WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

The OPTICS Cordillera

Nonparametric Assessment of Clusteredness

Outline

1. Problem Motivation
2. Clusteredness
   - Clusteredness Concept
   - Clusteredness Aspects
3. OPTICS Cordillera
   - Goal
   - Definition
   - Properties
   - Package
4. Example
5. Conclusion and Outlook

This is joint work with Kurt Hornik (WU) and Patrick Mair (Harvard).

Problem Motivation

"I'm a Republican, because ..." from Mair et al. (2014)

- Supporters of the Republican Party have been asked why they are Republican (254 statements)
- Natural language data that was scraped and processed $\Longrightarrow$ Sparse data matrix (document term matrix)
- Objects are the words (we use only 37 words that appeared at least 10 times)
- We look for themes in the statements: "Mantras" (words that occur often together)

Problem Description

- A common situation in exploratory research with unstructured data sets:
  - Purely exploratory and descriptive motivation
  - Little idea nor theory about what we might find
  - Let the data speak for themselves
- In this situation one often uses tools that are suggestive to the eyes of the beholder, e.g.,
  - Reduce dimensionality of the data set (here 254×37) and plot the result
  - We use a cosine distance for word co-occurrences and apply standard least squares MDS for representation.

Multidimensional Scaling (MDS)

- Popular method for representing multivariate high-dimensional proximities in some lower-dimensional space
- Provides an optimal map into continuous space $\mathbb{R}^M$ and looks for directions of spread in the low-dimensional space
- MDS utilizes a loss function, e.g., a least squares one

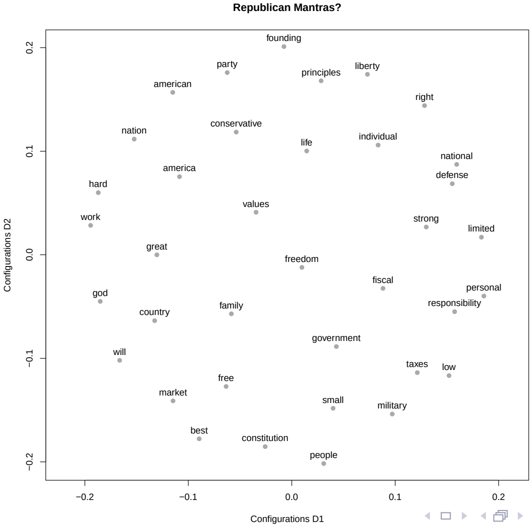
$$\sigma_{MDS}(X) = \sum_{i<j} w_{ij} \left[\hat{d}_{ij} - d_{ij}(X)\right]^2$$

which is minimized to find the configuration X

$$\arg\min_X \sigma_{MDS}(X)$$

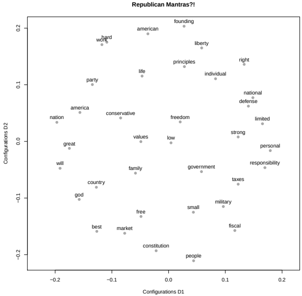
$\hat{d}_{ij} = f(\delta_{ij})$ ... disparaties, δ_{ij} ... proximities
$d_{ij}(X)$ ... fitted distances
$f(\cdot)$ ... transformation function
w_{ij} ... finite weights

The Solution is a Problem



The Solution is a Problem

- Optimal configuration lacks clusteredness (does not appear clustered).
- What if we fit another model, say, an MDS with power transformation by setting $\hat{d}_{ij} = \delta_{ij}^2$?



- Is this result more clustered? Equally clustered?

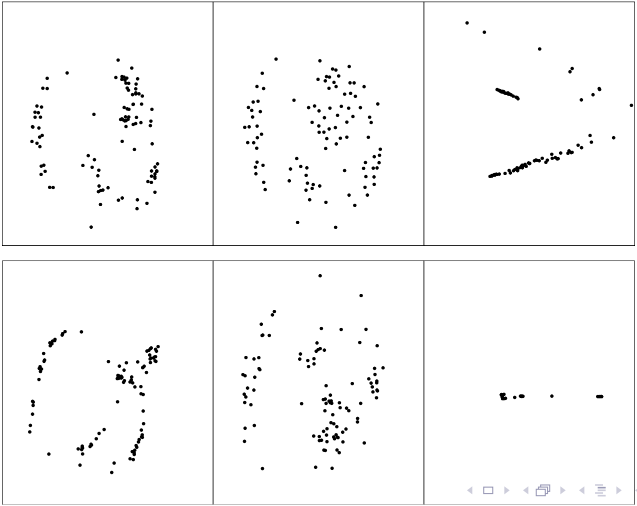
Clusteredness

- We need a way to assess how clustered a data representation result appears.
- How about just looking at things? That’s what most do.

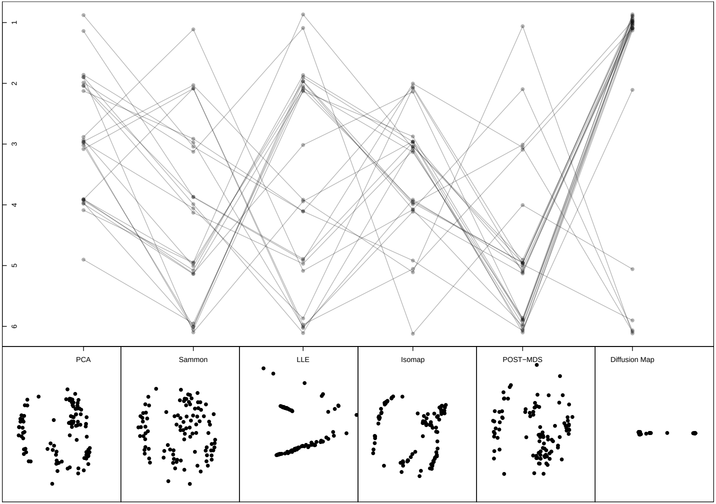
Photo by Pete Souza

Clustered Plots

How well does that work in general?



Plot Rankings by Clusteredness



Clusteredness

- Clusteredness is difficult to infer
- Simply eyeballing the result is usually not enough
  - Different subjective definitions of clusteredness
  - Different perception of observers
  - Difficulty of comparing differences
  - Often not reproducible
- We aim for a principled way to assess the obtained degree of clusteredness in the representation that quantifies how clustered the result is with some objectivity.

Concept of Clusteredness

Clusteredness: The appearance of how clustered a representation is (supervague definition).

It ...

- ... is a property of the representation
- ... says how well clusters can be perceived/formed
- ... is open to as many forms of appearances as possible
- ... shares conceptual similarity with hierarchical clustering

Aspects of Clusteredness

To make this more concrete we conducted a mixed-method study and did some thinking ourselves. We could derive aspects of clusteredness:

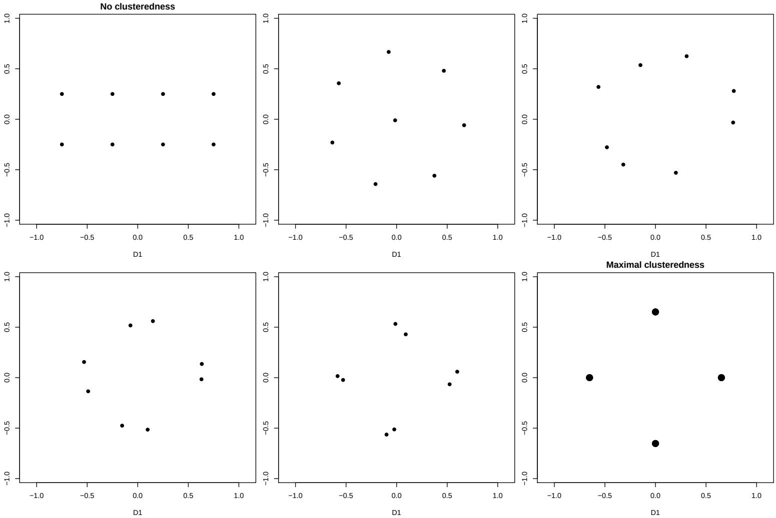
- Clear definition of extremes (no clusteredness and maximal clusteredness)
- Clusteredness is a continuum between the extremes
- Arbitrary cluster shapes should be detectable
- Nonparametric assessment (as little assumptions as possible)
- Sensible behaviour, e.g., higher clusteredness if
  - Higher compactness of clusters
  - Stronger separation of clusters
  - More clusters are visible

Clusteredness: Extremes

- No Clusteredness: The distance of each point to its neighbours is constant (a matchstick graph embedding of points is possible e.g., points lie on a regular grid or lattice).
- Maximal Clusteredness: Points are evenly distributed over the clusters and all points in a cluster coincide exactly at the cluster centers and all clusters are equally far away from their neighbouring clusters (matchstick embedding of the cluster centers is possible).

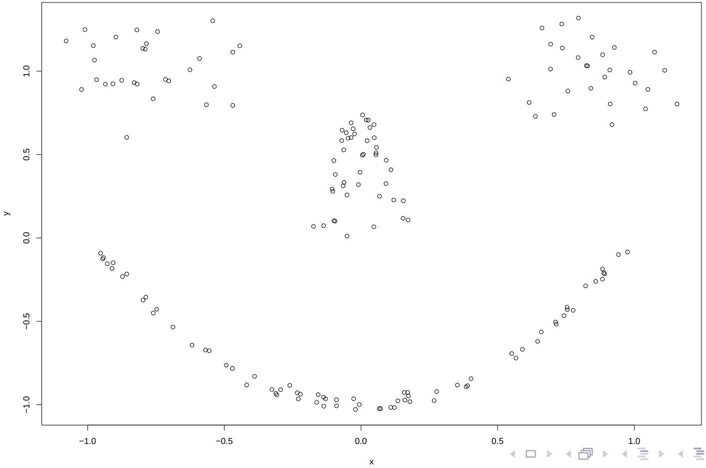
Note that the extremes are in some way regular. Also note we need a definition of neighbourhood so we must at least assume a number of observations that must comprise a cluster.

Clusteredness: Continuum

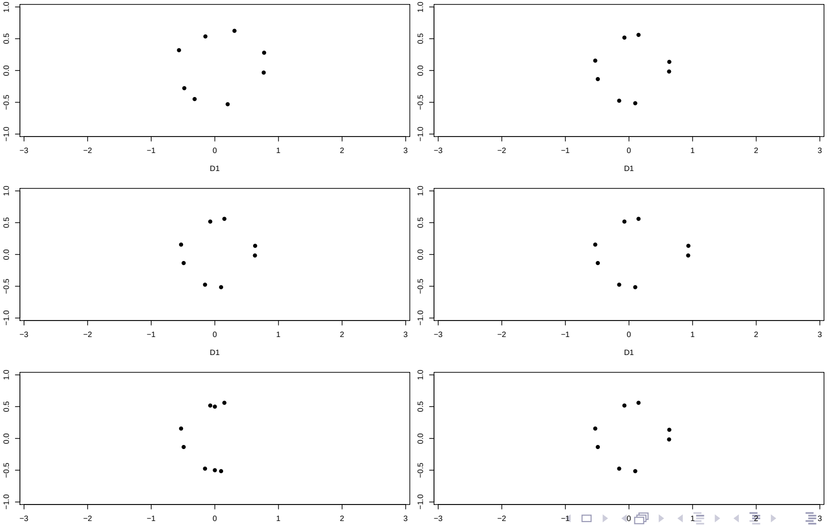


Clusteredness: Arbitrary Shapes

We want to detect arbitrary shapes.



Clusteredness: Sensible Behaviour



Requirement in Nutshell

To measure this concept of clusteredness we need a statistic that

- Operates only on the representation
- Is minimal/maximal in case of no clusteredness/maximal clusteredness
- Makes as little assumptions as possible
- Adheres to the clusteredness aspects from before

How to Measure Clusteredness?

The literature did not help much to find a statistic that meets these requirements:

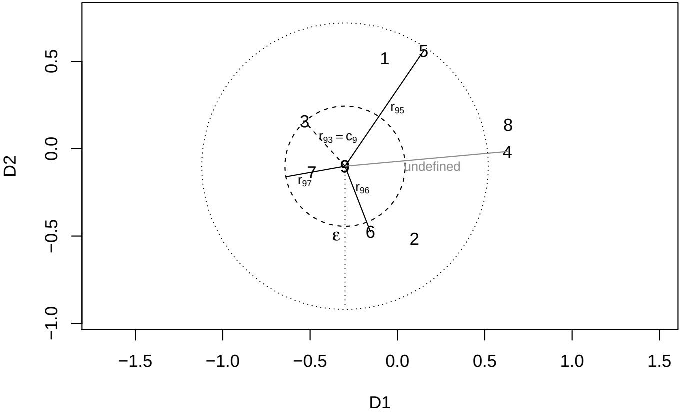
- Partitional clustering indices (Silhouette, Calinski-Habasz, ...)
  - Need concrete partition (number of clusters assumed, algorithm dependency)
  - Do not depend on the representation alone (no invariance to cluster assignment)
- Hierachical clustering indices (ultrametric based VAF, DAF)
  - Much closer conceptually
  - Do not match all clusteredness aspects

OPTICS Cordillera - I

- Cluster concept is density-based (ϵ as maximum radius)
- Only the minimum k number of observations that must comprise a cluster is specified
- Utilizes only minimum reachabilities $r^*_{(i)}$ of all points $x_{(i)}$ (essentially pairwise distances) and an ordering R of these points, $R = \{x_{(i)}\}_{i=1,\ldots,N}$.
- Ordering is obtained by OPTICS (Ankerst et al., 1999) with metaparameters k, ϵ. k is mandatory, ϵ is optional (needs only be “sufficiently large”).
- R and $r^*_{(i)}$ encode the clustering structure. We aggregate it to an index $OC(X)$ by defining (for metaparameter $q > 0$)

$$\mathrm{OC}(X) = \left(\sum_{i=2}^{N} \left| r^*_{(i)} - r^*_{(i-1)} \right|^q \right)^{1/q}$$

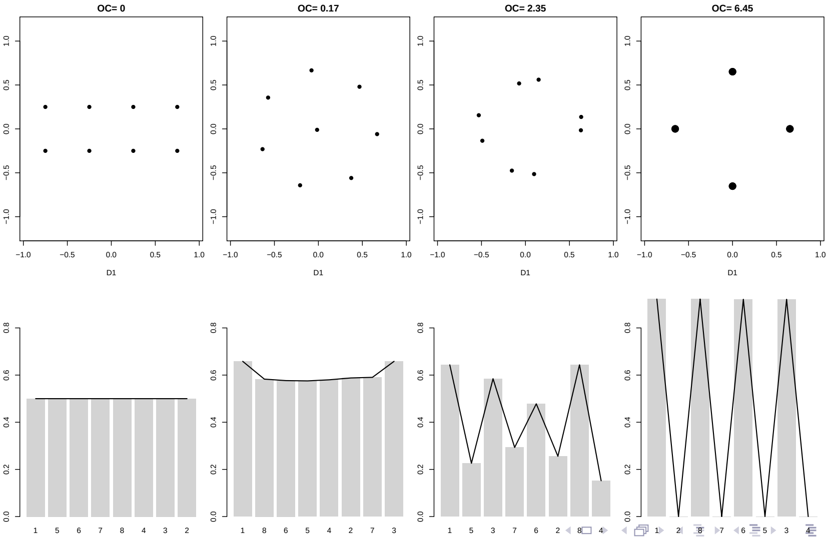
Reachabilities



OPTICS Algorithm

http://www.dbs.informatik.uni-muenchen.de/Forschung/KDD/Clustering/OPTICS/Demo/

OPTICS Cordillera - II



Bounds and Normalization

For given metaparameters ϵ, k, q the following holds

- Lower bound for $OC(X)$ is 0. This is the case of no clusteredness.
- Upper bound for $OC(X)$ in the maximal clusteredness case is

$$C^*(X; d_{max}, \epsilon, k, q) = d_{max}^q \cdot \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)$$

- d_{max} is an (optional) distance beyond which winsorizing happens.
- We can use this to normalize $OC(X)$ to $[0, 1]$

$$OC'(X) = \frac{OC(X)}{C^*(X; d_{max}, \epsilon, k, q)}$$

- Note this all depends on k so we can accommodate different opinions on the behaviour for different number of clusters

Properties of the OPTICS Cordillera

- Representation only: Utilizes only special pairwise distances between points and the point ordering (no cluster assignment or *a priori* defined number needed)
- Arbitrary Shapes: Density-based definition is open to any cluster shape (incl. nested clusters)
- Nonparametric: Only needs k specified, the rest is optional
- Sensible behaviour: $OC(X)$ typically increases when for given k, ϵ, R.
  - Distances between clusters increase (Emphasis Property)
  - Points are more densely clustered (Density Property)
  - Number of clusters increases (Tally Property)
  - Does not pick up unbalancedness in the number of points in a cluster as a sign of clusteredness (Balance Property)

R Package cordillera

All of this is implemented in the R package cordillera

- `cordillera()` ... Function to calculate the OPTICS Cordillera.
- `e_optics()` ... An interface to OPTICS reference implementation in ELKI
- S3 methods: `plot, summary, print`

Example: Republican Representations

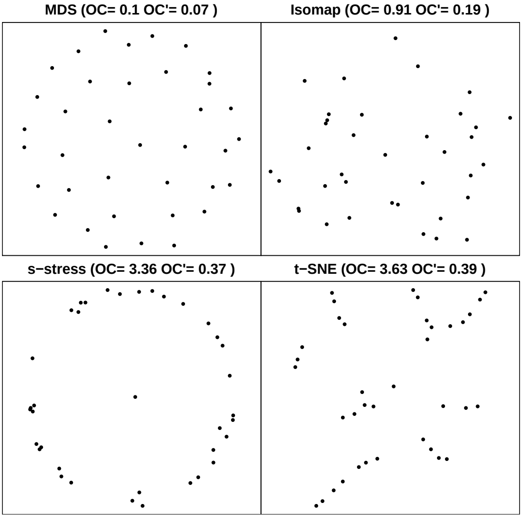
Some low dimensional representations for the "I'm a Republican, because..." data set:

- Standard MDS (Smacof): `powerStressMin(dt.dist)`
- S-Stress (Alscal): `powerStressMin(dt.dist,kappa=2,lambda=2)`
- t-SNE: `tsne(dt.dist,perplexity=3,max_iter=10000)`
- Isomap: `isomap(dt.dist,k=3)`

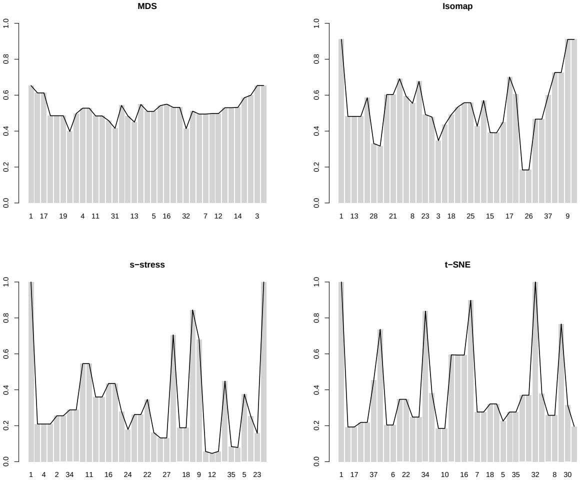
and look at the $OC(X)$ with $k = 3, \epsilon = 10, q = 2$ and $d_{max} = 1$

```
R> cordillera(X,epsilon=10,minpts=3,q=2,rang=c(0,1))
```

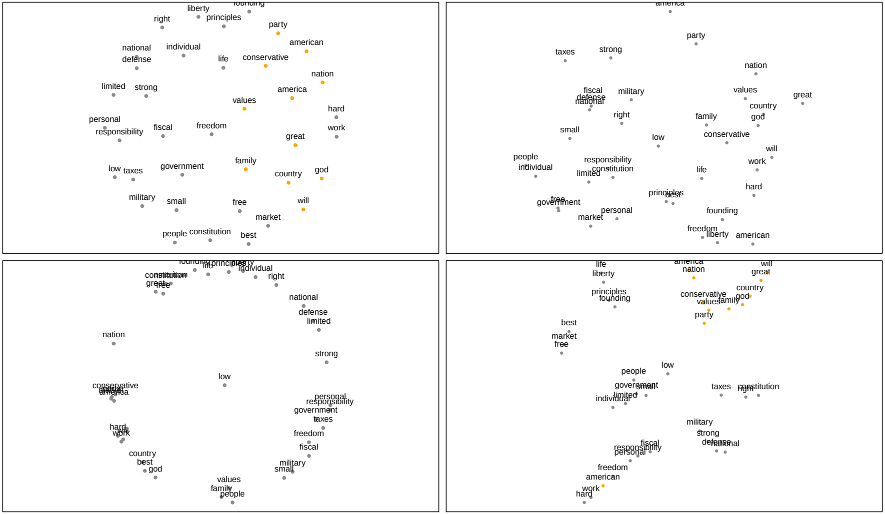
Republican Representations - I



Republican Representations - II



Republican Representations - III



Conclusion and Outlook

Summary

- We provided a concept of clusteredness and defined it
- We suggested the OPTICS Cordillera to assess clusteredness
- It is a measure of goodness-of-clusteredness that has appealing properties for pure exploratory data analysis

Outlook

- Next time we use the *OC* to guide us in dimension reduction
- This leads to Cluster Optimized Proximity Scaling (COPS; Rusch et al. 2015)

References

- Ankerst, M., Breunig, M., Kriegel, H.-P. & Sander, J. (1999) OPTICS: Ordering points to identify the clustering structure, ACM Sigmod Record 28, 49–60.
- Mair, P., Rusch, T. & Hornik, K. (2014) The grand old party - A party of values? SpringerPlus, 3:697.
- Rusch, T., Hornik, K., Mair, P. (2016) Assessing and quantifying clusteredness: The OPTICS Cordillera. Report 2016/1, Discussion Paper Series / Center for Empirical Research Methods, 2016/1. WU Vienna University of Economics and Business, Vienna.
- Rusch, T., Mair, P. & Hornik, K. (2015) COPS: Cluster optimized proximity scaling. Report 2015/1, Discussion Paper Series / Center for Empirical Research Methods, WU Vienna University of Economics and Business, Vienna.

Backup Slides

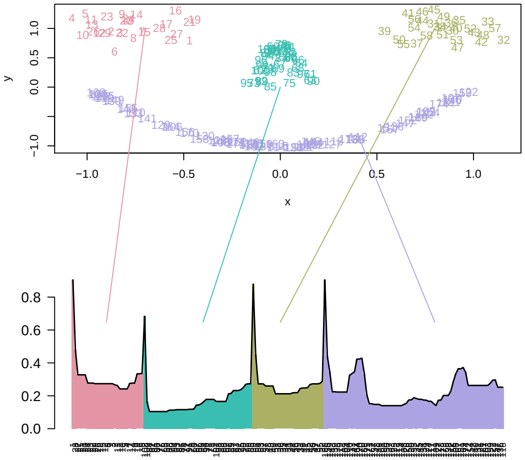
OPTICS algorithm-I

```
OPTICS(Data, epsilon, k)
    empty ordered list
    FOR i FROM 1 to N of Data
       x=x_i
    IF (processed(x) == FALSE)
       S = neighbors(x, epsilon)
       set x as processed
       x.reachability-distance = UNDEFINED
       x.core-distance = core-distance(S,epsilon,k)
       output x to ordered list
       IF (x.core-distance != UNDEFINED)
          OrderSeeds = empty priority queue
          update(OrderSeeds, S, x)
          WHILE (empty(OrderSeeds)==FALSE) DO
             y = next(OrderSeeds)
             S'= neighbors(y, epsilon)
             set y as processed
             y.core-distance = core-distance(S',epsilon,k)
             output y to the ordered list
             IF (core-distance(y, epsilon, k) != UNDEFINED)
                update(OrderSeeds, S',y)
END
}
```

OPTICS algorithm-II

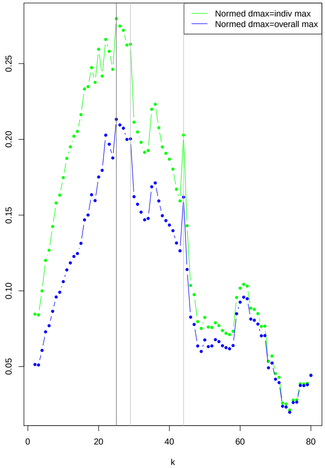
```
update(OrderSeeds, S, x)
    coredist = x.core-distance
    FOR EACH y IN S
       IF (processed(y) == FALSE)
          new-reach-dist = max(coredist, distance(x,y))
          IF (y.reachability-distance == UNDEFINED)
              y.reachability-distance = new-reach-dist   //y not in OrderSeeds
              insert(OrderSeeds, y, new-reach-dist)
          ELSE               // y is in OrderSeeds, check for improvement
              IF (new-reach-dist < y.reachability-distance)
                 y.reachability-distance = new-reach-dist
              moveup(OrderSeeds, y, new-reach-dist)
END
```

Arbitrary Shapes



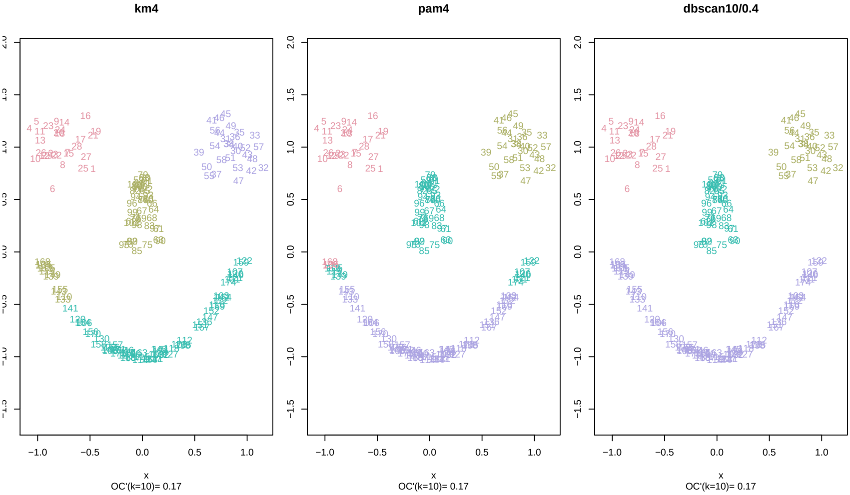
OC for k

Raw
Normed dmax=indiv max
Normed dmax=overall max
k
0
2
4
6
8
10
12
0
20
40
60
80

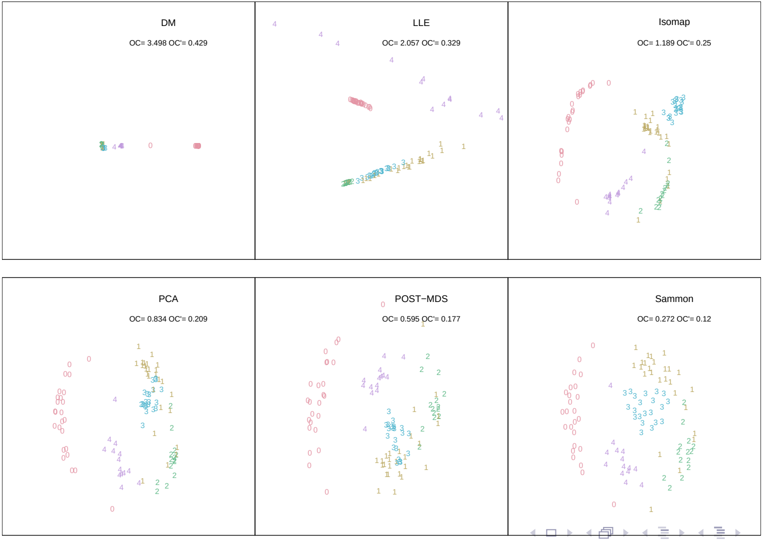


Normed dmax=indiv max
Normed dmax=overall max
k
0.05
0.10
0.15
0.20
0.25
0
20
40
60
80

Assignment Invariance



Diffplots


DM
OC= 3.498 OC'= 0.429
LLE
OC= 2.057 OC'= 0.329
Isomap
OC= 1.189 OC'= 0.25
PCA
OC= 0.834 OC'= 0.209
POST−MDS
OC= 0.595 OC'= 0.177
Sammon
OC= 0.272 OC'= 0.12

Thank You for Your Attention

Thomas Rusch

Competence Center for Empirical Research Methods

email: thomas.rusch@wu.ac.at

URL: http://wu.ac.at/methods/team/dr-thomas-rusch

WU Vienna University of Economics and Business

Welthandelsplatz 1, 1020 Vienna

Austria

License