COPS: Cluster optimized proximity scaling

Rusch, Thomas; Mair, Patrick; Hornik, Kurt

Published: 01/01/2015

Document Version
Publisher's PDF, also known as Version of record

COPS: Cluster Optimized Proximity Scaling

Thomas Rusch
WU (Wirtschaftsuniversität Wien)

Patrick Mair
Harvard University

Kurt Hornik
WU (Wirtschaftsuniversität Wien)

Abstract

Proximity scaling methods (e.g., multidimensional scaling) represent objects in a low dimensional configuration so that fitted distances between objects optimally approximate multivariate proximities. Next to finding the optimal configuration the goal is often also to assess groups of objects from the configuration. This can be difficult if the optimal configuration lacks clusteredness (coined c-clusteredness). We present Cluster Optimized Proximity Scaling (COPS), which attempts to solve this problem by finding a configuration that exhibts c-clusteredness. In COPS, a flexible scaling loss function (p-stress) is combined with an index that quantifies c-clusteredness in the solution, the OPTICS Cordillera. We present two variants of combining p-stress and Cordillera, one for finding the configuration directly and one for metaparameter selection for p-stress. The first variant is illustrated by scaling Californian counties with respect to climate change related natural hazards. We identify groups of counties with similar risk profiles and find that counties that are in high risk of drought are socially vulnerable. The second variant is illustrated by finding a clustered nonlinear representation of countries according to their history of banking crises from 1800 to 2010.

Keywords: multidimensional scaling, nonlinear dimension reduction, clusteredness, data visualization, exploratory data analysis.

1. Introduction

Proximity scaling (PS) describes a family of data analysis techniques which are based on multidimensional scaling (MDS; Torgerson 1958), used to represent high-dimensional proximities of N objects in a space of dimensionality $M, M < N$. This representation, the configuration, is found so that fitted distances in the configuration optimally approximate the proximities. Overviews of different types can be found in Kruskal and Wish (1978); Cox and Cox (2001); Borg and Groenen (2005). We distinguish between "pure" MDS utilizing only proximities and fitted distances in the loss function and the wider group of proximity scaling procedures that augment the pure MDS loss function in some way.

The ultimate goal in PS is often not only to find the optimal configuration but also to make statements about discrete structures (clusters) of objects based on the relative fitted distances in the target space. This is often done visually ex post, so a pre-requisite is a visual appearance of the configuration as clustered ("clusteredness"). The concept of clusteredness is used inconsistently in the literature in different contexts, so we coin the term "c-clusteredness" specifically for the degree of clusteredness of a configuration. By adopting the definition of clusteredness from Rusch, Hornik, and Mair (2016) we can informally describe c-clusteredness

as a property of the appearance of configurations where, starting from a result with no discernable c-clusteredness, c-clusteredness increases if in the configuration (a) a (specified) minimum number of represented objects cluster close to each other, (b) the represented objects cluster increasingly closer together, (c) the distances between the clusters increases (d) the number of clusters increases and (e) the observations and clusters are more spread out.

This paper and the proposed method are motivated by the observation that the two goals of optimal representation and finding discrete structures in PS are somewhat at odds as the optimal representation in the continuous target space does not necessarily consider producing appreciable discrete structures. In other words the resulting optimal configuration may show little discernable c-clusteredness. A prime example is the case where there is little variability in the proximities, for which a standard MDS solution will result in a configuration where the represented objects lie with low density and almost equidistantly on or close to one or more (concentric) $(M-1)-$spheres in $\mathbb{R}^M$ (De Leeuw and Stoop 1984; Buja, Logan, Reeds, and Shepp 1994; Buja and Swayne 2002).

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 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); 1 if so and 0 if otherwise. Greece and Hungary show an identical time series. We explore the similarities of countries based on these data of banking crises history with the Jaccard distance measure which measures how rarely banking crises occur in two countries in the same year with a distance of 1 being maximally different (two countries share no banking crisis in any given year). A standard SMACOF MDS leads to a rather concentric scaling of countries with little c-clusteredness (see left panel of Figure 1) due to little variability in the proximities.

Two approaches can and have been taken to alleviate such a problem of little discernable c-clusteredness. The first is to use a "strong transformation" on the original proximities and/or to fit a nonlinear transformation of the distances (Borg and Groenen 2005) possibly parametrized by a parameter vector θ. Different values of θ typically change the c-clusteredness to the overall solution. One example is the POST-MDS solution in the right panel of Figure 1 obtained after applying our subsequent suggestions. The result indeed appears more clustered but notice that the fit got worse (0.362 vs. 0.344).

The second approach (e.g., Heiser and Groenen 1997; Kiers, Vicari, and Vichi 2005; D'Enza,

[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.

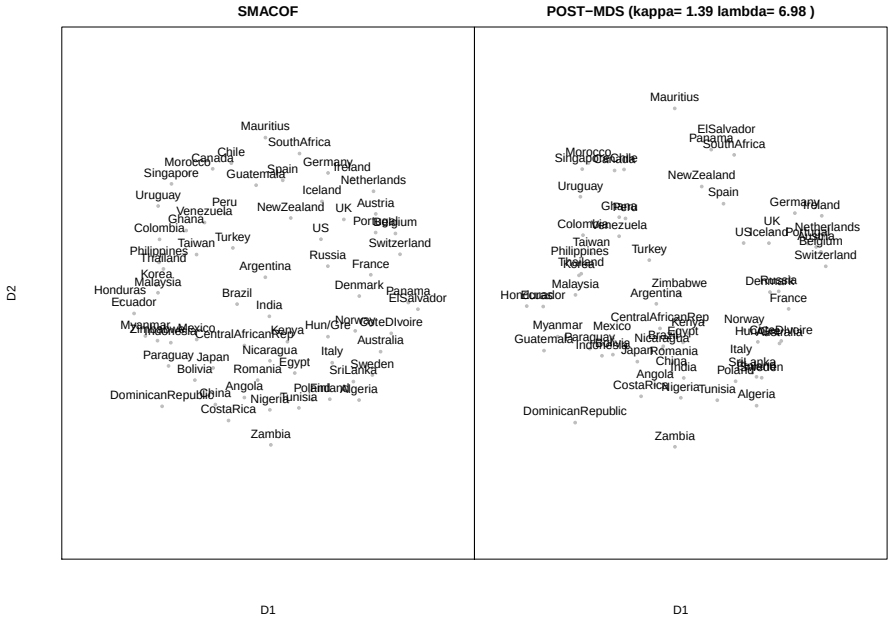

Figure 1: A SMACOF MDS solution (left panel) and a POST-MDS from COPS Variant 2 (right panel) for the banking crises data set from Reinhart and Rogoff (2009). The data set consists of binary entries of whether a banking crises was observed in a given year from 1800 to 2010 for 70 countries.

Van de Velden, and Palumbo 2014) is to augment the MDS loss function with a criterion that captures how clustered a result is and find the optimal configuration from the combined loss. This explicitly and simultaneously considers the two objectives of finding an optimal configuration and producing discrete structures.

In this paper we combine both approaches. We suggest an extension of MDS that uses a flexible stress function utilizing parametric transformations on proximities, fitted distances and weights as in the first approach and augment it with a nonparametric criterion for assessing c-clusteredness, the OPTICS Cordillera, which allows to represent the global c-clusteredness on a unidimensional scale. This version of PS we coin COPS (for *C*luster *O*ptimized *P*roximity *S*caling). It allows to find a continuous representation that has a clustered appearance. Depending on how the COPS loss function is used we will distinguish two variants: The first is to use the augmented loss function to find the optimal configuration for given parameters. The second is to provide a way to trade-off fit and c-clusteredness to select parameters for the transformations.

The article is organized as follows: It starts with a description of proximity scaling and elaboration on the idea of using strong transformations (particularly power transformations) in Section 2. In Section 3 the notion of c-clusteredness is discussed and the OPTICS Cordillera, an index that captures c-clusteredness, gets introduced. These ideas will be combined into COPS in Section 4. Two variants of COPS will be presented, one for finding either the optimal configuration or transformation parameters. Subsequently the use of both variants of COPS wil be illustrated on real data sets in Section 5. Concluding remarks can be found in Section 6. Appendix A describes the supplementary material.

2. Proximity Scaling

Let Δ be an $N \times N$ matrix of observed nonnegative proximities between objects i, j with elements δ_{ij}. We assume values closer to 0 stand for closer proximity, as in a dissimilarity measure. The main diagonal of Δ is 0. For proximity scaling we use a matrix $\Delta^* = f(\Delta)$ with elements δ^*_{ij}. Δ^* is symmetric. We call $f : \delta_{ij} \mapsto \delta^*_{ij}$ a proximity transformation function. This function can be parametrized. The problem that proximity scaling solves is to locate an $N \times M$ matrix X (the configuration) with row vectors (object representations or points) $x_i, i = 1, \ldots, N$ in low-dimensional space $\mathbb{R}^M$ ($M < N$) in such a way that transformations $d_{ij}(X)^* = g(d_{ij}(X))$ of the fitted distances $d_{ij}(X) = d(x_i, x_j)$ approximate the δ^*_{ij} as closely as possible, optionally subject to some other conditions. We may abbreviate $d_{ij}(X)$ with d_{ij} if X is given. We call $g : d_{ij}(X) \mapsto d^*_{ij}(X)$ a distance transformation function. This function can again be parametrized. In other words, proximity scaling means finding X so that $d^*_{ij}(X) = g(d_{ij}(X)) \approx \delta^*_{ij} = f(\delta_{ij})$.

Such X yielding an approximation $D^*(X)$ with elements $d^*_{ij}(X)$ to the matrix Δ^* can be found by optimizing (augmented) criterion functions $\sigma_{PS}(X) = L(X|\Delta^*, \Gamma)$, which provide an aggregate measure of how closely $D^*(X)$ approximates Δ^* and (optionally) a suitable "structural quality level" $\Gamma(X)$ of X (e.g., a measure of clusteredness).

In standard MDS the stress loss function (Kruskal 1964) is often used. Here the loss is quadratic and is a special case of $\sigma_{PS}(X)$ without making use of a $\Gamma(X)$

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

The distance fitted in the configuration is usually some type of Minkowski distance ($p > 0$)

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

typically the Euclidean norm, so $p = 2$. The w_{ij} are finite weights, with $w_{ij} = 0$ if the entry is missing and $w_{ij} \neq 0$ otherwise.

The $w_{ij}, g(\cdot)\text{and}f(\cdot)$ in $\sigma_{MDS}(X)$ enable one to express a rich class of popular stresses: In standard stress $g(\cdot)$ and $f(\cdot)$ are the identity function $I(\cdot)$. Setting $w_{ij} = (\sum_{ij} d^{*2}_{ij}(X))^{-1}$ leads to stress-1 (Kruskal 1964), $w_{ij} = (\sum_{ij} \delta^{*2}_{ij})^{-1}$ to explicitly normalized stress, $w_{ij} = \delta^{-1}_{ij}$ to Sammon stress (Sammon 1969), $w_{ij} = \delta^{-2}_{ij}$ to elastic scaling (McGee 1966). Specific choices for $f(\cdot)$ and $g(\cdot)$ in (1) further lead to s-stress (Takane, Young, and De Leeuw 1977) with $\delta^*_{ij} = \delta^2_{ij}$ and $d^*_{ij}(X) = d^2_{ij}(X)$, MULTISCALE stress (Ramsay 1977) with $\delta^*_{ij} = \log(\delta_{ij})$ and $d^*_{ij}(X) = \log(d_{ij}(X))$, generalized stress (Groenen, De Leeuw, and Mathar 1996) with $\delta^*_{ij} = f(\delta^2_{ij})$ and $d^*_{ij} = f(d^2_{ij})$ or r-stress with $\delta^*_{ij} = \delta_{ij}$ and $d^*_{ij} = d^{2r}_{ij}$ (De Leeuw 2014).

The loss function is then typically minimized to find the vectors $x_1, \ldots, x_N$, i.e.,

$$\arg\min_X \ \sigma_{PS}(X). \tag{3}$$

This can be achieved with e.g., majorization (De Leeuw 1977), gradient descent algorithms (Buja and Swayne 2002) or other techniques.

2.1. Inducing C-Clusteredness By Transformations

When faced with solutions with little c-clusteredness, one can apply transformations to the proximities and/or the fitted distances as a remedy (Borg and Groenen 2005). We define such a transformation as any monotonic transformation function with parameter vector θ. For proximity transformation functions this 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). In this situations the loss function then depends also on θ, so $\sigma_{PS}(X) = \sigma_{PS}(X|\theta)$.

Such transformations have been considered by, e.g., Ramsay (1977); Takane *et al.* (1977); Groenen *et al.* (1996); Buja and Swayne (2002); Borg and Groenen (2005); Buja, Swayne, Littman, Dean, Hofmann, and Chen (2008); Chen and Buja (2009, 2013); Mair, Rusch, and Hornik (2014); De Leeuw (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.

One class of transformations that meets these criteria is power transformations applied to the proximity transformation function and the distance transformation function and the weights simultaneously. This leads to a stress essentially already introduced by Buja *et al.* (2008) where θ is a three-dimensional parameter vector, $\theta = (\kappa, \lambda, \nu)^\top$ with $\lambda, \nu \in \mathbb{R}, \kappa \in \mathbb{R}_+$ and the transformations are

$$g(d_{ij}(X, \theta)) = d^*_{ij}(X) = d_{ij}(X)^\kappa \tag{4}$$

$$f(\delta_{ij}, \theta) = \delta^*_{ij}(\theta) = \delta^\lambda_{ij} \tag{5}$$

$$w_{ij}(\nu) = w^\nu_{ij} \tag{6}$$

This stress type loss measure we call *p-stress* (for power stress)

$$\sigma_{PS}(X) = \text{p-stress}(X|\theta) = \sum_{i<j} w^\nu_{ij} \left[d_{ij}(X)^\kappa - \delta^\lambda_{ij} \right]^2, \tag{7}$$

p-stress encompasses many popular stress functions including Kruskal's stress ($\kappa = \lambda = \nu = 1$), Sammon stress ($w_{ij} = \delta_{ij}$,$\kappa = \lambda = 1, \nu = -1$), elastic scaling stress ($w_{ij} = \delta_{ij}$,$\kappa = \lambda = 1, \nu = -2$), s-stress (or ALSCAL stress) ($\kappa = \lambda = 2, \nu = 1$), r-stress ($\kappa = 2r, \lambda = 1, \nu = 1$) and MULTISCALE stress ($\kappa \to 0, \lambda = 1, \nu = 1$) and can also be used to yield any combination of those. Minimizing (7) for given θ can be achieved, e.g., with nested majorization as in De Leeuw (2014). We call the resulting MDS variant using this loss POST-MDS (for Power Stress MDS). Gradients and mathematical properties for p-stress can be deduced from Groenen *et al.* (1996).

Applying a nonlinear, parametrized transformation to fitted distances or proximities to increase c-clusteredness can be done in other ways, see for example the suggestion of Chen and Buja (2013) who use Box-Cox type transformations on fitted distances and observed proximities.

3. Quantifying C-Clusteredness

Proximity scaling procedures often serve two different purposes. The first (which we covered so far) is to project objects from a multidimensional space into a continuous lower-dimensional target space based on some measure of proximity. The second (often implicit) goal is to use the scaling result to infer the existence of discrete accumulations ("clusters") of observations from the scaling result. This is often done by investigating and visually judging the existence of clusters in the target space. The latter only works when the projection is so that the configuration shows a number of clearly appreciable distinct structures of accumulations of points, in other words if it looks "well clustered". This appearance of clusteredness of the configuration is what we refer to as *c-clusteredness*. In this section we discuss the notion of clusteredness as defined by Rusch *et al.* (2016) in the context of a configuration and present an index that captures the c-clusteredness of the result. Eventually we augment the MDS from the previous section with the c-clusteredness index for a proximity scaling method that considers both goals simultaneously.

3.1. C-Clusteredness as a Property of a Configuration

By c-clusteredness we mean the appearance of how clustered the objects in a configuration are. This includes observing whether there are arbitrarily shaped structures into which objects accumulate, whether the objects accumulate in these structures compactly, whether these structures are well separated or not, how many such structures we find, and how spread out the structures or observations are.

C-clusteredness of a proximity scaling result is a property of the configuration and thus solely of the appearance of the pairwise distances between the objects in the configuration. To be that it must be invariant for different partitionings of the objects or the assignment of observations to clusters obtained from one and the same configuration. Therefore c-clusteredness is a concept independent of assumptions and decisions associated with obtaining a clustering such as how many clusters there must be, the assignment of observations to clusters, the choice of centroid, allowance of cluster shapes, distribution of objects in clusters or the chosen clustering method. This invariance property distinguishes c-clusteredness from similar ideas such as the concept of internal cluster validity as used in the Silhouette measure and others (Rousseeuw 1987; Kim and Billard 2011; Wu 2011), the concept of cluster stability (Hennig 2007) or previous combinations of classification/clustering with MDS (Heiser and Groenen 1997; Kiers *et al.* 2005; D'Enza *et al.* 2014).

To make the notion of c-clusteredness concrete we adopt the concept of clusteredness as defined by Rusch *et al.* (2016) and employ it in the context of a fitted configuration from a proximity scaling procedure. By specifying the minimum number of points that must comprise a cluster with k we can define no clusteredness and maximal clusteredness as follows:.

No c-clusteredness is given when the row vectors in configuration X can be represented as the vertices of a matchstick graph, i.e., a graph $G = (X, E)$ where the edges in E solely connect vertices with their nearest neighbours and there exists a planar embedding of X such that every edge is of constant length and no two edges cross (cf. top left panel of Figure 2).

Maximal c-clusteredness is the situation when the following conditions are met: i) There are $\lceil N/k \rceil$ discrete structures, ii) for all k objects x_i in the same cluster their distance to each other is zero and, iii) for all clusters each cluster is the same constant distance away from the

closest neighbouring cluster, so the clusters are evenly distributed (cf. the bottom left plot of Figure 2).

The observed c-clusteredness for a given configuration is its position on the continuum spread between no c-clusteredness and maximal c-clusteredness. It is a monotonic function of the distances between points and clusters and the number of discrete structures in the configuration, specifically if i) the distances between objects of accumulation increase, ii) the objects accumulate more densely, iii) the number of points of accumulation increase and iv) in the configuration the objects spread out more.

The left column of Figure 2 illustrates c-clusteredness with a toy example of 8 labeled data points. The top plot shows no c-clusteredness and the second plot from the top shows little c-clusteredness (a case of low variability in proximities). C-clusteredness increases from the top to the bottom and the bottom plot shows maximal c-clusteredness with $N = 8$ and $k = 2$; a configuration where at each of the four positions there are two points coinciding and all four positions are equally far away from the closest other group.

3.2. A C-Clusteredness Index

For the above definition of c-clusteredness, Rusch *et al.* (2016) suggest an index—the OPTICS Cordillera (or Cordillera)—that allows to quantify the position of a configuration on the continuum by adhering to the principles outlined above. The index is monotonically nondecreasing and typically increasing as a function of more c-clusterednes. We use the normalized version which lies between 0 in case of no clusteredness and 1 in case of maximal c-clusteredness.

The index is derived from the OPTICS algorithm (Ordering Points To Identify The Clustering Structure; Ankerst, Breunig, Kriegel, and Sander 1999), which produces an ordering with "respect to the density-based clustering structure containing the information about *every* clustering level of the data" (p. 51 Ankerst *et al.* 1999) up to a maximum radius ϵ. Let us denote the ordering by R and the position of object x_i in R by $x_{(s)}, s = 1, \ldots, N$. OPTICS augments R with each object's minimum reachability distance $r^*_{(s)}$ which is for point $x_{(s)}$ the maximum of $d_{(s)(s-1)}$ and the distance to the k-th neighbour of x_i.

The ordering R and is so that if the minimum reachability for $x_{(s)}$ is small then $x_{(s)}$ und $x_{(s-1)}$ are close. If it is large then $x_{(s)}$ is far away from $x_{(s-1)}$. Therefore points in the ordering that are subsequent or close and have small minimum reachability likely belong to the same discrete structure whereas points that are far away from each other in the ordering or have some large reachability between them likely belong to different structures.

The OPTICS Cordillera. The OPTICS Cordillera unidimensionally summarizes the information about the clusteredness of the fitted configuration X. It aggregates the pairwise differences of the minimum reachabilities over the OPTICS ordering. The larger the index is, the more c-clusteredness we find in the solution. The index is highly nonparametric as to obtain the index very little weak assumptions need to be made.

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. With c_i we denote the maximum distance to the $k-$th neighbour. Let $d_{max}, \min d_{ij} \leq d_{max} \leq \epsilon$ denote the maximum reference distance between clusters for maximal clusteredness. $r^*_{(s)}$ be the minimum

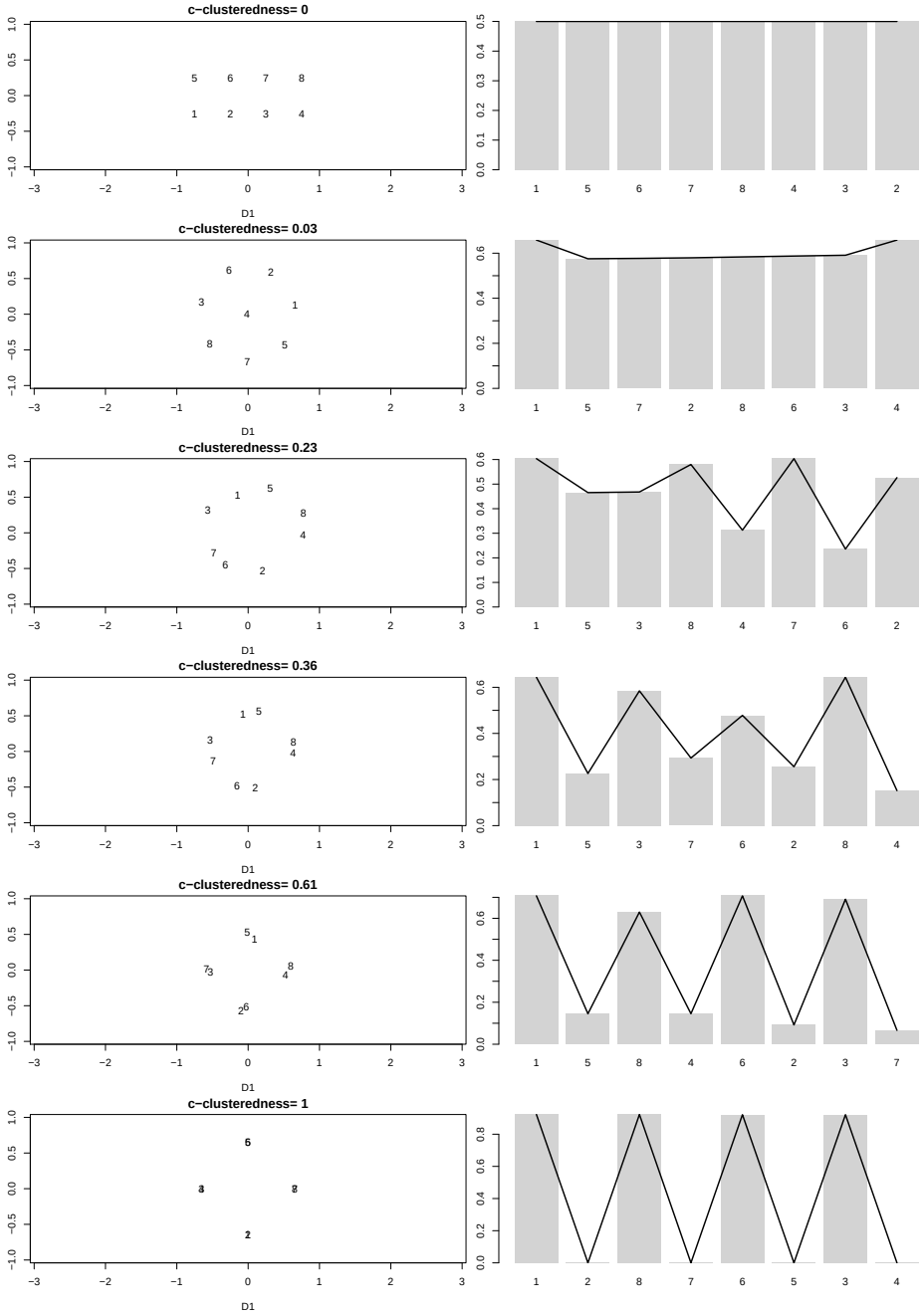

Figure 2: Differently clustered 2D configurations of 8 points and their OPTICS Cordillera. In the left column we find different configurations. The top left plot shows a case of no c-clusteredness, the second plot shows an MDS solution that appears for very little variability in the proximities, the bottom left panel shows maximal c-clusteredness for $N = 8$ and $k = 2$. The other three panels show configurations between these extremes. The c-clusteredness increases from top to bottom. In the right column we find the corresponding OPTICS reachability plots and with the black line an illustration of the OPTICS Cordillera (which is here accurately displayed up to a constant). The plots are labeled with the numeric value for the OPTICS Cordillera. It has been calculated with $k = 2, \epsilon = 2, q = 1$. After Rusch *et al.* (2016).

reachability for $x_i = x_{(s)}$ which is $r^*_{(s)} = \min(\max(c_i, d_{(s)(s-1)}), d_{max})$ or d_{max} if there are less than k neighbours within a radius of ϵ around x_i.

Then the (normalized) OPTICS Cordillera is

$$\mathrm{OC}'_{\epsilon,k,q}(X) = \frac{\left(\sum_{s=2}^{N} |r^*_{(s)} - r^*_{(s-1)}|^q\right)^{1/q}}{d^q_{max}\left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor\right)}, \tag{8}$$

where $q > 0$ is an optional metaparameter. The d_{max} caps the reachability distance and for a series of configurations $X^{(1)}, \ldots, X^{(G)}$ is set to be the same for all G results. It can be chosen so that the index is robust to large outliers or to $d_{max} = \max r_i^{*(G)}$.

Properties of the OPTICS Cordillera. The OPTICS Cordillera fullfills properties that are desirable for a measure of c-clusteredness. We consider these properties to be very important in the exploratory, unsupervised setting where PS is typically used and to the best of our knowledge they are not met in any other measure. We reproduce the properties here:

1. OC$'$ is parsimonious with respect to the mandatory parameters of which there is only one, namely k. The parameters ϵ, d_{max} and q are free but optional to make the index flexible for different data situations.
2. Clusters are defined solely as an accumulation of at least k objects that is more dense then the surrounding area. The geometrical shape of the clusters and distribution of objects within the clusters can thus be completely arbitrary.
3. The OC$'$ is invariant to any specific cluster assignment of observations.
4. The OC$'$ does not need the notion of centroids or prototypes.
5. There is no need to specify any number of clusters *a priori*.
6. Nested clusters at different densities are represented simultaneously.
7. It exhibits the desirable c-clusteredness properties:
   - Emphasis property: If the clusters are separated more clearly, OC$'$ typically increases.
   - Density property: If the clusters are more compact, OC$'$ typically increases.
   - Tally property: Up to the maximal number of possible clusters, for an increase in the number of clusters the OC$'$ typically increases.
   - Balance property: For a given number of clusters and distances within the cluster, the OC$'$ does not decrease as a function of the number of observations $> k$ in the cluster.
   - Spread property: If we shift objects in such a way that the distances to all other points increase sufficiently much, then the OC$'$ increases. This is then interpreted as an increase in c-clusteredness rather than decrease in compactness.

For details and proofs we refer to the propositions in Rusch *et al.* (2016).

4. Cluster Optimized Proximity Scaling

In this section we combine scaling with a loss function that includes c-clusteredness introducing transformations with maximizing the normalized OPTICS Cordillera. This leads to *cluster optimized proximity scaling (COPS)*. We distinguish two variants by how the loss function of COPS is used: First, we follow the tradition of augmenting the dimensionality reduction loss function with an additional criterion and solve the overall dimensionality reduction problem for given metaparameter in a multi-objective optimization similar to, e.g., Heiser and Groenen (1997); Kiers *et al.* (2005); Vichi and Saporta (2009); Timmerman, Ceulemans, Kiers, and Vichi (2010); Rocci, Gattone, and Vichi (2011); Vichi, Rocci, and Kiers (2007); D'Enza and Palumbo (2013); D'Enza *et al.* (2014). Second, following the tradition of, e.g., Akkucuk and Carroll (2006) and Chen and Buja (2009, 2013) we use the OPTICS Cordillera as a stress independent criterion to select metaparameters for the stress loss which is then used for finding the configuration without reference to the metacriterion.

In both variants, we address gaps in the prior approaches. In the first variant, using p-stress as the stress loss and the OPTICS Cordillera as the c-clusteredness measure allows for a flexible scaling with linear and nonlinear projections relying on only weak assumptions about the clustered appearance by utilizing the highly nonparametric nature of OC'. For the second variant, our contribution is that we define the combination of stress function and metacriterion as a nested multi-objective optimization procedure and conduct systematic metaparameter search.

4.1. Cluster Optimized Loss

The objective function at the heart of COPS, which we call *cluster optimized loss (coploss)*, is a weighted combination of the $\theta-$parametrized loss function to measure badness-of-fit, $\text{p-stress}\,(X|\theta)$, and the c-clusteredness measure, $\Gamma(X) = \mathrm{OC}'(X)$. More formally, coploss is then

$$\text{coploss}_{v_1,v_2,\gamma}(X|\theta) = v_1 \cdot \text{p-stress}\,(X|\theta) - v_2 \cdot \mathrm{OC}'_\gamma(X) \tag{9}$$

with $v_1, v_2 \in \mathbb{R}$ controlling how much weight should be given to the stress loss measure and the c-clusteredness respectively and γ being shorthand for the metaparameters controlling the Cordillera. In general v_1, v_2 are either *a priori* determined values or may be used to trade-off fit and c-clusteredness in a way for them to be commensurable. Note that v_1 and v_2 are complementary and having two weights is redundant but we deliberately allow this flexibility in our formulation so it can easily be used to account for different scales on which c-clusteredness and the loss function may lie or to set weights according to some hypothesis (including negative weights). If that is not necessary, removing redundancy in the weighting is possible by, e.g., a convex combination with setting $v_2 = 1 - v_1$ with $0 \leq v_1 \leq 1$.

This combined loss function can then be used in two ways:

Variant 1: Finding a Configuration based on Cluster Optimized Loss

We look at optimizing (9) directly with given metaparameter vector θ

$$\sigma_{PS}(X) = \text{coploss}(X|\theta) = v_1 \cdot \text{p-stress}(X|\theta) - v_2 \cdot \mathrm{OC}'_\gamma(X) \tag{10}$$

with $v_1, v_2 \in \mathbb{R}$ the weights. For this variant, we suggest to use the convex combination $v_2 = 1 - v_1$ with $0 \leq v_1 \leq 1$. For a given θ if $v_2 = 0$ then the result of (10) is the same

as solving the respective p-stress problem. Thus minimizing coploss in this variant pushes the obtained configuration increasingly towards a maximally c-clustered arrangement, the strength of the push is governed by the values of v_1, v_2.

We then need to find

$$X^*_{\text{coploss}}(\theta) = \arg\min_X \text{coploss}(X|\theta) \tag{11}$$

Computational Strategy. The OPTICS Cordillera is based on an ordering, so optimizing (10) difficult. We solve it by first finding an initial good solution by majorization of (10) with $v_2 = 0$ and then to improve upon this solution by the trust-region method NEWUOA (Powell 2006). This can be expected to work well if the initial solution is not too far from the optimum (Rios and Sahinidis 2013), e.g., for relatively more weight on the stress part of coploss (say $v_1/v_2 > 3$).

Variant 2: Metaparameter Optimization based on Cluster Optimized Loss

Let us write $X^*_{\text{p-stress}}(\theta) := \arg\min_X \text{p-stress}(X|\theta)$ for the optimal configuration obtained from minimizing p-stress for a transformation parameter θ. In this variant we first find a configuration by minimizing the stress part only and then use the obtained stress value in combination with OC′ to conduct metaparameter search over θ. We therefore use coploss in a profiling method and may call this COPS profile to distinguish it from Variant 1.

The objective function then becomes a profile version of coploss (p-coploss)

$$\text{p-coploss}(\theta) = v_1 \cdot \text{p-stress}\left(X^*_{\text{p-stress}}(\theta)|\theta\right) - v_2 \cdot \text{OC}'_\gamma\left(X^*_{\text{p-stress}}(\theta)\right) \tag{12}$$

We stress that in this case $\text{p-stress}(X^*_{\text{p-stress}}(\theta)|\theta)$ employed should be scale and unit free, e.g., normalized to lie between 0 and 1.

As default weighting, we suggest taking the fit function value as it is ($v_1 = 1$) and fixing the scale such that p-coploss = 0 for the scaling result with no transformation ($\theta = \theta_0$), i.e.,

$$v_1^0 = 1, \quad v_2^0 = \frac{\text{p-stress}\left(X^*_{\text{p-stress}}(\theta_0)|\theta_0\right)}{\text{OC}'_\gamma\left(X^*_{\text{p-stress}}(\theta_0)\right)}, \tag{13}$$

with $\theta_0 = (1,1,1)^\top$. Thus an increase of 1 in p-stress (i.e., perfect fit to worst fit) 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 selecting a nonlinear projection whose configuration exhibits a more clustered appearance relative to the starting solution.

The optimization problem for metaparameter optimization is then to find

$$\theta^* = \arg\min_\theta \text{p-coploss}(\theta) \tag{14}$$

by finding

$$\text{p-coploss}^* = \text{p-coploss}(\theta^*) \tag{15}$$

Here if $v_2 = 0$ than the result of (15) will minimize the loss over configurations obtained from using different θ.

Computational Strategy. Similar considerations as before apply with respect to the difficulty of optimizing a loss with the OPTICS Cordillera. In the formulation above, however, the problem can be considered as a nested optimization problem where we first solve for $X^*_{\text{p-stress}}(\theta)$

for a given θ based on the stress part only and then repeat this to find an optimal θ^* with p-coploss(θ). This enables us to utilize tailored algorithms for finding the $X^*_{\text{p-stress}}(\theta)$ and using metaheuristic to optimize over θ. An outline for an algorithm is thus:

1. Start with an initial θ.
2. Given θ, do $\arg\min_X$ p-stress$(X|\theta)$ to obtain $X^*_{\text{p-stress}}(\theta)$. Record the stress value of p-stress$(X^*_{\text{p-stress}}(\theta)|\theta)$.
3. Compute $\text{OC}'_\gamma(X^*_{\text{p-stress}}(\theta))$ for $X^*_{\text{p-stress}}(\theta)$ from Step 2 and plug it into (12) together with p-stress$(X^*_{\text{p-stress}}(\theta)|\theta)$.
4. Use a metaheuristic to repeat Steps 2 and 3 for different θ to find the θ^* that minimizes (12).

As metaheuristics 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) can be used. One problem of minimizing coploss in Variant 2 is that the inner minimization (Step 2) can be very costly. Thus the metaheuristic should need as little a number of evaluations of Step 2 as possible, which puts population-based strategies at a disadvantage. Considering that the dimensionality of the outer step is small (at most three) arguably a heuristic that may fail to find a global optimum but needs much less evaluations of Step 2 is good enough for most purposes.

For this we developed a variant of the Luus-Jaakola procedure (LJ; Luus and Jaakola 1973) to be used in Step 3 that usually converges in less than 200 iterations to an acceptable solution. It is displayed in pseudocode as Algorithm 1 where `lower`, `upper` denote upper and lower box constraints, $0 <$ `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.

5. Examples

In this section we illustrate the two variants of COPS. First we use COPS Variant 1 to find a more clustered configuration than the standard MDS solution. we use an example of scaling Californian counties with respect to climate change hazards. The second example uses COPS Variant 2 (COPS profile) for metaparameter selection for the p-stress loss function on the already introduced banking crises data set.

5.1. Climate Change in California—Finding a Configuration with COPS Variant 1

Variant 1 of COPS is used for the analysis of the similarity of the 58 counties in California with respect to a number of indicators for climate change related natural hazards and compared to the results to a standard MDS. The data set is a compilation of observed and projected climate change indicators from three sources, aggregated to the county level. The projected indicators were derived under two different scenarios (A2, the high emission scenario and B1,

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

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 Version 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 one 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 indicators was California Energy Commission (2008).

We normalized all variables to the interval [0, 1] and then used Euclidean distance between the normalized indicators as the proximity measure. In the models we use $\kappa = \lambda = \nu = 1$ and look for clusters of at least three counties ($k = 3$), use absolute differences in the index ($q = 1$) and set ϵ to a sufficiently large value to consider all points as possible neighbours, here $\epsilon = 10$. For robustness d_{max} is set to 1.2, which is 1.5 times the maximum r_i^* for the standard MDS configuration.

We fit a standard MDS (ie.e, COPS with $v_1 = 1, v_2 = 0$) and two COPS models, one with $v_1 = 0.9, v_2 = 0.1$ and the second with $v_1 = 0.8, v_2 = 0.2$. The results are displayed in Figure 3. The left column displays the COPS models with $v_1 = 1, 0.9, 0.8$ and $v_2 = 0, 0.1, 0.2$ respectively. The stress-1 values are 0.124, 0.141, 0.16 and the according Cordillera values are 0.115, 0.34 and 0.357 respectively. The second column in Figure 3 shows Shepard plots of the fitted distances $d_{ij}(X)$ vs. proximities δ_{ij} for the models.

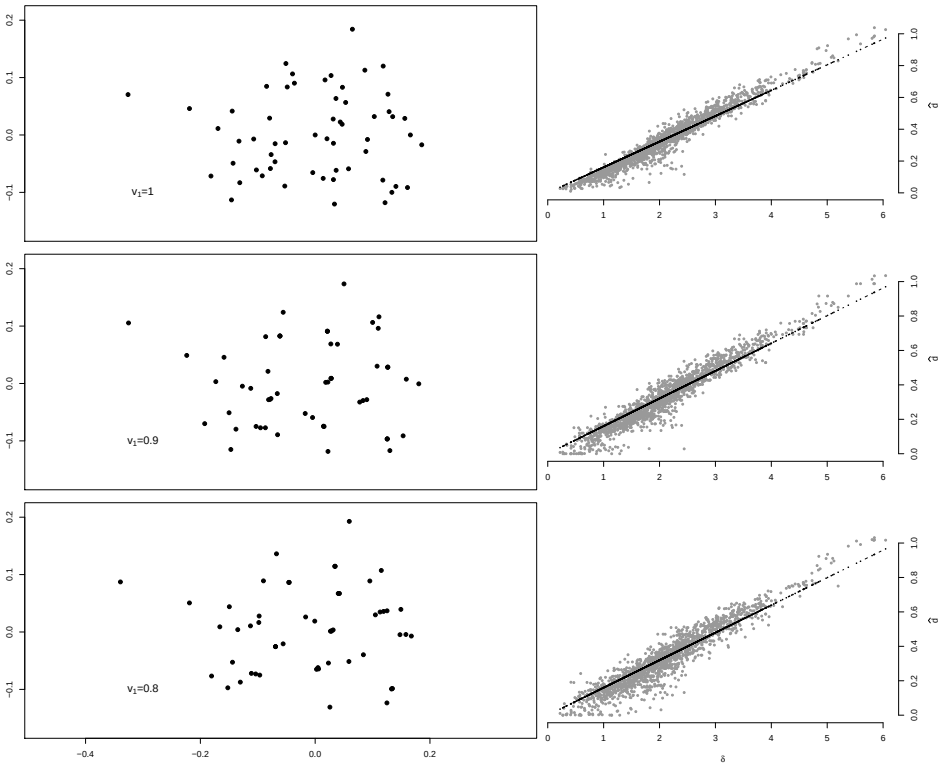

Figure 3: Configurations (left column, procrustes adjusted) and Shepard plots (right column) obtaied from of three different COPS models for the Californian climate change data. The weights were $v_1 = 1, v_2 = 0$ for the top row (standard MDS), $v_1 = 0.9, v_2 = 0.1$ for the middle row and $v_1 = 0.8, v_2 = 0.2$ for the bottom row.

Substantively, the plots for the different weights illustrate differences in the embedding produced by COPS versus the standard MDS. Since $k = 3$, COPS looks for clusters of at least triples of counties. With that in mind, the COPS results show a more clustered configuration with increasing weight on clusteredness illustrated for example by a number of counties accumulating very closely (the density property), e.g., Glenn, San Benito, San Luis Obispo and Colusa or Orange, Santa Clara and Alameda. Note that the accumulations can have different shapes and cluster variances or varying accumulation density which is easily seen with in the bottom left cluster(s) around Tuolumne. Overall the counties accumulate less closely in the standard MDS configuration. Also, between any accumulation of points in the COPS solution, the distances are larger than for the equivalent accumulation in the standard MDS (emphasis property), illustrated by the clearer separation of the accumulations. The COPS configuration also appears to show a higher tally of distinct discrete structures of at least three observations (tally property). All of this is reflected in the smaller OPTICS Cordillera value for the standard MDS (0.115) over the COPS models (0.34 and 0.357 respectively).

The most clustered configuration is displayed with the county labels in Figure 4. The x and y axis, Dimension 1 and Dimension 2, 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 Dimension 1 roughly represent increasing risk for drought, whereas Dimension 2 gives some indication of the risk of flooding. Dimension 1 defines a continuum of low precipitation and many days in extreme heat for higher values. The extreme ends are Kings, Merced, Kern and San Joaquin as the driest counties and Del Norte as the least dry. Dimension 2 divides the counties along aspects such as soil moisture and risk of flooding, with counties like San Francisco and Orange topping the list of flood risk.

In the COPS space there are some clear groups discernable: In the positive half of the $x-$ and $y-$axis we have a cluster of three counties with Santa Clara, Orange County, Alameda. They have a risk profile characterized by a relatively low risk of extreme heat and temperature, an increasingly lower evaporisation, baseflow and precipitation, average to low soil moisture and starting from a higher risk, a decreasing risk for wildfires in the coming 50 years. They are also susceptible to coastal flooding by 2100. Similar are Ventura, Monterey, Contra Costa but with less flood risk. Another cluster features San Diego, Los Angeles, Sacramento, Sutter and San Joaquin. They tend to show a higher drought risk profile, with increasing temperature, little precipitation, low baseflow, low soil moisture but a low risk of wild fires and flooding. The "direction of increasing drought risk" continues in the positive half of the first principal coordinate (Kings, Merced and Kern) which show even less projected precipitation, baseflow or soil moisture and higher temperatures but little risk of flooding or wild fires. This general pattern of high temperatures and low precipitation continues with a cluster in the low right corner of the configuration which comprises Imperial, Riverside, San Bernardino and Inyo, showing high projected temperatures, extremely low projected baseflow, soil moisture and precipitation and a moderate risk of wild fires. The area in the configuration spanned from Los Angeles county towards Inyo and Fresno can be interpreted as having increasing drought risk the further to the bottom right corner they lie.

Towards positive D2 on top of the Tulare cluster we find a cluster consisting of San Benito, San Luis Obispo, Glenn and Colusa. This cluster is closest to the configuration's origin and is thus characterized by a relatively average risk profile, with projected precipitation

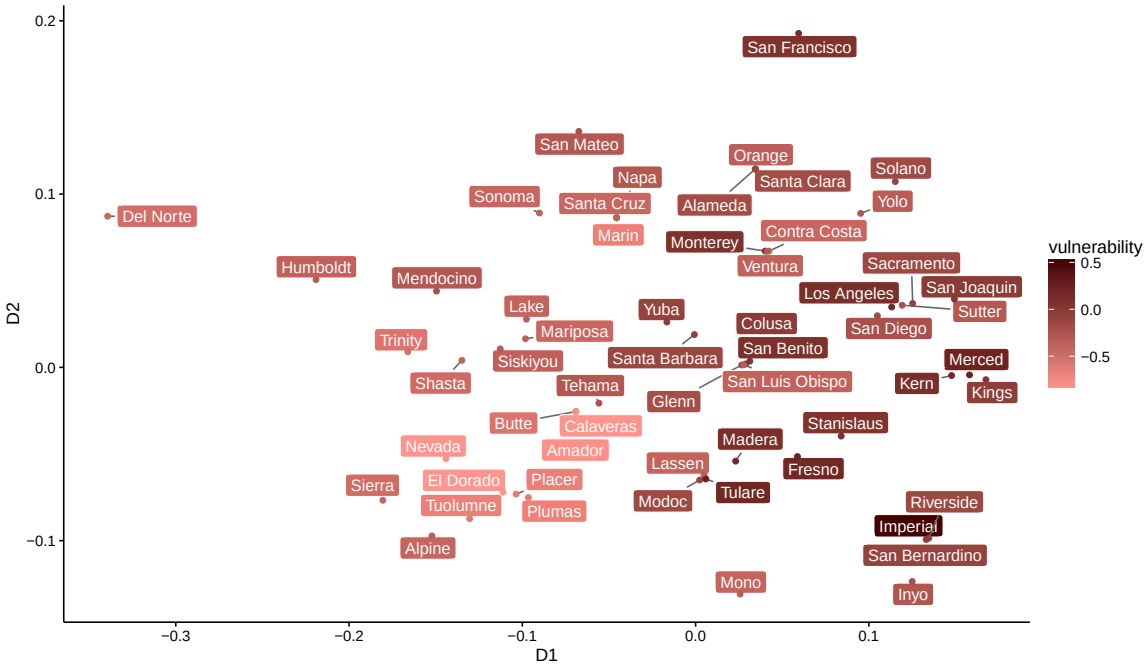

Figure 4: Optimal configuration of scaling median climate change risk indicators for Californian counties with COPS Variant 1 and $v_1 = 0.8, v_2 = 0.2$. The similarities between the counties are based on 50 normalized indicators of climate change such as temperature, precipitation, coastal flooding, wildfire risk either observed or derived from downscaled climate model projection for the years 2000-2099, aggregated to county level. Superimposed is a color gradient derived from the California social vulnerability index (the darker the higher the vulnerability). The latter was not used for scaling. The c-clusteredness index value was $\mathrm{OC}'(X; \epsilon, k, q)$=0.357.

at the moderate to lower end and an accordingly moderately high wild fire risk. Counties positioned further towards the positive part of D2 are generally associated with low projected temperature changes, a decrease in drought and fire risk and an increase in flood risk.

We also find three relative outliers: San Francisco as a county with high risk of flooding, high projected soil moisture and virtually no projected temperature changes, Del Norte county having the maximum projected soil moisture, precipitation, evaporisation and baseflow as well as risk of wild fires and Mono county with the highest projected temperature changes and high wild fire risk.

Figure 4 also shows a luminance gradient for the counties, based on the California vulnerability index (Cooley *et al.* 2012) averaged over each counties' census tracts. The scale is anchored at the minimum and maximum, so lighter means less vulnerable and darker means more vulnerable (the median vulnerability is -0.23). The vulnerability index is derived from 19 demographic variables (Cooley *et al.* 2012). The most socially vulnerable counties are located in the South of California and the San Joaquin Valley as well as the large cities, counties with less vulnerability are particularly the counties in the North and the East. The first latent dimension Dimension 1 separates the vulnerable counties from the resilient ones rather well.

With respect to climate change hazards the picture in Figure 4 is striking: Social vulnerability is strongly associated with risk of drought. The counties with the highest social vulnerability are often similar to each other with respect to the projected risks of climate changes and particularly indicators of drought which shows itself increasingly along the axis from Los Angeles towards Imperial. These counties make up the bottom right quadrant in Figure 3. Together with the population in these counties this means that a large part of the population of California is at high risk of facing increasing drought risk but may not be very resilient to deal with the challenges that this brings with it.

5.2. Banking Crises—Meta parameter optimization with COPS Variant 2

We apply the second COPS variant (profile COPS) for finding meta parameters for a POST-MDS model for the banking crises data set, i.e., choosing the nonlinear projection with coploss.

For the Cordillera $d_{max} = 2.5, q = 1, k = 2, \epsilon = 10$. The lower bound of the search space was set to $\theta = (1, 1, 1)^\top$ and the upper bound to $\theta = (3, 9, 1)^\top$. The resulting configuration can be found in the right panel of Figure 1, the standard MDS is in the left panel. The values for θ found were $\theta = (1.39, 6.978, 1)^\top$ and it took 112 iterations of the outer minimization. The stress-1 value was 0.362. To check whether the POST-MDS with $\theta = (1.39,6.978,1)^\top$ was chasing noise in a constant dissimilarity matrix, we followed suggestions in Mair, Borg, and Rusch (2016) and used a permutation test. The average stress obtained from the permutations was 0.543, the minimum was 0.531 both of which are much higher than our obtained value. Thus the associated one sided p-value is 0, indicating that the result is not an artifact.

Using COPS for metaparameter selection leads to a POST-MDS with a clearly clustered configuration (OC' of 0.21) with axes 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 opposite 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 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 1990s, 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 highest prevalence of banking crises was not in the 1990s. Clusters in between these two crude directions are formed by co-occurences of crises at specific timepoints, e.g, the United States 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.

The right panel in Figure 5 shows a "tranformation plot" of the COPS model, a plot of the fitted distances $d_{ij}(X)$ on the abscissa versus the transformed (darker) and untransformed (lighter) proximities on the ordinate (δ_{ij} and δ^*_{ij} respectively). Superimposed are the fitted nonlinear regression lines. The transformation plots illustrates how the nonlinear projection utilized in the p-stress model operates. From this plot one can see that the proximity transformation shrinks the proximities. Since all proximities are between 0 and 1 it indicates a $\lambda > 1$ and since the shrinking is rather substantial, λ must be rather high (here ist around 7). We can also see that the distance transformation function is curved slightly upwards but not much, indicating a $1 < \kappa < 2$. The (linearized) Shepard plot of δ^*_{ij} vs. $d_{ij}(X)^*$ is in the left panel.

6. Conclusions

In this paper we presented an approach for scaling to increase clusteredness of 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 something about discrete groups of objects from the optimal continuous representation. The latter objective, however, is not considered in standard MDS. To balance these two objectives we introduced COPS, an extension of MDS that considers the clusteredness of a configuration and likely leads to more appreciable grouping of observations in the target space and increases the c-clusteredness enabling the discovery of any type of density-based discrete structures. This is achieved by augmenting the standard MDS loss with power transformations of proximities, fitted distances and weights as well as with a c-clusteredness index. This index has the advantage of being highly nonparametric with respect to assessing the clustered appearance needing only one mandatory parameter. The loss function behind COPS, coploss, can be used in two ways: Either as the loss for directly finding a configuration for given power transformations (Variant 1) or as a loss function to select power transformations for nonlinear dimensionality reduction (Variant 2). We discussed optimization for both variants of COPS and illustrated the use with a data set on climate change risk in California and of banking

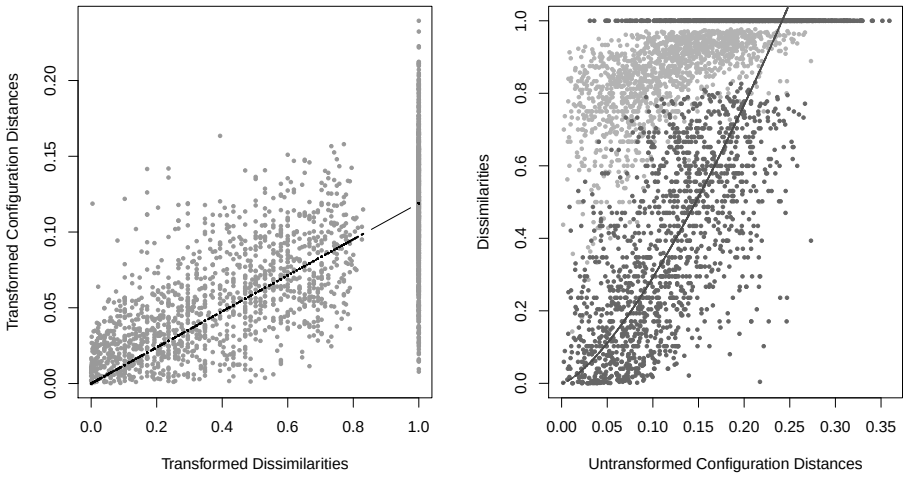

Figure 5: Diagnostic plots for the banking crises data. The left panel shows (linearized) Shepard plots of transformed fitted distances $d^*_{ij}(X)$ on the ordinate vs. transformed proximities δ^*_{ij} on the abscissa for the optimal p-stress model with parameters obtained from COPS Variant 2. Superimposed is a least squares regression line. The right panel shows the "tranformation plot" of the p-stress model, a plot of the fitted distances $d_{ij}(X)$ on the abscissa versus the transformed (darker) and untransformed (lighter) proximities on the ordinate (δ_{ij} and δ^*_{ij} respectively). Superimposed are fitted nonlinear regression lines.

crises. We find that both variants of COPS increase the c-clusteredness present in the solution as opposed to a standard scaling. We note that while we discussed power transformations only, the idea of metaparameter optimization with COPS is not limited to them and can be used for many other parametrized stress functions such as repulsion/attraction based stress functions or Isomap. Additionally, the idea of metaparameter selection based on structural quality can be extended to not only clusteredness but many different aspcts of what one might be interested in a configuration (Rusch 2015).

COPS can also be used to gauge how close a configuration obtained by untransformed scaling is to possible extreme cases of no clusteredness and allows to push the configuration towards a more clustered appearance. This way COPS also addresses the pertinent issue of having low variability in the dissimilarity matrix as called for by Groenen and Borg (2014) and accordingly will have the strongest effect on data with little variability in the proximities. This was illustrated with the banking crises data set. But even for data with already strong c-clusteredness it may be helpful to tease out subtle similarities as in the California data set.

7. Software

Dedicated functions for conducting both variants of COPS are available in the R package **stops** (Rusch, De Leeuw, and Mair 2015). They rely on an implementation of OPTICS, one of which is also available in **stops**. For finding a configuration the `coplossMin` function can be used. Metaparameter optimization can be carried out with `pcops` for models with power transformation and restricted special cases such as Kruskall's stress with symmetric distance matrices or for projection onto a sphere, Sammon stress, s-stress, elastic scaling, r-stress, powermds, powerstress and elastic scaling as well as Sammon mapping with power transformations. There also is a wrapper function `cops` that lets one choose the variant. We also provide two implementations to minimize p-stress, one with majorization and one with NEWUOA. An implementation of ALJ can also be found. The package also includes the data files used in this paper.

A. Supplementary Materials

Code: The code file to reproduce the results of the paper.

References

Akkucuk U, Carroll JD (2006). "PARAMAP vs. Isomap: A tomparison of two nonlinear mapping algorithms." *Journal of Classification*, **23**(2), 221–254.

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 (2013). "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 JR Barra, F Brodeau, G Romier, BV Cutsem (eds.), *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, Stoop I (1984). "Upper bounds for Kruskal's stress." *Psychometrika*, **49**(3), 391–402.

D'Enza A, Palumbo F (2013). "Iterative factor clustering of binary data." *Computational Statistics*, **28**(2), 789–807.

D'Enza A, Van de Velden M, Palumbo F (2014). "On joint dimension reduction and clustering of categorical data." In *Analysis and Modeling of Complex Data in Behavioral and Social Sciences*, pp. 161–169. Springer.

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. IEEE Press, Picataway, NJ.

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

Graves S (2014). "Countries in canking 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.

Groenen PJ, Borg I (2014). "Past, present, and future of multidimensional scaling." In J Blasius, M Greenacre (eds.), *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 W Gaul, D Pfeifer (eds.), *From Data to Knowledge: Theoretical Perspectives and Practical Aspects of Classification*, pp. 177–185. Springer, Berlin.

Heiser WJ, Groenen PJ (1997). "Cluster differences scaling with a within-clusters loss component and a fuzzy successive approximation strategy to avoid local minima." *Psychometrika*, **62**(1), 63–83.

Hennig C (2007). "Cluster-wise assessment of cluster stability." *Computational Statistics & Data Analysis*, **52**(1), 258 – 271.

Kiers HA, Vicari D, Vichi M (2005). "Simultaneous classification and multidimensional scaling with external information." *Psychometrika*, **70**(3), 433–460.

Kim J, Billard L (2011). "A polythetic clustering process and cluster validity indexes for histogram-valued objects." *Computational Statistics & Data Analysis*, **55**(7), 2250 – 2262.

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. Kluwer Academic Publishers, Boston.

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, Borg I, Rusch T (2016). "Goodness-of-fit assessment in multidimensional scaling and unfolding." *Technical report*, Harvard University. In preparation.

Mair P, Rusch T, Hornik K (2014). "The Grand Old Party: A 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.

Powell MJ (2006). "The NEWUOA software for unconstrained optimization without derivatives." In *Large-scale nonlinear optimization*, pp. 255–297. Springer.

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.

Rios LM, Sahinidis NV (2013). "Derivative-free optimization: a review of algorithms and comparison of software implementations." *Journal of Global Optimization*, **56**(3), 1247–1293.

Rocci R, Gattone SA, Vichi M (2011). "A new dimension reduction method: Factor discriminant k-means." *Journal of Classification*, **28**(2), 210–226.

Rousseeuw PJ (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis." *Journal of Computational and Applied Mathematics*, **20**, 53–65.

Rusch T (2015). "A tutorial on Structure Optimized Proximity Scaling (STOPS)." R Package Vignette, Available at *http://stops.r-forge.r-project.org/*.

Rusch T, De Leeuw J, Mair P (2015). *stops: STructure Optimized Proximity Scaling*. R package version 0.0-17, Available at *http://r-forge.r-project.org/projects/stops/*.

Rusch T, Hornik K, Mair P (2016). "Assessing and quantifying clusteredness: The OPTICS Cordillera." *Technical Report Paper 2016/1*, WU Vienna University of Economics and Business, Vienna, Austria.

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.

Timmerman ME, Ceulemans E, Kiers HA, Vichi M (2010). "Factorial and reduced k-means reconsidered." *Computational Statistics & Data Analysis*, **54**(7), 1858–1871.

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

Vichi M, Rocci R, Kiers HA (2007). "Simultaneous component and clustering models for three-way data: within and between approaches." *Journal of Classification*, **24**(1), 71–98.

Vichi M, Saporta G (2009). "Clustering and disjoint principal component analysis." *Computational Statistics & Data Analysis*, **53**(8), 3194–3208.

Wu HM (2011). "On biological validity indices for soft clustering algorithms for gene expression data." *Computational Statistics & Data Analysis*, **55**(5), 1969 – 1979.

Affiliation:

Thomas Rusch
Competence Center for Empirical Research Methods
WU (Vienna University of Economics and Business)
Welthandelsplatz 1, D4
1020 Wien, Austria
E-mail: Thomas.Rusch@wu.ac.at