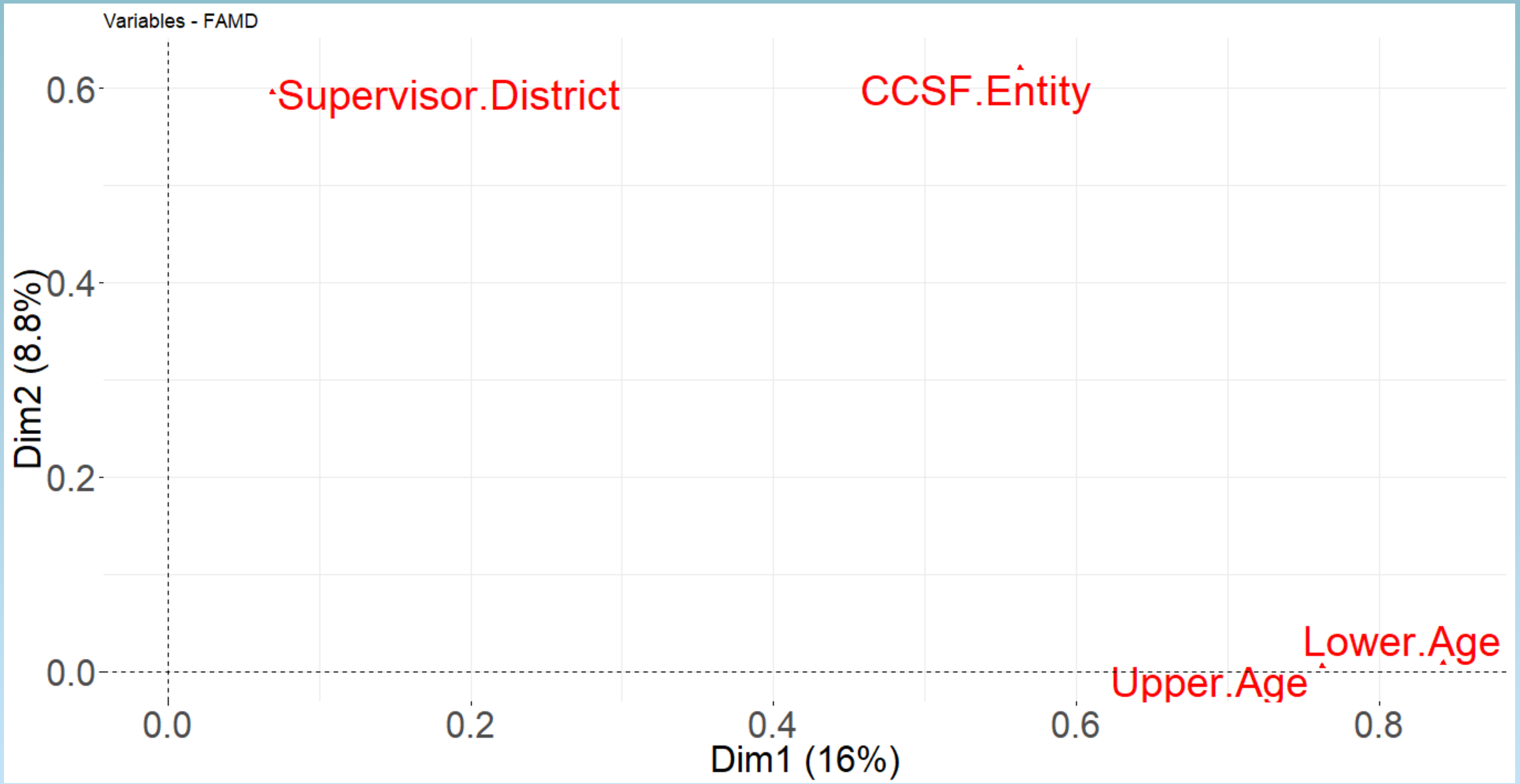
ARI: 44%
AvgSil: 0.35

Schools

- Features of 445 public and private schools for infant, Pre-K, and K-14 students in San Francisco, California
- 4 variables were selected out of 16 variables.
- 2 categorical variables:
  - o CCSF Entity: City College of San Francisco entities
    - ◎ Private
    - ◎ SFCCD = San Francisco Community College District
    - ◎ SFUSD = San Francisco Unified School District
  - o Supervisor District: City and County Supervisor District number
    - ◎ 1-9 (9 levels)
- 2 numerical variables:
  - o Lower Age: Lower bound of generic age of the education program
  - o Upper Age: Upper bound of generic age of the education program

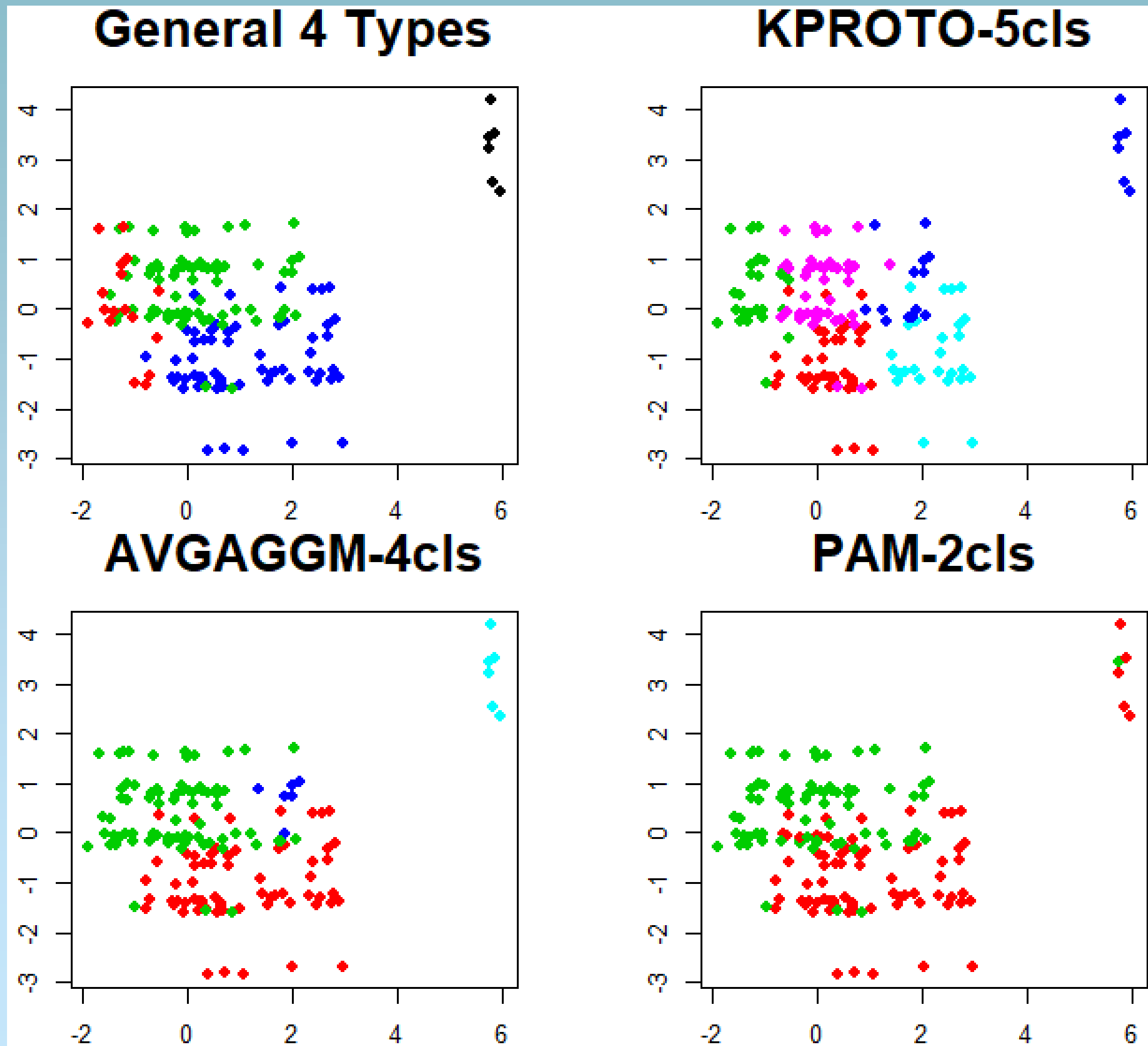
- General Type: Broad category of schools
  - ◎ CC = Community College
  - ◎ CDC = Child Development Center
  - ◎ IND = Independent / Private
  - ◎ PS = Public School

School: CCSF Entity, Supervisor District, Lower Age, Upper Age



School: CCSF Entity, Supervisor District, Lower Age, Upper Age

Black =CC
Red=CDC
Green=IND
Blue=PS



ARI: 71%
AvgSil: 0.43

ARI: 53%
AvgSil: 0.37

ARI: 48%
AvgSil: 0.41

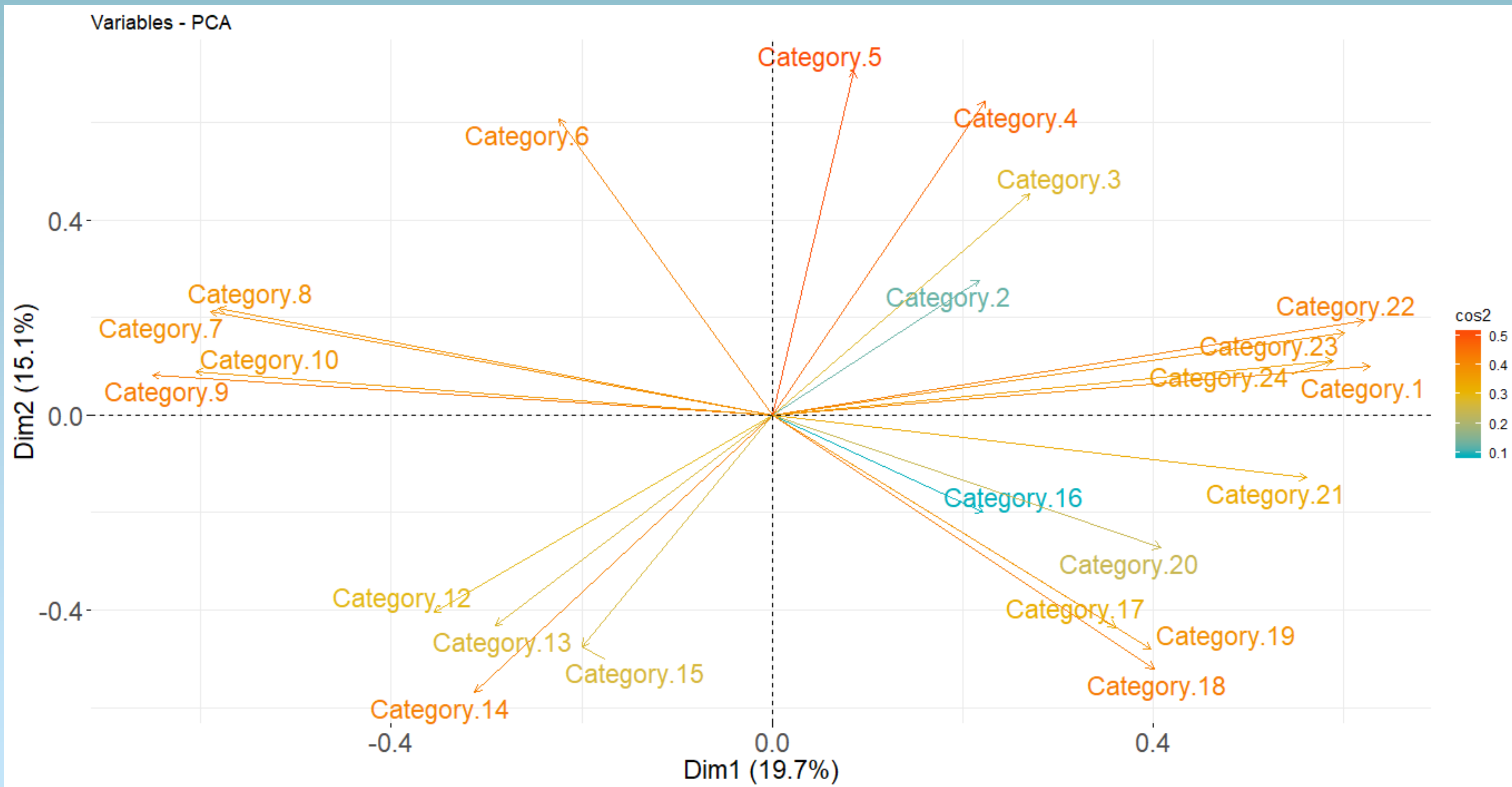
Travel Review Ratings

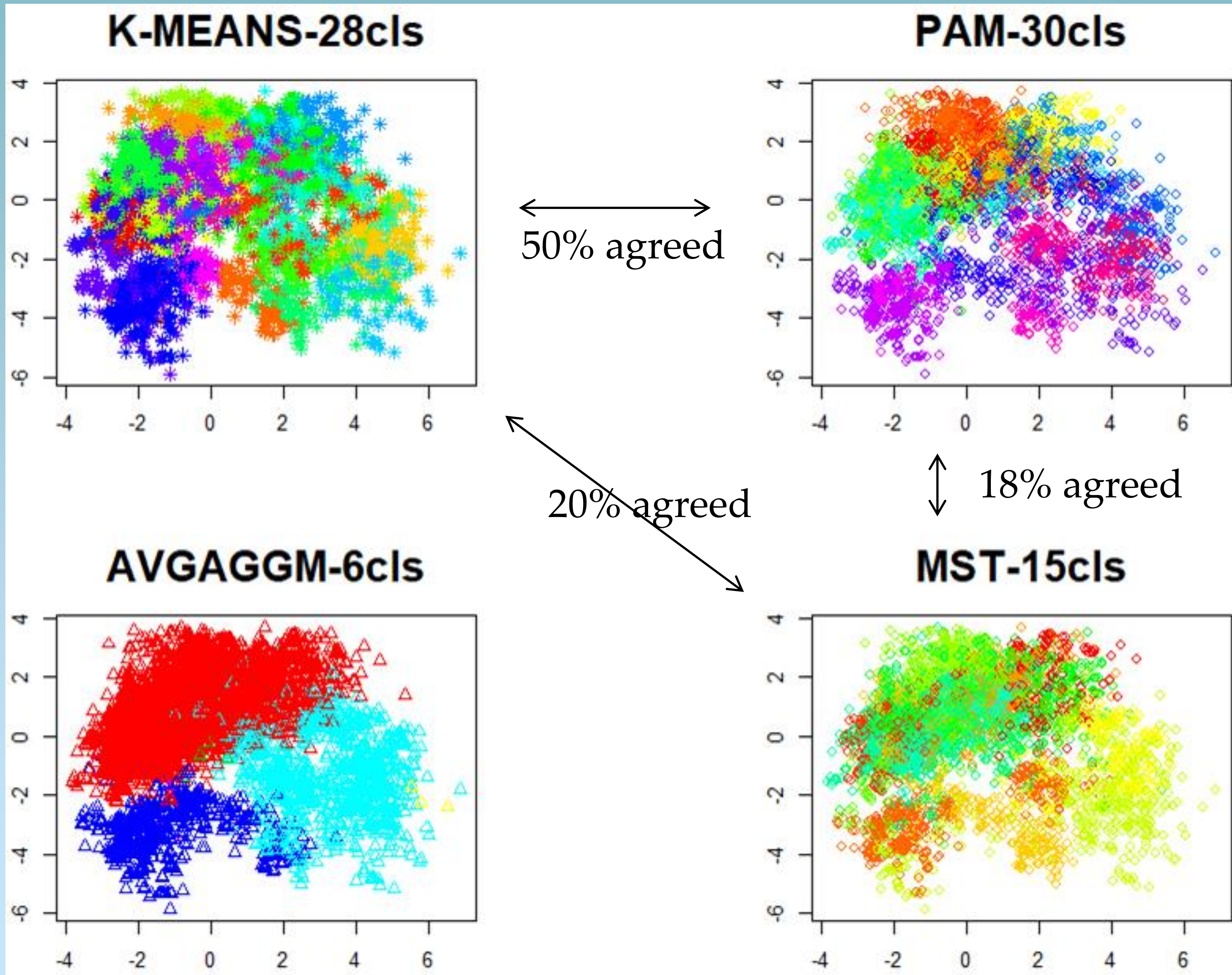
- 5454 Google ratings on attractions from 23 categories across Europe.

The rating ranges from 1 to 5.

- 23 numerical variables: Average user rating per category

Travel: 23 numeric variables





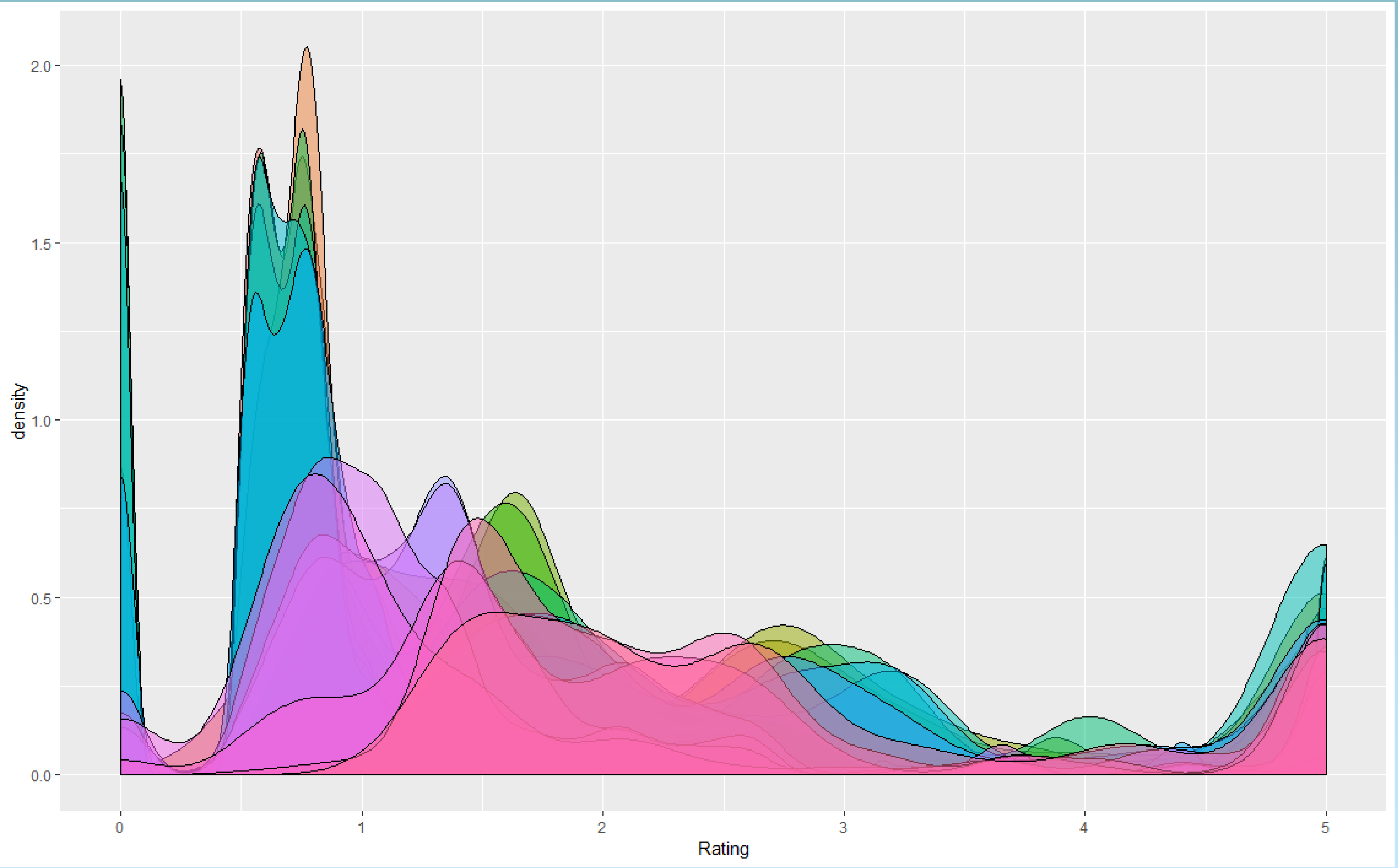
AvgSil: 0.18

AvgSil: 0.15

AvgSil: 0.12

AvgSil: 0.02

23 Distributions



Conclusion

- Mixture model is robust in overlapping clusters with different shapes
- DBSCAN detects non-spherical shapes with skewness
  - Sensitive to parameter settings
- K-prototypes more robust than PAM and Agglomerative clustering using lambda
  - Assumes spherical shapes and sensitive to initial prototypes
- Future: Mixture model for mixed data, proper parameter setting for DBS CAN, proper weights for K-prototypes

Thank you! Questions?