COPS: Cluster optimized proximity scaling

Thomas Rusch
WU (Wirtschafts-
universität Wien)

Patrick Mair
Harvard University

Kurt Hornik
WU (Wirtschafts-
universität Wien)

Abstract

Proximity scaling (i.e., multidimensional scaling and related methods) is a versatile statistical method whose general idea is to reduce the multivariate complexity in a data set by employing suitable proximities between the data points and finding low-dimensional configurations where the fitted distances optimally approximate these proximities. The ultimate goal, however, is often not only to find the optimal configuration but to infer statements about the similarity of objects in the high-dimensional space based on the the similarity in the configuration. Since these two goals are somewhat at odds it can happen that the resulting optimal configuration makes inferring similarities rather difficult. In that case the solution lacks "clusteredness" in the configuration (which we call "c-clusteredness"). We present a version of proximity scaling, coined cluster optimized proximity scaling (COPS), which solves the conundrum by introducing a more clustered appearance into the configuration while adhering to the general idea of multidimensional scaling. In COPS, an arbitrary MDS loss function is parametrized by monotonic transformations and combined with an index that quantifies the c-clusteredness of the solution. This index, the OPTICS cordillera, has intuitively appealing properties with respect to measuring c-clusteredness. This combination of MDS loss and index is called "cluster optimized loss" (coploss) and is minimized to push any configuration towards a more clustered appearance. The effect of the method will be illustrated with various examples: Assessing similarities of countries based on the history of banking crises in the last 200 years, scaling Californian counties with respect to the projected effects of climate change and their social vulnerability, and preprocessing a data set of hand written digits for subsequent classification by nonlinear dimension reduction.

Keywords: proximity scaling, multidimensional scaling, nonlinear dimension reduction, similarity, OPTICS cordillera, c-clusteredness, c-structuredness, clusteredness index, cluster optimized loss.

1. Introduction

Proximity scaling (PS), an umbrella term for multidimensional scaling (MDS; Torgerson 1958) and related approaches, is a versatile and popular family of methods used in data analysis to represent high-dimensional proximities in lower-dimensional space. Many variants have

been developed over the years for diverse purposes. They can roughly be divided into two groups with respect to assumptions and their goals: On the one hand, there are techniques that assume that the high-dimensional data actually are measurements of "intrinsic data" that exist on a low-dimensional manifold embedded in a high dimensional space. They primarily aim at recovering and representing the embedded manifold by filtering out noise and removing correlated bits of information or extra dimensionality. This group often uses non-linear mappings between fitted and observed distances and has gained much interest in recent years. It includes techniques such as Sammon mapping (Sammon 1969), Isomap (Tenenbaum, De Silva, and Langford 2000), Locally Linear Embedding (Roweis and Saul 2000), Hessian Eigenmaps (Donoho and Grimes 2003) and Local MDS (Chen and Buja 2009). On the other hand, there are techniques that assume that there is some global, multidimensional relation in the data that is difficult to grasp in the original high-dimensional space. They primarily aim at representing and scaling of the high-dimensional data, preserving important global high-dimensional structure in a low-dimensional space and inferring patterns of similarity of objects from the representation. These include among others classical scaling (Torgerson 1958), metric and non-metric MDS (Kruskal and Wish 1978; Cox and Cox 2001; Borg and Groenen 2005), maximum likelihood MDS (Ramsay 1977), PGMDS (Mair, Rusch, and Hornik 2014) and various extensions. We stress that the dividing line between these two groups is blurry.

The general idea in proximity scaling is to reduce the multivariate complexity in a data set by employing suitable proximities between the data points and finding a low-dimensional representation (the configuration) where the distances optimally approximate the proximities with respect to some geometry. When using proximity methods for scaling, similarity assessment and visualisation, the ultimate goal is often not only to find the optimal configuration in lower-dimensional space but also to infer statements about discrete structures of and similarity of objects in the high-dimensional space from the spatial arrangement ("clusteredness") in the resulting configuration. Since these two goals are somewhat different, it can happen that the resulting optimal configuration shows little discernable clusteredness from which to infer discrete structures. A prime example is the case when there is little to no variability in the proximities, for which a standard PS solution will result in a configuration where in the respective geometry each point lies equidistant on or close to a single line in $\mathbb{R}^1$ (de Leeuw and Stoop 1984), on a sharply defined circular disk with low point density in the area close to or on one of a number of concentric circles with points lying close to or on the same circles being almost equidistant to each other in $\mathbb{R}^2$ (Buja and Swayne 2002) and in $\mathbb{R}^d, d > 2$, they lie close to or on a sphere with the points on or close to each sphere being almost equidistant (Buja and Swayne 2002; Buja, Logan, Reeds, and Shepp 1994). This lack of clusteredness in the configuration means that the points are more or less exchangeable, i.e., could be permuted to achieve similar fit which makes it difficult to infer statements about similarities in the original space solely based on the configuration.

Clusteredness is a somewhat elusive concept in statistics, because—while it has been discussed (e.g., in Greenacre 2011)—the definition remains vague. This also applies to clusteredness in relation to proximity scaling procedures. Informally by clusteredness often some inherent, possibly unknown property of the relation between the elements that make up the original data is meant. In this sense, proximity scaling methods—particularly from the second group—try to represent the original data so that this clusteredness is preserved if possible. For the purpose of scaling and similarity assessment, the relationship properties that consti-

tute clusteredness are the relative proximities between all the objects in the original space. Since these are difficult to grasp directly, it is inferred visually based on the relative fitted distances between all represented objects in the target space or, put differently, by the *visual appearance of clusteredness of the configuration*. To single out this distinction we call the latter property "c-clusteredness". Informally, under c-clusteredness we understand the following: There is a continuum of appearances of configurations where, starting from a result with no discernable clusteredness (e.g., equidistance between points), c-clusteredness increases if in the configuration (a) a (specified) number of represented objects accumulate close to each other, (b) the object representations are accumulating increasingly closer together, (c) the distances between locations of accumulation are increasing and (d) the number of accumulation locations increases. In modern data analysis a lack of c-clusteredness can easily occur, particularly when there are many points to scale simultaneously and the proximity calculation depends on the number of observations, when points are very close to each other in the original space or when certain proximities are used, particularly proximities for categorical data. The suggested approaches to alleviate this in PS is to use a "strong transformation" on the original proximities and/or to fit a nonlinear transformation of the distances (Borg and Groenen 2005).

Certain transformations applied to the fitted distances or to the proximities can lead to more c-clusteredness into the optimal solution. These transformations may be parametrized by a parameter vector θ that controls the strength of the transformation. It can be thought of as a dilation of shrinkage factor of the proximities and/or the fitted distances of a standard MDS. Then different values of θ typically change the c-clusteredness to the overall solution.

For illustration consider the banking crises data set (Graves 2014) used in Chapter 10 of Reinhart and Rogoff (2009). It is a panel data set of banking crisis history from 1800 to 2010 for 70 present-day independent states. It has been compiled by Reinhart and Rogoff from a number of sources (see A. 3. and A. 4. of Reinhart and Rogoff 2009, for a detailed explanation)[1]. The observations are binary entries for each year in which the present-day state experienced a banking crises as defined by Reinhart and Rogoff (2009)—which may be simplified as a situation that leads the public sector to intervene at at least one bank; 1 if so and 0 if otherwise. Greece and Hungary show an identical time series, so we combined their labels. We explore the similarities of countries based on these data of banking crises history, using the Jaccard distance measure, which basically measures how rarely banking crises occur in two countries in the same year; a distance of 1 means that there is no year in which two countries share a banking crisis. Note that these similarities are subtle in their meaning and

[1]We note that these are secondary data and for present-day countries not having existed as independent bodies before a certain year the data entries leave some room for interpretation. It appears as if the data represent a judgement call on the then prevalent fiscal and banking ties of the countries in question. For example, for present-day Austria and Hungary—which were double-monarchy Austria-Hungary from 1867 to 1918 and so the "same country"—the time series is equal during that period, with one exception: The "panic of 1873" (*"Gründerkrach"*) which was sparked by a crash of the Vienna stock exchange. The data set contains an entry of banking crises for Austria in that year but not for Hungary. The data source is Conant (1915), in which he writes e.g., "the rate of discount of the National Bank [of Austria] varied between 1817 and 1862 [...], and from 1863 to the fusion with the Austro-Hungarian bank in 1878 [...]". We believe he perceives two seperate banking entities for the two independent parts of Austria-Hungary. His assessment of crisis prevalence may thus point to an interpretation that at that time the financial hub for the Austrian part of Austria-Hungary was Vienna with the National Bank of Austria which was involved in the crash but that in the Hungarian part the role was played mostly by the Austro-Hungarian bank, and therefore this part may be interpreted as not having been affected.

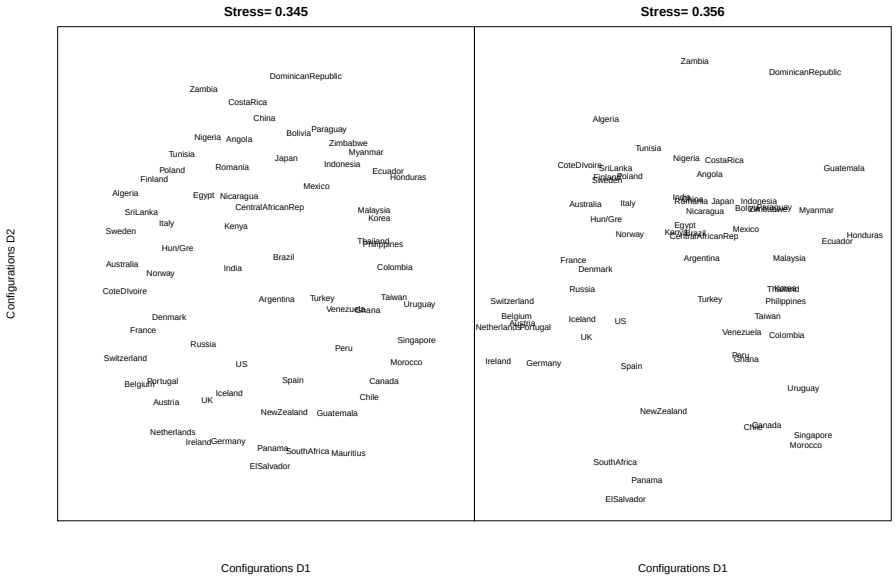

Figure 1: SMACOF MDS solutions for the banking crises data set from Reinhart and Rogoff (2009). The data set consists of binary entries of whether a banking crises was observed in a given year from 1800 to 2010 for 70 countries. The left panel is using the original Jaccard distance, the right plot uses the Jaccard distance to the power of 10.

reflect not only shared temporal occurence of banking crises between the countries but also geopolitical facts and mutual history. A standard SMACOF MDS leads to a rather concentric scaling of countries with little c-clusteredness (see left panel of Figure 1). Here this is due to little variability in the proximities as many countries have large Jaccard distances from each other. To alleviate this we can follow suggestions in, e.g., Mair *et al.* (2014); Buja and Swayne (2002); Buja, Swayne, Littman, Dean, Hofmann, and Chen (2008) and take the proximities to some power, say 10, to improve the visual display with respect to the clusteredness of the configuration. The resulting configuration is displayed in the left panel of Figure 1. The result appears more "clustered" in the configuration as compared to the original MDS. Put differently, it is easier to derive judgements about similarities of the objects based on the configuration obtained from the transformed proximities. Here this comes from emphasizing the original proximities differently: Larger ones are enlarged and smaller proximities are shrunk. Thus the (dis)-similarity of objects gets amplified.

There is one drawback to this procedure however: the fit of the configuration can get worse. For this example, the metric normalized stress value (stress-1) of SMACOF on the transformed observations is 0.36 which is worse than the fit for the original SMACOF (0.34). This evinces that the two objectives of finding an optimal configuration (in the sense of badness-of-fit to be as small as possible) and inferring discrete structures from the configuration's clusteredness can work in opposite directions. While using a power transformation does indeed improve the visual representation and c-clusteredness in the plots, it comes at the prize of a higher stress value. In this paper we suggest a way to balance both objectives and provide an optimal trade-off between stress and c-clusteredness.

Along these lines, the contribution of this paper is manifold: The main contribution is a

generic version of proximity scaling, which we call COPS (for Cluster Optimized Proximity Scaling), that incorporates an arbitrary MDS loss function, arbitrary strong transformations and a clusteredness index into a single loss function. For this we (re-)introduce power transformations of proximities and fitted distances as parametrized, flexible and general strong transformations that can increase the clustered appearance of a configuration. The former can actually be used in any generic proximity analysis variant. The latter will be tied more strongly to specific types of MDS. We also (re-)introduce a stress measure based on these ideas, powerStress. We define the notion of clusteredness of the configuration (c-clusteredness) rigorously and suggest an index that quantifies how much c-clusteredness we find. This index allows us to represent the global clusteredness property of the configuration, based on regions and distances of both close and far neighbouring points, in a unidimensional measure. While we present the procedure mainly for scaling and similarity analysis, the idea of incorporatiing an index into the loss function and the optimization for fit and structure is quite general and applicable to other related techniques (Rusch, Mair, and Hornik 2015b).

This article is organized as follows: We start with a description of proximity scaling and some related methods. We then turn to discuss the notion of clusteredness of the configuration, define c-clusteredness and suggest the OPTICS cordillera, an index that captures c-clusteredness unidimensionally. Subsequently, we elaborate on the idea of using transformations and describe power transformations as a general class of such transformations. This is followed by combining the ideas of using the transformations, minimizing loss and maximizing the c-clusteredness index into COPS, a variant of proximity scaling. In Section 5.2 we discuss optimization to find the optimal configuration and transformation for COPS. We then illustrate the use of COPS on three data sets and finish with concluding remarks in Section 7.

2. Proximity Scaling

For proximity scaling (PS) the input is typically an $N \times N$ matrix $\Delta^* = f(\Delta)$, a matrix of proximities with elements[2] δ^*_{ij}, that is a function of a matrix of observed non-negative dissimilarities Δ with elements δ_{ij}. Δ^* usually is symmetric (but does not need to be). The main diagonal of Δ is 0. We call $f(\cdot)$ the "proximity transformation function". The problem that proximity scaling solves is to locate an $N \times M$ matrix X (the "configuration") with row vectors $x_i, i = 1, \ldots, N$ in low-dimensional space ($\mathbb{R}^M, M \leq N$) in such a way that a transformation $g(d_{ij}(X))$ of the fitted distances $d_{ij}(X) = d(x_i, x_j)$—i.e., the distance between different x_i, x_j—approximates the (transformed) proximities δ^*_{ij} as closely as possible. We call $g(\cdot)$ the "distance transformation function". In other words, this means finding X so that $d^*_{ij}(X) = g(d_{ij}(X)) \approx \delta^*_{ij} = f(\delta_{ij})$.

The imperative in MDS is to find the approximation $D^*(X)$ to the matrix Δ^* which is in some sense optimal. It is found by defining a sensible fit criterion (the loss function), $\sigma_{MDS}(X) = L(\Delta^*, D^*(X))$, that is used to measure how close the approximation of $D^*(X)$ is to the observed proximity matrix Δ^*. Various such fit criteria have been suggested in the literature. Usually, they are closely related to the quadratic loss function. A general

[2]In the MDS literature these δ^*_{ij} are often called *dhats* or *disparities*. In the rest of the paper we choose a subtely different approach than is chosen in MDS with dhats, so we also call them differently: transformed proximities or δ^*_{ij}

formulation of a loss function based on a quadratic loss is

$$\sigma_{MDS}(X) = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left[d_{ij}^*(X) - \delta_{ij}^*\right]^2 = \sum_{i=1}^{N}\sum_{j=1}^{N} w_{ij}\left[g\left(d_{ij}(X)\right) - f(\delta_{ij})\right]^2 \tag{1}$$

Here, the w_{ij} are finite and allow for different ways of weighting the residuals by e.g., controlling the degree to which certain δ_{ij}^* contribute to the overall fit. The two most popular types of MDS loss functions, the *stress* and the *strain* loss functions, can be expressed as a special case of (1). They differ in how fitted distances and the proximities are defined.

The most popular type of MDS is based on the loss function type *stress*. This uses some type of Minkowski distance ($p > 0$) as the distance fitted to the points in the configuration,

$$d_{ij}(X) = ||x_i - x_j||_p = \left(\sum_{m=1}^{M} |x_{im} - x_{jm}|^p\right)^{1/p} \quad i, j = 1, \ldots, N. \tag{2}$$

Typically, the norm used in Equation (2) is the Euclidean norm, so $p = 2$ and representation happens in Euclidean space. In standard metric MDS $g(\cdot) = f(\cdot) \equiv I(\cdot)$, the identity function.

The w_{ij} play an important role in stress type loss functions as allowing for different types of weighting enables one to express a rich class of MDS variants as (1). For example one could be using w_{ij} that depend on the δ_{ij}^*, $w_{ij}(\delta_{ij}^*)$, for example as is used in explicitly normalized stress (Borg and Groenen 2005) with $w_{ij}(\delta_{ij}^*) = (\sum_{ij} \delta_{ij}^{*2})^{-1}$ or Sammon (1969) stress with $w_{ij}(\delta_{ij}^*) = \delta_{ij}^{*-1}$ or elastic scaling (McGee 1966) with $w_{ij}(\delta_{ij}^*) = \delta_{ij}^{*-2}$. Another is using w_{ij} as a function of the fitted distances, $w_{ij}(d_{ij}^*(X))$ which then allows to express e.g., stress-1 (Kruskal 1964) in (1) by setting $w_{ij}(d_{ij}^*(X)) = (\sum_{ij} d_{ij}^{*2}(X))^{-1}$. Often one also uses w_{ij} to incorporate some given additional information, e.g., derived from theory or hypothesis or using it in presence of missing values where $w_{ij} = 1$ if δ_{ij}^* is known and $w_{ij} = 0$ if δ_{ij}^* is missing. The simplest version is using $w_{ij} = 1$ (raw stress, Kruskal 1964). More complicated versions and all combinations of these are also possible.

With a specific choice for $f(\cdot)$ and $g(\cdot)$ in (1) one can also derive s-stress (with $\delta_{ij}^* = \delta_{ij}^2$ and $d_{ij}^*(X) = d_{ij}^2(X)$, Takane, Young, and de Leeuw 1977), multiscale stress (with $\delta_{ij}^* = \log(\delta_{ij})$ and $d_{ij}^*(X) = \log(d_{ij}(X))$, Ramsay 1977), generalized stress ($\sigma_G(X)$; with $\delta_{ij}^* = f(\delta_{ij}^2)$ and $d_{ij}^* = f(d_{ij}^2)$, Groenen, de Leeuw, and Mathar 1996) or the recent r-stress (with $\delta_{ij}^* = \delta_{ij}$ and $d_{ij}^* = d_{ij}^{2r}$, de Leeuw 2014).

The other popular type of MDS is based on the loss function type *strain*. Here the Δ^* are a transformation of the Δ, $\Delta^* = f(\Delta)$ so that $f(\cdot) = (h \circ l)(\cdot)$ where l is any function and $h(\cdot)$ is a double centering operation, $h(\Delta) = \Delta - \Delta_{i.} - \Delta_{.j} + \Delta_{..}$ where $\Delta_{i.}, \Delta_{.j}, \Delta_{..}$ are matrices consisting of the row, column and grand marginal means respectively. These then get approximated by (functions of) the inner product matrices of X

$$d_{ij}(X) = \langle x_i, x_j \rangle \tag{3}$$

In what follows, in the context of *strain* (but not in stress) we always assume that f is a composite function of the doubly centering function and some other function and can thus express classical scaling as a special case of (1) with $d_{ij}(X)$ as in (3), $g(\cdot) = I(\cdot)$ and $f(\cdot) = (h \circ I)(\cdot)$.

The loss function used, e.g., equation (1), is then minimized to find the vectors $x_1, \ldots, x_N$, i.e.,

$$\arg\min_X \ \sigma_{MDS}(X). \tag{4}$$

There are a number of optimization techniques one can use to solve the problem in (4). For example, for most of the stresses in use there exist theoretically well-understood optimization algorithms to minimize them. Often we have a majorization algorithm (de Leeuw 1977) or standard iterative gradient decent algorithms (Buja and Swayne 2002). Strain losses can be solved analytically. We note that the suggestions we make here are aimed at being applicable beyond specific fit functions and optimization techniques.

3. C-Clusteredness and a C-Clusteredness Index

We motivated this paper by the observation that for certain data, proximity scaling might lead to solutions where the configurations are not very clustered. In this section we first formalize the notion of clusteredness of the configuration (coined c-clusteredness) and then present an index that captures the c-clusteredness of the result.

3.1. Structure and C-Clusteredness

For this paper, we broadly assume that there exists some real, *unknown*, qualitatively defined relation between observations in the original sample space. We call this the "structure" in the data set. One type of structure that often is of particular interest are similarity relations between observations, e.g., discrete groupings of observations or closeness and mapping of observations in an underlying continuous space. This type of structure can be considered to be called clusteredness in a continuous space. Since it is unknown or difficult to grasp in the original sample space, scaling procedures are employed to preserve or unveil it. This is usually done by explicitly solving for the continuous representation and then deriving the discrete structure[3]. At any rate, the real, unknown structure of interest has to be inferred from the result of the scaling procedure. This is usually done by equating the clusteredness in the continuous space spanned by the configuration with the real structure. This property, the appearance of a clustered result in the low-dimensional representation found by the scaling procedure, is what we refer to as c-clusteredness.[4]

To make this notion concrete and quantitatively accessible, we need to be clear what constitutes a clustered result in the configuration, i.e., a result that has c-clusteredness. By complete lack of clusteredness of the configuration, we mean that all points fall on the vertices of a regular tesselation and all points can be connected to each other with non-crossing lines of constant length. Figure 3.1 shows some examples of no c-clusteredness. They all have in common that they can be described by unit distance graphs if the edges connect nearest neighbours, are not allowed to cross and the edge length is proportional to an integer constant.

Conversely, by perfect clusteredness of the configuration (highest possible c-clusteredness) we mean the following: First, it is possible to evenly distribute the N data points into N/k

[3] As opposed to clustering where the grouping is searched for directly.

[4] Clusteredness is only one type of structure that one may aim at preserving. We coin the set of these types of structures preserved in the configuration as "c-structuredness" and refer to our upcoming paper (Rusch *et al.* 2015b). c-clusteredness is an (often particularly interesting) instance of c-structuredness.

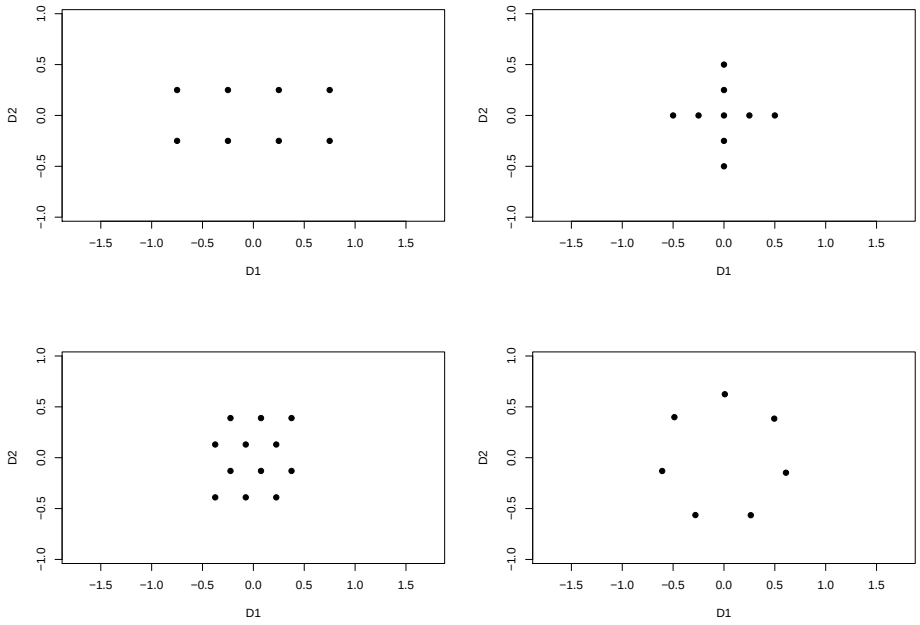

Figure 2: Examples of 2D configurations that have little c-clusteredness. Accordingly, they all have a normalized cordillera of < 0.001.

clusters each of size k. The parameter k is therefore also the minimum number of closest points that must form a cluster. More formally let k be the minimum number of points that form a cluster and let us assume it is possible that $N \equiv 0 \pmod{k}$. Let $d_{ij}(X) := d_{ij}$ be the distance between points x_i, x_j and $N_k(x_i) = \{x_j : \sum_{x_s} \mathbb{1}(d_{is} < d_{ij}) \leq k-1\}$ denote the set or neighbourhood of k closest points to and including x_i (a k-cluster). Second, we then define maximal c-clusteredness to be achieved if for all x in the same cluster their distance to each other is zero and each cluster is some constant distance d_c away from the closest other cluster, which is also the maximum distance between any two points in these two clusters, or for point x_i,

$$d_{ij} \begin{cases} = 0 \text{ if } x_j \in N_k(x_i) \\ = d_c > 0 \text{ if } x_j \notin N_k(x_i) \wedge x_j \in N_k(x_s) : d_{is} = \max(0, \min d_{it}) \quad \forall\ i \neq s, t \\ \geq d_c \text{ otherwise.} \end{cases} \tag{5}$$

where d_c is some constant (positive) distance.Thus the k points in the same cluster have no distance to each other, $d_{ij} = 0$, and all the positions at which k points coincide are some constant, minimal distance d_c away from each other, equidistant to their closest neighbouring cluster. See also the bottom right plot of Figure 3.

The observed c-clusteredness is now to be understood as the position of a configuration on a continuum between no c-clusteredness and maximal c-clusteredness as given above, subject to the following desirable properties: c-clusteredness increases i) if in the configuration the object representations are more spaced out, ii) if the distances between groups of points increase, iii) if the object representations are clustered more densely and iv) if the number of clusters

increase. In the next section we suggest an index that exhibits these properties among others and allows us to quantify c-clusteredness.

The left column of Figure 3 illustrates c-clusteredness with a toy example of 8 data points. In the top tow panels we have examples with low c-clusteredness, i.e., all points being close to equidistant to their closest neighbours. The top plot shows points lying on a regular tesselation and in the second plot the differences to the closest neighbours on the circles are equidistant but the radius of the circles are not constant. Note that the latter is very close to the case for standard MDS with equal proximities. In the next two rows we have examples of configurations showing higher c-clusteredness and in the bottom two rows we find highly clustered results where it is clear which two points accumulate and the locations of accumulation are rather distinct (high c-clusteredness). In the most bottom plot this is pushed to the extreme as at each of the four positions there are two points coinciding (ilustrated with the twice magnified dots) and all four positions are equally far away from the closest other group. This is our definition of maximal c-clusteredness.

3.2. A C-Clusteredness Index: The OPTICS cordillera

We now define an index which quantifies how close a given configuration X is to the definition in (5) and which fullfills the desirable properties laid out earlier. Our index is derived from OPTICS (Ordering Points To Identify The Clustering Structure; Ankerst, Breunig, Kriegel, and Sander 1999), an algorithm that outputs a unidimensional ordering of input points based on a matrix of distances. The algorithm assigns each input point a single linkage distance ("minimum reachability distance") and effectively orders points in such a way that points that get ordered in sequence are close to each other in the input space unless a point's minimum reachability distance is large. This ordering-reachability combination is appealing for our purpose: to map the clusteredness of a configuration to a univariate scale. Our index is an aggregation over the reachabilities of the OPTICS ordering for the points in the fitted configuration X and summarizes the information of clusteredness contained in both the ordering and an adapted definition of reachability. It exhibits a number of desirable properties for this taks, which we will discuss below.

Distance definitions and the OPTICS algorithm OPTICS allows to use two parameters: The mandatory parameter k (in OPTICS called minpts) which for our purpose is the minimum number of points needed to comprise a cluster, and a parameter ϵ which stands for the maximum radius of a neighbourhood around a point in which the algorithm looks for points that may form a cluster. The latter is optional and can be used in OPTICS to make the procedure robust to outliers and for improving the runtime. The parameter k has a smoothing effect and needs to be set *a priori*.[5]

The distances used in OPTICS are defined the following way: Let $N_\epsilon(x_i) = \{x_j : d_{ij} < \epsilon\}$ be the set of neighbouring points to and including x_i within a radius of ϵ. Let $S_k(x_i; \epsilon)$be the subset of $N_\epsilon(x_i)$ that contains the $k-$th closest neighbouring points to x_i, $S_k(x_i; \epsilon) \subseteq N_\epsilon(x_i)$. Note that this set will usually contain a single element, but may have more than one. If $\text{card}\,(N_\epsilon(x_i)) < k$, then $S_k(x_i; \epsilon) = \emptyset$. There is the "core distance", which is the distance of a

[5]There may be a value that makes sense in light of the application. Barring that, we made good experiences with simply setting it to 2.

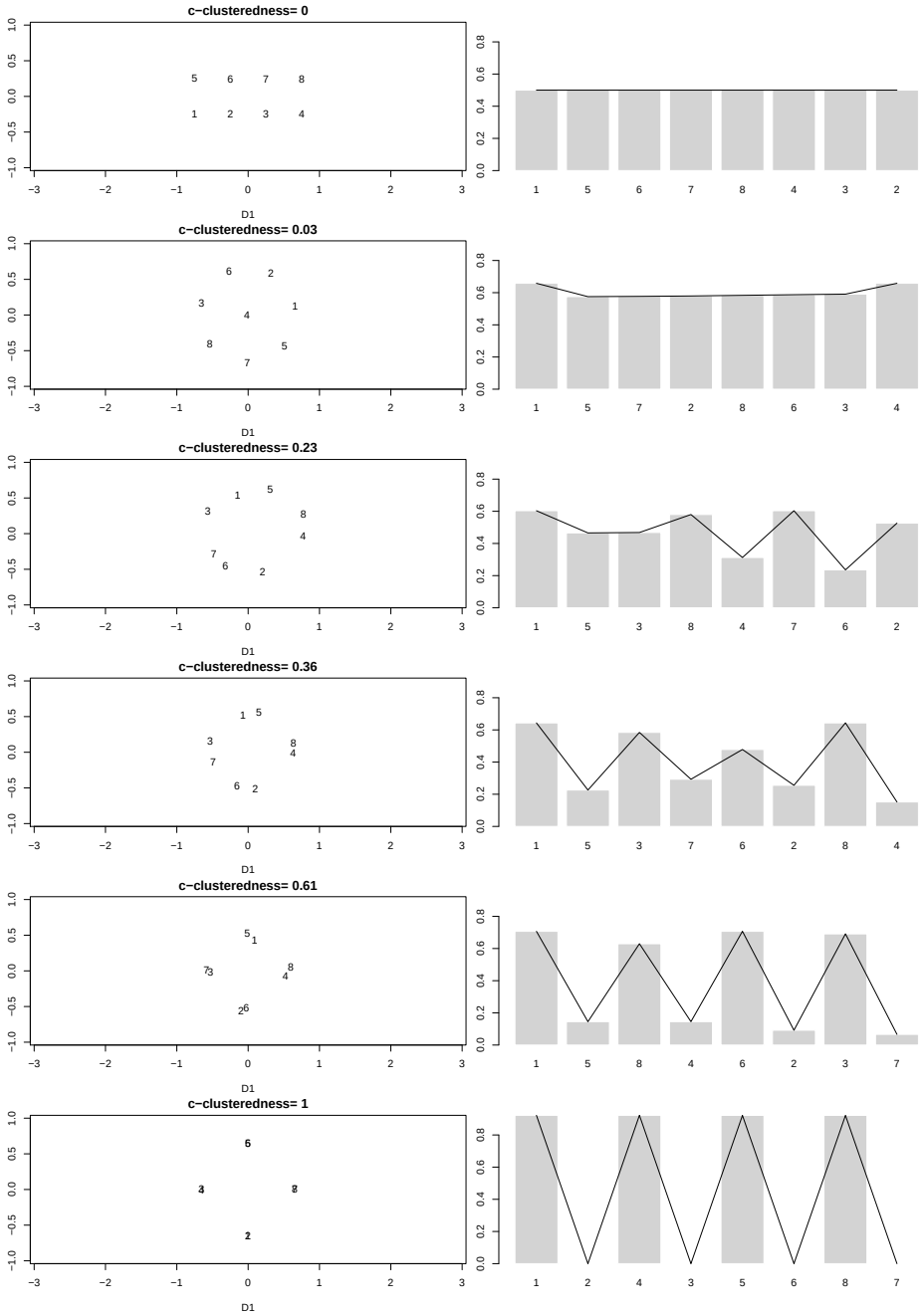

Figure 3: Differently clustered 2D representation of 8 points. In the left column we find different configurations. Here the top left plot shows a regular tesselation with little c-clusteredness, the second plot shows an MDS solution that appears for very little variability in the proximities, the bottom left panel shows extreme structure (at each point there lies the same number of points and between each group the distances are equally large). The other three panels show realistic versions between these extremes. The c-clusteredness increases from top to bottom. In the right column we find the corresponding OPTICS reachability plots and the derived c-clusteredness index, the raw OPTICS cordillera. The plots are labeled with the numeric value for the normalized OPTICS cordillera. It has been calculated with $k = 2, \epsilon = 2, p = 1$.

vector x_i to (any of) the $k-$th closest vector (if it exists)

$$c_i = c(x_i; \epsilon, k) = \begin{cases} \max(d_{ij}) : j \in S_k(x_i; \epsilon) & \text{if } S_k(x_i; \epsilon) \neq \emptyset \\ \text{undefined} & \text{if card}\left(N_\epsilon(x_i)\right) < k \end{cases} \tag{6}$$

and the "reachability distance" between two points x_i and x_j, which is either the distance d_{ij} between x_i and x_j or $c(x_i; \epsilon, k)$, i.e,

$$r_{ij} = r(x_i, x_j; \epsilon, k) = \begin{cases} \max\left(c_i, d_{ij}\right) & \text{if } S_k(x_i; \epsilon) \neq \emptyset \\ \text{undefined} & \text{if card}\left(N_\epsilon(x_i)\right) < k \end{cases} \tag{7}$$

Based on these distances the OPTICS algorithm now orders the points and outputs that ordering R together with the smallest reachability distance of the point x_i ("minimum reachability"), $r_i^* = \min_{j:i\neq j} r_{ij}(x_i, x_j; \epsilon, k)$, of a vector x_i. Later we will need also the position of point x_i in the ordering R, $s = s(x_i, R) = \text{position}(x_i, R)$, so when we refer to an x_i in R, we will call it $x_{(s)}, s = 1, \ldots, N$ with corresponding minimum reachability $r^*_{(s)}$. We will switch between both notations to emphasize whether we talk about points in the configuration X or in the OPTICS ordering R.

The ordering itself is created by a priority queue algorithm that is difficult to express non-algorithmically, so we refer to the OPTICS for details Ankerst *et al.* (1999). The principle is the following: A point gets visited and the neighbours are recorded. Then all the neighbours get pushed into a priority queue which is iteratively updated for the "reachability distance" based on the ϵ-neighbourhood of the point and the neighbours in the queue. Then the queue gets processed iteratively. This way, points in the ordering that are subsequent and have small minimum reachability correspond to points close to each other and may belong to the same cluster whereas points that are far away from each other in the ordering or have some large reachability between them belong to different clusters.

For our purpose we adapt the original distance definitions by setting

$$\begin{aligned} c_i = c(x_i; \epsilon, k) &= d_{max} \text{ if } c_i \text{ is undefined} \\ r_{ij} = r(x_i, x_j; \epsilon, k) &= d_{max} \text{ if } r_{ij} \text{ is undefined } \vee\ r_{ij} > d_{max} \end{aligned} \tag{8}$$

This assures that we have numeric values for undefined points as well and that we can include these in the computation of the index. Also, d_{max} caps the "reachability distance", so this can be used to make the index robust. The choice of d_{max} has different implications. In many cases one would want to set d_{max} to $\max_{i,j} d_{ij}$. This will assign the maximum observed reachability to the observations with undefined distances. This choice may make the index below susceptible to large outliers, so setting d_{max} to some hard threshold makes the index more robust. Another sensible choice would be $d_{max} = \epsilon$ if the parameter is actually used for the OPTICS result (and not just set to some large value).

The OPTICS Cordillera We can then use the ordering of points and the reachability distances to fashion a c-clusteredness index. Let $R = \{x_{(s)}\}_{s=1,\ldots,N}$ be the ordered set of the original points $x_i, (i = 1, \ldots, N)$ as output by the OPTICS algorithm, so $x_{(1)}$ is the x_i at the first position in R. Let $r^*_{(s)} = r_i^* = \min_{j\neq i} r_{ij}$ be the minimum reachability as defined in (7) and (8) of point $x_{(s)} = x_i$. Then by using the q-norm of the finite difference of the

minimum reachbilities over the ordering of points, we define a c-clusteredness index—the OPTICS cordillera as—

$$\mathrm{OC}(X;\epsilon,k,q) = \left(\frac{\sum_{s=2}^{N} |r^*_{(s)} - r^*_{(s-1)}|^q}{C}\right)^{1/q} \tag{9}$$

where C is some (optional) normalizing constant and metaparameter $q > 0$. If $C = 1$ we call (9) the raw cordillera. Often letting C depend on X and the parameters in (9), $C = C(X, \epsilon, k, q)$, is sensible.

Properties One can show (see the propositions in Appendix A) that this index has several intuitively appealing properties—which in turn could be used as axioms for measures of c-clusteredness. These are no concept of cluster assignment or cluster numbers, a definition of how many observatiosn must make up a cluster as well as a number of properties we call the shape, density, emphasis, tally and balance property respectively. It further exhibits a property (the spread property) that can be a double-edged sword. More detailed, for a given ϵ, k, q it holds that

- This index does not need any cluster assignment of observations, nor an *a priori* defined number of clusters nor any labels of real cluster membership. This is a very important property in the exploratory, unsupervised setting where MDS is typically used.
- At least k points with distance of at most ϵ must make up a cluster. An accumulation of less than k points with distances less than ϵ does not count as a cluster. This follows from the definition of $N_\epsilon(x_i)$ and $S_k(x_i)$.
- **Shape Property:** The geometrical shape of the cluster can be arbitrary. This carries over directly from the properties of the OPTICS algorithm, which picks up clusteredness based on density considerations (Ankerst *et al.* 1999).
- **Emphasis Property:** All else equal, for increasing distances between different clusters or groups of points (if possible), the index is non-decreasing and typically increasing (Proposition 1).
- **Density Property:** All else equal, for points that are close in together in a cluster, shrinking points monotonically in the cluster towards the center point x_i will lead to an nondecreasing and typically increasing index. (Proposition 2). In essence the index increases if the points are clustered more densely.
- **Tally Property:** All else equal, for an increase in the number of clusters, the index is non-decreasing and typically increasing (Proposition 3).
- **Balance Property:** All else equal, for a given number of clusters the index is non-increasing in the number of observations $> k$ in a bin. Thus it will not pick up unbalancedness in the number of points in a cluster as a sign of c-clusteredness (Proposition 4).
- **Spread Property:** All else equal, for a sufficiently large increase in distances between points (if possible), the index is nondecreasing (Proposition 5). In essence, the index is nondecreasing when points are so spread out that it appears sensible to assume there

are points that are qualitatively different in the sense of being very far from the rest. This property is related to the density property: If a point is so far away from other points that they are no longer recognized to be the same discrete structure, the spread property eventually takes over. In this case OPTICS tells us that it is no longer seeing a decreasing density in a cluster but a sign of something qualitatively different, or an outlier. We note that this can make the index susceptible to outliers which can be combatted by setting ϵ and k so that the index is robust against undue influence of outliers.

To summarize, the OPTICS cordillera as an index is basically looking for the difference in minimum reachability of the $x_{(s)}$ over R as a means of capturing the global appearance of similarity between observations. The OPTICS ordering is so that if the reachability for $x_{(s)}$ is small then $x_{(s)}$ und $x_{(s-1)}$ are close, if it is large then $x_{(s-1)}$ is far away from $x_{(s)}$ and $x_{(s-2)}, x_{(s-3)}$ and so on are also far away from $x_{(s)}$. How far away (at least) is quantified by $r^*_{(s)}$. Thus, if two points are close to each other in the ordering and the differences in reachabilities of all points between these two points are small, the group of points can be considered to belong to the same cluster. If there is some large reachability between these points then the points likely belong to different cluster. The OPTICS cordillera aggregates this information as the sum of the differences of minimum reachabilities for the different $x_{(s)}$ over R and the larger this sum, the larger the index is and the more c-clusteredness we typically find in the solution and vice versa.

Normalizing the OPTICS Cordillera The constant C in (9) can be used to normalize the index to the scale on which the proximity scaling target function operates. For example in many commonly used loss functions such as SMACOF and Sammon mapping, the normed stress lies in $[0, 1]$ and we can choose $C = C(X, \epsilon, k, q)$ so as to ensure that $OC(X; \epsilon, k, q) \mapsto [0, 1]$ as well. One way of achieving this is to use the definition of a maximally clustered configuration from (5) for normalizing. A non-trivial upper bound for the cordillera in the maximal c-clusteredness case is given by (Proposition 6 in Appendix B)

$$C^*(X, d_{max}, \epsilon, k, q) = d^q_{max} \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right) \tag{10}$$

Note $C^*(X, d_{max}, \epsilon, k, q)$ is monotonically nonincreasing in k, so setting $k = 2$ will serve as a general upper bound independent of k.

When choosing a value for d_{max} in $C^*(X, d_{max}, \epsilon, k, q)$, it is useful to distinguish between optimizing for c-clusteredness relative to the largest possible distance for a given configuration versus for a series of configurations or a constant. This will control the interpretation of the index: It can be given the interpretation as the amount of c-clusteredness attained relative to the most c-clusteredness achievable for a given configuration X. This is when $d_{max} = d_{max}(X) = \max_{i,j} d_{ij}(X)$. This will then give the index an interpretation of "goodness-of-clusteredness", conceptually similar to an R^2. It can also be given the interpretation of an absolute index for comparing a series of configurations $X^{(1)}, \ldots, X^{(G)}$ with respect to c-clusteredness. In this case $C^*(X, d_{max}, \epsilon, k, q)$ should be the same for all G results as d_{max} can be different for different solutions, so one might either set $d_{max}(X^{(1)}, \ldots, X^{(G)}) = \max_g d_{max}(X^{(g)})$ for solutions $g = 1, \ldots, G$. The third possibility is to set d_{max} to an *a priori* constant value, e.g., ϵ or some other distance that must be attained at least. That constant can also be used

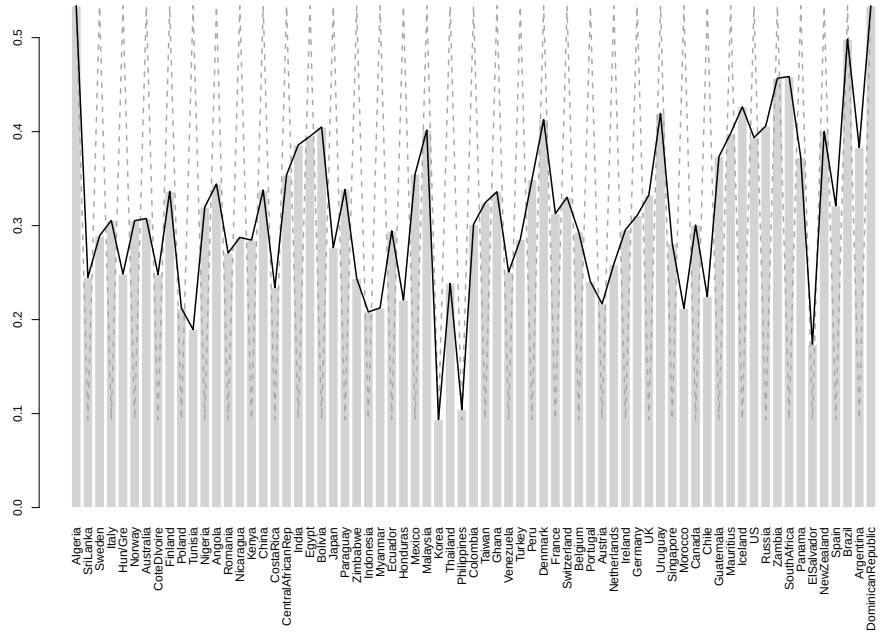

Figure 4: A visualisation of the OPTICS cordillera for the banking crises data on the untransformed proximities. The barplot is a reachability plot as output by OPTICS with the adaptions described in the text. The x axis are the points and the y axis is the minimum reachability distance for each point. The cordillera (black line) and the normalizing constant (darkgrey dashed line) is also drawn. The raw index is now simply the length of the black line (which is 5.2 and can be normalized by the dark grey line (which leads to a normed c-clusteredness of 0.14 in this example). Note that due to the bar width the displayed lengths are only proportional to the real length. The parameter of the cordillera were $q = 1, k = 2$ and $\epsilon = 10$.

to make the index robust against outlier points (by, e.g., setting it to the 0.9 quantile of the distribution of distances).

Illustration Figures 4 and 3 illustrate the concepts. Figure 4 does this for the banking data example. Figure 3 shows in the right column the OPTICS cordillera and the reachability plots for the configurations in the left column. We used $q = 1$ here. The grey barplot shows the minimum reachability on the y-axis for the ordering $x_{(i)}$ on the x-axis. The raw cordillera is the black line. It holds that the larger this line is, the more structured the configuration is. The cordillera reaches a minimum if all points have equal minimum reachability. The upper bound for the cordillera (10) is illustrated for the banking data in Figure 4 as the dashed line. The bottom right raw corrdilera is also the upper bound for all the cordilleras (with $d_{max} = \max_g d^{(g)}_{max},\ g = 1, \ldots, 6$) configurations in Figure 3. We clearly see the ever increasing cordillera with ever increasing c-clusteredness.

4. C-Clusteredness Inducing Transformations

When presented with unclustered solutions as described earlier, one can apply transformations to the dissimilarities ("strong transformations", see Borg and Groenen 2005, p. 272) or the fitted distances to alleviate the scaling problems. We define such a transformation as any monotonic transformation function, perhaps parametrized with vector θ. For proximities we call these proximity transformation functions (sometimes called representation functions), which means $f : (\delta_{ij}, \theta) \mapsto \mathbb{R}$ for which it holds that for the proximities $\delta^*_{ij} = f(\delta_{ij}; \theta)$ in (1).

For the fitted distances these are distance transformation functions, $g : (d_{ij}(X), \theta) \mapsto \mathbb{R}_+$ for which we then have $d^*_{ij}(X) = g(d_{ij}(X); \theta)$ in (1). Using such a transformation leads to a loss function value that depends on θ, so $\sigma_{MDS}(X, \theta)$.

Many such transformation can and have been considered, e.g., (Buja and Swayne 2002; Ramsay 1977; Takane *et al.* 1977; de Leeuw 2014; Borg and Groenen 2005; Buja *et al.* 2008; Groenen *et al.* 1996; Chen and Buja 2014; Mair *et al.* 2014). For the problem of having a result with little c-clusteredness for the original proximities or original distances, we are particularly interested in transformations that allow to (de)-emphasize proximities/distances differently by enlarging or shrinking proximities/distances relative to their magnitude and thus pronouncing a more clustered appearance in the configuration. It should also include the "worst case" of equal proximities/distances and the original proximities/distances as special cases and should be parametrized to switch easily between different solutions. Below we discuss a class of simple transformations that meet these criteria, namely power transformations. To illustrate we continue with the banking crises data.

4.1. Transforming Observed Proximities

A simple and flexible way of adding c-clusteredness to the solution is by (non-)linearly transforming the input proximities used in (1). In the MDS literature such transformed proximities are often called *dhats* and the resulting MDS is called metric MDS.

Following, e.g., Ramsay (1977); Buja and Swayne (2002); Buja *et al.* (2008); Mair *et al.* (2014) we can use power functions as the proximity transformation functions, i.e., taking δ_{ij} to the power of λ, with $\lambda \in \mathbb{R}$. Thus $\theta = \lambda$ is scalar and in (1)

$$f(\delta_{ij}, \theta) = \delta^*_{ij}(\theta) = \delta^{\lambda}_{ij} \qquad \text{(for stress)} \tag{11}$$

$$f(\delta_{ij}, \theta) = \delta^*_{ij}(\theta) = h(\delta^{\lambda}_{ij}) \qquad \text{(for strain)} \tag{12}$$

with δ_{ij} being the original, untransformed dissimilarities and $\delta_{ij} = \delta^*_{ij}(1)$ and $h(\cdot)$ being the doubly centering function. Note that if λ also gets estimated by minimizing the stress function, this is equivalent to metric MDS with power transformations for the *dhats*. This transformation's effect on a loss measure as in (1) is similar to differently weighting the proximities in the scaling process. The effect of $\lambda > 1$ is that there is a stronger relative emphasis on larger distances so clustered parts are more emphasized. For $\lambda \in (0, 1)$ the transformation is a root transformation and the effect would be to relatively shrink large proximities and increase small proximities. For $\lambda = 1$ all proximities stay as they are and thus the case of original proximities is represented in our parametrization. For $\lambda < 0$ the effect of the transformation is the opposite as now the reciprocal of the power transformed proximities is used. A special case is when $\lambda = 0$. There any variability in proximities gets nullified and a non-clustered solution is forced. This is important to be included as a special

parameterization as we can introduce clusteredness relative to this solution, the problem that motivated this approach in the first place. Notice that in case of a low c-clusteredness solution for the original proximities, the values of the cordillera for $\lambda = 0$ and $\lambda = 1$ should be very close.

A special case arises here when $\delta_{ij} = 0$ and $\lambda = 0$, which is somewhat problematic as this means two points coincide which is normally important information in MDS. Nevertheless, we decided to be treating this case as $0^0 = 1$ as the case of $\lambda = 0$ is only of interest if one wants to destroy any information from the proximity matrix and force an unclustered solution, and then having two points at the same place can be considered unimportant.

The advantage of these transformations is that they can easily be implemented for and applied with all metric proximity scaling approaches as they only needs to have the δ_{ij} transformed.

4.2. Transforming Fitted Distances

Another way of approaching the problem is by applying a distance transformation function to the fitted distances. Again such a transformation may be given by the power transformation, so taking $d_{ij}(X)$ to the power of κ, with $\kappa \in \mathbb{R}_+$. Thus here $\theta = \kappa$ and

$$g(d_{ij}(X), \theta) = d^*_{ij}(X, \theta) = d_{ij}(X)^\kappa, \tag{13}$$

gets plugged into (1) with the objective to solve (1) so that $d^*_{ij}(X, \kappa) \approx \delta_{ij}$. This transformation is consistent with our definition of a strong transformation because we assume $d_{ij}(X) \geq 0 \; \forall \; i, j$. A number of stress versions are special cases of this type, including raw stress ($\kappa = 1$, Kruskal 1964) and r-stress ($\kappa = 2r$, de Leeuw 2014). This parametrization includes the untransformed solution as a special case. As a function of κ this is a parametric form of nonmetric scaling with powers. The effect of a larger κ is that there is a stronger relative emphasize on smaller distances. Thus for increasing κ, points that are closer together when using untransformed distances are shrunk together wheras points that are further away from each other get pushed away further. Note that for this approach other than with transformed proximities, the fitting procedure must usually be adapted for different transformations. Gradients for this type of stress can be found in Groenen *et al.* (1996).

4.3. Transforming Observed Proximities and Fitted Distances

It is also possible to apply a proximity transformation function and a distance transformation function simultaneously. When using different power functions as the transformation, this leads to a stress version already introduced by Buja *et al.* (2008), which we call *powerStress*. Here one takes $d_{ij}(X)$ to the power of κ and δ_{ij} to the power of λ, with $\lambda \in \mathbb{R}, \kappa \in \mathbb{R}_+$, so θ is a two-dimensional parameter vector, $\theta = (\kappa, \lambda)^\top$ and the transformations are

$$\begin{aligned} g(d_{ij}(X, \theta)) &= d^*_{ij}(X, \theta) = d_{ij}(X)^\kappa, \\ f(\delta_{ij}, \theta) &= \delta^*_{ij}(\theta) = \delta^\lambda_{ij}. \end{aligned} \tag{14}$$

Substituted into (1) this leads to the stress measure

$$\text{powerStress}(X; \theta) = \sum_{i=1}^{N} \sum_{j=1}^{N} w_{ij} \left[d_{ij}(X)^\kappa - \delta^\lambda_{ij} \right]^2. \tag{15}$$

Minimizing (15) for given θ can be achieved, e.g., with nested majorization as in de Leeuw (2014). In case of $\kappa, \lambda = 1$ this is again the standard stress, with $\lambda, \kappa = 2$ this is s-stress (Takane *et al.* 1977). This loss function includes the untransformed solution as a special case. As a function of θ this is a parametric form of nonmetric scaling with powers. Gradients and mathematical properties for powerStress can again be deduced from Groenen *et al.* (1996). This is a general way of using simple power transformations for multidimensional scaling. The effects are similar to the behaviors described previously, only that now the two parameters can compensate each other. In our experience using powerStress with sensible values of θ leads to well fitting and clustered configurations, with the power transformations on the fitted distances allowing for a fit measure close to 0 and the power transformations of the proximities leading to high c-clusteredness.

Applying a nonlinear, parametrized transformation to fitted distances or proximities can be extended in myriad other ways. We want to specifically point out a recent approach by Chen and Buja (2014), who suggest to use Box-Cox transformations on fitted and observed distances. This constitutes an interesting alternative to powerStress should the case of $\kappa = \lambda = 0$ or $\kappa = \lambda = 1$ play a less prominent conceptual role.

5. Cluster Optimized Proximity Scaling (COPS)

In this section we combine the ideas of scaling with c-clusteredness introducing transformations and the OPTICS cordillera. We propose a variant of proximity scaling, coined COPS (for Cluster Optimized Proximity Scaling), that fits a configuration with optimal parameters θ to a proximity matrix based on the trade-off between the fit of the distances in the configuration to the proximities and the c-clusteredness introduced by using a θ-parametrized loss function.

5.1. Cluster Optimized Loss

Let us write $X(\theta) = \arg\min_X \sigma_{MDS}(X, \theta)$ for the optimal comfiguration for transfomration parameter θ., The overall objective function, which we call "cluster optimized loss" (coploss), is simply a weighted combination of a $\theta-$parametrized loss function, $\sigma_{MDS}\left(X(\theta), \theta\right)$, and the c-clusteredness measure, $OC(X(\theta); \epsilon, k, q)$ to be optimized as a function of θ. We stress that the $\sigma_{MDS}\left(X(\theta), \theta\right)$ employed here needs to be scale and unit free or different values of θ, so that less loss means a relatively better fit.

More formally, coploss is then

$$\text{coploss}(\theta) = v_1 \cdot \sigma_{MDS}\left(X(\theta), \theta\right) - v_2 \cdot \text{OC}\left(X(\theta); \epsilon, k, q\right) \tag{16}$$

with $v_1, v_2 \in \mathbb{R}$ controlling how much weight should be given to the stress and the c-clusteredness. In general v_2, v_2 is either an *a priori* determined value that makes sense for the application or may be used to trade-off fit and c-clusteredness in a way for them to be commensurable. In the latter case v_1, v_2 basically can be used to account for different scales on which c-clusteredness and the loss function may lie, for example, compensate different normalizing constants used in the loss function or the cordillera or similar.

When no *a priori* weight is known, we suggest taking the loss function value as it is ($v_1 = 1$) and either fixing the scale such that coploss $= 0$ for the scaling result with no transformations

($\theta = \theta_0$; this will be the default setup used subsequently), i.e.,

$$v_1^0 = 1, \quad v_2^0 = \frac{\sigma_{MDS}\left(X(\theta_0), \theta_0\right)}{OC\left(X(\theta_0); \epsilon, k, q\right)}, \tag{17}$$

with $\theta_0 = (1,1)^\top$ in case of loss functions derived from (11)-(14). This has the effect that an increase of 1 in the MDS loss measure can be compensated by an increase of v_1^0/v_2^0 in c-clusteredness. Selecting $v_1 = 1, v_2 = v_2^0$ this way is in line with the idea of pushing the configurations towards a more clustered appearance relative to the initial solution.

Another possibiliy is to choose them in such a way that coploss = 0 in the optimum value, i.e., choosing v_1^{opt}, v_2^{opt} so that

$$v_1^{opt} \cdot \sigma_{MDS}\left(X(\theta^*), \theta^*\right) - v_2^{opt} \cdot OC\left(X(\theta^*); \epsilon, k, q\right) = 0 \tag{18}$$

with $\theta^* := \arg\min_\theta \text{coploss}(\theta)$. This is in line with having $\text{coploss}(\theta) > 0$ for $\theta \neq \theta^*$ and allows to optimize over v_1, v_2 too.

One can also to scale the structure index to the range the stress value can take. Note that the c-clusteredness part of the optimization problem is independent of the original stress function, so there is a clear distinction between fit and c-clusteredness. If $v_1 = 0$ optimizing coploss is equivalent to optimizing the c-clusteredness as a function of θ. This can lead to degenerate MDS solutions. COPS has no mechanism in place to avoid such solutions other then setting different v_1, v_2. If $v_2 = 0$ coploss reduces to a loss function without a c-clusteredness penalization. In that case, only transforming the proximities will lead to solving metric MDS for power transformations (albeit inefficiently). This is the subtle difference we referred to earlier: We only solve the metric MDS problem if $v_2 = 0$, in every other case we look for a θ that is not necessarily optimal in metric MDS. Also note that there is a certain redundancy to v_1, v_2 and C in (9). Say we set C so that $OC \mapsto [0,1]$ and we use explicitly normalized stress, then we could simply set v_1, v_2 to 1. To equal effect, one may use raw stress, set $C = 1$ and set $v_1 = 1, v_2 = \sum \delta_{ij}^*$.

5.2. Optimization

The optimization problem in COPS is then to find

$$\arg\min_\theta \text{coploss}(\theta) \tag{19}$$

by doing

$$v_1 \cdot \sigma_{MDS}\left(X(\theta), \theta\right) - v_2 \cdot \text{OC}\left(X(\theta); \epsilon, k, q\right) \to \min_\theta! \tag{20}$$

For a given θ if v_2 is zero than the result of (19) is the same as solving the respective original MDS problem. Letting θ be variable, $v_2 = 0$ will minimize the loss over configurations obtained from using different θ (metric MDS with a power model for the dhats or the fitted distances).

We illustrate COPS for the banking crisis data set. We do it once for transforming the proximities—so $\delta_{ij}^* = \delta_{ij}^\lambda$ (Figure 5)—once for transforming the fitted distances utilizing r-stress—so, $d_{ij}^*(X) = d_{ij}(X)^\kappa$ (Figure 6)— and once for the combination of both, so $d_{ij}^*(X) = d_{ij}(X)^\kappa$ and $\delta_{ij}^* = \delta_{ij}^\lambda$ (powerStress, Figure 7). We use $k = 2, \epsilon = 10, q = 1$. We set

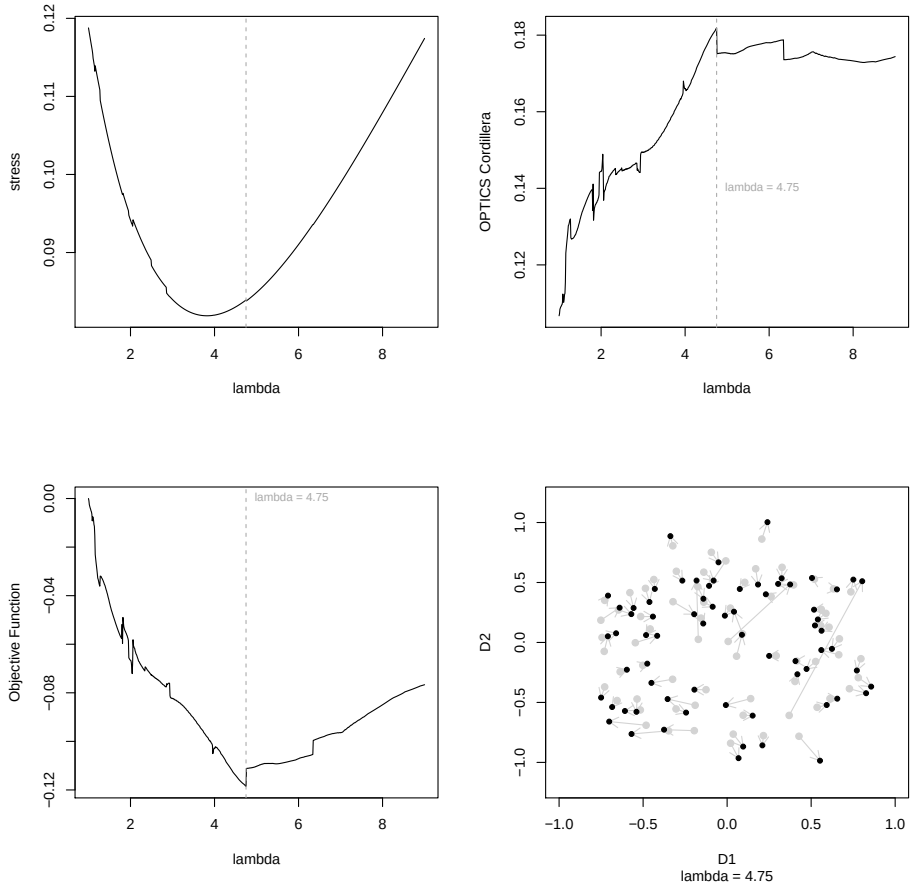

Figure 5: An illustration of a grid search for a λ transformation of the proximities on the banking crises data with explicitly Kruskal's stress and fitted with SMACOF. The optimal parameter is labeled. The topleft panel shows the stress value as a function of λ. The top right panel shows the OPTICS cordillera as a function of λ (parameters were $\epsilon = 10, q = 1, d_{max} = 0.7, v_1 = 1, v_2 = 1.11$ and $k = 2$). The bottom left panel shows the target function, the cluster optimized stress value as a function of λ and the bottom right panel shows the configuration obtained for the λ that had the minimal cluster optimized loss value in the grid search ($\lambda = 4.75$) as well as the change in configurations (procrustes adjusted) compared to using the original proximities (lightgrey labels and arrows).

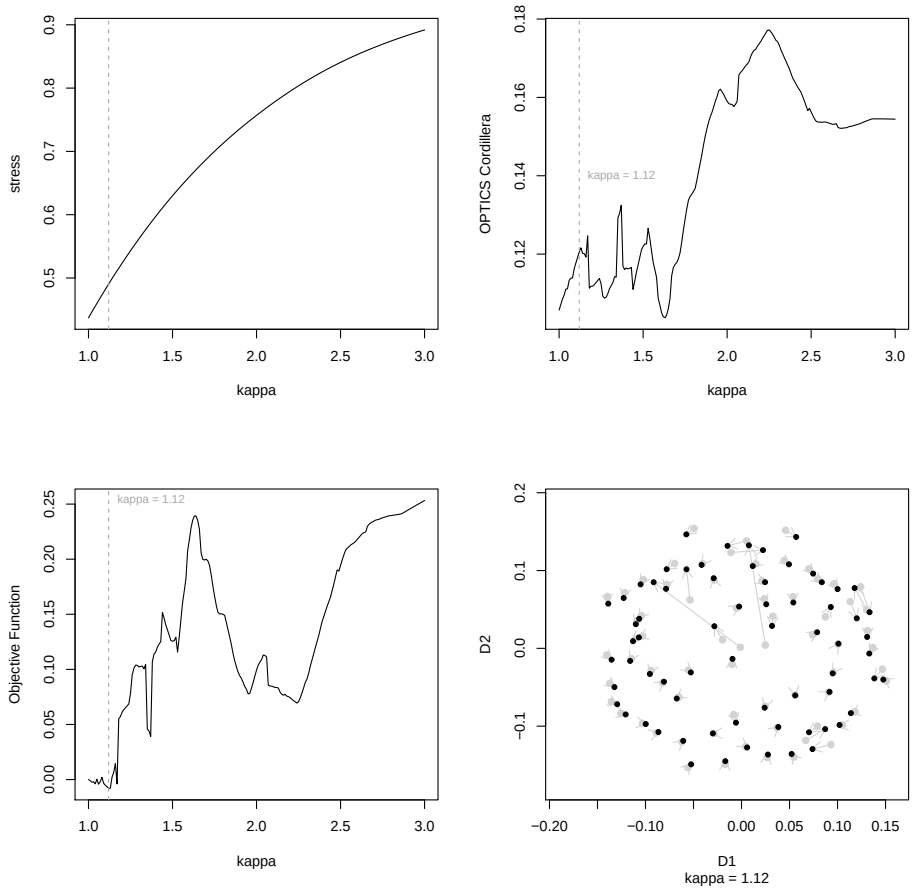

Figure 6: An illustration of a grid search for a κ transformation of the fitted distances (rStress) on the banking crises data. The optimal parameter is labeled. The topleft panel shows the explicitly normalized stress value as a function of κ. The top right panel shows the OPTICS cordillera as a function of κ (parameters were $\epsilon = 10, d_{max} = 0.7, q = 1, v_1 = 1, v_2 = 4.14$ and $k = 2$). The bottom left panel shows the target function, the cluster optimized stress value as a function of κ and the bottom right panel shows the configuration obtained for the κ that had the minimal cluster optimized stress value in the grid search ($\kappa = 1.12$) as well as the change in configurations (procrustes adjusted) compared to using the original proximities (lightgrey labels and arrows).

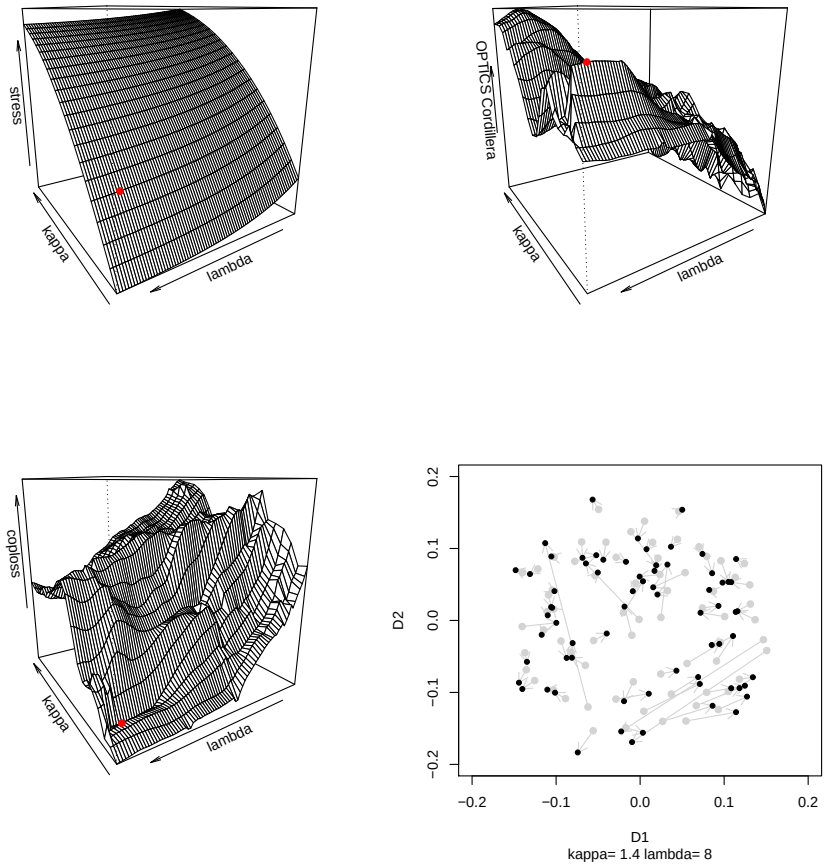

Figure 7: An illustration of a grid search for optimizing powerStress over $\theta = (\kappa, \lambda)^\top$ on the banking crises data. The optimal parameter combination is displayed as a red dot. The topleft panel shows the explicitly normalized stress value as a function of θ. The top right panel shows the OPTICS cordillera as a function of θ (parameters were $\epsilon = 10, q = 1, d_{max} = 0.7, v_1 = 1, v_2 = 4.14$ and $k = 2$). The bottom left panel shows the target function, the cluster optimized stress value as a function of θ and the bottom right panel shows the configuration obtained for the θ that had the minimal cluster optimized stress value in the grid search ($\kappa = 1.4, \lambda = 8$) as well as the change in configurations (procrustes adjusted) compared to using the original proximities (lightgrey labels and arrows).

$v_1 = 1, v_2 = v_2^{(0)}$.Figures 5, 6, and 7 shows line/grid searches for near-optimal $\theta = \lambda$, $\theta = \kappa$ and $\theta = (\kappa, \lambda)^\top$ respectively.

We see that the not very clustered original result becomes more clustered when minimizing coploss in all three cases, for transformed proximities (Figure 5), transformed fitted distances (Figure 6) or both (Figure 7). The grouping structure is quite clear. This grid search returns a $\lambda = 4.75$ as near-optimal for the transformed proximities and a $\kappa = 1.12$ for the coploss with rStress and $\theta = (1.4, 8)^\top$ for the coploss with powerStress.

In the plots in Figures 5-7 we also see a property of coploss, namely that it is not necessarily a convex nor even smooth function of the parameters in θ. The cordillera is based on an ordering, so will in general only be concave in θ if the ordering does not change and the reachabilities increase as a function of θ. In the case of transformed proximities this eventually happens for large λ (e.g., in Figure 5 this is the case for $\lambda > 9$) but this does not hold generally as the function in Figure 6 indicates. This property has implications for optimizing (16).

Computational Strategies for Minimization of coploss

Due to the nature of the cordillera as a function with possible discontinuities, the optimization problem (19) is difficult. In the most general case (19) is a nonlinear, non-smooth objective function with jumps. For the power transformations that we suggested and given X, this optimization problem is either uni- or bivariate. From a practical point of view, the elements of θ will almost always be bounded from above and below, which leads to a constrained optimization problem.

A nested algorithm We propose using a nested, two-stage algorithm combining optimization for finding a near-optimal X with a metaheuristic for discontinuous, nonlinear, constrained optimization to find good values for θ. Our suggestion for the double minimization problem in (19) involves using a nested algorithm that internally first solves (1) for X given θ, $\arg\min_X \sigma_{MDS}(X, \theta)$ and then optimize (19) over θ, so we actually solve (19) by

$$\left\{ v_1 \cdot \left[\arg\min_X \sigma_{MDS}(X, \theta) \right] - v_2 \cdot \mathrm{OC}\left(\left[\arg\min_X \sigma_{MDS}(X, \theta) \right] \right) \right\} \rightarrow min_\theta! \qquad (21)$$

The outline of an algorithm is thus

1. (Optional) Find an inital θ or set it to $(1, 1)^\top$.
2. Given θ, do $\sigma_{MDS}(X; \theta) \rightarrow \min_X$ to get $\arg\min_X \sigma_{MDS}(X, \theta) := X(\theta)$. This is equivalent to optimizing only the first part of the objective function (21). Note that this is usually the most costly step in optimizing (19).
3. Compute $\mathrm{OC}(X(\theta); \epsilon, k, q)$ and (19) for $X(\theta)$ from Step 2.
4. Use a general purpose metaheuristic to repeat Steps 2 and 3 for different θ to find the θ^* that minimizes (19).

Simulated annealing or population based strategies like genetic algorithms (Goldberg and Holland 1988), particle swarm optimization (Eberhart and Kennedy 1995) or estimation of distribution algorithms (Larrañaga and Lozano 2002), are general purpose metaheuristics

that can in principle be used in Step 3. However, the problem of minimizing coploss has a dimensionality of the outer optimization problem that is typically quite small while the inner minimization (Step 2) can be very costly. Thus the metaheuristic should have a small number number of evaluations of Step 2. Arguably a heuristic that may fail to find a global optimum but needs less evaluations of Step 2 is good enough for most purposes as the procedure aims at pushing the MDS solution towards more c-clusteredness.

We thus developed a variant of the Luus-Jaakola procedure (LJ; Luus and Jaakola 1973) (Algorithm 1) to be used in Step 3 that usually converges in less than 200 iterations to an acceptable solution. Let lower, upper denote upper and lower box constraints, $0 < \text{red} < 1$ be a factor for search space width reduction, accd denote the minimum search space width, acc the absolute tolerance for convergence between successive iterations and maxiter the maximum number of iterations.

Algorithm 1 Adaptive Luus-Jaakola Algorithm (ALJ)

procedure ALJ(θ, lower, upper, accd, acc, maxiter, red)
  $\theta^{(0)} \sim U_t(\text{lower}, \text{upper})$ ▷ θ is $t-$dimensional
  $d \leftarrow \text{upper} - \text{lower}$
  $i \leftarrow 1$
  repeat
    $a^{(i)} \sim U_t(-d, d)$
    $\theta^{(i)} \leftarrow \theta^{(i-1)} + a^{(i)}$
    if $\theta^{(i)} < \text{lower}$ **then** ▷ Violates the lower box contraint
      $\theta^{(i)} \leftarrow \text{lower} + U(0,1) \cdot d$
    end if
    if $\theta^{(i)} > \text{upper}$ **then** ▷ Violates the upper box contraint
      $\theta^{(i)} \leftarrow \text{upper} - U(0,1) \cdot d$
    end if
    if $\text{coploss}(\theta^{(i)}) < \text{coploss}(\theta^{(i-1)})$ **then**
      $\theta^{(opt)} \leftarrow \theta^{(i)}$
    else
      $m \leftarrow \min\left(\left\lfloor \frac{\log(\text{accd}) - \log(\max(\text{upper} - \text{lower})}{\log(\text{red})} \right\rfloor, \text{maxiter}\right)$ ▷ Weight for shrinkage
      $s \leftarrow \text{red} \cdot \frac{m+1-i}{m}$ ▷ Shrinkage factor adaptive in i
      $d \leftarrow d \cdot s$
    end if
    $i \leftarrow i + 1$
  until $(d < \text{accd})$ or $(i > \text{maxiter})$ or $|\text{coploss}(\theta^{(opt)}) - \text{coploss}(\theta^{(i)})| < \text{acc}$
  return $\theta^{(opt)}, \text{coploss}(\theta^{(opt)})$ ▷ The best θ found
end procedure

We apply this algorithm to the banking crises data set using coploss with powerStress (15) (optimized with nested majorization) in the next section.

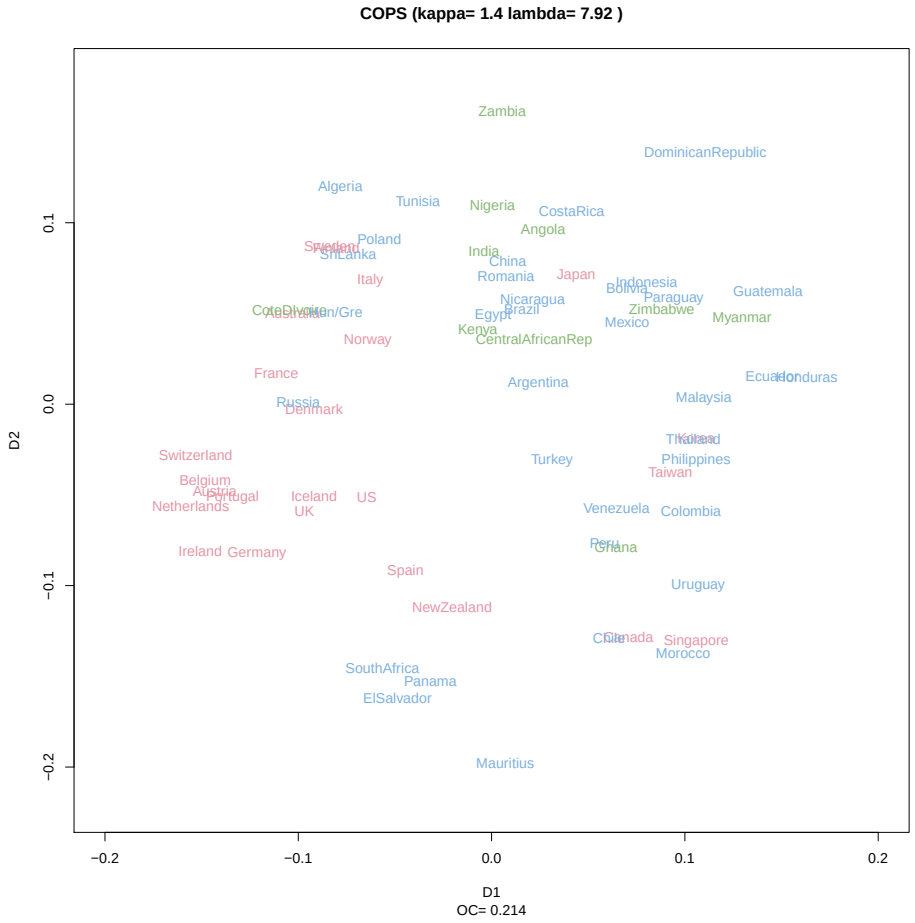

Figure 8: Results of scaling with COPS of the banking crises data (powerStress loss). The values for $\theta = (1.396, 7.919)^\top$ were found by optimization over θ with a random pattern search. The c-clusteredness is OC=0.214. The normalizing constant d_{max} was set to 0.07. High-income countries as defined by Reinhart and Rogoff (2009) are labeled red, middle-income countries green and low-income countries are blue.

6. Examples

6.1. Banking Crises

To illustrate, we apply the COPS procedure to the banking crises data set using coploss with powerStress (15) as the fit measure. We set the lower bound to $\theta = (1, 1)^\top$ and the upper bound to $\theta = (3, 9)^\top$. The resulting configuration can be found in Figure 8. It coincides with the values for θ found in the grid search (here: $\theta = (1.396, 7.919^\top$, there: $\theta = (1.4, 8)^\top$) and took 93 iterations of the outer minimization.

The COPS result leads to a clearly clustered configuration with an OC of 0.21 with axis representing time intervals and clusters representing specific additional shared crises prevalence patterns. The D1 axis represents a continuum of high prevalence of banking crises in the late 2000s (2008-2010) vs. in the late 1990 to early 2000s. Countries with negative values on D1

had crises in the years 2008 to 2010, for increasing values of D1 crises were more prevalent towards the late 1990 early 2000. Among the former is the group of Austria, Switzerland, Belgium, Netherlands, Portugal, Ireland, Germany all of which had their main streak of crises in the late 2000. On the oposite end we find clusters of countries like Guatemala and Myanmar, Ecuador and Honduras, Thailand and Philippines, Korea and Taiwan, all of which had main streaks of crises in the late 1990s to early 2000s. This dimension can also be crudely interpreted as an axis seperating high-income countries from low- to middle incomes countries as about 80% of high income countries (red labels) have a location on D1 of less than -0.01. High-income countries with a positive D1 value are—with the exception of Canada—Asian. D2 has a similar interpretation but for different time periods. It represents roughly the per country percentage of years in banking crises that happened in the 1990s (positive values on D2) or 1980s (values around 0 to negative on D2). Positive values of D2 are found for countries for whom a high percentage of banking crises years fell into the 1990ies, with 24 of them having had a crisis in 1995. One example is Japan, which had a banking crisis in each year from 1992 to 2001 but few crises outside that time period with the 1990s accounting for 50% of all the years in banking crises. For countries with negative values on D2 the peak prevalence of banking crises was not in the 1990s. Most countries had crises in the 1980s, for example the cluster of South Africa, El Salvador and Panama has in common to show a cirises in 1989. Clusters in between these two crude directions are formed by co-occurences of crises at specific timepoints. The United States, for example, fit neatly into the axis description by having had crises in the 1980s (small negative value on D2) but also the late 2000s (small negative value on D1) which places them toward the mid point of the configuration. But it also showed a streak of banking crises in the late 1830s which it has in common with similar streaks in the United Kingdom and Iceland and also in the late 1920s to early 1930s, a pattern which is also (less pronounced) shared by Spain, which explains the positioning close to but in between those other three countries. A comparable similarity to US–Spain is found in the other direction by US–Denmark, which reflects the co-occurence of banking crises in the 1990s between Denmark and the US.

6.2. Natural Hazards in California

We further illustrate the proposed method with an analysis of the similarity of the 58 counties in California with respect to a number of observed and projected indicators for climate change related natural hazards. We compiled a data set of observed and projected data from three sources and aggregated them to the county level. The projected data were derived under two different scenarios (A2, the high emission scenario and B1, the moderate emission scenario Nakićenović and Swart 2000). Overall we had 50 indicators which were:

- County average 95th percentile daily maximum temperature from May 1 to September 30 over the historical period (1971-2000) under the two climate scenarios A2 and B1. These are averaged values for 4 different climate models. The source was Table 7 of Cooley, Moore, Heberger, and Allen (2012).

- Projected average number of days where the daily maximum temperature exceeds the high-heat threshold (see above) over periods 2010-2039, 2040-2069 and 2070-2099. Projections are based on the A2 and B1 scenarios and are averaged for four downscaled climate models. The source was Table 7 of Cooley *et al.* (2012).

- The percentage of a county's census block area vulnerable to unimpeded coastal flooding under baseline conditions (2000) and with a 1.4-meter (55-inch) sea-level rise (projected for 2100). The raw data were obtained from Pacific Institute (2009). From the census block areas we computed an area-weighted percentage for each county.
- The median aggregated Community Climate System Model v. 3 (CCSM3) projected annual actual evapotranspiration for years 2000, 2049 and 2099 under scenarios A2 and B1 by county.
- The median aggregated CCSM3 projected annual baseflow for years 2000, 2049 and 2099 under scenarios A2 and B1 by county.
- The median aggregated Centre National de Recherches Meteorologiques (CNRM) projected annual wildfire risk (observing 1 or more fires in the next 30 years). For years 2020 and 2085 under scenarios A2 and B1 by county.
- The median aggregated CCSM3 projected annual fractional moisture in the entire soil column for years 2000, 2049 and 2099 under scenarios A2 and B1 by county.
- The median aggregated CCSM3 projected annual precipitation for years 2000, 2049 and 2099 under scenarios A2 and B1 by county.

The source of the raw data for the last five items was California Energy Commission (2008).

We use Euclidean distance between the indicators as our dissimilarity measure. For COPS, we used normalized powerStress as the loss function, standardized the columns of X and set $q = 1$, the number of minimum points to 2 (so we aim at at least pairs of counties), ϵ to 10 and d_{max} to 1.2. The COPS configuration can be found in Figure 6.2. The c-clusteredness values were 0.11 for the standard SMACOF configuration and 0.15 for the COPS configuration. We see that the SMACOF configuration is already quite structured, but COPS with powerStress improves on that in terms of adding c-clusteredness to groups of observations and separating the observations more clearly.

The similarity of counties can be inferred. The x and y axis, D1 and D2, correspond roughly to the geography of California with the x-axis distinguishing along the lines of the North-South divide (higher values on x are more south) and the y-axis distinguishing coastal versus inland counties (higher values are more coastal). Accordingly, higher values on D1 roughly represent increasing risk for drought, whereas D2 gives some indication of the risk of flooding. In that space there are some clear groups discernable: In the positive half of the $x-$ and $y-$axis Santa Barbara, Monterey, San Luis Obispo, San Benito, Ventura, Contra Costa, Santa Clara are very similar with respect to the used indicators. These are the ones with a moderate risk profile: a relatively low risk of extreme heat and temperature, low evaporisation and moderate baseflow, average soil moisture but little precipitation and average risk for wildfires in the coming 50 years. They are susceptible to coastal flooding but not extremely so. Similar to these counties and to each other further are Los Angeles, San Diego, Orange County. They tend to have a higher risk profile with less precipitation and lower baseflow and lower soil moisture than the previously discussed counties. This pattern continues with Stanislaus, Sacramento, San Joaquin, Merced and Fresno, which all show increasingly higher projected temperatures, less precipitation and soil moisture. San Francisco is lying in the opposite direction and shows a low risk profile for heat and drought, high precipitation and base flow

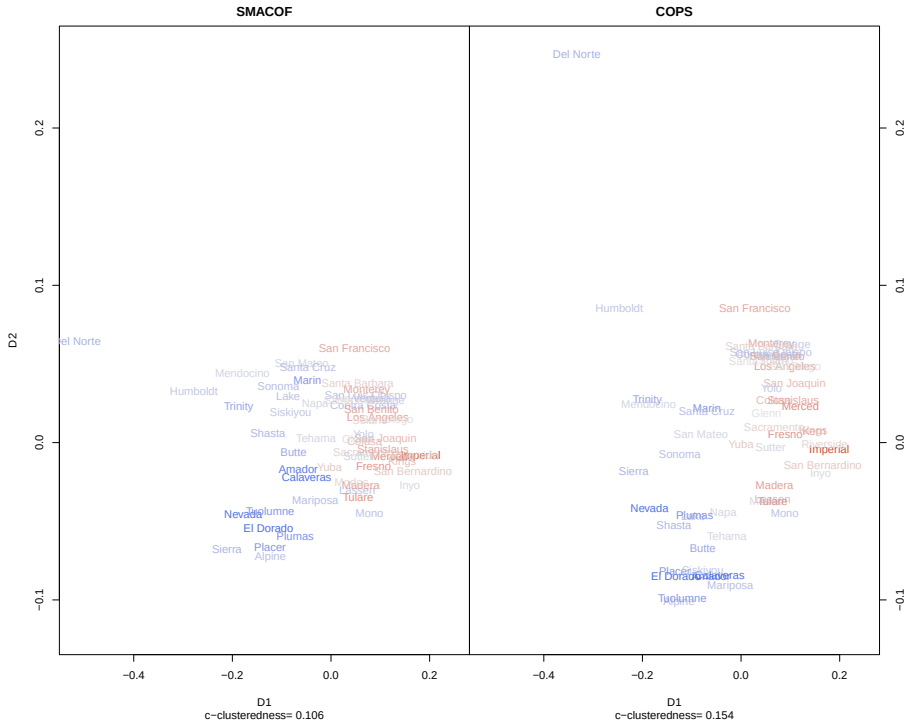

Figure 9: Results of scaling median climate change risk indicators for Californian counties with COPS. The similarities between the counties are based on 50 indicators of climate change such as temperature, precipitation, coastal flooding, wildfire risk derived from downscaled climate model projection for the years 2000-2099, aggregated by the median to county level. Superimposed is the color gradient of the California social vulnerability index (the redder the higher the vulnerability). The latter was not used for scaling. The left plot shows COPS with $\theta = (1,1)^\top$ (a standard SMACOF solution, $\mathrm{OC}(X(1,1);\epsilon,k,q)$=0.11). The right plot shows the configuration from coploss utilizing the powerStress loss function (the optimization over θ led to $\theta = (1.46,0.96)^\top$.The c-clusteredness is $\mathrm{OC}(X(1.46,0.96);\epsilon,k,q)$=0.15. The normalizing constant d_{max} was set to 1.2. The pictures are procrustes adjusted to be comparable.

and soil moisture, but is similar to the Santa Barbara group by having a relatively high susceptibility to coastal flooding. Another group of counties can be identified in the negative half of D2 but positive half of D1 consisting of Kings and Kern, Riverside and Imperial, San Bernadino and Inyo. These are counties with no susceptibility to coastal flooding but low precipitation, high projected temperatures, very low soil moisture and thus high susceptibility to drought. The direction from Los Angeles county towards Imperial can be interpreted as an axis of increasing drought risk. Counties that are similar but less prone to drought are Tulare, Modoc, Lassen, Madera and Mono, who have in common a relatively high number of projected days in extreme heat, no susceptibilty to coastal flooding but otherwise relatively average profile. In the quadrant negative D1 and negative D2, El Dorado, Placer, Alpine, Tuolumne and Siskiyou, Amador, Calaveras, Mariposa which show very similar risk profile. They are mostly counties at high risk of wildfires and many days of extreme heat with at the same time having relatively high projected precipitation and baseflow. The cluster around Amador distinguishes itself from the other by higher projected average temperature and less precipitation. Counties like Sierra, Nevada, Shasta, Plumas fall inline with a high risk for wildfires and many days of extreme heat but show more projected precipitation than the previous group. In the negative D1 but positive D2 quadrant we find Marin, Santa Cruz, San Mateo and Trinity, Mendocino being similar. They display high susceptibility to coastal flooding, relatively high evaporisation, high precipitation, average soil moisture and wildfire risks. Del Norte county is interesting here as it is particularly different from the rest. This is mainly due to it having a much higher projected soil moisture and precipitation than all the other counties.

Additionally, in Figure 6.2 we colored the counties based on the counties average vulnerability index for California (Cooley *et al.* 2012). This index is derived from 19 demographic variables such as age composition, percentages of different ethnic groups, education level, income, employment status, number of births, property and infrastructure variables. A higher index stands for higher vulnerability and makes up the red end of our color palette. The most socially vulnerable counties are located in the South of California and the San Joaquin Valley as well as the large cities. A negative index stands for social resilience and comprises the blue spectrum. These are particularly the counties in the North and the East. The higher the luminance of a color the higher the vulnerability or resilience. When coloring the configuration this way, the picture is striking: The counties with the highest vulnerability or resilience are also the ones that are most similar with respect to the projected risks of climate changes. The first latent dimension D1 separates the vulnerable counties from their resilient counterparts rather well. The counties with the socially most vulnerable population are also the counties that are very much in risk of drought, which make up the bottom right quadrant in Figure 6.2. The axis from Los Angeles towards Imperial is increasingly in danger of drought and features the counties that are on average the most socially vulnerable. In the direction towards more susceptibility to coastal flooding we also find some counties with relatively high social vulnerability like Monterey or San Francisco, but in general the areas that are susceptible to coastal flooding are resilient. This also holds for the counties that have a high risk of wildfires.

In conclusion, this COPS analysis suggests that the greatest challenge that California faces on the county level with respect to climate change is extreme heat, high temperatures, low precipitation, low soil moisture—all are indicators of a high risk of drought. Unfortunately the counties that are at high risk of drought are also the ones most socially vulnerable, which in effect means that it consists of a population that may not be able to deal with the

consequences of the increasing drought.

6.3. Classifying Handwritten Digits

So far we considered the application of cluster optimized proximity scaling in terms of classic scaling, multivariate analysis of similarities and as a tool for displaying multidimensional data for descriptive purposes and visual clustering. In this section we show COPS in another popular use case, namely to reduce the dimensionality of data which is then to be used in further analyisis, say classification or regression. Our approach lends itself well to reduce data in this way as it will try to emphasize similarities in higher dimensional space in the projection, which in turn might lead to clearer classification or regression. As an example we employ COPS with a Sammon stress function and consider a random snapshot of the pendigits data from Alimoglu (1996). Following Izenman (2009), the original data were from 44 writers who handwrote 250 times the digits $0, \ldots, 9$. The digits were written inside a rectangular box with a resolution of 500×500 pixels and the first 10 per writer were ignored for further analysis. This led to 10992 digits. They were recorded in small time intervals by following the trajectory of the pen on the 500×500 grid and then normalized. From the normalized trajectory 8 points (x and y axis position) were randomly selected for each handwritten digit, leading to 16 predictors variables.

We look at a random sample of 500 of these 10992 digits. The lighter points in Figure 10 shows the 2-dimensional Sammon mapping of the Euclidean distances between the 16 input variables, together with the label of the digit. Sammon mapping leads to a rather clear separation of the clusters of digits. Still, overlap between points is common. As a preprocessing step for, e.g., a classification analysis, we want the dimensionality reduction to possibly give us an even stronger separation and to preserve as much of the highdimensional separability as possible. To achieve this we use COPS with Sammon stress and put a high weight on the c-clusteredness part (in this case of 100 times the suggestion of (17)). We used the pattern search version with $q = 1$, $k = 5$ and ϵ set to 10. Each column per configuration was standardized and d_{max} was set to 0.6. The optimal λ was found at 8 and the Sammon stress and the c-clusteredness value were 449.969 and 0.087 (as opposed to 0.151 and 0.048 for the original Sammon mapping). The configurations for the optimal COPS solution is shown in darker shade in Figure 10. We also included arrows to illustrate the change for the individual points. There is much change in the positioning. Most observations of the same kind are pulled tighter together to their respective group. For certain observations the change is very large. We see that in general the clusters of digits are better separated and overlap has been reduced a lot.

The later claim can actually be quantified. Since we have labels, we use the configurations in the projected space to classify the digits based on the sampled trajectory inputs. In case of the ordinary Sammon mapping solution, a classification tree achieves an in-sample classification accuracy of 0.73 (95% CI from 0.69 to 0.77) and a κ of 0.7. Using the configurations of the cluster optimized version of Sammon mapping, we get an accuracy of 0.86 (95% CI from 0.83 to 0.89) and a κ of 0.84. The tree using the 16 input variables directly has an accuracy of 0.89 (95% CI from 0.85 to 0.91) and 0.87. The COPS result leads to higher classification accuracy than the ordinary solution, exceeding the upper bound of the 95% confidence interval. When COPS is applied, the dimension reduction from 16 to 2 dimensions leads to a loss of about 3 percentage points compared to the accuracy obtained without dimension reduction. In standard Sammon mapping 15 percentage points are lost.

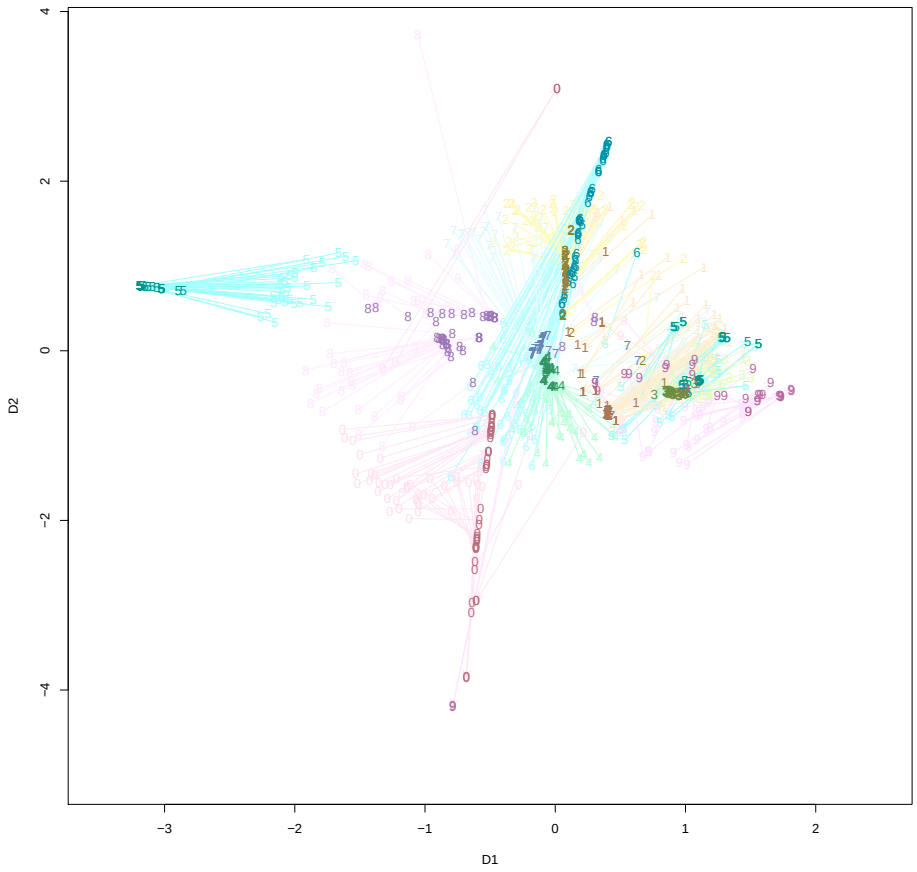

Figure 10: Twodimensional representations of an ordinary (lighter) and cluster optimized (darker) Sammon mapping for a random sample of 500 of the digits data set. The colors and plotting characters highlight the written digit. The arrows between points show the change in configuration (procrustes adjusted) from standard Sammon mapping to COPS with Sammon stress.

7. Conclusions

In this paper we presented a general approach for scaling to increase clusteredness in the configuration (c-clusteredness). The rationale behind this was that while scaling procedures like multidimensional scaling solve for an optimal continuous representation of a proximity matrix in low dimensional space, the reason many data analysts use this technique is to also be able to infer the "real" similarity in form of discrete groups of objects in the high-dimensional space. The latter objective, however, is usually not considered when finding the optimal configuration. To balance these two objectives we introduced COPS, a variant of MDS that selects (power) transformations of proximities or fitted distances based on a c-clusteredness index, that may likely lead to clearer separation or grouping of observations in the target space and thus hopefully reproduce high-dimensional structure more faithfully. This is achieved by balancing stress with a c-clusteredness index derived from OPTICS. We discussed optimization for COPS and illustrated the use with three data sets. The experiments are promising. In all examples we used COPS on, we found that COPS either increases the c-clusteredness present or allows to qualify the statement that untransformed scaling already achieves an acceptable degree of c-clusteredness. The latter comes from the property of the power transformation that untransformed scaling is a special case with $\theta = c(1,1)^\top$. It is therefore entirely possible that the untransformed scaling is chosen as the one with optimal c-clusteredness and fit trade-off. Similarly, the parametrization subsumes an interesting edge case, namely the one where all proximities are equal (in this case 1). Hence COPS (and the c-clusteredness index) can also be used to gauge how close a configuration obtained by untransformed scaling is to this extreme case and allows to push the configuration towards a more clustered appearance. Thus, COPS also addresses the pertinent issue of having low variability in the dissimilarity matrix (Groenen and Borg 2014) and accordingly will have the strongest effect on data with little variablity in the proximities (such as the banking data example) but even for data with already strong c-clusteredness it may be helpful to tease out subtle similarities (as was the case for the Californian county data). When used for dimension reduction COPS can better preserve the high-dimensional cluster structure than untransformed scaling would, as we saw for the pen digits data set.

The biggest drawback we see so far is the possibly high cost for optimizing coploss over θ. This is mainly due to the costly task of finding the optimal configuration for a given parameter configuration. In the best case of our examples, it took 132 iterations for COPS to converge. This translates to carrying out 132 MDS optimizations. This becomes prohibitively costly even for data sizes of more than thousand observations. In case of using rStress or powerStress with majorization it becomes prohibitive for even a fraction of that. Future research could therefore be concerned with speeding up optimization in COPS.

8. Computational Details and Software

Dedicated functions for conducting cluster optimized proximity scaling are available in the R package **stops** (Rusch, de Leeuw, and Mair 2015a). They rely on an implementation of OPTICS which is also available in **stops**. Versions of COPS have been implemented for models with power transformation of the proximities only, by either utilizing a strain loss function (based on `cmdscale`, R Core Team 2014) or a stress type loss functions such as Kruskall's stress with symmetric distance matrices or for projection onto a sphere (based on

the R package **smacof**, de Leeuw and Mair 2009), Sammon stress (based on `sammon` from the R package **MASS**, Venables and Ripley 2002) or Takane et al.'s s-stress. Furthermore there is a COPS version for elastic scaling, for models with power transformations for the fitted distances only (de Leeuw's r-stress) as well as models with power transformations for both the fitted distances and the observed proximities, such as powerstress and elastic scaling as well as Sammon mapping with power transformations. They can all be accessed via the high level function `cops`. The function can also be used simply for fitting any of these MDS models without using coploss (by setting cordweight to zero).

A. Properties of the OPTICS Cordillera

In this section we establish the properties claimed in Section 3.2. Subsequently we assume that ϵ, k, q are given, so we drop them from $OC(X;\epsilon,k,q)$ and only write $OC(X)$. Let us assume we have g configurations, $X^{(g)}, g = 1, 2, \ldots$. Let $R^{(g)}$ denote the OPTICS ordering for the configuration $X^{(g)}$. In what follows it will be convenient to write $s^{(g)} = s^{(g)}(x_i^{(g)}, R^{(g)}) = \text{position}(x_i^{(g)}, R^{(g)})$. If it is clear from the context which configuration we refer to we also drop the superscript $^{(g)}$ from $s^{(g)}$ and only use s. Hence, when we refer to an $x_i^{(g)}$ in $R^{(g)}$ we denote the associated point with $x_{(s)}^{(g)}$, $s, i = 1, \ldots, N$. At each position $s^{(g)}(x_i^{(g)}, R^{(g)})$ we have a minimum reachability of point $x_i^{(g)} = x_{(s)}^{(g)} \in R^{(g)}$ denoted by $r_{(s)}^{*(g)} = r_i^{*(g)}$. Note that we choose to highlight on what level we operate by how the indices are used: if we use a simple subscript like in x_i or r_i^* we are talking about the configuration or the original data fed into OPTICS, the X. If we talk about the result returned from OPTICS, we use the parenthesized subscript, so we talk about $x_{(s)}$ or $r_{(s)}^*$. This choice does not have on influence on the actual values or observations but helps us work through the proofs of the properties as some can be made on the level of the original configuration but for some we need to work on the level of the "reachability plot", which has the OPTICS ordering of the x_i, $x_{(1)}, \ldots, x_{(N)}$ on the abscissa and the corresponding smallest reachabilities $r_{(s)}^*$ on the ordinate. We assume that any minimum reachability $r_{(s)}^{*(g)} \leq d_{max}, \forall s$. For some properties we need additional notation. Inputting the configuration X into the OPTICS procedure leads to a cluster in the configuration corresponding one-to-one to a "valley" in the reachability plot, which we will denote by $V(x_i^{(g)}) = V(x_{(s)}^{(g)})$. $V(x_i^{(g)})$ is the valley to which point $x_i^{(g)}$ belongs. Each valley has at least one "deepest" point, i.e., an $x_i^{(g)}$ for which $r_i^{*(g)} = \min_j r_j^{*(g)}, x_j^{(g)} \in V(x_i^{(g)})$ (so a point with smallest minimum reachability, a "bottom"). In the proofs that follow we need usually only consider a single valley/cluster, so we can skip without loss of generality any reference to what actual valley we look at. We therefore denote the bottom point by $x_b^{(g)}$ which is at position $b = s(x_b^{(g)}, R^{(g)}) = \text{position}(x_b^{(g)}, R^{(g)})$ in the ordering $R^{(g)}$ and so the point in the ordering is denoted by $x_{(b)}$. By this we actually mean the bottom of the valley we currently look at. By a valley we now mean a sequence of points in $R^{(g)}$ that have corresponding minimum reachabilities $r_{(b-t_1)}^{*(g)}, r_{(b+t_2)}^{*(g)}, t_1 = 0, 1, \ldots, T_1;\ t_2 = 0, 1, \ldots, T_2;\ T_1 + T_2 = k - 1$. In a valley it holds that the minimum reachabilities are monotonically nondecreasing the further away the position of $x_i^{(g)}$ is from $x_{(b)}^{(g)}$ in the ordering $R^{(g)}$, so $r_{(b-t_1-1)}^{*(g)} \geq r_{(b-t_1)}^{*(g)} \geq r_{(b)}^{*(g)}$ and $r_{(b+t_2)}^{*(g)} \geq r_{(b+t_2-1)}^{*(g)} \geq r_{(b)}^*, \forall t_1, t_2$. So, $x_{(b)}^{(g)}$ is the bottom of the valley $V(x_b^{(g)}) = V(x_{(b)}^{(g)})$. Each valley is bordered on by two points, $x_l^{(g)}$ and $x_u^{(g)}$, with position in the odering of

$u = s^{(g)}(x_u^{(g)}, R^{(g)}) = \text{position}(x_u^{(g)}, R^{(g)})$ and $l = s^{(g)}(x_l^{(g)}, R^{(g)}) = \text{position}(x_l^{(g)}, R^{(g)})$ so $x_u^{(g)} = x_{(u)}^{(g)} = x_{(b+T_2+1)}^{(g)}$ and $x_l^{(g)} = x_{(l)}^{(g)} = x_{(b-T_1-1)}^{(g)}$ for which the corresponding minimum reachabilities $r_{(l)}^{*(g)}$ and $r_{(u)}^{*(g)}$ are locally maximal over the ordering $R^{(g)}$. They appear as peaks in the OPTICS reachability plot. Each point in $X^{(g)}$ belongs either to a single valley or is a peak.

The poperties in Section 3.2 follow from showing under which conditions the sum of the differences of smallest reachabilities in (9) do not decrease or are increasing. In particular the following properties hold:

"Emphasis Property" This property states that if the distances between the clusters increases, the index is non-decreasing and typically increasing. Thus if we take a cluster in the configuration and shift it away from the other clusters, the index does not become smaller and usually gets larger.

Proposition 1. *Let $X^{(1)}$ be a configuration that produces OPTICS ordering $R^{(1)}$. Let $x_j^{(1)}$ be a row vector in $X^{(1)}$. Let $N_k^{(1)}(x_j^{(1)})$ be the cluster to which $x_j^{(1)}$ belongs. Here k is so that card$(N_k^{(1)}(x_j^{(1)})) = k$. Let us shift all vectors in $N_k^{(1)}(x_j^{(1)})$ by the same direction vector with length $a > 0$ away from all other points in $X^{(1)}$ so that $R^{(1)}$ does not change (if that is geometrically possible) and denote the resulting configuration by $X^{(2)}$. Let $R^{(2)}$ denote the corresponding OPTICS ordering of $X^{(2)}$. Given this, for shifting the cluster in $X^{(2)}$ so that the distances between clusters in $X^{(2)}$ are larger as compared to the distance between the corresponding clusters in $X^{(1)}$ it holds that $OC(X^{(2)} \geq OC(X^{(1)})$. Equality holds only if the shift takes no effect on the minimum reachabilities of the peaks in the valley corresponding to the shifted cluster, or $|r_{(l)}^{*(2)}| + |r_{(u)}^{*(2)}| = |r_{(l)}^{*(1)}| + |r_{(u)}^{*(1)}|$.*

Proof. Given the setup in Proposition 1, $X^{(1)}$ and $X^{(2)}$ are identical apart from the vector positions in cluster $N_k^{(1)}(x_j^{(1)})$ and $N_k^{(2)}(x_j^{(2)})$. The distances between the vectors within $N_k^{(2)}(x_j^{(2)})$ stay constant, so they are the same as in $N_k^{(1)}(x_j^{(1)})$. Since $N_k^{(1)}(x_j^{(1)})$ was shifted away from the other points, $R^{(1)} = R^{(2)}$. From the transformation of $X^{(1)}$ to $X^{(2)}$, the distance between points in non-overlapping k-cluster of the same configuration has not decreased, so for $g = 1, 2$ and $\forall\, x_s^{(g)}, x_t^{(g)} : x_s^{(g)} \in N_k^{(g)}(x_j^{(g)}), x_t^{(g)} \in N_k^{(g)}(x_i^{(g)}), N_k^{(g)}(x_j^{(g)}) \cap N_k^{(g)}(x_i^{(g)}) = \emptyset$ it holds that

$$d(x_s^{(2)}, x_t^{(2)}) \geq d(x_s^{(1)}, x_t^{(1)}), \tag{22}$$

We look only at a single shifted cluster and its corresponding valley. Let $x_b^{(g)}$ be the bottom point in the valley $V(x_b^{(g)})$ that corresponds to the shifted cluster $N_k^{(g)}(x_j^{(g)}) = N_k^{(g)}(x_b^{(g)})$. Let its position in the ordering be at (b) and denote by (l) and (u) the positions of the peaks $x_l^{(g)}$ and $x_u^{(g)}$ that border on $V(x_b^{(g)})$. Since $R^{(1)} = R^{(2)}$ and from the non-decreasing distance in (22) between points in non-overlapping cluster, it follows that the distances between the "peaks" and the "bottom" increase or stay constant when comparing the shifted cluster to its non-shifted counterpart,

$$\begin{aligned} |r_{(l)}^{*(2)} - r_{(b)}^{*(2)}| &\geq |r_{(l)}^{*(1)} - r_{(b)}^{*(1)}|, \\ |r_{(u)}^{*(2)} - r_{(b)}^{*(2)}| &\geq |r_{(u)}^{*(1)} - r_{(b)}^{*(1)}|. \end{aligned}$$

and therefore from the definition of the cordillera as a sum of differences of smallest reachabilities (9), it follows analogue to (31) that $\mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(X^{(1)})$. Strict equality is given only if $|r^{*(2)}_{(l)} - r^{*(2)}_{(b)}| + |r^{*(2)}_{(u)} - r^{*(2)}_{(b)}| = |r^{*(1)}_{(l)} - r^{*(1)}_{(b)}| + |r^{*(1)}_{(u)} - r^{*(1)}_{(b)}|$ or, since $r^{*(1)}_{(b)}$ is constant, $|r^{*(2)}_{(l)}| + |r^{*(2)}_{(u)}| = |r^{*(1)}_{(l)}| + |r^{*(1)}_{(u)}|$. □

"Density Property" If points in the same cluster shrink monotonically towards a central point it will lead to an non-decreasing and typically increasing index. Basically, the denser the clustering in a given cluster is, the higher the index usually becomes.

Proposition 2. *Let $X^{(1)}$ and $X^{(2)}$ be two configurations with the same number of observations that produce OPTICS orderings $R^{(1)} = R^{(2)}$. Let $N^{(1)}_k(x^{(1)}_j)$ and $N^{(2)}_k(x^{(2)}_j)$ be corresponding cluster around a point x_j in both configurations, with respective valleys in the reachability plot of $V(x^{(1)}_j), V(x^{(2)}_j)$. We look at only a single shifted cluster and its corresponding valley. The point $x^{(g)}_b$ is again the point with minimum smallest reachability in the valley and is at position (b), so it is the "bottom" point in the respective valley $V(x^{(g)}_b)$ and thus the point with lowest reachability of any point in the valley, $r^{*(g)}_{(b)} = \min_j r^{*(g)}_j, x^{(g)}_j \in V(x^{(g)}_b)$. Note that $V(x^{(g)}_b) = V(x^{(g)}_{(b)})$. We look at the case where the points in a cluster are shrunk together, which is the same as reducing the minimum reachability for each point in the valley. This reduction must be monotonic in such a way that it does not introduce a new valley. Formally we express this as letting points $x^{(2)}_s \in N^{(2)}_k(x^{(2)}_b), s \neq b$ be moved by positive increments $a_s > 0$ from their position in $X^{(2)}$ towards $x^{(2)}_b$ (if that is geometrically possible) and let these increments be monotonically related to the minimum reachability of $x^{(g)}_s$ and $x^{(g)}_t$, so that $r^{*(g)}_s - a_s \geq r^{*(g)}_t - a_t$ if $r^{(g)}_s \geq r^{(g)}_t$ and so that the ordering in does not change i.e., $R^{(2)} = R^{(1)}$.*
Given this, we have $OC(X^{(2)}) \geq OC(X^{(1)})$. Equality holds only if $|r^{*(2)}_{(b)} - r^{*(1)}_{(b)}| = |(r^{*(2)}_{(l)} + r^{*(2)}_{(u)}) - (r^{*(1)}_{(l)} + r^{*(1)}_{(u)})|$.

Proof. In the setup of Proposition 2, the distances of the points in $N^{(2)}_k(x^{(2)}_b)$ are reduced over these in $N^{(1)}_k(x^{(1)}_b)$ by positive amounts a_s, so

$$d(x^{(2)}_s, x^{(2)}_t) \leq d(x^{(1)}_s, x^{(1)}_t), \tag{23}$$

for $x^{(2)}_s, x^{(2)}_t \in N^{(2)}_k(x^{(2)}_b)$ and $x^{(1)}_s, x^{(1)}_t \in N^{(1)}_k(x^{(1)}_b)$ respectively. From the definition of core distance (6) and reachability distance (7) it follows that for points in this cluster and the corresponding valley in $R^{(1)} = R^{(2)}$, $r^{*(2)}_s \leq r^{*(1)}_s$. Let the indices of points in valley $V(x^{(g)}_b)$ in the ordering be $(b-T_1), (b-T_1+1), \ldots, (b), (b+1), \ldots, (b+T_2-1), (b+T_2)$ with $x^{(g)}_b = x^{(g)}_{(b)}$ and denote by $(b+T_2+1) = (u)$ and $(b-T_1-1) = (l)$ the order in $R^{(g)}$ of an $x^{(g)}_l$ and $x^{(g)}_u$ bordering on the valley (the peaks). Due to the conditions on the increments a_s, the distance between reachabilities of two successive points in the valley remains constant or shrinks, so

$$\begin{aligned} |r^{*(2)}_{(b-t_1)} - r^{*(2)}_{(b-t_1+1)}| &\leq |r^{*(1)}_{(b-t_1)} - r^{*(1)}_{(b-t_1+1)}|, \quad t_1 = 0, \ldots, T_1, \\ |r^{*(2)}_{(b+t_2)} - r^{*(2)}_{(b+t_2-1)}| &\leq |r^{*(1)}_{(b+t_2)} - r^{*(1)}_{(b+t_2-1)}|, \quad t_2 = 0, \ldots, T_2. \end{aligned} \tag{24}$$

To points outside the cluster, however, the distances stay constant or increase, so

$$\begin{aligned} |r^{*(2)}_{(l)} - r^{*(2)}_{(b-T_1)}| &\geq |r^{*(1)}_{(l)} - r^{*(1)}_{(b-T_1)}|, \\ |r^{*(2)}_{(u)} - r^{*(2)}_{(b+T_2)}| &\geq |r^{*(1)}_{(u)} - r^{*(1)}_{(b+T_2)}|. \end{aligned} \tag{25}$$

From the definition of the cordillera (9) as a sum of differences of r^*_js, what in effect counts for the numeric size of the index is the smallest reachability in the valleys and of the bordering peaks as well as their differences. We look only at a single valley, so this is $r^{*(g)}_{(b)}$ for the smallest reachability and the reachabilities of the bordering peaks are $r^{*(g)}_{(u)}$ and $r^{*(g)}_{(l)}$. Utilizing (23-25), for them it holds that

$$\begin{aligned} r^{*(2)}_{(b)} &\leq r^{*(1)}_{(b)}, \\ r^{*(2)}_{(l)} + r^{*(2)}_{(u)} &\geq r^{*(1)}_{(l)} + r^{*(1)}_{(u)}. \end{aligned}$$

and following from (25) and (9), this means $\mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(X^{(1)})$. Only when the difference between the minimum reachabilities of the lowest points in the valley exactly trades off the difference in minimum reachability of the peaks will strict equality hold, or only if $|r^{*(2)}_{(b)} - r^{*(1)}_{(b)}| = |(r^{*(2)}_{(l)} + r^{*(2)}_{(u)}) - (r^{*(1)}_{(l)} + r^{*(1)}_{(u)})|$. □

"Tally Property": For an increase in the number of cluster, the index is non-decreasing and typically increasing. This property tells us that if there are more clusters, the index gets larger.

Proposition 3. *Let $X^{(1)}$ and $X^{(2)}$ be two configurations with the same number of observations that produce OPTICS orderings $R^{(1)} = R^{(2)}$. Let $x^{(1)}_j, x^{(2)}_j$ be corresponding row vectors in $X^{(1)}, X^{(2)}$ respectively. Let $N^{(1)}_k(x^{(1)}_j)$ and $N^{(2)}_k(x^{(2)}_j)$ be corresponding clusters around a point $x^{(g)}_j$ in both configurations, with respective valleys on the reachability plot of $V(x^{(1)}_j), V(x^{(2)}_j)$. Let the number of observations per cluster/valley be k_e. Let us add E new observations to $X^{(1)}$ and $X^{(2)}$, $\widetilde{x}^{(g)}_e, e = 1, \ldots, E$. For $X^{(1)}$ the points are added to existing clusters so that $\widetilde{N}^{(1)}_k(x^{(1)}_j) = N^{(1)}_k(x^{(1)}_j) \cup \widetilde{x}_e$ and the distance of the new points to $x^{(1)}_b$ is not larger than all any other distances of points in $N^{(1)}_k(x^{(1)}_b)$ to $x^{(1)}_b$, i.e., $d(x^{(1)}_b, \widetilde{x}^{(1)}_e) \leq \max d(x^{(1)}_s, x^{(1)}_j), x_s \in N^{(1)}_k(x^{(1)}_b)$. We call the resulting new configuration $\widetilde{X}^{(1)}$, its ordering with $\widetilde{R}^{(1)}$. Let the point $x^{(1)}_b = x^{(1)}_{(b)}$ be the "bottom" point in the valley $V(x^{(1)}_b)$ to which the points were added and thus the point with lowest reachability of any point in the valley, $r^{*(1)}_{(b)} = \min_j r^{*(1)}_j, j : x^{(1)}_j \in V(x^{(1)}_b)$. For the added points, we denote the smallest minimum reachability over all added points $\widetilde{x}^{(1)}_e$ by $\min \widetilde{r}^{*(1)}_e$. For $X^{(2)}$ we add E new observations $\widetilde{x}^{(2)}_e, e = 1, \ldots, E$ so that they form a new cluster around one of the new observations, denoted by $\widetilde{N}^{(2)}_k(\widetilde{x}^{(2)}_b)$. The new cluster adds an extra valley $V(\widetilde{x}^{(2)}_b)$ to the reachability plot. Here, $\widetilde{x}^{(2)}_b$ is the point with minimal reachability $\widetilde{r}^{*(2)}_b$ in that extra valley. The resulting configuration is labeled with $\widetilde{X}^{(2)}$, its OPTICS ordering with $\widetilde{R}^{(2)}$. Given this, we have for an increase in the number of cluster $OC(\widetilde{X}^{(2)}) \geq OC(\widetilde{X}^{(1)})$ if $OC(\widetilde{X}^{(2)}) - OC(X^{(2)}) \geq OC(\widetilde{X}^{(1)}) - OC(X^{(1)})$. Equality holds only if the new cluster is at a distance of zero from points in any other cluster.*

Proof. Because of arguments similar to the ones in (2), namely that per valley only the difference between minimum reachability of the peaks and minimum reachability of the bottom counts, it holds that

$$OC(\widetilde{X}^{(1)}) \geq OC(X^{(1)}), \tag{26}$$

with $OC(\widetilde{X}^{(1)}) > OC(X^{(1)})$ if $\widetilde{r}_e^{*(1)} < r_{(b)}^{*(1)}$, so $\widetilde{r}_e^{*(1)}$ has smallest reachability in the cluster and equality holds otherwise because $\min_s r_s^{*(1)} \leq \widetilde{r}_e^{*(1)} \leq \max_s r_s^{*(1)}, s : x_s \in V(x_{(b)}^{(1)})$. For $\widetilde{X}^{(2)}$, by definition (6) and (7) the reachabilities for points in $\widetilde{X}^{(2)}$ are all ≥ 0, so

$$OC(\widetilde{X}^{(2)}) \geq OC(X^{(2)}). \tag{27}$$

From (26) and (27) we have $OC(\widetilde{X}^{(2)} \geq OC(\widetilde{X}^{(1)})$ if

$$OC(\widetilde{X}^{(2)}) - OC(X^{(2)}) \geq OC(\widetilde{X}^{(1)}) - OC(X^{(1)}). \tag{28}$$

Let the position of the new points in the new valley in $\widetilde{R}^{(2)}$ be $(N+1), \ldots, (N+E)$. Let the index of the bottom point in the new valley be $(N+b), 1 \leq b \leq E$. Utilizing arguments as in (2) then (28) holds if

$$|\widetilde{r}_{(N)}^{*(2)} - \widetilde{r}_{(N+b)}^{*(2)}| + |\widetilde{r}_{(N+b)}^{*(2)} - \widetilde{r}_{(N+E)}^{*(2)}| \geq |r_{(b)}^{*(1)} - \min \widetilde{r}_e^{*(1)}|. \tag{29}$$

This means that the sum of minimum reachability differences in the new valley in $\widetilde{X}^{(2)}$ must be larger than the difference between the two smallest reachabilities in the corresponding valleys in $\widetilde{X}^{(1)}$ and $X^{(1)}$. Since the minimal distances in a cluster will typically be much smaller than distances between clusters, Proposition 3 follows. □

"Balance Property": For a given number of cluster the index is non-increasing in the number of observations $> k$ in a cluster.

Proposition 4. *In what follows the point $x_b^{(g)}$ is the point with minimum smallest reachability in its valley $V(x_b^{(g)})$ and is at position $s^{(g)}(x_b^{(g)}, R^{(g)})$ which we denote in shorthand by (b) and it is the "bottom" point in the respective valley and thus the point with lowest reachability of any point in the valley, $r_{(b)}^{*(g)} = \min_j r_j^{*(g)}, x_j^{(g)} \in V(x_b^{(g)})$. Let $X^{(1)}$ be a configuration with N observations, $x_j^{(1)}$ be row vectors in $X^{(1)}$, let $N_k^{(1)}(x_j^{(1)})$ be a $k-$cluster around $x_j^{(1)}$ that corresponds to a given valley in the reachability plot $V(x_j^{(1)})$. Without loss of generality we look at a single cluster/valley. As outlined above $x_b^{(1)} = x_{(b)}^{(1)}$ is the point with the smallest reachability in the valley, with reachability $r_b^{*(1)} = r_{(b)}^{*(1)}$. Now assume a second configuration $X^{(2)}$ with $N+1$ observations, with $x_j^{(2)}$ being a row vector in $X^{(2)}$. $X^{(2)}$ is exactly like $X^{(1)}$, apart from having an additional data point $x_{N+1}^{(2)}$. Let $N_k^{(2)}(x_j^{(2)})$ be a cluster around $x_j^{(2)}$ in $X^{(2)}$ and $V(x_j^{(2)})$, the valley to which $x_j^{(2)}$ belongs. Again, $x_b^{(2)}$ has smallest minimum reachability in $V(x_b^{(2)})$, denoted by $r_{(b)}^{*(2)}$. Let us further assume x_{N+1} is a point added to the cluster $N_k^{(2)}(x_b^{(2)})$ with valley $V(x_b^{(2)})$ and that $N_k^{(1)}(x_b^{(1)}) = N_k^{(2)}(x_b^{(2)}) \setminus x_{N+1}$. The minimum reachability of x_{N+1} is denoted by r_{N+1}^*. Given this, we have that $OC(X^{(2)}) \leq OC(X^{(1)})$.*

Proof. With the above setup we note that $r_b^{*(2)} \leq r_{N+1}^{*(2)}$, so the point x_b still has the smallest reachability in its valley. Also, $r_{(b)}^{*(1)} = r_{(b)}^{*(2)}$. What in effect counts for the length of the index is the smallest minimum reachability in the valley, and the minimum reachabilities of the bordering peaks $r_{(u)}^{*(g)}, r_{(l)}^{*(g)}$, $g = 1, 2$, and their differences. As $r_{N+1}^{*(2)} \geq r_{(b)}^{*(2)} = r_b^{*(2)}$ it holds that these differences remain constant or shrink

$$|r_{(u)}^{*(2)} - r_{N+1}^{*(2)}| + |r_{(l)}^{*(2)} - r_{N+1}^{*(2)}| \leq |r_{(u)}^{*(2)} - r_{(b)}^{*(2)}| + |r_{(l)}^{*(2)} - r_{(b)}^{*(2)}| = |r_{(u)}^{*(1)} - r_{(b)}^{*(1)}| + |r_{(l)}^{*(1)} - r_{(b)}^{*(1)}|$$

and and so from the definition of the cordillera as a sum of differences of these reachabilities (9) we have $\mathrm{OC}(X^{(2)}) \leq \mathrm{OC}(X^{(1)})$. □

"Spread Property" This property basically says that if we shift points in such a way that the distances to all other points increases sufficiently much, then the index is also increasing. That is when points in the configuration are at some point very spread out that a density based clustering cannot be upheld, the index does no longer become smaller. In a sense this property works against the density property insofar that when points that are far away from each other and no longer appear likely to form a cluster the index treats this no longer as a decrease in density but as an increase in clusteredness. This property makes the index susceptible to outliers if large values of ϵ.

Proposition 5. *Let $s = s^{(g)}(x_j^{(g)}, R^{(g)})$. Let $X^{(1)}$ be a configuration which produces OPTICS ordering $R^{(1)}$. Let the vector $x_j^{(1)}$ be shifted by a positive increment $a > 0$ (relative to the minimum reachabilities of neighbouring points in the ordering points) in a direction away from all other points in $X^{(1)}$ so that $R^{(1)}$ does not change (if it is geometrically possible). Denote the shifted vector by $x_j^{(2)}$. The configuration with the shifted vector is called $X^{(2)}$ and has associated OPTICS ordering $R^{(2)}$. Then, if $x_j^{(1)}$ is a peak and $a > 0$ we have $OC(X^{(1)}) < OC(X^{(2)})$. If $x_j^{(1)}$ is not a peak, then we have $OC(X^{(1)}) < OC(X^{(2)})$ for $a > \max\left(|r_{(s)}^{*(1)} - r_{(s-1)}^{*(1)}|, |r_{(s)}^{*(1)} - r_{(s+1)}^{*(1)}|\right)$.*

Proof. Given the setup in Proposition 5, $X^{(1)}$ and $X^{(2)}$ are identical apart from the j-th row vector. The point $x_j^{(2)}$ was shifted away from the other points so that $R^{(1)} = R^{(2)}$. From the definitions of the core distance (6) and reachability distance (7), it follows that the shifted point $x_j^{(2)}$ has a equal or larger minimum reachability then the corresponding unshifted point $x_j^{(1)}$,

$$r_j^{*(1)} < r_j^{*(2)} \leq r_j^{*(1)} + a. \tag{30}$$

For simplicity let the index of $x_j^{(g)}$ in the ordering be (N). Let us set $r_{(N+1)}^{(*1)}$ to 0 (this point does not exist so its minimum reachability is 0). The shifting did not change the ordering for the points at positions $(1), \ldots, (N)$, so $R^{(1)} = R^{(2)}$. From the definition of the cordillera in (9) and from (30) we can write for different values of $a > 0$—the actual value depending of

the nature of $x^{(1)}_{(N)}$:

$$\begin{aligned} \sum_{s=2}^{N} |r^{*(1)}_{(s)} - r^{*(1)}_{(s-1)}| &= \left(\sum_{s=2}^{N-1} |r^{*(1)}_{(s)} - r^{*(1)}_{(s-1)}| \right) + |r^{*(1)}_{(N)} - r^{*(1)}_{(N-1)}| \\ \leq \sum_{s=2}^{N} |r^{*(2)}_{(s)} - r^{*(2)}_{(s-1)})| &< \left(\sum_{s=2}^{N-1} |r^{*(1)}_{(s)} - r^{*(1)}_{(s-1)}| \right) + |r^{*(1)}_{(N)} + a - r^{*(1)}_{(N-1)}| \end{aligned} \tag{31}$$

and so $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$. The values for a must be so that if $x^{(1)}_{(N)}$ is a peak, then $r^{*(1)}_{(N-1)}, r^{*(1)}_{(N+1)} \leq r^{*(1)}_{(N)}$ and $a > 0$ will suffice for (31) to hold. If $x^{(1)}_{(N)}$ is not a peak, (31) holds for $a \geq \max\left(|r^{*(1)}_{(N)} - r^{*(1)}_{(N-1)}|, |r^{*(1)}_{(N)} - r^{*(1)}_{(N+1)}|\right)$ (this would effectively turn $x^{*(2)}_{(N)}$ into a peak). In both of these cases $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$. □

B. An Upper Bound for the OPTICS Cordillera

In Equation (10) we suggest a normalization constant $C(X, d_{max}, \epsilon, k, q)$ that maps the raw cordillera $OC(X; \epsilon, k, q)$ to the interval $[0, 1]$. It is equivalent to the cordillera in the most clustered case and thus an upper bound. It depends on the number of observations N and the number of points that must make up a cluster, k, and is therefore reasonably tight. For $k = 2$ it is an absolute upper bound for all $k > 1$.

Proposition 6. *If* $d_{max} \geq \max_{ij} d_{ij}$ *then,*

$$OC(X; \epsilon, k, q) \leq C(X, d_{max}, \epsilon, k, q) \tag{32}$$

where

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

or, more compact,

$$C(X, d_{max}, \epsilon, k, q) = d^q_{max} \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right) \tag{33}$$

Proof. This can be shown by employing the definition of perfect structure as in (5). The corresponding cordillera must look like as in the last row of Figure 3 and thus we need to count the maximum possible number, s of cluster of observations with $r^*_i = 0$ as for each of these cluster there must be at most two jumps from and to an observation with $r^*_j > 0$. This must in the most perfectly structured case where $(N-1)/k$ is integer satisfy

$$\begin{aligned} N &\leq s(k-1) + t \\ s &\leq t \leq s + 1 \end{aligned}$$

with t being the number of observations with points with $r^*_j > 0$. Substituting the second equality into the first leads after algebraic manipulation to

$$\frac{N-1}{k} \leq s$$

If OPTICS cannot order the points for these identity to hold exactly, then the above identity is an upper bound. Since s must be integer this means the next closest s fullfilling this is

$$s = \left\lceil \frac{N-1}{k} \right\rceil$$

This means the number of jumps in the cordillera from a group of observations with $r_i^* = 0$ to $r_j^* > 0$ or back is at most

$$2 \left\lceil \frac{N-1}{k} \right\rceil$$

and since the maximum possible length of the jump is d_{max}^q, with perfect structure we have

$$OC(X; \epsilon, k, q) \leq d_{max}^q 2 \left\lceil \frac{N-1}{k} \right\rceil$$

This bound can be improved sligthly for the case where the last group has no last jump anymore by subtracting a single d_{max}^q. Overall this means therefore

$$OC(X; \epsilon, k, q) \leq d_{max}^q \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right) \qquad \square$$

References

Alimoglu F (1996). *Combining Multiple Classifiers For Pen-Based Handwritten Digit Recognition*. Master's thesis, Bogazici University, Istanbul, Turkey.

Ankerst M, Breunig MM, Kriegel HP, Sander J (1999). "OPTICS: Ordering points to identify the clustering structure." In *ACM SIGMOD International Conference on Management of Data*, volume 28, pp. 49–60. ACM Press.

Borg I, Groenen PJ (2005). *Modern multidimensional scaling: Theory and applications*. 2nd edition. Springer, New York.

Buja A, Logan B, Reeds J, Shepp L (1994). "Inequalities and positive-definite functions arising from a problem in multidimensional scaling." *The Annals of Statistics*, pp. 406–438.

Buja A, Swayne DF (2002). "Visualization methodology for multidimensional scaling." *Journal of Classification*, **19**(1), 7–43.

Buja A, Swayne DF, Littman ML, Dean N, Hofmann H, Chen L (2008). "Data visualization with multidimensional scaling." *Journal of Computational and Graphical Statistics*, **17**(2), 444–472.

California Energy Commission (2008). "Raster Downloads." [accessed July 14, 2014], URL http://cal-adapt.org/data/download/.

Chen L, Buja A (2009). "Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis." *Journal of the American Statistical Association*, **104**(485), 209–219.

Chen L, Buja A (2014). "Stress functions for nonlinear dimension reduction, proximity analysis, and graph drawing." *Journal of Machine Learning Research*, **14**, 1145–1173.

Conant CA (1915). *A history of modern banks of issue*. GP Putnam's Sons.

Cooley H, Moore E, Heberger M, Allen L (2012). "Social vulnerability to climate change in California." *Technical Report Publication Number:CEC-500-2012-013*, Pacific Institute, California Energy Commission. [accessed, July 16, 2014], URL http://pacinst.org/wp-content/uploads/sites/21/2014/04/social-vulnerability-climate-change-ca.pdf.

Cox TF, Cox MA (2001). *Multidimensional scaling*. CRC Press, Boca Raton, FL.

de Leeuw J (1977). "Applications of convex analysis to multidimensional scaling." In *Recent Developments in Statistics*, pp. 133–145. North Holland Publishing Company, Amsterdam.

de Leeuw J (2014). "Minimizing r-stress using nested majorization." *Technical report*, UCLA Statistics Preprint Series.

de Leeuw J, Mair P (2009). "Multidimensional Scaling Using Majorization: SMACOF in R." *Journal of Statistical Software*, **31**(3), 1–30. URL http://www.jstatsoft.org/v31/i03/.

de Leeuw J, Stoop I (1984). "Upper bounds for Kruskal's stress." *Psychometrika*, **49**(3), 391–402.

Donoho DL, Grimes C (2003). "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data." *Proceedings of the National Academy of Sciences*, **100**(10), 5591–5596.

Eberhart RC, Kennedy J (1995). "A new optimizer using particle swarm theory." In *Proceedings of the Sixth International Symposium on Micro Machine and Human Science*, volume 1, pp. 39–43. New York.

Goldberg DE, Holland JH (1988). "Genetic algorithms and machine learning." *Machine Learning*, **3**(2), 95–99.

Graves S (2014). "Countries in Banking Crises [data set]." Obtained from the R package Croissant, Y. (2014) Ecdat: Data sets for Econometrics, version 0.2-5, URL http://CRAN.R-project.org/package=Ecdat.

Greenacre M (2011). "A simple permutation test for clusteredness." *Technical report*, University Pompeu Fabra, Barcelona, Spain.

Groenen PJ, Borg I (2014). "Past, Present, and Future of Multidimensional Scaling." In *Visualization and Verbalization of Data*, pp. 95–117. CRC Press, Boca Raton, FL.

Groenen PJ, de Leeuw J, Mathar R (1996). "Least squares multidimensional scaling with transformed distances." In *From Data to Knowledge*, pp. 177–185. Springer, New York.

Izenman AJ (2009). *Modern multivariate statistical techniques: regression, classification, and manifold learning*. Springer, New York.

Kruskal JB (1964). "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis." *Psychometrika*, **29**(1), 1–27.

Kruskal JB, Wish M (1978). *Multidimensional scaling*. Sage, New York.

Larrañaga P, Lozano JA (2002). *Estimation of distribution algorithms: A new tool for evolutionary computation*, volume 2. Springer, New York.

Luus R, Jaakola T (1973). "Optimization by direct search and systematic reduction of the size of search region." *American Institute of Chemical Engineers Journal (AIChE)*, **19**(4), 760–766.

Mair P, Rusch T, Hornik K (2014). "GOP - The party of values?" *Springer Plus*, **3**(697). doi:10.1186/2193-1801-3-697.

McGee VE (1966). "The multidimensional analysis of 'elastic' distances." *British Journal of Mathematical and Statistical Psychology*, **19**(2), 181–196.

Nakićenović N, Swart R (2000). "Special report on emission scenarios." *Intergovernmental Panel on Climate Change*.

Pacific Institute (2009). "Census Blocks, Percent Flooded under Sea Level Rise Scenarios [CSV data file]." [accessed July 9, 2014], URL http://pacinst.org/reports/sea_level_rise/files/Blk_fld.zip.

R Core Team (2014). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org/.

Ramsay JO (1977). "Maximum likelihood estimation in multidimensional scaling." *Psychometrika*, **42**(2), 241–266.

Reinhart C, Rogoff K (2009). *This Time Is Different: Eight Centuries of Financial Folly*. Princeton University Press, New Jersey.

Roweis ST, Saul LK (2000). "Nonlinear dimensionality reduction by locally linear embedding." *Science*, **290**(5500), 2323–2326.

Rusch T, de Leeuw J, Mair P (2015a). *stops: STructure Optimized Proximity Scaling*. R package version 0.0-9, URL http://r-forge.r-project.org/projects/stops/.

Rusch T, Mair P, Hornik K (2015b). "Structuredness Indices and Augmented Nonlinear Dimension Reduction." In preparation.

Sammon JW (1969). "A nonlinear mapping for data structure analysis." *IEEE Transactions on Computers*, **18**(5), 401–409.

Takane Y, Young F, de Leeuw J (1977). "Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features." *Psychometrika*, **42**(1), 7–67.

Tenenbaum JB, De Silva V, Langford JC (2000). "A global geometric framework for nonlinear dimensionality reduction." *Science*, **290**(5500), 2319–2323.

Torgerson WS (1958). *Theory and methods of scaling.* Wiley.

Venables WN, Ripley BD (2002). *Modern Applied Statistics with S.* Fourth edition. Springer, New York. ISBN 0-387-95457-0, URL http://www.stats.ox.ac.uk/pub/MASS4.

Affiliation:

Thomas Rusch
Competence Center for Empirical Research Methods
WU (Wirtschaftsuniversität Wien)
Welthandelsplatz 1, D4
1020 Wien, Austria
E-mail: Thomas.Rusch@wu.ac.at