Scaling for Clusters with COPS Cluster Optimized Proximity Scaling

Rusch, Thomas; Mair, Patrick; Hornik, Kurt

Published: 06/12/2014

Document Version
Unknown

WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

Scaling for Clusters with **COPS**

Cluster Optimized Proximity Scaling

Outline

This is joint work with Patrick Mair and Kurt Hornik.

Multidimensional Scaling (MDS)

- Popular method for representing multivariate high-dimensional proximities in some lower-dimensional space
- MDS utilizes a stress function, e.g., a least squares one

$$\text{stress}(X) = \sum_{i<j} w_{ij} \left[f(\delta_{ij}) - g(d_{ij}(X)) \right]^2$$

- and minimizes it to find the configuration X

$$\arg\min_X \text{stress}(X)$$

$d_{ij}(X)$ ... fitted distances
δ_{ij} .. proximities
w_{ij} ... finite weights
$g(\cdot), f(\cdot)$ ... transformation functions

Multidimensional Scaling (MDS)

- Provides an optimal map into continuous space $\mathbb{R}^M$ and looks for directions of spread in the low dimensional space (objective 1)
- But often one is also interested in discrete structures of similarity between objects (“clusters”; objective 2)
- MDS does solve objective 1 but not objective 2. The latter is often inferred from the former by how it looks
- It can happen that what is optimal for objective 1 is not very useful for objective 2

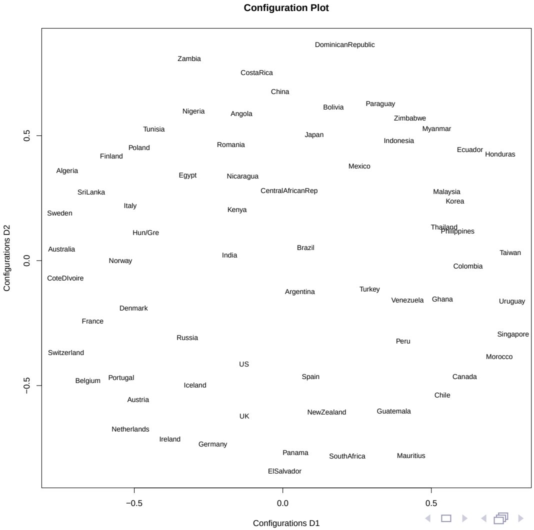
Illustration

Banking crises data from Reinhart & Rogoff (2009) (compiled by Graves, 2014):

- A panel data set of banking crisis history
- Time frame: 1800 to 2010
- Objects: 70 present-day independent states
- Binary entries (had crisis yes/no)

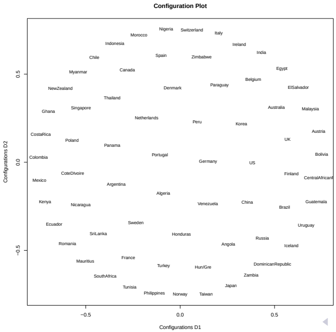
We use a binary asymmetric distance between the objects (Jaccard distance) and apply standard least squares MDS (SMACOF) for representation.

Illustration



Illustration

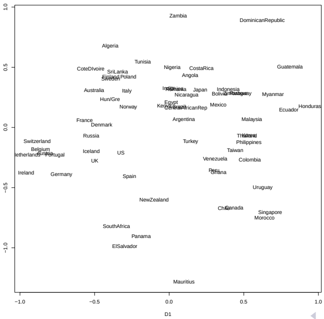
- Optimal configuration does not reveal a nice clustering structure.
- This is because of little variability in the proximities
- Known problem: MDS on data with little/no variability in proximities generates a configuration that resembles a sparsely populated sphere in $\mathbb{R}^M$ (projected to a disc in $\mathbb{R}^2$)



Illustration

Is there a way out?

- Following e.g., Mair et al. (2014) fit metric MDS with power transformation by setting e.g. $f(\delta_{ij}) = \delta_{ij}^{10}$)
- Clusters are much clearer but the fit is now worse (0.119 versus 0.127)



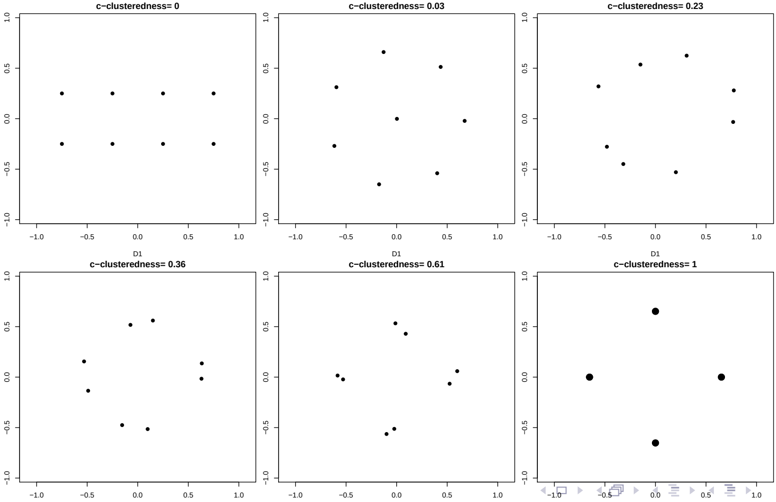
COPS for the Rescue

We propose a general solution to this problem that consists of the following steps:

- Use a stress with θ-parametrized strictly monontonic nonlinear transformations of either proximities or fitted distances or both e.g., power transformations (powerStress, $g(d_{ij}(X)) = d_{ij}(X)^\kappa$ and $f(\delta_{ij}) = \delta_{ij}^\lambda$, so $\theta = c(\kappa, \lambda)$)
- Use an index of the obtained degree of clusteredness in the configuration (c-clusteredness) to quantify the clusteredness
- Combine the stress function, the transformations and the clusteredness index into a single target function and optimize over the parameters
- We call this COPS (*C*luster *O*ptimized *P*roximity *S*caling; Rusch et al., 2014)

C-Clusteredness

C-Clusteredness: The amount of clusteredness of a configuration



OPTICS Cordillera

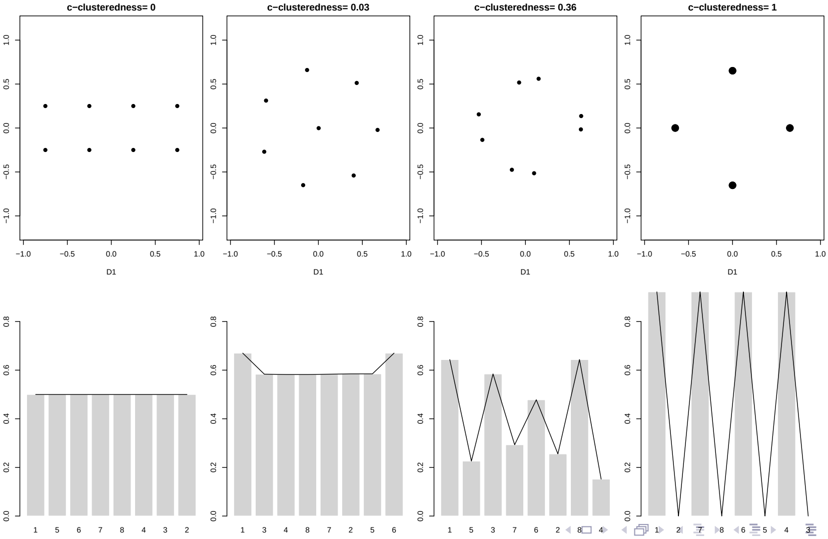
Index for clusteredness: OPTICS cordillera

- Employs OPTICS (Ankerst et al., 1999) with metaparameters k, ϵ on the configuration distances. For row vectors x_j of X returns an ordering R of these points, $R = \{x_{(i)}\}_{i=1,\ldots,N}$. So, $x_{(1)}$ is the x_j that is at position 1 in the ordering.
- OPTICS also returns a reachability plot (dendrogram of minimum reachabilities $r^*_{(i)}$ of point $x_{(i)}$)
- Ordering and reachability represents the cluster structure. We aggregate that to an index $OC(X)$ by defining (for metaparameter $q > 0$)

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

C... (optional) normalizing constant

OPTICS Cordillera



Properties of the OPTICS Cordillera

For given metaparameters ϵ, k, q the following applies (Rusch et al, 2014)

- Upper bound for the cordillera in the maximal c-clusteredness case (d_{max} is the maximum distance between any two points)

$$C^*(X, d_{max}, \epsilon, k, q) = \begin{cases} d_{max}^q 2\lceil \frac{N-1}{k} \rceil & \text{if (N-1)/k is integer} \\ d_{max}^q 2\lceil \frac{N-1}{k} \rceil - d_{max}^q & \text{if (N-1)/k is not integer} \end{cases}$$

- Cluster assignment or *a priori* defined number or shape of clusters not needed
- Index typically increases when
  - Distances between points increase
  - Distances between clusters increase
  - Points are denser clustered
  - Number of clusters increases
- Index does not pick up unbalancedness in the number of points

The Full COPS Procedure

Combine the $\theta-$ parametrized stress measure, stress $(X(\theta), \theta)$ and the OPTICS cordillera to cluster optimized stress (copStress):

$$\text{copStress}(\theta) = \text{stress}\left(X(\theta), \theta\right) - a \cdot \text{OC}\left(X(\theta)\right)$$

with $\arg\min_X \text{stress}(X, \theta) := X(\theta)$ and $a \in \mathbb{R}_+$ controlling how much weight should be given to the c-clusteredness, e.g,

$$a_0 = \frac{\text{stress}\left(X(\theta_0), \theta_0\right)}{OC\left(X(\theta_0)\right)}$$

with $\theta_0 = (1, 1)^\top$.

Optimization

We need to find

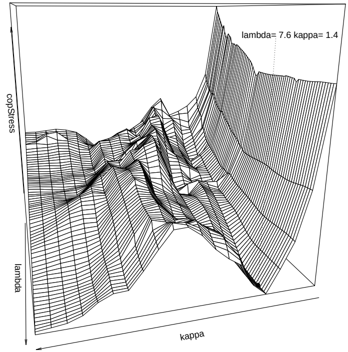
$$[\text{stress}\,(X(\theta),\theta) - a \cdot \text{OC}\,(X(\theta))] \to min_\theta!$$

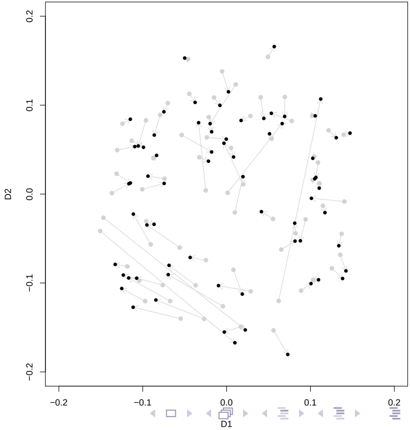
- We use an alternating algorithm that first solves for X and then minimizes over θ.
- The latter is non-smooth, so we employ random search or particle swarm algorithms or similar.
- The inner minimization to find X is extremely costly, so we need to have as small a number of outer steps.
- We had good experiences with an adapted Luus-Jaakola search (Luus & Jaakola, 1973; see Rusch et al., 2014).

This is implemented in the R package `stops`.

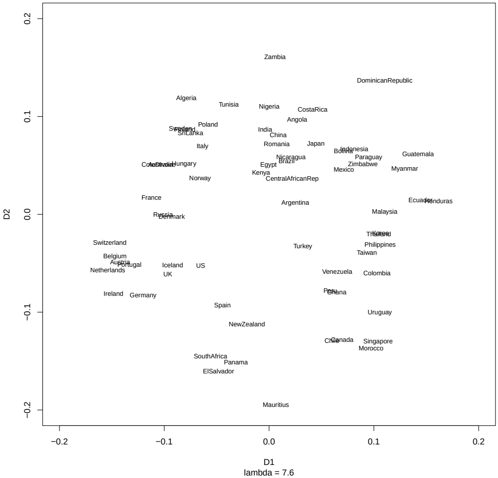
Example: Banking Crises

We use COPS on the banking data with power transformations of fitted distances and proximities (powerStress):





Example: Banking Crises



Conclusion

COPS

- COPS works well when the objective is to obtain a scaling and a clustering
- It is easily adaptable to many stress functions
- It is particularly useful when there is only little variability in the proximities

C-Clusteredness and OPTICS cordillera

- A concept and a measure of goodness-of-clustering in dimension reduction results that has appealing properties
- May be interesting beyond COPS

Outlook

TO DO

- The inner minimization is costly so COPS is not feasible even for a moderate number of objects
- Global optimality cannot be guaranteed
- Exploit the structure of the optimization problem
- Current implementation is still rudimentary

Beyond COPS (stay tuned)

- c-clusteredness is an aspect of a more general idea which we coin c-structuredness (Rusch et al., 2015)
- The idea of COPS can be generalized to STOPS (Structure optimized proximity scaling) (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.
- Graves, S. (2014) Countries in Banking Crises [data set]. From the R package Ecdat: Data sets for Econometrics, version 0.2-5
- Luus, R. & Jaakola, T. (1973) Optimization by direct search and systematic reduction of the size of search region, AIChE Journal, 19, 760–766.
- Mair, P., Rusch, T. & Hornik, K. (2014) The grand old party - A party of values? SpringerPlus, 3:697.
- Reinhart, C. & Rogoff, K. (2009) This Time Is Different: Eight Centuries of Financial Folly, Princeton University Press, New Jersey.
- 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.
- Rusch, T., Mair, P. & Hornik, K. (forthcoming) Structuredness indices and augmented nonlinear dimension reduction.

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/hum/trusch

WU Vienna University of Economics and Business
Welthandelsplatz 1, 1020 Vienna
Austria