Assessing and quantifying clusteredness: The OPTICS Cordillera

Rusch, Thomas; Hornik, Kurt; Mair, Patrick

Published: 01/01/2016

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

Assessing And Quantifying Clusteredness: The OPTICS Cordillera

Thomas Rusch, Kurt Hornik, Patrick Mair

Discussion Paper Series
Paper 2016/1, January 2016

Center for Empirical Research Methods
http://wu.ac.at/methods

WIRTSCHAFTS UNIVERSITÄT WIEN VIENNA UNIVERSITY OF ECONOMICS AND BUSINESS

EFMD EQUIS ACCREDITED

Discussion Paper Series of the
Center for Empirical Research Methods
WU Vienna
Welthandelsplatz 1, D4
1020 Vienna
Austria

Assessing and quantifying clusteredness: The OPTICS Cordillera

Thomas Rusch
WU (Wirtschafts-universität Wien)

Kurt Hornik
WU (Wirtschafts-universität Wien)

Patrick Mair
(Harvard University)

Abstract

Data representations in low dimensions such as results from unsupervised dimensionality reduction methods are often visually interpreted to find clusters of observations. To identify clusters the result must be appreciably clustered. This property of a result may be called "clusteredness". When judged visually, the appreciation of clusteredness is highly subjective. In this paper we suggest an objective way to assess clusteredness in data representations. We provide a definition of clusteredness that captures important aspects of a clustered appearance. We characterize these aspects and define the extremes rigorously. For this characterization of clusteredness we suggest an index to assess the degree of clusteredness, coined the OPTICS Cordillera. It makes only weak assumptions and is a property of the result, invariant for different partitionings or cluster assignments. We provide bounds and a normalization for the index, and prove that it represents the aspects of clusteredness. Our index is parsimonious with respect to mandatory parameters but also flexible by allowing optional parameters to be tuned. The index can be used as a descriptive goodness-of-clusteredness statistic or to compare different results. For illustration we use a data set of handwritten digits which are very differently represented in two dimensions by various popular dimensionality reduction results. Empirically, observers had a hard time to visually judge the clusteredness in these representations but our index provides a clear and easy characterisation of the clusteredness of each result.

Keywords: clusteredness, index, dimensionality reduction, clustering, unsupervised learning.

1. Introduction and Motivation

Visual representation of data in a low-dimensional space is an integral part of exploratory data analysis. This representation can either be made with the low-dimensional untransformed data or obtained by means of unsupervised dimensionality reduction like principal component analysis (PCA) (Pearson 1901; Hotelling 1933), correspondence analysis (Hirschfeld 1935; Benzécri 1973), multidimensional scaling (MDS; Torgerson 1958), locally linear embedding (Roweis and Saul 2000) and many others. The latter project high-dimensional multivariate data into a lower-dimensional subspace according to some criterion of optimality whose result can then be visualized, interpreted or otherwise used in a way infeasible for the original data set.

The usage or interpretation of such a data representation in low dimensional space is not always clearly defined. Often the representation is used to infer properties of the original data set *ex post* conditional on the appearance in the low-dimensional space. One type of

appearance that plays a major role in this *ex post* interpretation of results is to assess the arrangement of data points in the low-dimensional space with respect to whether there are discrete structures of accumuluations of data points ("clusters"), how many there are and how well they are separated. This can be called the "clusteredness" of the representation.

Clusteredness is a somewhat elusive concept in statistics. While it has been discussed (e.g., in Greenacre 2011) its definition remains vague. Clusteredness is usually assessed visually from how clustered the objects in a representation appear but often in an unclear and intransparent way. How to judge the appearance and how to interpret it depends largely on the subjective skills and implicit assumptions of the observer.

To illustrate our case we use a random subset of 100 cases of the example of handwritten digits from Alimoglu (1996). The subset uses the first 4 classes (digits 1 to 4, chosen so that it is not too cluttered overall). We use six dimension reduction techniques to find a two-dimensional representation of the data set, principal component analysis (PCA), Sammon mapping (Sammon 1969), locally linear embedding (LLE, with number of neighbours of 10), Isomap (Tenenbaum, De Silva, and Langford 2000, with number of neighbours of 10), power stress multidimensional scaling (POST-MDS; Buja, Swayne, Littman, Dean, Hofmann, and Chen 2008; Rusch, Mair, and Hornik 2015, with parameters $\kappa = 1.5, \lambda = 7.5, \nu = 1$), and a diffusion map (DM; Coifman, Lafon, Lee, Maggioni, Nadler, Warner, and Zucker 2005). The plots of the points in these two dimensions are shown in Figure 1.

We see that the obtained representations are different with respect to how clustered they appear in two dimensions. All projections provide some clusteredness but the degree is different for all of them. In a small empirical study, we asked 24 people to rank order the plots according to the perceived clusteredness of the results giving the instruction that all plots show exactly the same data (including the same number of data points). The people were diverse with respect to their academic background and the possible experiences with these type of results: sociologists (qualitatively and quantitatively oriented), psychologists, statisticians, economists, mathematicians, nutritionists, and computer scientists. The patterns of ranking, their frequencies and the consensus rankings (Emond and Mason 2002) of the plots are given is available in Table 1.

The picture is striking: The 24 people made 20 different rankings of the plots with respect to the perceived clusteredness. They judge the clusteredness of the plots very differently—at most three people agreed with each other on a common ranking. Furthermore a consensus ranking is difficult to derive: Five patterns show the same maximum average rank correlation $\tau_X = 0.428$ (Emond and Mason 2002) with the individual rankings. It seems likely that different observers have different implicit definitions of what clusteredness is and how to judge it, and therefore might get to different assessments. Also, the assessment of clusteredness present was possible in terms of giving a ranking but asking for a quantification of how clustered the representations appear was reported as very difficult. The visual *ex post* interpretation of clusteredness in a given representation therefore appears to be highly subjective and leads to diverse outcomes, yet this is what is usually done.

In this paper we aim at the following: We first define and formalize the notion of clusteredness in a (low dimensional) data representation. Informally, by clusteredness we define a continuum of appearances of a representation where, starting from a result with no discernable clusteredness, clusteredness increases up to the point of maximal clusteredness. We then discuss aspects of clusteredness relevant to determine its degree. These aspects relate to

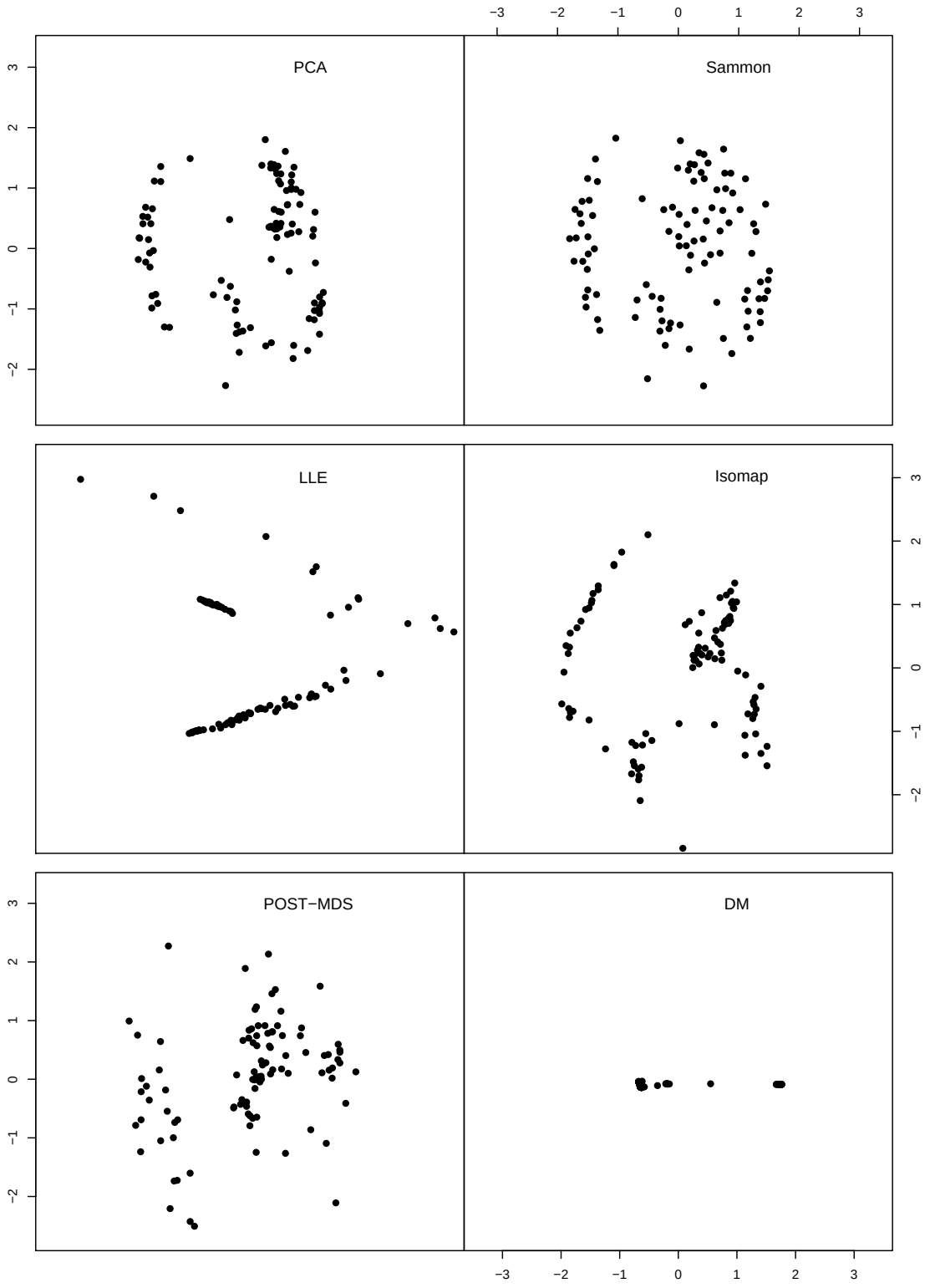

Figure 1: 2D projections of the digits data subset for six different dimension reduction techniques. Top left is principal component analysis (PCA), top right is Sammon mapping, middle row left is Locally Linear Embedding (LLE) with $k = 10$, middle row right is Isomap with $k = 10$, bottom row left is powerStress MDS (POST-MDS with $\kappa = 1.5, \lambda = 7.5, \nu = 1$) and bottom right is a diffusion map (DM). All the dimensions have been scaled to mean=0 and sd=1 and put on the same plotting region.

	Ranks						
Pattern	PCA	Sammon	LLE	Isomap	POST-MDS	DM	Frequency
1	1	3	4	2	5	6	1
2	1	4	5	2	3	6	1
3	2	3	1	6	4	5	1
4	2	3	4	5	6	1	1
5	2	3	6	4	5	1	1
6	2	4	5	3	6	1	2
7	2	4	6	3	5	1	1
8	2	6	4	3	5	1	1
9	3	1	5	4	2	6	1
10	3	2	4	5	1	6	1
11	3	2	6	4	5	1	1
12	3	4	6	2	5	1	1
13	3	5	2	4	6	1	1
14	3	6	2	4	5	1	2
15	4	2	6	5	3	1	1
16	4	5	1	3	6	2	1
17	4	5	2	3	6	1	3
18	4	5	3	2	6	1	1
19	4	6	2	3	5	1	1
20	5	6	2	4	3	1	1
C1	2	5	4	3	6	1	
C2	2	4	4	3	5	1	
C3	2	4	5	3	6	1	
C4	2	4	3	3	5	1	
C5	2	5	3	4	6	1	

Table 1: Ranking patterns and their frequency of the ordering of the plots in Figure 1 with respect to the perceived clusteredness for $n = 24$ observers. The table produces five consensus rankings (C1 to C5) all with the same maximum average rank correlation τ_X of 0.428.

(i) a (specified) number of represented objects accumulating close to each other in discrete structures (clusters), (ii) how close the represented objects accumulate together, (iii) how well accumulations are separated, (iv) the number of accumulations and (v) the spread of represented observations in the target space.

We suggest an index that assesses how much clusteredness we find in a data representation. The index quantifies the clusteredness adhering to the aspects of clusteredness. The index allows us to represent the global clusteredness property of the representation, based on regions, density and distances of both close and far neighbouring points, in a unidimensional measure. The index makes no assumptions on the shape of the clusters, it is independent of a partitioning of the objects, assignment of observation to clusters, centroids or similar concepts are not needed and nested clustering structures can be appropriately considered by the index.

2. Clusteredness

We motivated this paper by the observation that *ex post* interpretation of clusteredness of the very same data representation can be quite different for different people. We believe a major factor for this are an observer's implicit assumptions which are not necessarily transparent. In other words the common usage of such results is subjective and the interpretation may not be replicable. In order to objectively assess the clusteredness of such results, in this section we first formalize the notion of clusteredness and then present an index that captures clusteredness for a given representation.

First we need to make an important distinction: The concept of clusteredness is different from the concept of goodness-of-clustering (or internal cluster validity). For the latter, assessment and quantification in given results has received considerable attention and a number of indices have been proposed over the years: the Calinski-Harabasz index (Caliński and Harabasz 1974), the Silhouette measure (Rousseeuw 1987), prediction strength (Tibshirani and Walther 2005), the Theoretical Clustering Index (TCI; Huang, Liu, Yuan, and Marron 2014) and many more (see e.g., Liu, Li, Xiong, Gao, and Wu 2010, for an overview). The concept behind these indices is the assessment of *how well a given partitioning or clustering of the data* works internally usually by, e.g., looking at separation between clusters and compactness within clusters.

Our aim is different: We pursue a concept (and ultimately an index) that mimics what an observer would do when judging the clusteredness in a data representation solely from the appearance *independent of a given partitioning*. We derived five aspects by which clusteredness is judged (described in detail below). Some aspects have a counterpart in the concept of goodness-of-clustering but not all do.

In a nutshell, clusteredness is the appearance of (many) more than one appreciable structures of arbitrarily shaped accumulations of observations in a representation as a function of distances between accumulations of observations, whether and how separated the accumulations are, how dense the accumulations are, of the number of such accumulations and of the spread of observations overall. It is a property of the representation's appearance, which is invariant to any actual decision on a partitioning. For one and the same representation, clusteredness is therefore constant for ever possible clustering, irrespective of which partitioning is chosen and to which cluster an observation is assigned. This also applies to nested cluster structures,

different cluster definitions, and different cluster numbers.

Take for example the PCA result in 1 which we cluster with a Clara algorithm (where the centroids are medoids; Kaufman and Rousseeuw 2009) and a k-means algorithm (where the centroids are means; MacQueen *et al.* 1967)), with 3, 5 and 10 clusters. The average Silhouettes derived from the clusterings are 0.478 for the 5 cluster solution, 0.574 for the 3 cluster solution and 0.515 for the 10 cluster solution. For k-means it is 0.511, 0.57 and 0.441 respectively. The decision on how many clusters there are, the assignment of observations to clusters, the choice of centroid/cluster shape changes the value of the index and the ordering of best to worst goodness-of-clustering. But the appearance of how clustered the configuration is does obviously not change at all.

Clusteredness as we define it must capture the appearance irrespective of any of these choices as the appearance does not change. The above mentioned indices of goodness-of-clustering do not fulfill these aspects as they usually also change as a function of at least one of the decisions on how clusters are formed, what the clusters are, what the cluster centroids are and what cluster an observation is assigned to.

Let $x_1, \ldots, x_N \in \mathbb{R}^p$. The $x_1, \ldots, x_N$ are row vectors of a matrix X, the configuration. Let $d_{ij}(X) = d(x_i, x_j)$ denote the distance between the observations x_i and x_j, for example a p-norm distance $d_{ij} = ||x_i - x_j||_p$ with $p \geq 1$. Let k be the minimum number of points that must comprise a discrete structure we might be willing to call a cluster (so $2 < k < N$). The set or *accumulation* of k closest points to and including x_i is denoted by $C_k(x_i)$, a k-accumulation (or a k-cluster). Note that while we use accumulation and cluster interchangeably, not every accumulation needs to correspond to a cluster in reality or to one from a clustering, so the two can be different depending on the definition of what is a cluster.

We interpret clusteredness as the departure from no clusteredness. No clusteredness is characterized by points falling on any subset of vertices of a matchstick graph where any two closest points are equidistant. One can also say that vertices fall onto a regular tesselation (e.g., a lattice) in $\mathbb{R}^p$. Not every possible points of such a tesselation needs to be realized, giving rise to the graph structure mentioned above. No clusteredness is the situation where the distances of points in the configuration to their closest neighbours are the same for all points and thus we cannot speak of accumulation of points. See also the top left plot of Figure 2.

Definition 1 (No clusteredness)**.** No clusteredness is given when the row vectors in X can be represented as vertices of 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. That is, following Gervacio, Lim, and Maehara (2008), the graph $G = (X, E)$ has an injection $l : X \mapsto R^2$ satisfying that $x_i, x_j \in E \Rightarrow ||l(x_i) - l(x_j)|| = c$ for $x_j \in C_2(x_i), x_j \neq x_i$ and is planar.

Clusteredness is the deviation from Definition 1 so that we have an embedding in $\mathbb{R}^p$ that not all points are projected onto the vertices of a matchstick graph. The extreme case is a maximally clustered appearance. This happens if for a maximum number of distinct points of accumulation, the observations belonging to an accumulation coincide exactly and the distance between any two closest accumulation points is constant. The maximum number of accumulation points is defined with respect to how the observations can be distributed under the conditions of having at least k observations per accumulation (so at most N/k accumulation points).

Definition 2 (Maximal clusteredness)**.** Let the maximum number of accumulation points that could theoretically be formed be n. The maximum number of possible accumulations of at least k points for given k and N is only achieved if it is possible to evenly distribute the N data points into N/k accumulations of size k, i.e., $N \equiv 0 \pmod{k}$. If the points cannot be distributed evenly into N/k accumulations, some points may not be part of an accumulation of size k but the counterfactual $n = \lceil N/k \rceil$ is still maximal. Maximal c-clusteredness is now defined as the situation of $n = \lceil N/k \rceil$ accumulations and for all k points x_i in the same accumulation the distance to each other is zero and each accumulation is some constant minimal distance $d_{\max}$ away from the closest neighbouring accumulation. The $d_{\max}$ represents some constant (positive) distance between two neighbouring accumulations, e.g., the maximum distance between any two points from these two accumulations. Or for point x_i,

$$d_{ij} \begin{cases} = 0 \text{ if } x_j \in C_k(x_i) \\ = d_{\max} > 0 \text{ if } x_j \notin C_k(x_i) \wedge x_j \in C_k(x_s) : d_{is} = \max(0, \min d_{it}) \quad \forall\, i \neq s, t \\ \geq d_{\max} \text{ otherwise.} \end{cases} \tag{1}$$

Thus the k points in the same cluster have no distance to each other, $d_{ij} = 0$, and all the positions of accumulation are some constant distance $d_{\max}$ away from each other, so there is equidistance among all the closest neighbouring accumulations. See also the bottom right plot of Figure 2. The choice of defining maximal clusteredness for $n = \lceil N/k \rceil$ accumulations means we treat the situation of a balanced distribution as unambiguously ideal for clusteredness. A situation with extra accumulations of points of a size smaller than k or less accumulations of points of a size larger than k, and possible "noise" points is therefore always less than ideal.

Figure 2 illustrates clusteredness with a toy example of 8 labeled data points. The top plot shows no clusteredness (a unit distance drawing is possible) and the second plot from the top shows some clusteredness. Note that point 6 is deliberately placed so the plot does not appear perfectly regular but we still have a matchstick graph (the regular case would also be perfectly non-clustered). Clusteredness increases from the top to the bottom and the bottom plot shows a maximal 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.

The clusteredness of a given representation is now positioned on a continuum between no clusteredness and maximal clusteredness as given above, subject to a number of aspects of the appearance which can overlap with similar considerations in goodness-of-clustering. These aspects are that clusteredness increases if:

- Distances between points of accumulation increases. We call this the emphasis aspect as the clusters are better emphasized and they are better separable. This overlaps with separation in goodness-of-clustering.
- Points accumulate more densely. We call this the density aspect as a closer packed accumulation appears more clustered. In goodness-of-clustering this is usually referred to as cohesion or compactness.
- The number of points of accumulation increase. We call this the tally aspect as a representation with five accumulations appears more clustered than one with only two or one.

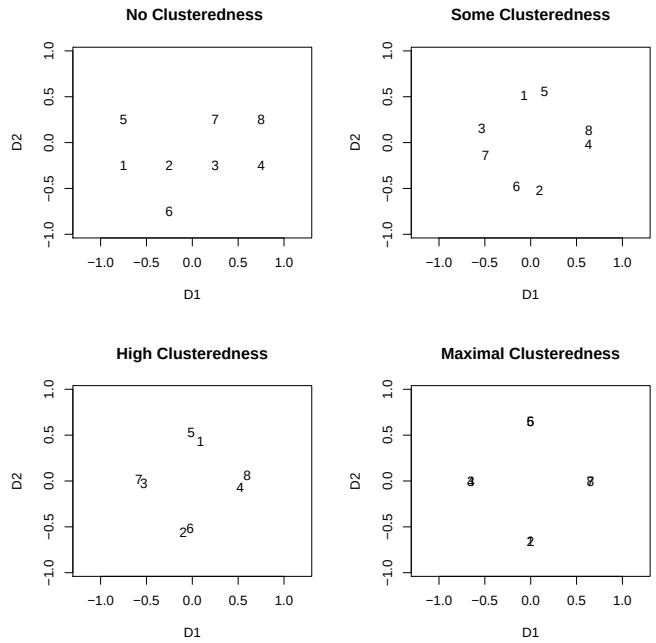

Figure 2: Differently clustered 2D representations of 8 points. The top left plot shows a case of no clusteredness, the bottom right panel shows maximal clusteredness for $N = 8$ and $k = 2$. The other two plots show cases between these two extremes.

- In the configuration the observations are spread over a larger space. We call this property the spread aspect. The idea is that if points (or accumulations) are distributed over a larger portion of the target space the appearance of clusteredness increases.

A clusteredness index should represent these aspects numerically. Note that the spread aspect represents the decision to introduce a cut-off at which a distance between points is no longer treated as a decrease in cluster density but as an increase in clustered appearance. This property can have the consequence of being susceptible to outliers. A clusteredness index should have a mechanism to deal with that.

3. A Clusteredness Index: The OPTICS Cordillera

In this section we propose an index that assesses and quantifies the departure of the result from a lack of clusteredness and objectivizes the process of *ex post* interpretation of clusteredness in data representations like from unsupervised dimension reduction results. Since these methods are usually used in an exploratory fashion, say for discovery of possible clusters or to visualize relationships, we need to be as impartial with respect to what should be discovered and as restrained with making assumptions about the nature of results as possible. This particularly applies to the following: the number of clusters that result, the shape of the clusters, the size of the clusters, that clusters can be nested. Additionally, independence of the decision which observation belongs to which cluster or the choice of a centroid to relate the points in a cluster to each other is necessary. In fact, the only decision we are willing to make is to define

the minimum number of observation that can comprise a cluster and that these observations are in a way connected by being close. Everything else shall remain free. In other words we aim for an index that is much more restrained with respect to the assumptions necessary as goodness-of-clustering indices are.

The index has the following properties:

- It is bounded, i.e., have unique minimum and maximum corresponding to no clusteredness and maximal clusteredness.
- It is a function of the distance function used in the dimension reduction.
- It is broadly applicable to any clustered appearance.
- It makes only very weak assumptions on what constitutes an accumulation or cluster.
- It represents the clusteredness aspects in such a way that the more clusteredness a result shows, the higher the index and vice versa.

Our proposal, the OPTICS Cordillera (or simply Cordillera), quantifies the degree of clusteredness. It is derived from the OPTICS algorithm (Ordering Points To Identify The Clustering Structure; Ankerst, Breunig, Kriegel, and Sander 1999) that outputs a unidimensional ordering of input points based on a matrix of distances. OPTICS is not a clustering algorithm *per se*, it only produces an augmented ordering. 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. The ordering and reachability can be displayed in the "reachability plot". We derive a mapping of the clusteredness of a representation to a univariate scale from the OPTICS' ordering-reachability combination by aggregating the reachabilities over the OPTICS ordering for the points in the configuration X. The name "cordillera" derives from an analogy of the reachability plot with its "peaks" and "valleys" to a mountain range and that the index in a sense measures its raggedness. It is 0 in case of no-clusteredness as defined in Definition 1 and quantifies how far away a given configuration X is from a situation as in Definition 1, or how close it is to (1) in Definition 2 for given $N, k, d_{\max}$.

This index only needs the number of points that comprises an accumulation to be specified, but has additional optional parameters to control the outlier influence and the weighting of the clusteredness properties. It fulfills the desirable properties of allowing to represent arbitrary accumulations shapes and nested accumulations, being invariant under different partitionings, cluster assignments of the points, does not include a notion of centroids and fulfills the clusteredness properties of being monotonically nondecreasing and typically increasing as a function of increasing distances between accumulations of points, as a function of increasing cluster density, as a function of an increasing number of accumulations, and as a function of more spread-out points. Here we first only describe the index and lower and upper bounds for it; a discussion and proofs of the properties can be found in the Appendix.

Distance definitions and the OPTICS algorithm. OPTICS allows to use two parameters: The mandatory parameter k (in OPTICS called minpts), the minimum number of points needed to comprise a cluster ($k > 1$), and an optional parameter ϵ, the maximum radius of

the neighbourhood around a point to look for another point. The latter is optional insofar as it does not need to be tuned but can be simply set very large, e.g., the maximum distance between any two points. It can be used for refinement, e.g., to make the procedure robust to outliers. The parameter k has a smoothing effect and needs to be set *a priori*. Based on these parameters, OPTICS calculates special distances for each point and iteratively processes them to produce an ordering of the points augmented with each point's associated minimum reachability distance.

The distances used in OPTICS are defined the following way (Ankerst *et al.* 1999): 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, $S_k(x_i;\epsilon) \subseteq N_\epsilon(x_i)$. If $\text{card}\,(N_\epsilon(x_i)) < k$, then $S_k(x_i;\epsilon) = \emptyset$.

Definition 3 (Core Distance and Core Point)**.** The "core distance" c_i is the distance of a vector x_i to (any of) the $k-$th closest points

$$c_i = c(x_i;\epsilon,k) = \begin{cases} \max(d_{ij}) : j \in S_k(x_i;\epsilon) & \text{if } S_k(x_i;\epsilon) \neq \emptyset \\ \text{undefined} & \text{if } \text{card}\,(N_\epsilon(x_i)) < k \end{cases} \tag{2}$$

If $\text{card}\,(N_\epsilon(x_i)) \geq k$, x_i is called a "core point".

Definition 4 (Reachability Distance)**.** The "reachability distance" r_{ij} between two points x_i and x_j is the maximum of d_{ij} or $c(x_i;\epsilon,k)$, so

$$r_{ij} = r(x_i,x_j;\epsilon,k) = \begin{cases} \max(c_i,d_{ij}) & \text{if } S_k(x_i;\epsilon) \neq \emptyset \\ \text{undefined} & \text{if } \text{card}\,(N_\epsilon(x_i)) < k \end{cases} \tag{3}$$

Definition 5 (Minimum Reachability Distance)**.** The smallest or minimum reachability distance, r_i^*, for point x_i is

$$r_i^* = \min_{j:i\neq j} r_{ij}(x_i,x_j;\epsilon,k) \;\forall\; r_{ij} \neq \text{ undefined} \tag{4}$$

A graphical representation of these distances is shown in Figure 3. The parameters are $k = 3$ and $\epsilon = 0.82$, x_9 is the reference point. The region around x_9 in which to look for neighbours is defined by ϵ, the dotted circle. The core distance of x_9, c_9, is the distance to the $k-$th closest point which is point x_3. The core distance is roughly 0.34, the length of the dashed line. It is also r_9^*. At $k = 3$ and $\epsilon \geq c_9$ all points including x_9 within the distance c_9 around x_9 are core points and are directly density reachable from x_9. The core region around x_9 is illustrated by the dashed circle. The set of these points is $S_3(x_9;\epsilon) = \{9,7,3\}$ and the core distance is the maximum distance to any of the points in $S_3(x_9;\epsilon)$. The reachability distance between x_9 and any other point x_j, r_{9j}, is now the maximum of core distance of x_9 or direct distance of d_{9j} or is undefined if the point falls beyond the ϵ radius. This is illustrated as the length of the solid lines for a few examples. For example for r_{97} it is $\max(c_9,d_{97})$ which is c_9, for r_{96} it is $\max(c_9,d_{96})$ which is d_{96} and for r_{94} it is undefined as x_4 is more than $\epsilon = 0.8$ distant from x_9. This illustrates the function and optionality of ϵ: Any ϵ will contain the defined distances of any smaller ϵ (i.e., denser accumulation) which enables the simultaneous characterisation of many accumulations with different densities between objects up to ϵ. Thus, ϵ needs not necessarily be set but can be just large. Setting ϵ will lead to treating points further away as "noise" instead of a neighbour (here, x_8 and x_4).

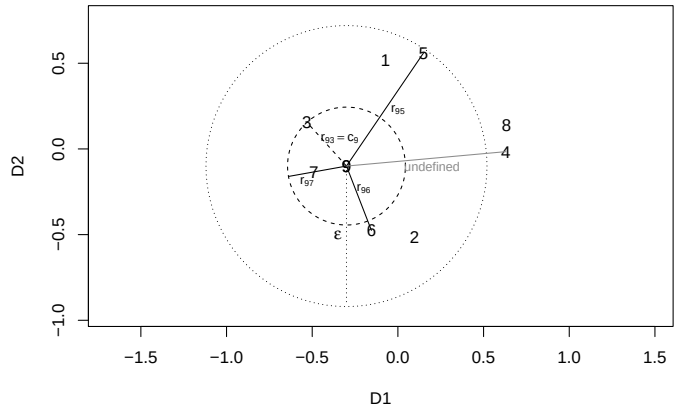

Figure 3: Illustration of the distances c_i and r_{ij} used in OPTICS. The parameters are $k = 3$ and $\epsilon = 0.82$. We use point x_9 as the reference. The region around x_9 in which to look for neighbours is defined by ϵ and is illustrated by the dotted circle. The core distance of x_9, c_9 is the distance to the $k-$th closest point (x_3). The core distance is roughly 0.34 (dashed line). The core region around x_9 is illustrated by the dashed circle. The reachability distance between x_9 and any other point x_j, r_{9j}, is the maximum of core distance or direct distance or is undefined if the point falls beyond the ϵ radius. The illustrated as solid lines for a few examples, $r_{93} = c_9$, $r_{97} = c_9$, $r_{96} = d_{96}$, $r_{95} = d_{95}$ and $r_{94} = undefined$.

It might be illuminating to state the definition of what constitutes an accumulation in this setting. Let a point x_j be called "directly density reachable" from point x_i if $x_j \in (N_\epsilon(x_i))$ and x_j is a core point. Let it be called "density reachable" from point x_i if there is a chain of directly density reachable points x_i to x_j. Let a point x_j be called "density connected" to x_i if there is a point x_k that is density reachable from both x_i and x_j. Then we define an accumulation of points as (cf. Ankerst *et al.* 1999):

Definition 6 (An Accumulation)**.** For given ϵ and k an accumulation is a (non-empty) set C of at least k points satisfying

$$\begin{aligned} \forall x_i, x_j &: \text{If } x_i \in C \text{ and } x_j \text{ is density reachable from } x_i \implies x_j \in C \\ \forall x_i, x_j \in C &: x_i \text{ and } x_j \text{ are density connected} \end{aligned} \tag{5}$$

Every point not in an accumulation is noise.

Based on (3) the OPTICS algorithm orders the points and outputs that ordering R of the vectors x_i together with the r_i^*. The R and r_i^* are what we need subsequently. We will also need the position of point x_i in the ordering R, $s = s(x_i, R) = \text{position}(x_i, R)$. 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 utilizing a priority queue with three operations implemented, `insert()`, `next()` and `moveup()` (Ankerst *et al.* 1999). A pseudo-code representation of the algorithm to produce the R and the r_i^* is listed as Algorithm 1.

The principle is the following: A point gets visited and the neighbours are recorded. Then its core distance is calculated (if defined, else the next point is used). Then the directly density reachable neighbours get inserted into a priority queue sorted by reachability distance to the closest core point. This queue is iteratively updated for the reachability distance based on the ϵ-neighbourhood of the point and the neighbours in the queue. The queue gets processed so that the point with smallest reachability distance is selected, the neighbours get recorded and core distance gets determined. If the current point is again a core point the above is is repeated until no unprocessed points are left. This way the OPTICS ordering is so that if the minimum reachability for $x_{(s)}$ is small then $x_{(s)}$ und $x_{(s-1)}$ are close together. If it is large then $x_{(s)}$ is far away from $x_{(s-1)}$. Therefore, points in the ordering that are subsequent in the ordering and have small minimum reachability appear clustered together whereas points that are far away from each other in the ordering or have some large reachability between them appear spread out. Note that this is irrespective of any cluster assignment of observations.

3.1. The OPTICS Cordillera

We can use the ordering R of points and the respective minimum reachabilities r_i^* to fashion a clusteredness index. Let $R = \{x_{(s)}\}_{s=1,\dots,N}$ be the ordered set of the original points $x_i, (i = 1, \dots, N)$ as output by the OPTICS algorithm, so $x_{(1)}$ is the x_i at the first position in R. Let $r_{(s)}^* = r_i^*$ the minimum reachability as defined in (4) of point $x_{(s)} = x_i$. Let us further set

$$r_{(s)}^* = d_{\max} \text{ if } r_{(s)}^* \text{ is undefined } \vee \; r_{(s)}^* > d_{\max} \tag{6}$$

This $d_{\max}$ caps the reachability distance and 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 $\max r_i^*$ for the defined r_i^*. 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).

Then by using the q-norm of the finite difference of the minimum reachabilities (6) over the ordering of points, we define the (raw) OPTICS cordillera as

$$\mathrm{OC}(X; \epsilon, k, q) = \left(\sum_{s=2}^{N} |r_{(s)}^* - r_{(s-1)}^*|^q \right)^{1/q}. \tag{7}$$

The parameter $q > 0$ is a meta-parameter and controls the relative strength with which distances are involved in the index calculation and can be used to weigh large reachabilities (usually distances between points in different accumulations) and small reachabilities (usually distances between points in the same accumulation) differently.

The OPTICS Cordillera looks for the pairwise differences of $r_{(s)}$ over R as a means of capturing the global clusteredness of a data representation. It aggregates this information as the norm of the differences of minimum reachabilities and the larger this norm, the larger the index is and the more clusteredness we typically find in the solution and vice versa.

Upper and lower bounds for the OPTICS Cordillera. The OPTICS Cordillera in (7) is bounded. The lower bound is 0 and an upper bound for the observed OPTICS Cordillera

Algorithm 1 A pseudo code representation of the main OPTICS algorithm (upper part) and the update function (after Ankerst *et al.* (1999) and Wikipedia (2015)).

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

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

in the maximal clusteredness case is given by the value of the OPTICS Cordillera in the most clustered case. this upper bound depends on the number of observations N and the number of points k that must make up a cluster.

Proposition 1 (Bounds of the OPTICS Cordillera.)**.** If $d_{\max} \geq \max r_i^*$ for the defined r_i^* then,

$$0 \leq \mathrm{OC}\left(X; \epsilon, k, q\right) \leq \mathrm{C}\left(X, d_{\max}, k, q; \epsilon\right) \tag{8}$$

where

$$\mathrm{C}\left(X, d_{\max}, k, q; \epsilon\right) = d_{\max}^q \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right) \tag{9}$$

The proof can be found in Appendix A.

Normalizing the OPTICS Cordillera. We can use Proposition 1 to normalize the OPTICS Cordillera to lie between 0 and 1. The normalized OPTICS Cordillera is given by

$$\mathrm{OC'}(X; \epsilon, k, q) = \frac{\mathrm{OC}(X; \epsilon, k, q)}{\mathrm{C}\left(X, d_{\max}, k, q; \epsilon\right)} = \frac{\left(\sum_{s=2}^{N} |r_{(s)}^* - r_{(s-1)}^*|^q \right)^{1/q}}{d_{\max}^q \left(\left\lceil \frac{N-1}{k} \right\rceil + \left\lfloor \frac{N-1}{k} \right\rfloor \right)}. \tag{10}$$

Note that $d_{\max}$ in (6) and (9) are the same. When choosing a value for $d_{\max}$ it is useful to distinguish between clusteredness relative to the largest possible distance for a given representation versus for a series of representations, or a constant. This will control the interpretation of the index: It can be given the interpretation of goodness-of-clusteredness—conceptually similar to a pseudo R^2 and a goodness-of-clustering index—as the amount of clusteredness attained relative to the most clusteredness achievable for a given maximum distance in a representation when $d_{\max} = \max_i r_i^*$. It also allows to generate an index for comparing a series of representations $X^{(1)}, \ldots, X^{(G)}$ with respect to their clusteredness. In this case $\mathrm{C}(X, d_{\max}, k, q; \epsilon)$ should be the same for all G results, e.g., 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.

Illustration. Figure 4 illustrate the concepts involved. It shows in the right column the OPTICS cordillera and the reachability plots for the representation in the left column. We use $q = 1$ here. The grey barplot shows the minimum reachability on the y-axis for the ordering $x_{(i)}$ on the x-axis. The OPTICS Cordillera is the black line (displayed up to a constant). It holds that the larger this line is, the more clustered the representation is. The Cordillera reaches a minimum if all points have equal minimum reachability. The length of the bottom right raw OPTICS Cordillera is also the upper bound for all the other Cordilleras (with $d_{\max} = \max_g d_{\max}^{(g)},\ g = 1, \ldots, 3$) of representations in Figure 4. We clearly see the ever increasing Cordillera with ever increasing clusteredness.

Properties of the OPTICS Cordillera. The index in (7) has appealing properties meeting the requirements of measuring the aspects of clusteredness as laid out earlier. For readability, we only paraphrase the properties here but an in-depth characterization can be found in Appendix B.

Some properties of the raw OPTICS Cordillera carry over directly from OPTICS (Ankerst *et al.* 1999). These properties are:

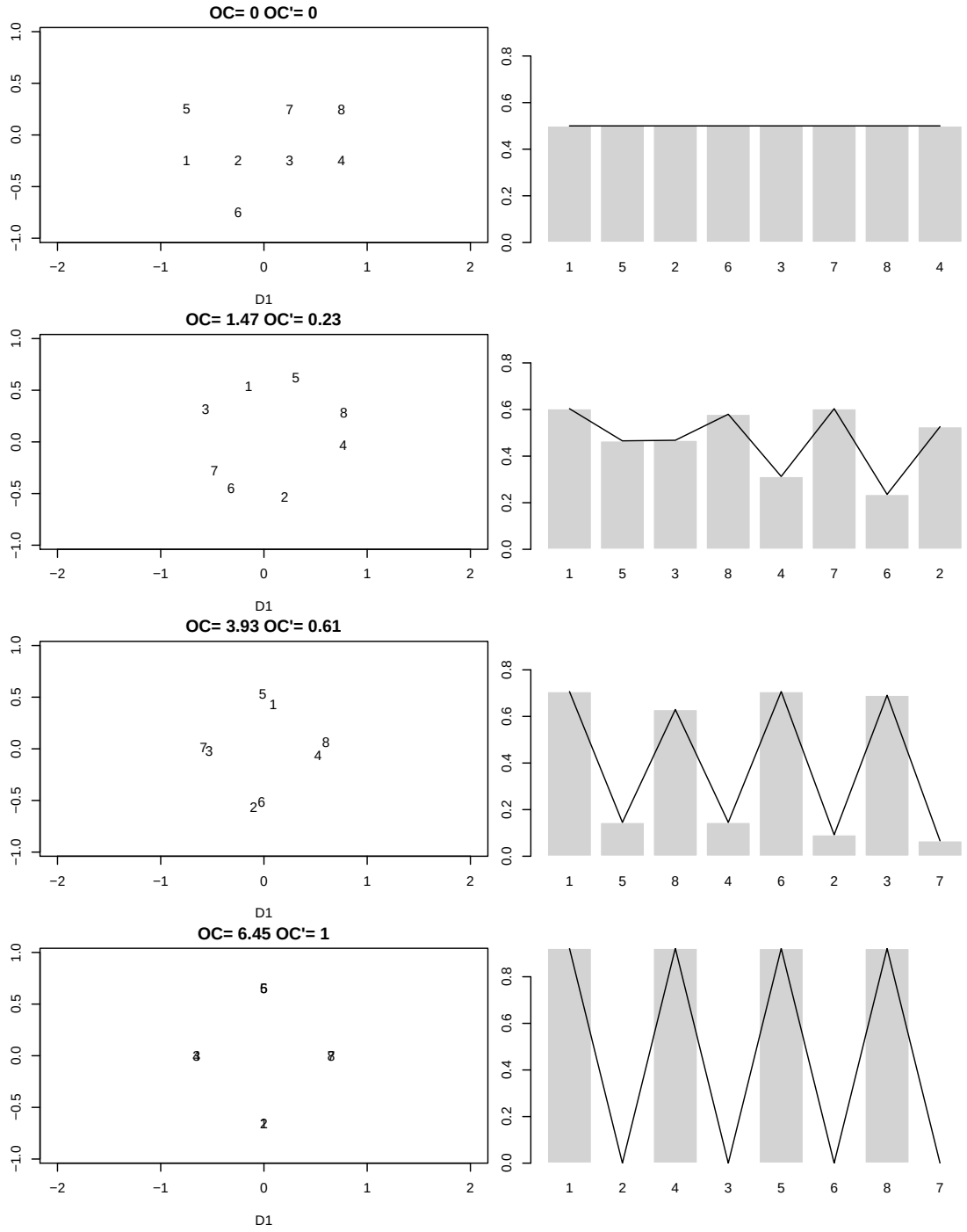

Figure 4: Differently clustered 2D representations of 8 points and their OPTICS Cordillera. In the left column we find the representations. The top left plot shows a case of no clusteredness, the bottom left panel shows maximal clusteredness for $N = 8$ and $k = 2$. The other two panels shows representations between these extremes. 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 derived clusteredness index, the raw OPTICS Cordillera (which is here proportional to the real value). The plots are labeled with the numeric value for the normalized OPTICS Cordillera. It has been calculated with $k = 2, \epsilon = 2, q = 1$. The black line in the bottom plot is also the normalizing constant (which is an upper bound to the OC in all other plots).

- The raw OPTICS Cordillera does not need any specific cluster assignment or real label of observations. It simply is an aggregation of differences of distances over an ordering. Thus the OPTICS Cordillera is invariant to any cluster assignment or membership of points as longs as the augmented ordering does not change.
- The result is independent of any partitioning of the data or decision on the number of clusters prior or afterwards. For a given k and ϵ, the raw OPTICS Cordillera is solely a property of the data representation X and therefore constant for every possible clustering of X of any number of clusters from 1 to $\lceil N/k \rceil$.
- The OPTICS Cordillera does not need the notion of centroids or prototypes.
- Nested accumulations of varying density are considered simultaneously.
- Accumulations are defined solely by the minimum number k of points that must make up an accumulation and by points in an accumulation being mutually density reachable and density connected, which is governed by ϵ. This abstracts the OPTICS Cordillera from making any stronger assumptions about the clustering.
- Due to an accumualtion being defined as Definition 6, the geometrical shape of the accumulations and distribution of objects within the accumulation can be completely arbitrary (beyond the effects of the used distance measure).

Other properties are specific to the OPTICS Cordillera and match the aspects of clusteredness. They follow from showing under which conditions the norm of the differences of smallest reachabilities in (7) does not decrease.

- Emphasis property: If the distances between the accumulations increases, the index does not decrease and typically increases. Thus if we take an accumulation from the representation and shift it away from all the other accumulations the index does not become smaller and usually gets larger.
- Density property: Shrinking points comprising an accumulation monotonically towards a central point (i.e., point with minimal reachability) will lead to a nondecreasing and typically increasing index. In essence the index usually increases if the points in a accumulation are accumulating more densely.
- Tally property: For an increase in the number of accumulations, the index does not decrease and typically increases.
- Balance property: For a given number of accumulations the index does not decrease as a function of the number of observations $> k$ comprising an accumulation. It will not pick up unbalancedness in the number of points in a accumulation as a sign of clusteredness and thus enable a unique maximum for maximal clusteredness.
- Spread property: If we shift a number of points in such a way that the distances to all other points increase sufficiently much, then the index increases. The index increases when points are so spread out that it appears sensible to assume there are points that are qualitatively different in the sense of being far from others. This property may run counter to the density property. In this case a density based clustering cannot be

upheld, so 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 which can be combatted by setting ϵ and k so that the index is more robust.

These properties are formalized and established as Propositions 2-6 and proven in Appendix B.

4. Example and Practical Usage

We return to the motivating example from Figure 1. The data are a random subset of a data set of handwritten digits, namely the digits 0, 1, 2, 3, 4, which we subjected to six different dimensionality reduction techniques for representation in two dimensions. The absolute frequencies of the digits were 25 0's, 26 1's, 14 2's, 21 4's and 14 4's. For Isomap and LLE we used 10 as the parameter for the neighbourhoods, POST-MDS was fitted with $\kappa = 1.5, \lambda = 7, \nu = 1$. Figure 5 shows the same plots as in Figure 1, but this time with the digits as the label and the OPTICS Cordillera for $k = 10$ and $q = 1$. For all situations $d_{\max} = 1$. They are ordered decreasingly based on the OC value. The maximal possible OC in this situation is 19 for all plots. Note that because we use $k = 10$ the core distance is calculated as the distance to the $10-$th neighbour.

Figure 6 lists the according reachability plots. The highest clusteredness we find for the diffusion map with its 3-4 extremely dense accumulations ($OC = 4.488$, $OC' = 0.236$). This is a very clustered result, but the normed cordillera is not extremely high relative to the maximal clusteredness because with $N = 100$ and $k = 10$ we would need ten accumulations for that. Also those accumulations would have to be equidistant which is clearly not the case as the 2's and 3's are very close to each other. The result is in line with what we would expect however. If we have around three accumulations, we would expect a normalized Cordillera value of around 0.3 in the perfect case for this given k.

The next clustered result is LLE (OC =4.189, OC' =0.22) with three appreciable discrete structure. Note that the LLE clusters are not spherical but the Cordillera picks them up nonetheless.

Isomap leads to the third clustered representation (OC =4.045, OC' =0.213). The reason for the Isomap representation not ranking higher is that there are bridges between the clusters of 0's, 4's, 2's and 1's which makes the real clusters close to density connected in the mapping. This is reflected by the very broad valley in the right half of the reachability plot. Notably, this is the first representation that hints at the real five cluster solution in the reachability plot.

Similar things can be said about the PCA result, which is slightly less clustered than the Isomap result according to the Cordillera (OC =3.845, OC' =0.202), mainly because density within and separation between accumulations is stronger for the Isomap representation (the emphasis and density properties are at work here). Again the reachability plot hints at five clusters.

Next is the POST-MDS result (OC =3.1, OC' =0.163). The lower values compared to the PCA result can be explained by less density within accumulations and that there are two clearly appreciable accumulations with one of them having two more nested accumulations found. This nestedness with a higher density is also what distinguishes the POST-MDS result from the least clustered result, the Sammon mapping (OC =2.548, OC' =0.134).

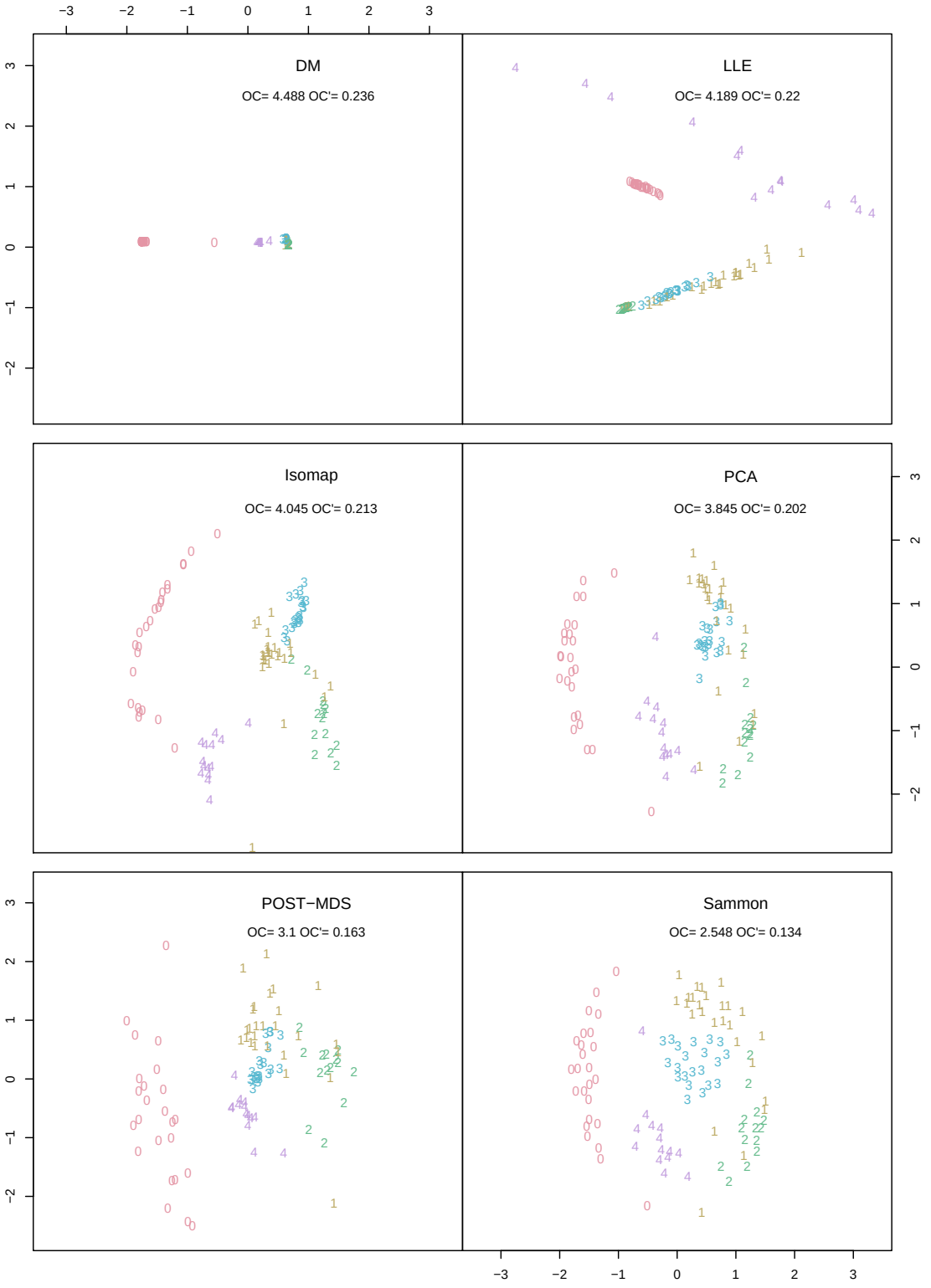

Figure 5: 2D representation of the digits data subset obtained from six different dimensionality reduction techniques from Figure 1 with the digits label and the OPTICS Cordillera values (raw OC and normalized OC', calculated with $k = 10, \epsilon = 10$ and $d_{\max} = 1$). The plots are ordered decreasingly based on the OC' value.

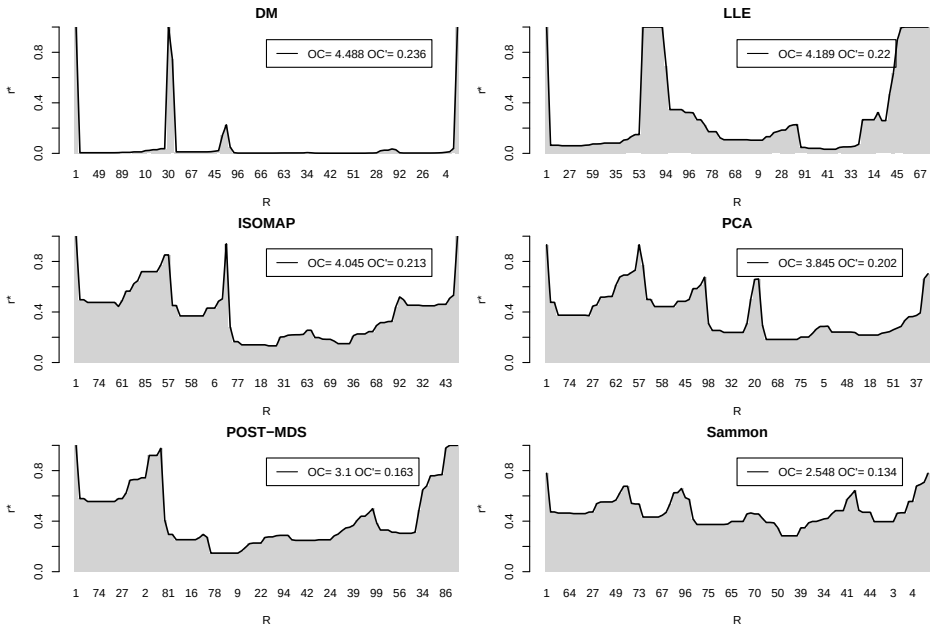

Figure 6: Reachability plots (grey bars) and illustration of the OPTICS Cordillera (black line) for the 2D representations of the digits data subset obtained for six different dimensionality reduction techniques from Figure 5. The OPTICS Cordillera has been calculated with $k = 10, \epsilon = 10$ and $d_{\max} = 1$.

The Sammon mapping produces a configuration that shows little density in the accumulations even though they are more or less clearly appreciable. The core distances with $k = 10$ are rather high in this result, as can be seen in the reachability plot and thus the valleys are not very deep. This reduces the Cordillera value. An important thing to note when looking at the reachability plot is that this is the representation that shows the strongest indication that there are indeed five clusters. However, this is not reflected in the Cordillera value which is independent of any external cluster membership labels.

Table 2 lists values for the OPTICS Cordillera for different parameters k and $q = 1, 2$ to develop some intuition for the change in behavior of the Cordillera for the data representations in Figure 1. For instance, ϵ controls the neighbourhood in which to search for and would not be set to a small value unless in a very noisy setting where one would expect that some represented objects (or clusters) are pure noise. When setting it to 0.5 for our examples (i.e. all objects farther away than 0.5 from the current object would be set to undefined) we find a large reduction in the Cordillera values throughout. In this situation the LLE solution is identified as the most clustered when clusters must comprise at least $k = 10$ points, PCA when accumulations must comprise at least $k = 5$ point and POST-MDS when $k = 2$ suffices. The diffusion map is considered to be unclustered because every accumulation is seen as noise from every other one as they are further apart than 0.5 and all observations coincide in each accumulation. This low ϵ does not allow a chaining of observations beyond the same accumulation. In the LLE and PCA situation there are many more points that are density connected at such a small ϵ.

For a more realistic setting where all points are considered to be real, the first six rows of

Table 2 list the OC and OC' values for different k. We see that DM is indeed most clustered when k is quite high (20, 14 or 10). We use 14 because that is the minimum frequency of any of the digits. As soon as k is reduced however the fact that DM produces 3-4 accumulations whereas other methods produce many more makes the tally property take over—after all with this setup we would be able to produce 20 clusters at most. With $k = 5$ the PCA result identifies around ten accumulations. In general, if k is reduced the OC will increasingly favor representations that show more accumulations. For very low $k = 2, 3$ LLE fares very well because it produces the most accumulations (around 15) of 2-3 points each. These are mainly the 4's at the top of Figure 5 which fall into groups of two to three points.

Changing the q parameter (rows 7–9 in Table 2) also has an effect. The q parameter basically controls how the distances are weighted by means of controlling the norm of the aggregation. A $q > 1$ will give relatively more weight to larger distances whereas a $q < 1$ to the smallest distances. This is why for $q = 2$ and $k = 5$ the DM result is still seen as most clustered and PCA as next to least clustered (the opposite of the situation with $q = 1$) because the valleys in the OPTICS result for DM are very deep and $q > 1$ exaggerates this. For $q = 0.5$ the situation is reversed for both $k = 5$ and $k = 10$. This parameter allows to control how emphasis, density and spread property are combined to produce the index.

The $d_{\max}$ parameter can be used to make the normalized version of the index robust, as illustrated with the rows 10 and 11 in Table 2. The POST-MDS result does produce a number of more outlying points (judged by their minimum reachability) then, for example, the Sammon result. When using a $d_{\max} = 5$ these outliers get more or less full bearing in the normalization and the normalized Cordillera gets larger for POST-MDS. When $d_{\max} = 0.7$ this outlying minimum reachability gets trimmed at 0.7 and thus ignores the distance of these outlying points beyond 0.7. This caps the effect of the spread property on the normalized version and may be easier to use than ϵ which can be used to similar effect.

Lastly, using the Cordillera as a goodness-of-clusteredness measure relative to the largest reachability between any two points can produce different effects. In this case an individual $d_{\max}$ for each representation is used. In Table 2 these are the rows 12–14. Here POST-MDS and DM are the farthest away from being maximally clustered relative to the largest minimum reachability attainable. For $k = 10$ and $k = 5$ the PCA result comes closest to the maximal clusteredness possible given its highest minimum reachability and for $k = 2$ the Sammon mapping shines. Note that when using different $d_{\max}$ the results are not comparable between configurations.

5. Computational Details and Software

We implemented the functionality described above in R. The OPTICS Cordillera can be calculated with the function `cordillera` and takes as arguments a matrix of coordinates. This function relies on an implementation of OPTICS that provides an interface to the OPTICS implementation in ELKI. It further takes the arguments, `minpts` the minimum number of points (k, defaults to 2), `epsilon` (ϵ, which is optional and by default set to the range of the data matrix), `q` (defaulting to 1), a flag `scale` controlling whether the columns of the matrix should be standardized and `rang` which gives the range to which to normalize as $(0, d_{\max})$. For `rang=NULL`, the maximum defined reachability in R is used for $d_{\max}$. There are a `print`, `summary` and `plot` method for objects returned from `cordillera`.

Parameters				PCA		Sammon		LLE		Isomap		Power Stress MDS		Diffusion Map	
k	$d_{\max}$	q	ϵ	raw	normed	raw	normed	raw	normed	raw	normed	raw	normed	raw	normed
10	1	1	10	3.845	0.202	2.548	0.134	4.189	0.22	4.045	0.213	3.1	0.163	4.488	0.236
14	1	1	10	2.972	0.198	2.049	0.137	3.854	0.257	3.196	0.213	1.88	0.125	4.199	0.28
20	1	1	10	1.342	0.149	0.75	0.083	3.208	0.356	1.391	0.155	1.189	0.132	3.942	0.438
5	1	1	10	5.469	0.14	3.785	0.097	4.802	0.123	5.453	0.14	4.583	0.118	4.518	0.116
3	1	1	10	6.743	0.102	5.205	0.079	7.819	0.118	6.688	0.101	6.492	0.098	4.608	0.07
2	1	1	10	10.892	0.11	8.338	0.084	11.856	0.12	9.415	0.095	12.759	0.129	4.713	0.048
10	1	2	10	0.834	0.209	0.272	0.12	2.057	0.329	1.35	0.267	0.716	0.194	3.498	0.429
5	1	2	10	1.291	0.182	0.549	0.119	2.328	0.244	1.702	0.209	1.364	0.187	3.523	0.301
2	1	2	10	2.987	0.174	1.372	0.118	6.403	0.254	2.768	0.167	4.36	0.21	3.536	0.189
10	5	1	10	3.845	0.04	2.548	0.027	6.399	0.067	4.063	0.043	4.834	0.051	12.73	0.134
10	0.7	1	10	3.146	0.237	2.39	0.18	2.989	0.225	2.663	0.2	1.948	0.146	3.288	0.247
10	indiv.	1	10	3.845	0.217	2.548	0.172	6.399	0.204	4.063	0.212	4.834	0.136	22.587	0.12
5	indiv.	1	10	5.469	0.155	3.785	0.141	6.72	0.104	5.471	0.139	6.317	0.087	22.616	0.058
2	indiv.	1	10	10.892	0.122	8.338	0.129	13.881	0.099	9.433	0.094	14.493	0.078	22.812	0.023
10	1	1	0.5	1.547	0.081	0.999	0.053	2.055	0.108	1.523	0.08	1.256	0.066	1.393	0.073
5	1	1	0.5	3.739	0.096	2.942	0.075	2.085	0.053	3.703	0.095	2.218	0.057	1.423	0.036
2	1	1	0.5	9.273	0.094	7.488	0.076	6.598	0.067	8.224	0.083	9.935	0.1	1.618	0.016

Table 2: Different values of the OPTICS Cordillera for different parameter setups for the representations from Figure 1.

6. Conclusion and Discussion

Representations of data like results from unsupervised dimensionality reduction methods are often visually interpreted with respect to if and how clusters of observations form and how well these clusters of observations are visible. To be able to do this demands that the appearance is somehow clustered, i.e., there are some appreciable accumulations of observations. We observed that the judgement of what makes such a result appear clustered hinges on implicit assumptions which can be highly diverse for different people. Therefore, the interpretation ultimately lies in the eyes of the beholder.

To make this task more objective, in this paper we introduced and defined a concept of clusteredness. Clusteredness was defined as a continuum of appearances between a definition for no clusteredness and for maximal clusteredness characterized by a number of aspects used to assess how clustered such results appear. These aspects are that for a number of objects that accumulate together clusteredness increases when the objects cluster closer together, the distances between accumulations increases, the number of accumulations increases and the observations are spread out more.

For this operational definition of clusteredness we suggested an index that quantifies clusteredness. This index, the OPTICS Cordillera, is appealing for measuring clusteredness in data representations. It makes very weak assumptions on what can be considered to be a cluster including no assumptions on cluster number, cluster shapes, cluster centroids and is independent of a specific clustering or cluster assignment of observations. We proof that the index represents aspects of clusteredness. The index is parsimonious with the number of mandatory parameters but also includes optional parameters that allow to tune the index to different needs including making the index robust to noise points or weighting the aspects of clusteredness differently. We further derived bounds for the index and use them to normalize the index.

For a single data representation the index can be used as a descriptive goodness-of-clusteredness statistic, e.g., to assess and quantify how close the result is to displaying no clusteredness or maximal clusteredness, or to assess the change of clusteredness relative to different cluster sizes or cluster density specifications. For more representations, the index can be used to compare them with respect to their clustered appearance, e.g., ranking different results or assessing the change in ranking for different specification of cluster size or density. The OPTICS Cordillera can be also be used in augmenting dimensionality reduction loss functions to improve the clusteredness of the result or for structure-based parameter selection in parametrized dimensionality reduction methods like Isomap or POST-MDS. The OC also has its limitations. It was developed for use in exploratory settings and in conjunction with unsupervised procedures, so particularly when a decision on what the actual clusters are has to be made or when the real cluster labels are available, different measures can be more appropriate.

The OPTICS Cordillera was developed and presented primarily for use with data representations obtained from unsupervised dimensionality reduction results but is not necessarily limited to that. It is a versatile and flexible index to gauge the structure of clusteredness which might be of interest in other contexts as well where the tendency of vectors to accumulate in a space should be assessed. One such case could be astronomy, where one would want to assess the arrangement of stars in galaxies, or in neuroscience, where it might be of interest to find out how the activation pattern of neurons in a brain is clustered for different tasks.

A. Proofs of Bounds for the OPTICS Cordillera

Proof of Proposition 1 (Bounds for the OPTICS Cordillera). Proposition 1 can be shown by establishing the upper and lower bound of $\mathrm{OC}(X;\epsilon,k,q)$.

For the lower bound observe that if $d_{ij}(X) \geq 0$ then $\mathrm{OC}(X;\epsilon,k,q) \geq 0$. This follows directly from the fact that if $d_{ij}(X) \geq 0$ then the definitions (2), (3) and (6) mean that $r_i^* \geq 0 \forall i$. From the definition of $\mathrm{OC}(X;\epsilon,k,q)$ in (7) the left hand side in Proposition 1 follows. □

For the upper bound of $\mathrm{OC}(X;\epsilon,k,q)$ we can use the definition of maximal c-clusteredness as in (1). For that the corresponding Cordillera must look like in the last row of Figure 4 and thus we need to count the maximum possible number, s, of accumulations of observations with $r_i^* = 0$ as for each of these accumulations 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 maximal c-clusteredness we have

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

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

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

B. Details on Properties of the OPTICS Cordillera

The OPTICS Cordillera index in Definition 7 has several intuitively appealing properties for measuring clusteredness corresponding to the aspects that make up the visual appearance of

c-clusteredness. They can be shown by looking at conditions under which the norm of the differences of the minimum reachabilities in (7) does not decrease or increases. We establish them as propositions with proofs.

Notation. We need some (non-standard) notation that we introduce here. We subsequently 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. 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 refer to the data matrix X. If we talk about the result returned from OPTICS, the ordering $x_{(1)}, \ldots, x_{(N)}$ on the abscissa and the corresponding smallest reachabilities $r_{(s)}^*$ on the ordinate, we use the parenthesized subscript $x_{(s)}$ or $r_{(s)}^*$.

We assume that any minimum reachability $r_{(s)}^{*(g)} \leq d_{\max}, \forall s$. Processing $X^{(g)}$ by Algorithm 1 leads to the augmented ordering $R^{(g)}$ which can graphically be displayed as a reachability plot with "valleys" and "peaks" (see Figure 6). From Definition 6 each valley can be interpreted as an accumulation (or cluster). A valley is a set of sequences of points in $R^{(g)}$ that have corresponding minimum reachabilities $r_{(b-t_1)}^{*(g)}$ and $r_{(b+t_2)}^{*(g)}, t_1 = 0, 1, \ldots, T_1;\ t_2 = 0, 1, \ldots, T_2;\ T_1 + T_2 = k - 1$. We will denote a valley 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"). We denote the bottom point of the valley we currently look at 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 $x_{(b)}$. In what follows we need usually only consider a single valley, so we can skip without loss of generality any reference to what actual valley we look at. 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 $r_{(b+t_2)}^{*(g)} \geq r_{(b+t_2-1)}^{*(g)} \geq r_{(b)}^*, \forall t_1, t_2$. Thus $x_{(b)}^{(g)}$ 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 ordering of $u = s^{(g)}(x_u^{(g)}, R^{(g)})$ and $l = s^{(g)}(x_l^{(g)}, R^{(g)})$ 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. These are the "peaks". Each point in $X^{(g)}$ belongs either to a single valley or is a peak.

Then the OPTICS Cordillera exhibits the following properties:

Proposition 2 (Emphasis property)**.** Let $X^{(1)}$ be a configuration that produces OPTICS ordering $R^{(1)}$. Let $x_j^{(1)}$ be a row vector in $X^{(1)}$. Let $C^{(1)}(x_j^{(1)})$ be the cluster in $X^{(1)}$ to which $x_j^{(1)}$ belongs. Here k is so that $\text{card}\left(C^{(1)}(x_j^{(1)})\right) = k$. Let us shift all vectors in $C^{(1)}(x_j^{(1)})$ 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 $\mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(X^{(1)})$. 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 $\left|r^{*(2)}_{(l)}\right| + \left|r^{*(2)}_{(u)}\right| = \left|r^{*(1)}_{(l)}\right| + \left|r^{*(1)}_{(u)}\right|$.

Proof of Proposition 2 (Emphasis Property). Given the setup in Proposition 2, $X^{(1)}$ and $X^{(2)}$ are identical apart from the vector positions in cluster $C^{(1)}(x^{(1)}_j)$ and $C^{(2)}(x^{(2)}_j)$. The distances between the vectors within $C^{(2)}(x^{(2)}_j)$ stay constant, so they are the same as in $C^{(1)}(x^{(1)}_j)$. Since $C^{(1)}(x^{(1)}_j)$ 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-clusters of the same configuration has not decreased, so for $g = 1, 2$ and $\forall\ x^{(g)}_s, x^{(g)}_t : x^{(g)}_s \in C^{(g)}(x^{(g)}_j), x^{(g)}_t \in C^{(g)}(x^{(g)}_i), C^{(g)}(x^{(g)}_j) \cap C^{(g)}(x^{(g)}_i) = \emptyset$ it holds that

$$d(x^{(2)}_s, x^{(2)}_t) \geq d(x^{(1)}_s, x^{(1)}_t). \tag{11}$$

We look only at a single shifted cluster and its corresponding valley. Let $x^{(g)}_b$ be the bottom point in the valley $V(x^{(g)}_b)$ that corresponds to the shifted cluster $C^{(g)}(x^{(g)}_j) = C^{(g)}(x^{(g)}_b)$. Let its position in the ordering be at (b) and denote by (l) and (u) the positions of the peaks $x^{(g)}_l$ and $x^{(g)}_u$ that border on $V(x^{(g)}_b)$. Since $R^{(1)} = R^{(2)}$ and from the nondecreasing distance in (11) between points in non-overlapping clusters, it follows that the difference in reachabilities of the peaks and the bottom increase or stay constant when comparing the shifted cluster to its non-shifted counterpart,

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

and therefore from the definition of the Cordillera as a norm of differences of minimum reachabilities (7), it follows that $\mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(X^{(1)})$. 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, $|r^{*(2)}_{(l)}| + |r^{*(2)}_{(u)}| = |r^{*(1)}_{(l)}| + |r^{*(1)}_{(u)}|$. □

Proposition 3 (Density property)**.** Let $X^{(1)}$ and $X^{(2)}$ be two configurations with the same number of observations that produce OPTICS orderings $R^{(1)} = R^{(2)}$. Let $C^{(1)}(x^{(1)}_j)$ and $C^{(2)}(x^{(2)}_j)$ be corresponding clusters 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 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 get shrunk together towards $x^{(g)}_{(b)}$ 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 peak. Formally we express this as letting points $x_s^{(2)} \in C^{(2)}(x_b^{(2)}), s \neq b$ be moved by positive increments $a_s > 0$ from their position in $X^{(2)}$ towards $x_b^{(2)}$ (if that is geometrically possible) and let these increments be monotonically related to the minimum reachability of $x_s^{(g)}$ and $x_t^{(g)}$, so that $r_s^{*(g)} - a_s \geq r_t^{*(g)} - a_t$ if $r_s^{(g)} \geq r_t^{(g)}$ and so that the ordering in does not change i.e., $R^{(2)} = R^{(1)}$. Given this, we have $\mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(X^{(1)})$. Equality holds only if $\left|r_{(b)}^{*(2)} - r_{(b)}^{*(1)}\right| = \left|(r_{(l)}^{*(2)} + r_{(u)}^{*(2)}) - (r_{(l)}^{*(1)} + r_{(u)}^{*(1)})\right|$.

Proof of Proposition 3 (Density Property). In the setup of Proposition 3, the distances of the points in $C^{(2)}(x_b^{(2)})$ are reduced over these in $C^{(1)}(x_b^{(1)})$ by positive amounts a_s, so

$$d(x_s^{(2)}, x_t^{(2)}) \leq d(x_s^{(1)}, x_t^{(1)}), \tag{12}$$

for $x_s^{(2)}, x_t^{(2)} \in C^{(2)}(x_b^{(2)})$ and $x_s^{(1)}, x_t^{(1)} \in C^{(1)}(x_b^{(1)})$ respectively. From the definition of core distance (2) and reachability distance (3) it follows that for points in this cluster and the corresponding valley in $R^{(1)} = R^{(2)}$, $r_s^{*(2)} \leq r_s^{*(1)}$. Let the indices of points in valley $V(x_b^{(g)})$ in the ordering be $(b-T_1), (b-T_1+1), \ldots, (b), (b+1), \ldots, (b+T_2-1), (b+T_2)$ with $x_b^{(g)} = x_{(b)}^{(g)}$ and denote by $(b+T_2+1) = (u)$ and $(b-T_1-1) = (l)$ the order in $R^{(g)}$ of an $x_l^{(g)}$ and $x_u^{(g)}$ bordering on the valley (the peaks). Due to the conditions on the increments a_s, the difference between reachabilities of two successive points in the valley remains constant or shrinks, so

$$\begin{aligned} |r_{(b-t_1)}^{*(2)} - r_{(b-t_1+1)}^{*(2)}| &\leq |r_{(b-t_1)}^{*(1)} - r_{(b-t_1+1)}^{*(1)}|, \quad t_1 = 0, \ldots, T_1, \\ |r_{(b+t_2)}^{*(2)} - r_{(b+t_2-1)}^{*(2)}| &\leq |r_{(b+t_2)}^{*(1)} - r_{(b+t_2-1)}^{*(1)}|, \quad t_2 = 0, \ldots, T_2. \end{aligned} \tag{13}$$

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

$$\begin{aligned} |r_{(l)}^{*(2)} - r_{(b-T_1)}^{*(2)}| &\geq |r_{(l)}^{*(1)} - r_{(b-T_1)}^{*(1)}|, \\ |r_{(u)}^{*(2)} - r_{(b+T_2)}^{*(2)}| &\geq |r_{(u)}^{*(1)} - r_{(b+T_2)}^{*(1)}|. \end{aligned} \tag{14}$$

From the definition of the OPTICS Cordillera (7) as a norm of differences of r_j^*'s, 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 $r_{(b)}^{*(g)}$ for the smallest reachability and the reachabilities of the bordering peaks are $r_{(u)}^{*(g)}$ and $r_{(l)}^{*(g)}$. Utilizing (12-14), for them it holds that

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

and following from (14) and (7), 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 $|r_{(b)}^{*(2)} - r_{(b)}^{*(1)}| = |(r_{(l)}^{*(2)} + r_{(u)}^{*(2)}) - (r_{(l)}^{*(1)} + r_{(u)}^{*(1)})|$. □

Proposition 4 (Tally poperty)**.** Let $X^{(1)}$ and $X^{(2)}$ be two configurations with the same number of observations that produce OPTICS orderings $R^{(1)} = R^{(2)}$. Let $x_j^{(1)}, x_j^{(2)}$ be corresponding row vectors in $X^{(1)}, X^{(2)}$ respectively. Let $C^{(1)}(x_j^{(1)})$ and $C^{(2)}(x_j^{(2)})$ be corresponding clusters around a point $x_j^{(g)}$ in both configurations, with respective valleys in the reachability plot of $V(x_j^{(1)}), V(x_j^{(2)})$. Let us add E new observations to $X^{(1)}$ and $X^{(2)}$, $\widetilde{x}_e^{(g)}, e = 1, \ldots, E$. For $X^{(1)}$ the points are added to existing clusters so that $\widetilde{C}^{(1)}(x_j^{(1)}) = C^{(1)}(x_j^{(1)}) \cup \widetilde{x}_e$ and the distance of the new points to $x_b^{(1)}$ is not larger than any other distances of points in $C^{(1)}(x_b^{(1)})$ to $x_b^{(1)}$, i.e., $d(x_b^{(1)}, \widetilde{x}_e^{(1)}) \leq \max d(x_s^{(1)}, x_j^{(1)}), x_s \in C^{(1)}(x_b^{(1)})$. We denote the resulting new configuration with $\widetilde{X}^{(1)}$, its ordering with $\widetilde{R}^{(1)}$. Let the point $x_b^{(1)} = x_{(b)}^{(1)}$ be the bottom point in the valley $V(x_b^{(1)})$ to which the points were added and thus the point with smallest minimum reachability of any point in the valley, with $r_{(b)}^{*(1)} = \min_j r_j^{*(1)}, j : x_j^{(1)} \in V(x_b^{(1)})$. For $X^{(2)}$ we add E new observations $\widetilde{x}_e^{(2)}, e = 1, \ldots, E$ so that they form a new cluster around one of the new observations, denoted by $\widetilde{C}^{(2)}(\widetilde{x}_b^{(2)})$. The new cluster adds an extra valley $V(\widetilde{x}_b^{(2)})$ to the reachability plot. Here, $\widetilde{x}_b^{(2)}$ is the point with minimal reachability $\widetilde{r}_b^{*(2)}$ 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 clusters $\mathrm{OC}(\widetilde{X}^{(2)}) \geq \mathrm{OC}(\widetilde{X}^{(1)})$ if $\mathrm{OC}(\widetilde{X}^{(2)}) - \mathrm{OC}(X^{(2)}) \geq \mathrm{OC}(\widetilde{X}^{(1)}) - \mathrm{OC}(X^{(1)})$. Equality holds therefore only if the new cluster is at a distance of zero from points in any other cluster.

Proof of Proposition 4 (Tally Property). Let $\min \widetilde{r}_e^{*(1)}$ denote the smallest minimum reachability over all added points $\widetilde{x}_e^{(1)}$. Because of arguments similar to the ones in Proposition 3, namely that per valley only the difference between minimum reachability of the peaks and minimum reachability of the bottom counts, it holds that

$$OC(\widetilde{X}^{(1)}) \geq OC(X^{(1)}), \tag{15}$$

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 valley 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 (2) and (3) the reachabilities for points in $\widetilde{X}^{(2)}$ are all ≥ 0, so

$$OC(\widetilde{X}^{(2)}) \geq OC(X^{(2)}). \tag{16}$$

From (15) and (16) we have $OC(\widetilde{X}^{(2)}) \geq OC(\widetilde{X}^{(1)})$ if

$$OC(\widetilde{X}^{(2)}) - OC(X^{(2)}) \geq OC(\widetilde{X}^{(1)}) - OC(X^{(1)}). \tag{17}$$

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 Proposition 3 then (17) holds if

$$|\widetilde{r}_{(N)}^{*(2)} - \widetilde{r}_{(N+b)}^{*(2)}| + |\widetilde{r}_{(N+b)}^{*(2)} - \widetilde{r}_{(N+E)}^{*(2)}| \geq |r_{(b)}^{*(1)} - \min \widetilde{r}_e^{*(1)}|. \tag{18}$$

This means that the norm 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 smaller than distances between clusters, Proposition 4 follows. □

Proposition 5 (Balance property)**.** 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 $C^{(1)}(x_j^{(1)})$ be the cluster to which $x_j^{(1)}$ belongs. It corresponds to a given valley in the reachability plot, $V(x_j^{(1)})$. As in Proposition 4 $x_b^{(1)} = x_{(b)}^{(1)}$ 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 $C^{(2)}(x_j^{(2)})$ be the cluster in $X^{(2)}$ to which $x_j^{(2)}$ belongs and $V(x_j^{(2)})$ be the corresponding. 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 $C^{(2)}(x_b^{(2)})$ with valley $V(x_b^{(2)})$ and that $C^{(1)}(x_b^{(1)}) = C^{(2)}(x_b^{(2)}) \setminus x_{N+1}$. The minimum reachability of x_{N+1} is denoted by r_{N+1}^*. Given this, we have that $\mathrm{OC}(X^{(2)}) \leq \mathrm{OC}(X^{(1)})$.

Proof of Proposition 5 (Balance Property). With the setup in Proposition 5 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 so from the definition of the cordillera as a norm of differences of these reachabilities (7) we have $\mathrm{OC}(X^{(2)}) \leq \mathrm{OC}(X^{(1)})$. □

Proposition 6 (Spread property)**.** 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 $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$. If $x_j^{(1)}$ is not a peak, then we have $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$ for $a > \max\left(|r_{(s)}^{*(1)} - r_{(s-1)}^{*(1)}|, |r_{(s)}^{*(1)} - r_{(s+1)}^{*(1)}|\right)$.

Proof of Proposition 6 (Spread Property). Given the setup in Proposition 6, $X^{(1)}$ 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 (2) and reachability distance (3), it follows that the shifted point $x_j^{(2)}$ has a equal or larger minimum reachability then the corresponding unshifted point $x_j^{(1)}$,

$$r_j^{*(1)} < r_j^{*(2)} \leq r_j^{*(1)} + a. \tag{19}$$

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 can be set to 0 to no effect on the Cordillera). 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 (7) and from (19) we can write for different values of $a > 0$—the actual value depending of the nature of $x_{(N)}^{(1)}$:

$$\begin{aligned} \sum_{s=2}^{N} |r_{(s)}^{*(1)} - r_{(s-1)}^{*(1)}| &= \left(\sum_{s=2}^{N-1} |r_{(s)}^{*(1)} - r_{(s-1)}^{*(1)}| \right) + |r_{(N)}^{*(1)} - r_{(N-1)}^{*(1)}| \\ \leq \sum_{s=2}^{N} |r_{(s)}^{*(2)} - r_{(s-1)}^{*(2)})| &< \left(\sum_{s=2}^{N-1} |r_{(s)}^{*(1)} - r_{(s-1)}^{*(1)}| \right) + |r_{(N)}^{*(1)} + a - r_{(N-1)}^{*(1)}| \end{aligned} \tag{20}$$

and so $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$. The values for a must be so that if $x_{(N)}^{(1)}$ is a peak, then $r_{(N-1)}^{*(1)}, r_{(N+1)}^{*(1)} \leq r_{(N)}^{*(1)}$ and $a > 0$ will suffice for (20) to hold. If $x_{(N)}^{(1)}$ is not a peak, (20) holds for

$$a \geq \max \left(|r_{(N)}^{*(1)} - r_{(N-1)}^{*(1)}|, |r_{(N)}^{*(1)} - r_{(N+1)}^{*(1)}| \right)$$

(this would effectively turn $x_{(N)}^{*(2)}$ into a peak). In both of these cases $\mathrm{OC}(X^{(1)}) < \mathrm{OC}(X^{(2)})$. □

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.

Benzécri JP (1973). *Analyse des Données*. Dunod, Paris.

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.

Caliński T, Harabasz J (1974). "A dendrite method for cluster analysis." *Communications in Statistics-theory and Methods*, **3**(1), 1–27.

Coifman RR, Lafon S, Lee AB, Maggioni M, Nadler B, Warner F, Zucker SW (2005). "Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps." *Proceedings of the National Academy of Sciences of the United States of America*, **102**(21), 7426–7431.

Emond EJ, Mason DW (2002). "A new rank correlation coefficient with application to the consensus ranking problem." *Journal of Multi-Criteria Decision Analysis*, **11**(1), 17–28.

Gervacio SV, Lim YF, Maehara H (2008). "Planar unit-distance graphs having planar unit-distance complement." *Discrete Mathematics*, **308**(10), 1973 – 1984. ISSN 0012-365X. doi:http://dx.doi.org/10.1016/j.disc.2007.04.050. URL http://www.sciencedirect.com/science/article/pii/S0012365X07002841.

Greenacre M (2011). "A simple permutation test for clusteredness." *Technical Report 555*, University Pompeu Fabra, Barcelona, Spain.

Hirschfeld H (1935). "A connection between correlation and contingency." In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pp. 520–524. Cambridge Univ. Press.

Hotelling H (1933). "Analysis of a complex of statistical variables into principal components." *Journal of Educational Psychology*, **24**(6), 417.

Huang H, Liu Y, Yuan M, Marron JS (2014). "Statistical Significance of Clustering using Soft Thresholding." *Journal of Computational and Graphical Statistics*. doi:10.1080/10618600.2014.948179. Forthcoming.

Kaufman L, Rousseeuw PJ (2009). *Finding groups in data: an introduction to cluster analysis*, volume 344. John Wiley & Sons.

Liu Y, Li Z, Xiong H, Gao X, Wu J (2010). "Understanding of internal clustering validation measures." In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pp. 911–916. IEEE.

MacQueen J, *et al.* (1967). "Some methods for classification and analysis of multivariate observations." In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, 14, pp. 281–297.

Pearson K (1901). "On lines and planes of closest fit to systems of points in space." *Philosophical Magazine*, **2**(11), 559–572.

Rousseeuw PJ (1987). "Silhouettes: a graphical aid to the interpretation and validation of cluster analysis." *Journal of computational and applied mathematics*, **20**, 53–65.

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

Rusch T, Mair P, Hornik K (2015). "COPS: Cluster Optimized Proximity Scaling." *Technical Report 2015/1*, 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.

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

Tibshirani R, Walther G (2005). "Cluster validation by prediction strength." *Journal of Computational and Graphical Statistics*, **14**(3), 511–528.

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

Wikipedia (2015). "OPTICS algorithm — Wikipedia, The Free Encyclopedia." [Online; accessed 10-October-2015], URL https://en.wikipedia.org/wiki/OPTICS_algorithm.

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