ISilDR: Isometric Seriation-based Dimensionality Reduction for Visual Cluster Analysis

Rene Cutura, Sophie Sadler, Quynh Quang Ngo, Michaël Aupetit, and Michael Sedlmair

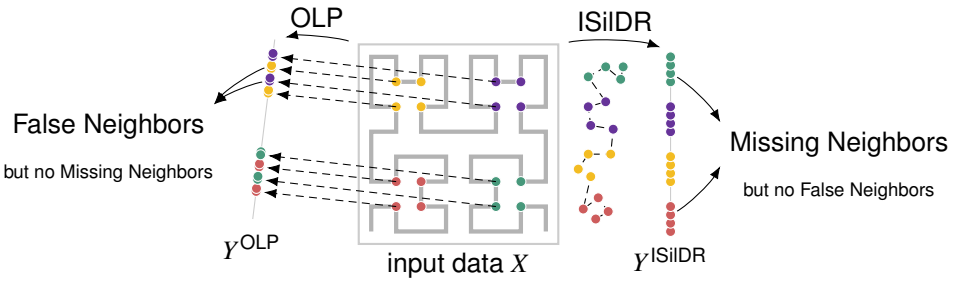


Fig. 1: Distortions occur when projecting multidimensional data to a lower-dimensional space. While orthogonal linear projections (OLP, like Principal Component Analysis or random linear orthogonal projections) can only produce false neighbor distortions, the proposed Isometric Seriation-based projections (ISilDR, like the Hilbert-based or greedy shortest path projections) can only produce missing neighbor distortions. This new category of projection techniques is identified for the first time and their properties studied in this work.

Abstract—Visual cluster analysis is a central task to explore multidimensional data. Dimensionality Reduction (DR) techniques support this task by spatializing multidimensional (MD) data similarities as point patterns in scatterplots. However, unavoidable false and missing neighbor distortions limit their accuracy. For instance, false neighbors make truly separated data clusters appear to overlap in the layout, while missing neighbors split true clusters into falsely separated groups. In general, both types of distortions exist in DR layouts except for orthogonal linear projections (OLP) that only generate false neighbors. In this work, we propose Isometric Seriation-based Dimensionality Reductions (ISilDR) that provably generate at most missing neighbors. We study how ISilDR and OLP together could be leveraged to discover true MD clusters. An ISilDR first creates a seriation of the MD data points, *i.e.*, an ordering along a one-dimensional projection axis, and then each pair of consecutive points along this axis is spaced by their MD distance. An mD ISilDR can be obtained by combining m 1D ISilDRs. We study the theoretical and empirical characteristics of different variants of ISilDRs and OLPs and propose a systematic and formal analysis based on ε-neighborhood graphs. From there, we derive rules to discover cluster patterns in MD data from interactive linking of ISilDR and OLP coordinated layouts. We then conduct case studies and illustrate scenarios for trustworthy visual cluster analysis using a combination of ISilDR and other classical DR techniques.

Index Terms—Dimensionality Reduction, Visual Clustering.

1 Introduction

Visualization can support the analysis of multidimensional data through graphical representations generated by dimensionality reduction (DR) techniques [27]. Existing DR techniques aim to preserve specific data characteristics from the multidimensional (MD) space, such as data density, neighborhoods, or pairwise distances, in the low-dimensional (LD), typically Cartesian, representation space graphically encoded as a scatterplot [28,50,56,62]. However, all DRs suffer from distortions that hinder accurate visual data clustering [53] and trustworthy knowledge discovery [60]. Each DR technique attempts to maintain similarities of the MD input data as distances in the LD projection by minimizing a stress function, i.e., reducing a certain kind of distortion at the expense of another, as in general, not all can be avoided [53]. Thus, increasing distances can lead to *missing neighbors* (MN), and decreasing distances can lead to *false neighbors* (FN).

There are three typical ways to overcome the residual distortions. One can represent the data with several coordinated DR layouts like in a scatterplot matrix (SPLOM) [33], or multiple DRs with different parameters [22, 33]. However, combining multiple distorted views cannot guarantee capturing faithful, distortion-free data characteristics. Another approach enriches the DR layouts with the amount of local distortions to help interpretation [45], but it leaves the user clueless if only a small portion of the layout appears trustworthy. Yet another approach is to enrich the DR layout with indicators of the characteristics of the original data of interest [5,19,49] to see some additional insights about the MD data through their projection. Still, these views are only partial, and capturing the global data patterns, like clusters, is difficult from trustworthy but scattered information.

Although in general non-linear DR layouts have both MN and FN distortions [18, 53, 64], Bessel's inequality [40] proves that orthogonal linear projections (OLP), such as Principal Component Analysis (PCA) [56], can only maintain or decrease the pairwise Euclidean distances of points compared to their MD distances, making their layout *MN-free*. In other words, if points are neighbors in the MD space, they are also neighbors in the PCA projection. But FNs commonly exist in PCA; not all neighbors in the projection are necessarily MD-neighbors.

However, no existing DR technique exhibits only MN distortions, *i.e.*, is *FN-free*, guaranteeing that all neighbors in the projection are

- *Rene Cutura is with University of Stuttgart. E-mail: rene.cutura@visus.uni-stuttgart.de.*
- *Sophie Sadler is with University of Stuttgart. E-mail: sophie.sadler@visus.uni-stuttgart.de.*
- *Quynh Quang Ngo is with University of Stuttgart. E-mail: quynh.ngo@visus.uni-stuttgart.de.*
- *Michaël Aupetit is with Qatar Computing Research Institute. E-mail: maupetit@hbku.edu.qa.*
- *Michael Sedlmair is with University of Stuttgart. E-mail: michael.sedlmair@visus.uni-stuttgart.de.*

Manuscript received xx xxx. 201x; accepted xx xxx. 201x. Date of Publication xx xxx. 201x; date of current version xx xxx. 201x. For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org. Digital Object Identifier: xx.xxxx/TVCG.201x.xxxxxxx

also neighbors in the MD space. Although it may be more challenging to interpret due to MN distortions, we hypothesize that such an FN-free DR could complement MN-free DR to explore and identify the neighborhood patterns of the original multidimensional data with high trustworthiness. For example, brushing and linking a PCA with an FN-free DR layout could help identify visually if the neighbors of a point are true neighbors in the MD space, or identify distortions in other non-linear DRs. It would provide a new way to support trustworthy visual analysis of MD data from DR layouts. Moreover, as cluster patterns depend on pairwise distances, we also hypothesize that overcoming distance-based distortions by linking FN-free with MN-free DR layouts could allow visual inference of the true cluster structure of the data.

Toward this goal, we propose Isometric Seriation-based Dimensionality Reduction (ISilDR) techniques, which can only maintain or increase the pairwise distances in the projection compared with their MD distances. We study their properties and show that they are the first DR techniques able to produce *FN-free* layouts. An overview of our approach is illustrated in Fig. 1. Our contributions are as follows:

- We first define ISilDR based on any ordering, and then orderings derived from the greedy shortest path (**GSP**) and the Hilbert space-filling curve (**HilDR**). As a base, ISilDR is a one-dimensional projection technique. We propose to generate mD projections by assembling m one-dimensional ISilDRs built from m independent subspaces or m different orderings of the MD space.
- We propose a mathematical foundation that provides definitions and theorems with theoretical proofs of concepts related to the *FN-free* characteristic of ISilDR and the *MN-free* characteristics of OLP. From the analysis of their distortion properties, we can derive that OLP can only merge clusters, and ISilDR can only split clusters from MD space to their LD representation; from there, we can determine if a cluster in LD is an actual cluster in MD using brushing and linking among these FN-free and MN-free complementary views.
- We provide an empirical comparative evaluation of the distortion level of OLP and ISilDR on various datasets based on visual quality metrics, such as trustworthiness and continuity [63], that confirm their MN-free and FN-free properties, respectively.
- We combine our theoretical findings and analysis to ground the design of an interactive visual cluster analysis tool. The tool allows us to reliably conclude about the MD data cluster structures by linking ISilDR and OLP layouts. We illustrate it with several case studies of synthetic and real-world datasets. The results of the case studies demonstrate the usefulness of ISilDR in combination with OLP to analyze the distortions of additional DR layouts and to support the identification of real clusters in the data space.
- The code to use ISilDR is publicly available at: `https://github.com/saehm/IsilDR`. See screenshots Appendix Sec. E.

2 Related Work

We focus on reviewing DR work regarding three aspects: (1) DR distortions, (2) Ordering-based and Curve-based DR, related to the seriation of MD points, and (3) DR-based visual cluster analysis.

2.1 Dimensionality Reduction distortions

Dimensionality reduction techniques are prone to distortions [53] that impair the accuracy of data analysis tasks based on DR layout. Different measures of distortions have been proposed. Lee and Verleysen [42] introduced rank-based distortions with intrusions (some points not belonging to the k-Nearest Neighbors (k-NN) of other points in the MD space become k-NN in the projection) and extrusions (some points lose their neighbors). Venna et al. [63] defined LMDS to control distortions based on normalized difference of distances, and introduced trustworthiness and continuity rank-based distortion measures to measure intrusions and extrusions. Rank-based measures are more robust to the shift of distances occurring in MD space due to the curse of the dimensionality [43]. Lee et al. [44] explored how the stress function of tSNE [62] controls the type of distortion at different distance scales. Venna et al. [64] extended the stress of tSNE to create NeRV following the LMDS approach. NeRV enables the control of intrusion (i.e., FN) and extrusion (i.e., MN), linking them to standard supervised learning false positive and false negative errors. Collange et al. [18] defined a DR stress function that controls how MN and FN are tolerated between or within classes of labeled data.

Although the DR stress function can penalize certain types of distortions, there is no guarantee that the resulting projection will be free from that specific type of distortions. The only exception are OLP DR techniques like PCA [56], feature-based scatterplots typical of Scatterplot matrices [33], and random OLP [11] in general, which guarantee to shrink distances from MD to LD space; therefore, they cannot produce missing neighbors and are thus MN-free. However, we are not aware of techniques that could prevent the generation of FN in general. Even NeRV or LMDS, whose stress can be controlled to maximally penalize FN, can still generate some FN as we show empirically in section 3.4. In search of symmetry to OLP, we propose ISilDR as a new family of DR techniques that produce no FN distortions and explore the benefits and limits of such approaches.

2.2 Ordering and Curve-based Dimensionality Reduction

Our approach to FN-free DR leverages the idea of seriation: putting MD points into a 1D order by techniques such as the Hilbert or Gosper curve [30, 34]. Raj and Whitaker [58] considered the MD points' halfspace depth to derive their ordering while in RankVisu [46], the relative orderings of data neighbors are computed. In both cases, the ranking information is used in a modified Multidimensional Scaling stress. Ngo and Linsen [52] used dynamic tSNE [59] for projecting temporal data as a 2D scatterplot, then the user interactively draws a polyline passing through visual cluster patterns formed in the first time step view. This curve serves as a backbone onto which all 2D points are projected at every further time step, forming a sequence of stacked 1D projections. Wulms *et al.* [68] addressed the same problem and used PCA to smooth the transition between 1D slices computed automatically from MD data. In all these approaches, the MDS [41], PCA or tSNE projections, and the orthogonal projection onto a polyline introduce FN distortions in the final layout.

Buchmüller *et al.* [14] proposed a visualization technique for capturing the spatio-temporal patterns of a set of 3D points moving over time. For each time step, the points are seriated into a temporal slice with Hilbert curves. Zhou *et al.* [70] derived Hamiltonian paths over multi-scale Hilbert curves by optimizing the similarity of 2D or 3D data attributes and their spatial location to achieve a 1D linearization of the MD data. Anders [3] derives a 1D projection by projecting the MD data onto an MD Hilbert curve [34], which is then reshaped into a 2D Hilbert curve to form a heatmap representation. In all these cases, the linearization is constructed by using the Hilbert integer indices as axis coordinates. In contrast to these space-filling curve approaches, we go beyond integer index, injecting MD pairwise distances between adjacent points into the 1D projection, leading to an FN-free DR.

2.3 Visual Cluster Analysis using DRs

Our approach's main benefit is its ability to support reliable cluster analysis of MD data. We can cluster MD data with automatic clustering techniques, but the results are difficult to evaluate [38]. Practically, end-users resort to visualizing the data using DR techniques to interpret clustering results or to initiate and control it using interactive approaches [67]. For instance, Darwish *et al.* [24] use mean-shift clustering on top of tSNE layouts to discover political stances. Bonakala *et al.* [12] propose an interactive visual clustering method using a class-based scatterplot matrix called ClassMat [7]. Other tools [13, 15, 65] combine DR techniques with automatic clustering.

An issue with hybrid approaches is that the perception of visual clusters does not match with automatic clustering results even for counting simple cluster patterns in 2D scatterplots [9], and common cluster validation indices (CVI) are not accurate for evaluating visual clustering quality [10]. Xia *et al.* [69] studied visual clustering-related tasks together with DR variants and concluded that each task requires different DR characteristics, with not all DR techniques being accurate

for all tasks, supporting the conclusion stemming from the typology of tasks with DR techniques proposed by Nonato and Aupetit [53]. Essentially, clustering distortions accumulate with the DR distortions in hybrid visual clustering approaches. To tackle this issue and find DR techniques more reliable for visual clustering tasks, Jeon *et al.* [39] built a quality measure that compares CVI in the MD space and the DR layout, and a brushing technique locally correcting DR distortions [37]. Topomap [26] uses a graph-drawing approach further studied by Paulovich *et al.* [55] to display the minimum-spanning tree (MST) of the MD data in the LD projection, preserving a topological invariant essential for clustering. Indeed, the MST is the data structure used by agglomerative clustering with single-linkage [31] to generate hierarchical clusters. We will investigate further in that direction by considering the ε-neighbor graph as a base structure to determine neighborhoods and compare clusters across different spaces and projections.

Other approaches consider multiple complementary projections to visualize the data through different angles and get insights into the MD data. For instance, scatterplot matrices (SPLOM) [33] display all combinations of feature-based scatterplots organized in a matrix where plots in each row or column share the same y-axis or x-axis respectively. In VisCoDeR [22], different DR layouts are linked and coordinated, while in Compadre [21], two DR layouts are linked to the MD data similarity heatmap for comparison and cross-validation of visual cluster patterns. However, in general, linking distorted DR layouts cannot guarantee that visual clusters correspond to MD clusters. We aim to tackle this issue by developing a reliable visual cluster analysis technique that links DR layouts standing at the two end-points of the distortion spectrum.

3 Isometric Seriation-based Dimensionality Reduction

In this section, we introduce our approach called Isometric Seriation-based Dimensionality Reduction (ISilDR), and its properties. Our idea is motivated by the characteristics of Orthogonal Linear Projections (OLP) methods. OLP [20] such as PCA [56] are popular DR techniques that are easy to interpret thanks to their linearity. A main characteristic of OLP is that pairwise distances never increase from MD to LD [40]. One could think of OLP as lighting with a flashlight from a direction and looking at the shadow, where the shadow is the outcome of the projection. This way, no input point can move outside of the neighborhood where it belongs, so that separate visual clusters come from separated MD areas. However, individual visual clusters may come from the overlap of multiple MD clusters. To overcome this issue, we look for a DR approach with the opposite characteristic of never decreasing distances from MD to LD.

We get inspiration from previous work using a greedy shortest path heuristic to seriate multidimensional data [6, 25], and from the locality-preserving property of Hilbert space-filling curves [47, 51]. We first present how we achieve 1D projections by using arbitrary orderings of the input data. Next, we show how to create 1D and mD projections with specific orderings and discuss their theoretical and empirical properties in terms of cluster preservation. Fig. 1 illustrates this approach.

Notations: Let $X \in \mathbb{R}^{N\times M}$ be N input points with M features, and $x_i \in \mathbb{R}^M$ the i^{th} row of X. The space spanned by X is X-space. We note $Y^{tech} \in \mathbb{R}^{N\times m}$ the data projected in the m-dimensional space by technique *tech*. When we consider feature spaces, we let X^F be the column subset of X corresponding to the subset of features $F \subset \{1,\dots,M\}$ and Y^F the projection of X^F. δ denotes the Euclidean distance. $\mathbf{X}_{i,j}$ and $\mathbf{Y}_{i,j}$ are the short forms for $\delta(x_i,x_j)$ and $\delta(y_i,y_j)$, respectively. Finally, $I = \{1,\dots,N\}$ is the index set, π is a permutation of I, and C_i^Z is the i^{th} partitioning cluster of set Z.

3.1 1D Isometric Seriation-based DR

A 1D ISilDR is obtained from an ordering π of points X^F called a seriation:

Definition 1 (1D-**ISilDR**)**.** Let π be any index permutation of the data X, where $x_{\pi(1)}$ is the first point of the seriation. Then, the 1D ISilDR of X relative to π and a column subset $F \subset \{1,\dots,M\}$, is the column vector $Y_\pi^F = [y^F_{\pi(1)},\dots,y^F_{\pi(N)}]^\top$ constructed as follows:

$$y^F_{\pi(1)} = 0, \quad \forall k > 1, y^F_{\pi(k)} = \sum_{j=2}^{k} \mathbf{X}^F_{\pi(j-i),\pi(j)}$$

We take points in the subspace X^F as per their permuted index π, and map them one by one on the real directed line. The first point is mapped to 0, and each following point $x_{\pi(j)}$ is mapped next to the preceding one $x_{\pi(j-1)}$ in the positive direction at a distance equal to their original distance $\mathbf{X}^F_{\pi(j-i),\pi(j)}$. The number of distinct permutations π noted $|\Pi^{ISilDR}|$ measures the expressiveness of an ISilDR.

3.2 Hilbert and Greedy-Shortest-Path Seriations

Any ordering of points can be used; for illustrative purposes, we demonstrate our technique with two ordering chosen for their conceptual clarity. We then discuss the characteristics of the resulting ISilDR.

Greedy shortest-path (GSP) With the greedy shortest-path [25] (Fig. 3a), the seriation starts with a random index $\pi(1)$ set to $i \in I$. Then, $\pi(j)$ $(j > 1)$ is the index of the nearest neighbor to $x_{\pi(j-1)}$ among yet unvisited data in X: $\pi(j) = \underset{k\in I\setminus\{\pi(1),\dots,\pi(j-1)\}}{\arg\min} \mathbf{X}_{\pi(j-1),k}$.

Hilbert Space-Filling Curve (HilDR) The Hilbert space-filling curve [34, 47, 51, 66] is a fractal curve that traverses every point of the M-dimensional unit hypercube using a recursive pattern preserving locality: two points with neighboring indices on the curve are always neighbors in the indexed MD space (see [47] for details). It starts with a pattern (in 2D: ⊓) that partitions the initial cell into 2^M cells.

A κ^{th}-order Hilbert curve is constructed by applying this recursion κ times, leading to a map of $[0,1]^M$ into an index $i \in \{1,\dots,(2^M)^\kappa\}$. The Hilbert index depends on the order of the input space dimensions (for example in 2D: ∃). We rescale the bounding box of the data to fit into the unit M-cube and can pick the order κ of the Hilbert curve as the minimum order such that any two distinct data points get a distinct index (we empirically set $\kappa = 8$ in our implementation for its sufficient fine-grained covering of the data points). We call the ISilDR (Def. 1) based on a Hilbert curve **HilDR** (Fig. 3b).

Note that all approaches we know of in the literature which are based on Hilbert or related space filling curves, use directly the Hilbert index as coordinate in the projection space. However, in our case, it is essential to use the distances in MD space. We demonstrate in Fig. 2 the importance of spacing the points by their MD distance to get a more faithful representation of the data with **HilDR**. Indeed, the Hilbert curve can zigzag between three points x_1, x_2, x_3, so that the Euclidean distance $\mathbf{X}_{1,2}$ is larger than $\mathbf{X}_{2,3}$ in the data space, but the Hilbert index-based length $\mathbf{X}'_{1,2}$ is shorter than $\mathbf{X}'_{2,3}$ in the projection. Hence, the relative distance ordering between adjacent pairs is not necessarily preserved with index-based projections. By contrast, using MD distance preserves this ordering for adjacent points in **HilDR**, allowing us to develop the definitions and theorems of Secs. 3.4 and 3.5.

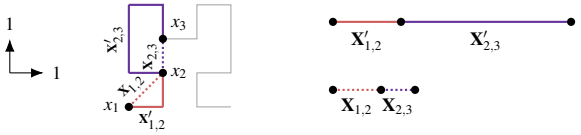


Fig. 2: $\mathbf{X}_{1,2} < \mathbf{X}'_{1,2}$ & $\mathbf{X}_{2,3} < \mathbf{X}'_{2,3}$, but $\mathbf{X}_{1,2} > \mathbf{X}_{2,3}$ & $\mathbf{X}'_{1,2} < \mathbf{X}'_{2,3}$.

This property holds also for random seriations that we call **R-ISilDR**, mirroring OLPs based on random orthogonal directions (**R-OLP**).

3.2.1 Qualitative Comparison of GSP and HilDR

We discuss here the qualitative aspects of **GSP** and **HilDR**.

Input Data Format: GSP can use a data similarity matrix $\mathbf{X}$ as input, while **HilDR** requires an MD feature space. Hence, **GSP** can also

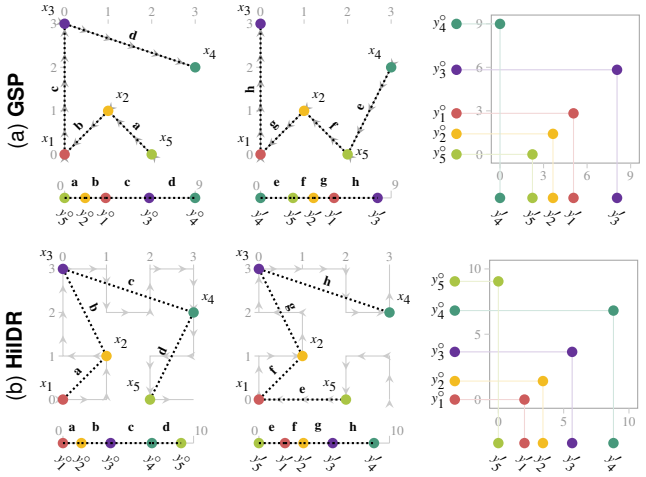


Fig. 3: **GSP** and **HilDR**: Creating two 1D-ISilDR Y° and Y' from 2D data X using different starting points x_1 (**GSP**) or permutations of the data dimensions (**HilDR**). Consecutive points y_j and y_{j+1} in the projection are spaced by their original distance $\mathbf{X}_{j,j+1}$ (a,b,c,d) or (e,f,g,h). The 2D scatterplot (right) is obtained by the Cartesian product of two 1D ISilDRs.

be used for cyclic, spherical or more complex data space topologies depending only on the distance metric used.

Seriation variants: Given a dataset X, **GSP** depends only on the starting point and, therefore, leads to $|\Pi^{\mathbf{GSP}}| = N$ possible distinct 1D ISilDRs. By contrast, **HilDR** is more expressive, leading to $|\Pi^{\mathbf{HilDR}}| = M!$ different 1D projections, largely exceeding typical N even for moderate dimensions ($10! > 3 \cdot 10^6$). While **GSP** seriation only depends on the starting point, **HilDR** seriation depends on the choice of the data box mapped to the unit m-cube including per-dimension scaling, translations, and global rotations.

Computational complexity: The time to assign an MD point to the Hilbert curve is $O(M)$, and to calculate the distance to its preceding point is $O(M)$ [34]. These steps are required for each of the N points, leading to a global $O(MN)$ for **HilDR**. By contrast, finding the next closest point among all other unvisited ones in **GSP** leads to $O(MN^2)$ computing time in total, making it less scalable than **HilDR**. Approximate nearest-neighbor search [48] can be used to reduce that time.

Distortions: **HilDR** has critical locations in the indexed space where neighbors can be torn apart (*e.g.*, the gap between the projection of points x_1 and x_5 in Fig. 3b exists in the Y^o seriation but not in the Y' one). These distortions depend on the relative position of the data in X-space. Since **GSP** is a greedy approach, the last indexed data are more likely far apart than the first ones [25] and not all seriations of the same technique are equally distorted [6], opening the way to seek optimal ones. Notably, in both approaches, consecutive points in the seriation are not necessarily nor likely nearest MD-neighbors.

3.3 mD Isometric Seriation-based DR

If we generate mD projections (Def. 1) using m distinct seriations from the full data X (mD-**ISilDR**$_{\text{FULL}}$), the unwanted gap between neighbors in X formed by one seriation could be bridged by another one. By contrast, m disjoint subspaces of X being mutually orthogonal, if there is no feature correlation between them, we expect their m subspace seriations (mD-**ISilDR**$_{\text{SUB}}$) to capture independent information like would do the m axes of an OLP. We present both options below.

Definition 2 (mD-**ISilDR**$_{\text{FULL}}$ $-$ m seriations of the full X-space)**.** We define the mD projection Y of X, constructed from Def. 1 with m distinct permutations $\pi_k \in \Pi^{tech}, k \in \{1,\ldots,m\}$ as the rescaled concatenation of m column vectors: $Y = m^{-1/2}\left[Y_{\pi_1}, \quad \ldots, \quad Y_{\pi_m}\right]$.

The rescaling factor $m^{-1/2}$ is the inverse diagonal length of the unit m-cube. It is used to achieve desired pairwise distances in the projection as detailed in the supplemental material (Sec. A.1.2, Theorem S2).

Definition 3 (mD-**ISilDR**$_{\text{SUB}}$ $-$ m seriations of disjoint subspaces of X-space)**.** Given m disjoint subsets $F_k \subset \{1,\ldots,M\}$ with their respective orderings π_k, we have $X^{F_k} = \left[X^{F_k(1)}, \quad \ldots, \quad X^{F_k(|F_k|)}\right] \in \mathbb{R}^{N\times|F_k|}$. We compose the m-dimensional projection of X with the corresponding m disjoint subspace projections $Y^{F_k}_{\pi_k}$: $Y = \left[Y^{F_1}_{\pi_1}, \quad \ldots, \quad Y^{F_m}_{\pi_m}\right]$.

We can visualize these mD-ISilDR seriations either as a single scatterplot for $m = 2$ (Figs. 3b, 4 and 8) or a SPLOM for larger m, or as a parallel juxtaposition of m rug plots (Figs. 10d to 10g). The scatterplot representation is counterintuitive, though, as distant points may be actual neighbors (See x_1 and x_5 in Fig. 3b). The parallel design of the m rug plots, instead, breaks the direct visual grouping between adjacent seriations, so we use coordinated view with linking and brushing to detect MD clusters as detailed in Sec. 5.2. As we shall see in the sequel, neighbors in 1D rug plots or 2D scatterplots of mD-ISilDR are always neighbors in the MD space.

3.4 Distance-based Distortions

Dimensionality reduction (DR) techniques alter pairwise distances between data points, resulting in distortions [53] called stretching when distances increase, and compression when distances decrease [5]. Due to the way OLP and ISilDR project data, we can conclude about the occurrence of stretching and compression.

Now, we provide several empirical observations using two common quality evaluation measures: the Shepard diagram [41] to show compression or stretching of all pairwise distances of a projection, and the Trustworthiness and Continuity [36, 39] quality metrics measuring the preservation of the rank of the k-Nearest Neighbors of the data in the projection (see Fig. 4). We compare: random-axis-based OLP (R-OLP); Principal Component Analysis OLP (PCA); LMDS [63] with different settings of λ: $\lambda = 0$ tends to penalize compressions like ISilDR while $\lambda = 1$ tends to penalize stretching like OLP; **GSP**; **HilDR**$_{FULL}$ and **HilDR**$_{SUB}$ (**HilDR** with full and subspace seriations, respectively), and random-seriation-based ISilDR (**R-ISilDR**). PCA, Random Linear Projection (**R-OLP**), and tSNE are computed using Druid$_{\text{JS}}$ [23], and the LMDS is computed with dredviz [54]. We present the results for the 5-dimensional Ecoli dataset [35] [1] in Figs. 4a to 4i. From Figs. 4a to 4i, we show scatterplots for the OLP projections of the dataset (first row), the Shepard diagram of the projection (second row), and the Trustworthiness and Continuity (T&C) measures [63] of the projection with increasing parameter k, as a line chart (third row). The same analysis done for the 4D Iris [29] dataset can be found in Fig. S2 in the supplemental material (Sec. B).

The Shepard diagrams confirm our Theorems 1 and 2 empirically. All of the points in the Shepard diagrams of OLP techniques, such as **R-OLP** and PCA, are below the main diagonal line, i.e., the distances of the original space are always larger than the ones in the projection space (compressions, red dots). Meanwhile, all of the points in the diagrams of ISilDR techniques, such as **GSP**, **HilDR**, and **R-ISilDR** are above the diagonal line, i.e., the distances of the original space are smaller than in the projection space (stretching, blue dots). The diagrams for LMDS include points above and below the main diagonal line, so LMDS stretches and compresses distances in the projection.

We observe a similar pattern with T&C quality measures. The trustworthiness across different neighborhood sizes k (purple line) is lower than continuity (green line) for OLP and higher for ISilDR (purple line above the green line).

OLP and ISilDR act as the two extreme families of techniques along the continuous spectrum of distortions (from left to right), with **R-OLP** and **R-ISilDR** being their most extreme instances. The T&C patterns hint that OLP DRs have a higher tendency (rankwise) to show "wrong" neighbors than misplacing "true" neighbors somewhere else in the projection. In contrast, ISilDR DRs have a higher tendency (rankwise) to misplace "true" neighbors somewhere else than showing "wrong" neighbors in the projection.

[1] We removed the 2 dimensions "lip" and "chg" that have very low variance.

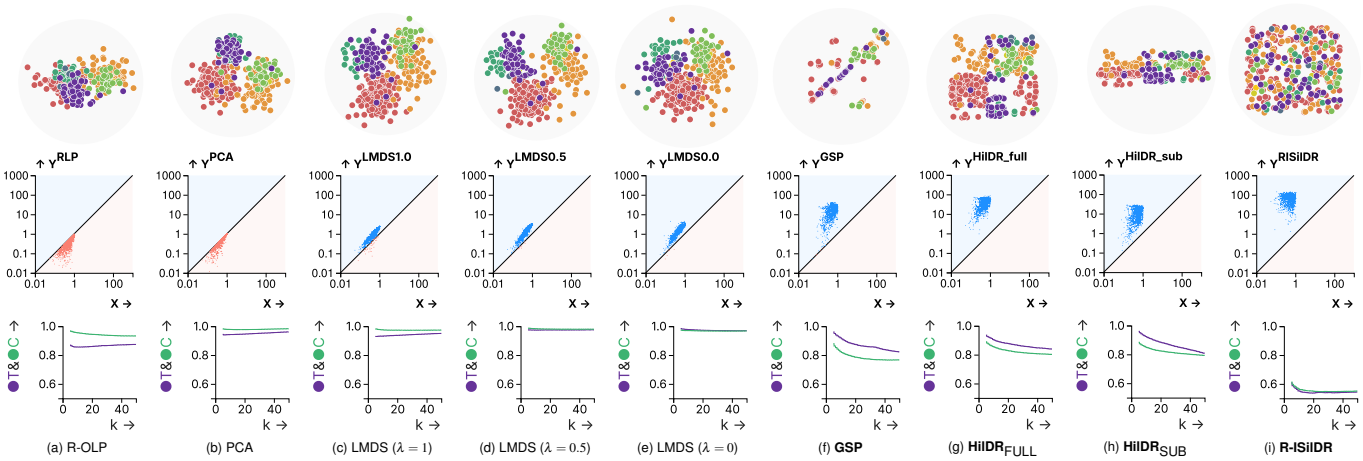


Fig. 4: Projections of the Ecoli [35] dataset (336 data points - 5D space), colored by their labels (Top row). ISilDR plots (**GSP**, **HilDR$_{\text{full}}$**, **R-ISilDR**) concatenate two ISilDR projections to make a 2D-ISilDR, with the exception of **HilDR$_{\text{sub}}$**, where we use 2 disjoint subspace projections. Points on the diagonal (black line) of the Shepard diagram (Middle row) represent unchanged distances, red/blue points mean the distance in Y (LD) is smaller/greater than in X (MD), respectively. Trustworthiness (purple line) and Continuity (green line) (Bottom row) measure the preservation of the k nearest neighbors between X and Y spaces across values of k. These projections cover the distortion spectrum from pure compression (left, 4a and 4b) to pure stretching (right, 4f to 4i, our main discovery).

These results verify empirically the following theorems, proved in the supplemental material (Sec. A).

Theorem 1 (**No stretching with OLP**). In an OLP projection Y^{OLP} of data X, no pairwise distance increases:

$$\forall i,j \in I : \mathbf{Y}^{OLP}_{i,j} \leq \mathbf{X}_{i,j} \tag{1}$$

Theorem 2 (**No compression with ISilDR**). In an ISilDR projection Y^{ISilDR} of data X, no pairwise distance decreases:

$$\forall i,j \in I : \mathbf{Y}^{ISilDR}_{i,j} \geq \mathbf{X}_{i,j} \tag{2}$$

We take advantage of these results to analyze the preservation of neighborhoods and clusters in the next section.

3.5 Interpretation for OLP and ISilDR Projections

Our definition of false and missing neighbor distortions, as well as clusters, stems from the determination of discrete sets from the distance δ and continuous point locations X and Y. Multiple options exist [8]. We consider the ε-neighborhood as a backbone of our framework to formalize OLP and ISilDR capacities regarding cluster preservation, as it is related to a notion of clustering well studied in the literature [26,31]. We first define DR distortions based on the ε-neighbors (Sec. 3.5.1), then the notion of ε-clusters (Sec. 3.5.2), the cluster distortions (Sec. 3.5.3) and how cluster distortions and patterns are linked between the data space and the projection space (Sec. 3.5.4).

3.5.1 False and Missing ε-Neighbor Distortions

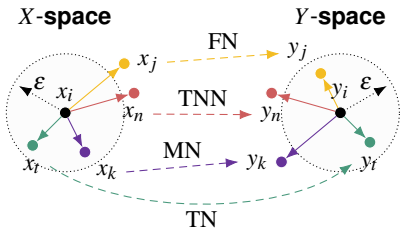


Fig. 5: ε-Neighbor Distortions Def. 4. Considering y_i, y_j is a false neighbor (FN), y_k a missing neighbor (MN), y_t a true neighbor (TN), and y_n a true non-neighbor (TNN).

Various definitions of false (FN) and missing (MN) neighbors are possible depending on the way we set the size of neighborhoods in both data and projection spaces, and the way we measure their difference. We focus on definitions based on ε-neighbors illustrated in Fig. 5:

Definition 4 (ε-**Neighborhood**). For $Z \in \{X,Y\}$, let $U^Z_\varepsilon(i)$ be the set of indices of points in set Z, whose distances to the point z_i are no greater than ε: $U^Z_\varepsilon(i) = \{k \in I | \mathbf{Z}_{k,i} \leq \varepsilon\}$.

Definition 4.1 (**TN**). y_j and y_i are **true neighbors**
$\Leftrightarrow j \in U^Y_\varepsilon(i) \wedge j \in U^X_\varepsilon(i) \Leftrightarrow \mathbf{X}_{i,j} \leq \varepsilon \wedge \mathbf{Y}_{i,j} \leq \varepsilon$.

Definition 4.2 (**TNN**). y_j and y_i are **true non-neighbors**
$\Leftrightarrow j \notin U^Y_\varepsilon(i) \wedge j \notin U^X_\varepsilon(i) \Leftrightarrow \mathbf{X}_{i,j} > \varepsilon \wedge \mathbf{Y}_{i,j} > \varepsilon$.

Definition 4.3 (**MN**). y_j is a **missing neighbor** of y_i
$\Leftrightarrow j \notin U^Y_\varepsilon(i) \wedge j \in U^X_\varepsilon(i) \Leftrightarrow \mathbf{Y}_{i,j} > \varepsilon \geq \mathbf{X}_{i,j}$.

Definition 4.4 (**FN**). y_j is a **false neighbor** of y_i
$\Leftrightarrow j \in U^Y_\varepsilon(i) \wedge j \notin U^X_\varepsilon(i) \Leftrightarrow \mathbf{Y}_{i,j} \leq \varepsilon < \mathbf{X}_{i,j}$.

From these definitions and the previous theorems, we conclude that:

OLP is MN-free Any OLP verifies Theorem 1, contradicting Def. 4.3, hence OLP cannot generate missing neighbors.

ISilDR is FN-free Any ISilDR verifies Theorem 2, contradicting Def. 4.4, hence ISilDR cannot generate false neighbors.

We present alternative definitions and the respective results in the supplemental material (Sec. C).

3.5.2 Cluster Definition Based on ε-Graph

In the sequel, we consider **all clusters are ε-neighbor-based clusters** generated by computing the connected components of the ε-graph [17] connecting any pairwise ε-neighbors. The ε-graph is undirected. Its edges and the ones of the Minimum Spanning Tree (MST) equal or shorter than ε form the same connected components (clusters) as the single-linkage hierarchical clustering [31] cut at ε level. Doraiswamy *et al.* [26] proposed TopoMap as a cluster-preserving DR projecting the connected components of the MST of the data built in the data space, using a custom graph layout technique. Thus, ε-neighbor-based clusters are well-established in the literature. Although they can differ from the clusters which are perceived visually [9,10], our definitions of false and missing neighbors require a distance-based clustering, making ε-neighbor-based clusters well-suited. By contrast, other clusterings (e.g. rank-based) are not compatible with these definitions as demonstrated in Sec. D of the supplemental material. Here, we study how the connectedness of the ε-graphs built in the data space and in the OLP and ISilDR projections match, given their specific properties.

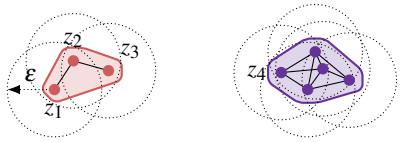


Fig. 6: ε-based Clusters (Sec. 3.5.2). With $Z \in \{X,Y\}$, z_2 is **directly ε-connected** (DεC) to z_1 and z_3 (solid lines), while z_1 and z_3 are **indirectly ε-connected** (IεC). All belong to the same cluster. There is no ε-connection between red and purple points; they are **ε-separated** (εS).

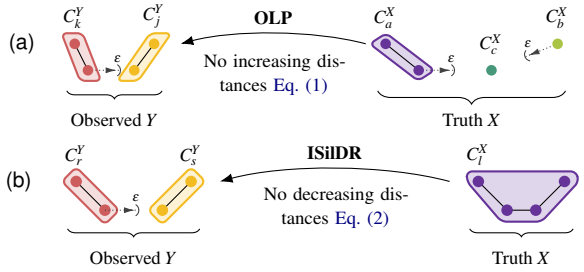


Fig. 7: (a) As the distances can only decrease in an OLP projection, multiple clusters in X may appear merged in the projection. Here, $C_j^Y = C_b^X \cup C_c^X$, is a **False Merge**. (b) As the distances can only increase in an ISilDR projection, a cluster in X may split in the projection. Here, $C_r^Y \cup C_s^Y = C_l^X$, is a **False Split**.

Definition 5 (Within-cluster characterization (DεC, IεC)). Points i and j in $Z \in \{X,Y\}$ are grouped in a cluster C_k^Z if they are either **directly ε-connected** (DεC), i.e. their distance is not longer than ε, or **indirectly ε-connected** (IεC) through an ε-path of multiple directly ε-connected points $p_1,\ldots,p_n$, (see Fig. 6):

$$\begin{aligned} &\forall i \in C_k^Z : \exists j \in C_k^Z, i \neq j : \mathbf{Z}_{i,j} \leq \varepsilon \\ &\text{or } \exists p_1,\ldots,p_n \in C_k^Z, (\mathbf{Z}_{i,p_1} \leq \varepsilon) \wedge (\mathbf{Z}_{p_1,p_2} \leq \varepsilon) \wedge \ldots \wedge (\mathbf{Z}_{p_n,j} \leq \varepsilon) \end{aligned} \tag{3}$$

Definition 6 (Between-cluster characterization (εS)). Points of $Z \in \{X,Y\}$ belong to separate clusters C_k^Z and C_l^Z if they are **ε-separated** (εS), *i.e.* neither directly nor indirectly ε-connected, (see Fig. 6):

$$\forall i \in C_k^Z, \forall j \in C_l^Z, k \neq l : \mathbf{Z}_{i,j} > \varepsilon \tag{4}$$

3.5.3 Clusters Connection and Separation

The ε-clusterings of the same data within two spaces are likely different. We can characterize the quality of the mapping from X to Y by how it preserves ε-clusters. In general, due to properties of OLP and ISilDR, visual (LD) clusters C^Y do not match with MD clusters C^X.

For instance, considering the separation between 2 visual clusters C_1^Y and C_2^Y, a **True Separation** occurs when they come from two disjoint sets of clusters $C_1^Y = C_{a_1}^X \cup \cdots \cup C_{a_n}^X$ and $C_2^Y = C_{a_{n+1}}^X \cup \cdots \cup C_{a_{n+k}}^X$ with $\forall i \neq j, a_i \neq a_j$, while a **True Connection** happens if the points forming a visual cluster C^Y are also ε-connected in X: $C^Y \subseteq C^X$. A **False Merge** (Fig. 7a) occurs when two or more MD clusters $C_1^X \ldots C_n^X$ overlap to form a single visual cluster C^Y ($\forall i, C^Y \supset C_i^X$) which is a super-cluster of any of those C_i^X. A **False Split** (Fig. 7b) happens if the two visual clusters come from a single cluster C^X ($C_1^Y \subset C^X$ and $C_2^Y \subset C^X$) where each of C_1^Y and C_2^Y is a sub-cluster of C^X.

3.5.4 Interpreting Cluster Patterns in OLP and ISilDR

We want to know what true cluster patterns in X can be discovered based on observed cluster patterns in the projection Y^{tech}.

Table 1a shows the 6 possible orderings of the true distance $\mathbf{X}_{i,j}$, the observed one $\mathbf{Y}_{i,j}$, and the clustering cut-off ε for two arbitrary points i and j (red shows compression, blue stretching, and grey mismatch between ε-connections in X and Y).

Table 1b shows that OLP matches with cases above the diagonal (Theorem 1) and ISilDR with cases below it (Theorem 2). It also shows that if we observe two separated points i and j (εS) in Y (first column) with OLP, they must be εS in X too (first row, dark red triangle), while the same pattern observed in Y^{ISilDR} can come from any of the three patterns in X (full column blue). Symmetrically, if we observe a direct connection (DεC) between i and j in Y (third column) with ISilDR, the only option is that the same pattern exists in X (last row, dark blue triangle), while OLP cannot help us decide (full column red). Observing an indirect connection (IεC) in either Y^{OLP} or Y^{ISilDR} (second column) does not help, as two options in X are possible for each.

Table 1: What ε-cluster pattern in X can we discover from observing ε-cluster patterns in Y^{OLP} or Y^{ISilDR} (See Sec. 3.5.4 for details).

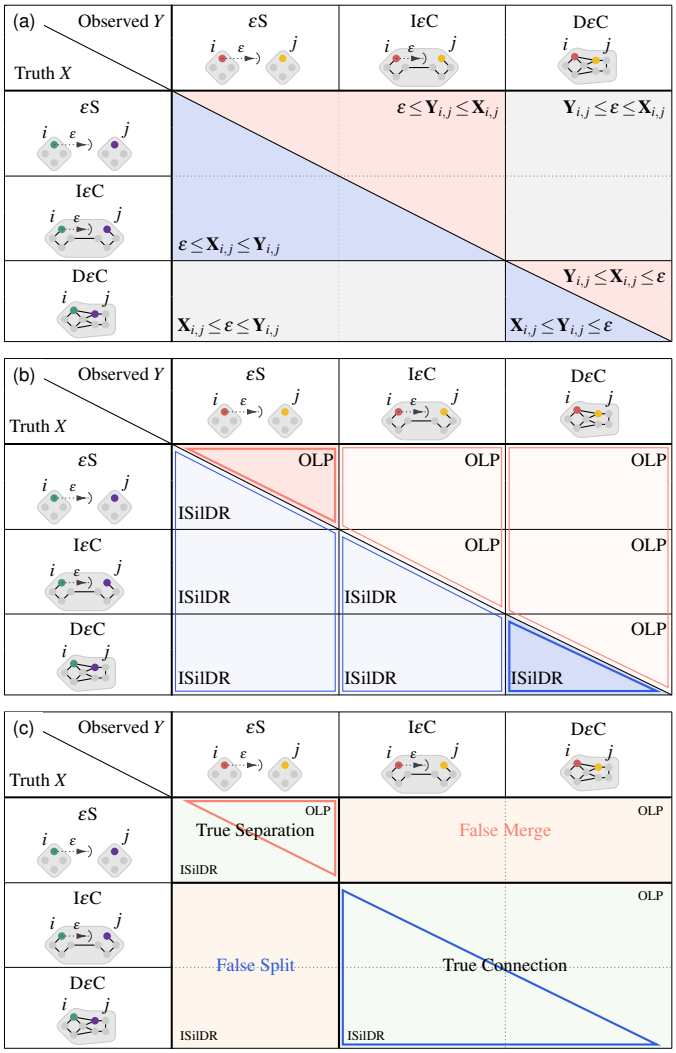


Table 1c considers ε-clusters. As two points indirectly connected (IεC) are actually connected through a path of directly connected points (DεC) within the same ε-cluster (Def. 5), we can be certain that if a DεC path exist in Y^{ISilDR}, this same path exists in X (blue triangle in (b)), so points forming a cluster in Y^{ISilDR} must be all ε-connected in X (True Connection), merging IεC and DεC cases. Two clusters appear εS (Def. 6) if any point i in one is ε-separated from any point j in the other. This pattern happens in Y^{OLP} only if the same pattern exists in X (red triangle) hence OLP can confirm True Separation.

However, True Separation and True Connection do not ensure an observed cluster is a true MD cluster containing exactly the same points.

4 The Interpretation-based Visual Cluster Analysis

We show how ISilDR can be used together with OLP for visual cluster analysis, first through an analytic process built on top of the previous definitions and observations, then through two use cases. We then

Table 2: We first select an ε-neighborhood matching the visual cluster (thick cluster border, purple in Case A, red in Case B) of interest in the master view (Select column), then we observe the corresponding color-coded clusters formed in the secondary view (Observe column), and infer the MD (hidden) ε-clusters (Infer column). The arrows's directions are from MD to LD and arrows's names are assigned with distortion types respectively.

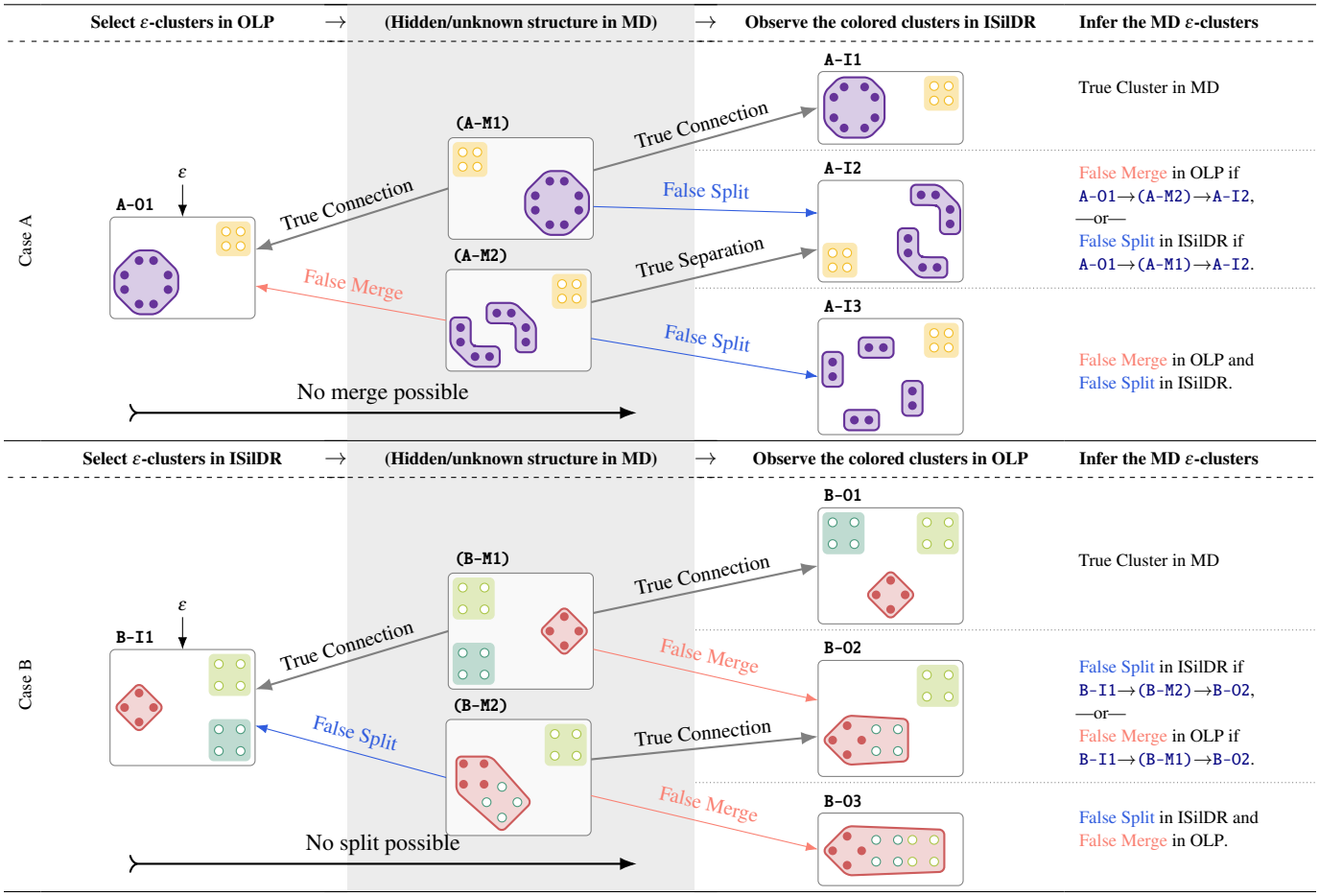


Table 3: Once the true MD ε-clusters have been identified using Case A or Case B, we can evaluate the clusters in another DR layout.

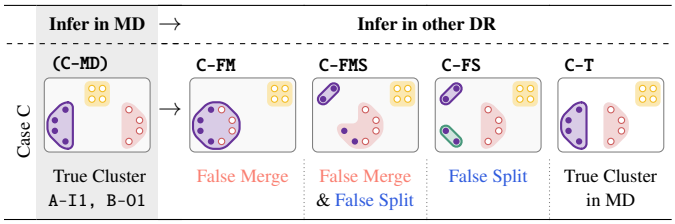


Table 4: Summary of the decision rules from Cases A and B in Table 2

Y^{ISilDR} \ Y^{OLP}	OLP: Two ε-Clusters	OLP: One ε-Cluster
ISilDR: Two ε-Clusters	True ε-Clusters	Uncertain
ISilDR: One ε-Cluster	Uncertain	True ε-Cluster

propose an interactive workflow to support the visual cluster analysis.

4.1 Visual Cluster Analytic Process

We elicit the observations in Table 1 for the ε-cluster analysis of MD data through their OLP and ISilDR visualizations, by proposing three analytic processes described in Table 2 and Table 3. Cases A and B (Table 2) show how visual cluster patterns in OLP and ISilDR can support the inference of the MD cluster structure. They are summarized in Table 4. Case C (Table 3) shows how the inferred MD cluster from Case A or Case B support the qualitative evaluation of other DR layouts.

Case A (Table 2): Spot True Clusters with OLP

A cluster observed in an OLP (filled purple points group in `A-01`) may arise either from a singular, distinct pure cluster `(A-M1)` or from a combination of multiple clusters (`(A-M2)`) (invisible) in the MD data. When this particular cluster first visually identified in OLP aligns with a pure cluster `A-I1` identified in ISilDR, it qualifies as a **True Cluster** within the context of the MD data (`A-01`→`(A-M1)`→`A-I1`). Because OLP can only merge clusters, and ISilDR can only split clusters , a pure cluster in ISilDR and OLP has to stem from a pure cluster in the MD data.

Case B (Table 2): Spot True Clusters with ISilDR

The points of a cluster observed in an ISilDR could stem from a pure `(B-M1)` or from a bigger cluster `(B-M2)` (invisible) in the MD data. When this particular cluster first visually identified in ISilDR aligns with a pure cluster identified in OLP, it qualifies as a **True Cluster** within the context of the MD data (`B-I1`→`(B-M1)`→`B-01`).

Cases A and B are summarized in Table 4. They differ only in which of the OLP or ISilDR view is used to first identify the cluster of interest. Notice that the value of ε used in OLP and ISilDR can be different as far as the points of the cluster of interest match in both spaces.

Case C (Table 3): Identify Wrong Clusters in Arbitrary DR Layout
If we select a true cluster (verified by an OLP and an ISilDR, Cases A & B), then we can find false or missing neighbors and thus **False Merge** or **False Split** in another DR layout.

4.2 Visual Encodings and Analysis Workflow

We propose an analysis workflow with linked views to derive the faithfulness of the clusters visually. We start the analysis with an OLP and an ISilDR projection, one of which will be the *master* view, and the other the *secondary* view. The sole purpose of the master view is to identify the clusters which will be subsequently colored in all views. Both an ISilDR and an OLP must be present to apply Cases A and B.

Cluster selection in master view The user changes ε interactively with a slider to identify an ε-cluster visualized with contour lines, that matches well with a visual cluster of interest in the master view. ε-clusters of the master view are color-coded to identify them in all the other views. For instance, Fig. 10a shows two clusters (red and purple dots) selected in the PCA master view of the Iris dataset with $\varepsilon = 1$.

Note that the same value of ε is used for both projections, to ensure coherent cluster analysis in the MD space. Trivial values such as 0 or a very large number result in identical clustering (respectively N single-point clusters, or one cluster containing all points). Thus, the optimal selection of ε is challenging and should be guided by the user's analytical goals and the specific structures they wish to investigate.

Cluster matching in secondary view ε-clusters are then built in the other views with the same ε and identified through contour lines, but using the colors assigned in the master view. To ease focus on interesting master-view clusters, we allow coloring of all other clusters in grey. The user relies on contours and colors to evaluate the matching between clusters in master and secondary views. The process can be repeated when another view is selected as the master view.

5 Case Studies

Our approach has been mathematically proven reliable to identify true clusters from two uncertain projections. Now, we showcase its usefulness in practice to infer valid MD cluster structures from the observed projections of a synthetic and a real-world dataset.

5.1 A Synthetic Dataset

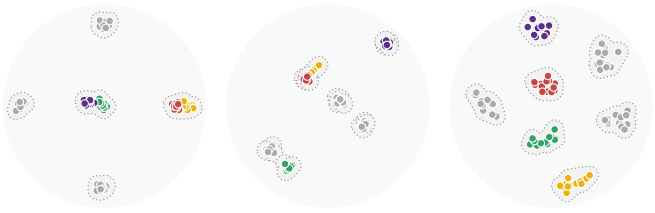


Fig. 8: Projections of a synthetic labeled dataset consisting of 6 Gaussian-like clusters in 3D, with 7 labels (4 are shown as colors, red and yellow actually belong to the same cluster made of two adjacent Gaussians). (a) ε is tuned to match the 5 clusters in PCA (contour lines); one is incorrectly formed from the green and purple points (False Merge). (b) The same ε in **GSP** shows 6 clusters with the correct colors (True Cluster). (c) tSNE (with arbitrary parameterization) shows 7 clusters, the red and yellow points being wrongly separated (False Split). The same ε-cluster (yellow and red) identified in both PCA and **GSP** ensures it is a True MD cluster. We cannot conclude anything certain from tSNE.

We created a 3D dataset with 7 labels for 6 Gaussian-like clusters, in which one cluster has two labels. Fig. 8 shows the outcome of the dataset with PCA, **GSP**, and tSNE. For **GSP** we use two projections with distinct starting points of the full space with a rescaling factor (Def. 2) to create a 2D-**ISilDR**$_{\text{FULL}}$. Meanwhile, for tSNE, we use an arbitrary parameterization; although the visual outcome can change, the faithfulness of the clustering does not depend on these parameters. For illustration, we use the 7 labels to color the points.

First, we select ε to match the 5 visual clusters appearing in the PCA master-view (contour lines). The cluster formed by red and yellow points is a **True Cluster** in the MD space because it appears as a single ε-cluster in both PCA and **GSP** following Case A (Table 2, `A-O1`→`(A-M1)`→`A-I1`). However, we cannot conclude about the cluster observed in PCA formed by purple and green points and split into two clusters in **GSP**: it is either a **False Split** (Table 2, `A-O1`→`(A-M1)`→`A-I2`)) or a **False Merge** (Table 2, `A-O1`→`(A-M2)`→`A-I2`)). False Merge is the correct answer according to the ground-truth, but we cannot infer it from the observations. By contrast, we cannot infer anything certain about MD clusters from tSNE, even paired with PCA or **GSP**. Finally, the discovery of the true yellow-red cluster with PCA and **GSP** ensures that the tSNE view describes the MD data incorrectly, showing a **False Split** of that cluster following Case C (Table 3, `(C-MD)`→`C-FS`).

5.2 Iris

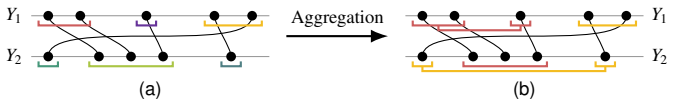
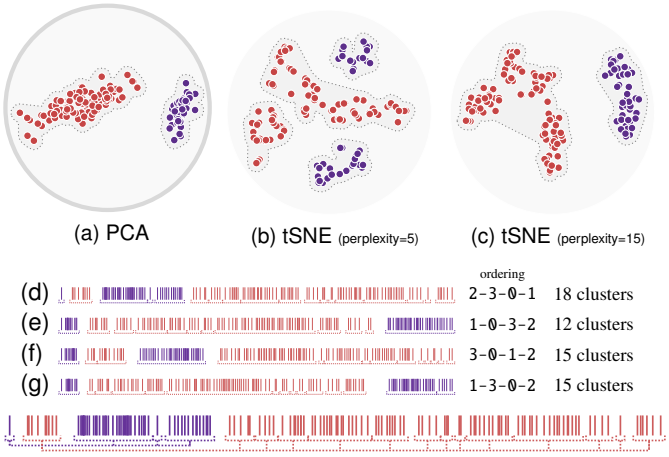


Fig. 9: Cluster aggregation across m ISilDR seriations, and their underneath bracket representation. Each black line connects two projections of the same data point.



(h) Aggregation of 10d to 10g on **HilDR**$_{\text{FULL}}$ projection 10d; 2 clusters.

Fig. 10: Projections of the Iris [29] dataset. We use an ε-graph for the clustering (Sec. 3.5.2) in PCA (10a) with $\varepsilon = 1$. 10d to 10g show individual **HilDR**$_{\text{FULL}}$ projections with their respective ε-clusters ($\varepsilon = 1$) represented as dashed brackets underneath. 10h shows the aggregation of all 4 individual rug plots, whose clusters (dashed connecting brackets) match with the red and purple ε-clusters identified in PCA 10a (full color in Fig. S1), confirming that those two clusters are true ε-clusters in MD.

Iris [29] is a well-known 4D dataset of 150 data points. With the help of **HilDR**$_{\text{FULL}}$ and PCA, we can identify all the true clusters (those shown in Figs. 10a and 10h). Note that although the Iris dataset has 3 classes, these form only 2 ε-clusters, since Versicolor and Virginica are ε-connected for $\varepsilon = 1$.

Using several individual 1D ISilDR projections Def. 1, it is possible to aggregate the cluster information from many seriation at once. Note that we have no way to know a priori which permutation of the dimensions will provide the "best" projection, and therefore we will always benefit from more information by combining several of them. For **GSP** these projections could come from different starting points, or from different permutations for **HilDR**. In the case of Figs. 10d to 10g, we can see **HilDR**$_{\text{FULL}}$ rug plots representing the projections generated from 4 different permutations of the Iris features. The data are colored as per the clusters selected in the PCA master view (Fig. 10a) with

$\varepsilon = 1$. This same ε produces an order of magnitude more ε-clusters (up to 18, see Fig. S1) in each **HilDR** projection (brackets underneath each rug plot in Figs. 10d to 10g), far more than the two (true) ε-clusters present in Iris. Most of these clusters come from False Split, but we know that two points are ε-neighbors in MD space if they belong to the same ε-cluster in *at least one* of the m $Y^{\mathbf{HilDR}}$ projections (Table 1c). Thus, we can aggregate all the pairwise (True) ε-connections from each **HilDR** projection, merging all the clusters falsely split by each seriation. For that, we connect each data point to its duplicates in all other projections (Fig. 9a) and compute the connected components of the union of the m ε-graphs with these added edges. We represent the aggregated clusters with additional connecting brackets underneath (Fig. 9b). Using this process, we aggregate the cluster information from all of Figs. 10d to 10g into the rug plot of Fig. 10d, leading to Fig. 10h, where the two aggregated clusters correspond perfectly to the color-coded clusters identified in PCA (Fig. 10a). Using Case A or B (Table 2, `A-O1`→`(A-M1)`→`A-I1`, or `B-I1`→`(B-M1)`→`B-O1`), we are certain that these two clusters are **True Clusters**, showing the benefit of aggregating several ISilDR to reconnect falsely split clusters.

The first tSNE plot (Fig. 10b) shows a **False Split** for purple and red labels (Table 3, `(C-MD)`→`C-FS`). The second tSNE plot using a different perplexity (Fig. 10c), shows two ε-clusters (contour lines) matching with the PCA ε-clusters (red and purple), but it cannot ensure these are two true MD clusters, and we cannot aggregate cluster information from both tSNE plots either, because of their mix of FN and MN distortions. Any number of tSNE plots cannot help us be certain about the true MD ε-cluster structure.

6 Discussion

Two wrongs making a right Rather counter-intuitively, the use of linked OLP and ISilDR views with extreme distortion patterns allows us to draw firm conclusions about the MD data clusters that even linking multiple views of more efficient DR like tSNE with low MN and FN distortions cannot guarantee despite its wide use by practitioners ("Two wrongs don't make a right"). Notice in particular that we do not need to compute the ε-neighbor graph of the MD data to draw a conclusion about it from the ε-neighbors computed in OLP and ISilDR LD spaces.

The interesting fact that can inspire further research is that we can get correct information (MD clusters) from distorted partial views (OLP and ISilDR), a very desirable property for visualization systems.

Counting ε-neighbor clusters Along this line, we can estimate the number of MD clusters from OLP and ISilDR. Indeed, from Table 2, the deductive reasoning paths (`A-O1`→`(A-M2)`→`A-I3`, or `B-I1`→`(B-M2)`→`B-O3`), show that we cannot conclude about the exact number of MD clusters from observing OLP and ISilDR views. However, as OLP can only merge MD clusters while ISilDR can only split them, the number of ε-clusters in OLP and ISilDR are respectively lower and upper bounds of the number of MD ε-clusters.

7 Limitations and Future Work

Type of False and Missing Neighbors We rely on a metric definition of the neighborhood [8, 53] of a data point (ε-neighborhood) on which our DR distortions and clustering definitions are based. This metric approach is the basis of many exploratory analysis techniques like the agglomerative clustering and the Euclidean Minimum Spanning tree [31], and is also the backbone of modern topological data analysis [16, 26]. We proposed two alternative definitions in the supplemental material, but they do not make OLP and ISilDR MN-free and FN-free, respectively, leaving open interesting research avenues.

Non-metric approaches and dimensional scalability The scalability of our approach to large or high-dimensional datasets is so far limited for the following reasons. Firstly, we use DR techniques (PCA, LMDS, **GSP**, **HilDR**) which minimize Euclidean distance-based stress functionals known to be sensitive to the curse of the dimensionality and the distance concentration phenomenon [43]. Our use cases are also limited to low-dimensional MD spaces because similar ε-neighborhoods are unlikely to capture exactly the same clusters in OLP and ISilDR projection spaces due to their large difference in scales. This is certainly the most challenging limitation that could be addressed by extending our framework to shift-invariant similarity metrics known to be more robust to data dimensionality, such as those used in NeRV [64] or UMAP [50].

Selection of ε Our current workflow incorporates manual, iterative tuning of ε until a clustering match is identified between OLP and ISilDR. In future work, this selection could be automated by computing the ε-clusters in both spaces and identifying a range of values for which the clusterings match exactly.

Soft cluster-matching We set a strong constraint on the clusters which must match between OLP and ISilDR spaces for our decision rules to be valid. However, in practice, clusters from two data representation spaces never match exactly. Could we extend this framework in a mathematically grounded way to soft or probabilistic measures of cluster-matching between OLP and ISilDR spaces? For instance, using external clustering scores [57] or Kulback-Leibler divergence [64]?

Visual Cluster Perception We consider a computational definition of clusters based on the ε-neighbors, allowing strong mathematical results, but the perception of visual clusters may depart from this mathematical definition [9, 10] and the DR distortions can also affect the visual clustering task [37, 39, 53, 69]. Some ways to make computational visual quality measures of cluster patterns closer to human perceptual judgments have been explored [1, 32], too. The graphical encoding of ISilDR seriations as scatterplots or rug plots, as well as the effect of the ε-neighbor-based cluster definition and our interactive visual clustering framework, are important aspects that need further user evaluations.

Optimal Projections A typical characteristic of a DR technique is its capacity to reduce distortions and to project new data points [27, 53]. As we propose a new family of DR techniques, it is natural to consider how ISilDR can implement such features. For instance, what are the consequences of inserting a new data point in an ISilDR seriation on the cluster patterns or in terms of computation? Computing mD-ISilDR requires selecting m axes among many, depending on the expressivity of the ISilDR. This problem has been studied for selecting features of DRs [4] and for optimizing the generation of space-filling curves to seriate voxel data [70]. Which axes shall we select to improve our layout? Which criterion shall we optimize?

8 Conclusion

We present ISilDR—an FN-free dimensionality reduction technique with a reliable visual cluster identification approach when used in combination with an OLP. We provide the mathematical foundation for our technique, including definitions, theorems, and proofs, along with related concepts, to support a trustworthy visual cluster analysis process for knowledge generation [60].

Our framework considers the spectrum of distortions generated by DR techniques, ranging from pure False ε-Neighbor to pure Missing ε-Neighbor distortions. OLPs have long been identified for generating the former, but the existence of the latter was unknown so far. In the pursuit of developing trustworthy visualizations of multi-dimensional data, we propose a family of DR techniques that only produce Missing ε-Neighbors. By exploring this concept and two of its instances (**GSP** and **HilDR**), we set the ground for the development of new visual analytic approaches in the spirit of lower and upper bounds and surrogate functions used in optimization theory and machine learning [2], i.e., solving complex problems using simpler functions enclosing the solution – here, simple linked 2D graphical representations that not only give hints but tell facts about the complex MD data cluster structure.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161, project A08. We acknowledge the contributors to earlier exploratory unpublished versions of this project, started in May 2018 at QCRI: Abdelkader Baggag, Ali Sheharyar, Safiya Jan, Almiqdad ElZein, Julia Ann Jose, Reuben Suju Varghese, and Yash Mathne. M. Aupetit wishes a special thanks to Grant Sanderson, whose `3Blue1Brown` popularization video on Hilbert Space Filling curves [61] sparked the idea of this work.

References

[1] M. M. Abbas, M. Aupetit, M. Sedlmair, and H. Bensmail. ClustMe: A Visual Quality Measure for Ranking Monochrome Scatterplots based on Cluster Patterns. *Computer Graphics Forum*, 38(3):225–236, 7 2019. doi: 10.1111/cgf.13684 9

[2] A. A. Alemi, B. Poole, I. Fischer, J. Dillon, R. A. Saurous, and K. Murphy. Fixing a Broken ELBO. In J. Dy and A. Krause, eds., *Int. Conf. Machine Learning*, vol. 80 of *Proceedings of Machine Learning Research*, pp. 159–168. PMLR, 2018. 9

[3] S. Anders. Visualization of Genomic Data with the Hilbert Curve. *Bioinformatics*, 25(10):1231–1235, 3 2009. doi: 10.1093/bioinformatics/btp152 2

[4] M. Ankerst, S. Berchtold, and D. Keim. Similarity Clustering of Dimensions for an Enhanced Visualization of Multidimensional Data. In *Information Visualization Symp.*, pp. 52–60. IEEE, 10 1998. doi: 10.1109/INFVIS.1998.729559 9

[5] M. Aupetit. Visualizing Distortions and Recovering Topology in Continuous Projection Techniques. *Neurocomputing*, 70(7):1304–1330, 3 2007. Advances in Computational Intelligence and Learning. doi: 10.1016/j.neucom.2006.11.018 1, 4

[6] M. Aupetit. Nearly Homogeneous Multi-partitioning with a Deterministic Generator. *Neurocomputing*, 72(7–9):1379–1389, 3 2009. doi: 10.1016/J.NEUCOM.2008.12.024 3, 4

[7] M. Aupetit, A. Ali, A. Baggag, and H. Bensmail. ClassMat: a Matrix of Small Multiples to Analyze the Topology of Multiclass Multidimensional Data. In *Topological Data Analysis and Visualization*, pp. 70–80. IEEE, 10 2022. doi: 10.1109/TopoInVis57755.2022.00014 2

[8] M. Aupetit and M. Sedlmair. SepMe: 2002 New Visual Separation Measures. In *Pacific Visualization Symp.*, pp. 1–8. IEEE, 5 2016. doi: 10.1109/PACIFICVIS.2016.7465244 5, 9

[9] M. Aupetit, M. Sedlmair, M. Abbas, A. Baggag, and H. Bensmail. Toward Perception-Based Evaluation of Clustering Techniques for Visual Analytics. In *Information Visualization Symp.*, pp. 141–145. IEEE, 10 2019. doi: 10.1109/VISUAL.2019.8933620 2, 5, 9

[10] G. Blasilli, D. Kerrigan, E. Bertini, and G. Santucci. Towards a Visual Perception-Based Analysis of Clustering Quality Metrics. In *Visualization in Data Science*, pp. 15–24. IEEE, 10 2024. doi: 10.1109/VDS63897.2024.00007 2, 5, 9

[11] A. Blum. Random Projection, Margins, Kernels, and Feature-Selection. In *Subspace, Latent Structure and Feature Selection*, pp. 52–68. Springer, Berlin, Heidelberg, 2006. doi: 10.1007/11752790_3 2

[12] S. Bonakala, M. Aupetit, H. Bensmail, and F. El-Mellouhi. A Human-in-the-Loop Approach for Visual Clustering of Overlapping Materials Science Data. *Digital Discovery*, 3:502–513, 2 2024. doi: 10.1039/D3DD00179B 2

[13] P. Bruneau, P. Pinheiro, B. Broeksema, and B. Otjacques. Cluster Sculptor, an Interactive Visual Clustering System. *Neurocomputing*, 150:627–644, 2 2015. doi: 10.1016/j.neucom.2014.09.062 2

[14] J. F. Buchmüller, D. Jäckle, E. Cakmak, U. Brandes, and D. A. Keim. MotionRugs: Visualizing Collective Trends in Space and Time. *Trans. Visualization and Computer Graphics*, 25:76–86, 1 2019. doi: 10.1109/TVCG.2018.2865049 2

[15] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided Visual Clustering Analysis. *Trans. Visualization and Computer Graphics*, 25(1):267–276, 1 2019. doi: 10.1109/TVCG.2018.2864477 2

[16] F. Chazal and B. Michel. An Introduction to Topological Data Analysis: Fundamental and Practical Aspects for Data Scientists. *Frontiers in Artificial Intelligence*, 4, 9 2021. doi: 10.3389/frai.2021.667963 9

[17] V. Chvátal and P. L. Hammer. Aggregation of Inequalities in Integer Programming. *Annals of Discrete Mathematics*, 1:145–162, 1975. doi: 10.1016/S0167-5060(08)70731-3 5

[18] B. Colange, J. Peltonen, M. Aupetit, D. Dutykh, and S. Lespinats. Steering Distortions to Preserve Classes and Neighbors in Supervised Dimensionality Reduction. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, eds., *Conf. Neural Information Processing*, vol. 33, pp. 13214–13225. Curran Associates, Inc., 2020. 1, 2

[19] B. Colange, L. Vuillon, S. Lespinats, and D. Dutykh. Interpreting Distortions in Dimensionality Reduction by Superimposing Neighbourhood Graphs. In *Visualization Conf. (VIS)*, pp. 211–215. IEEE, 10 2019. doi: 10.1109/VISUAL.2019.8933568 1

[20] J. P. Cunningham and Z. Ghahramani. Linear Dimensionality Reduction: Survey, Insights, and Generalizations. *J. Machine Learning Research*, 16(89):2859–2900, 2015. 3

[21] R. Cutura, M. Aupetit, J.-D. Fekete, and M. Sedlmair. Comparing and Exploring High-Dimensional Data with Dimensionality Reduction Algorithms and Matrix Visualizations. In *Int. Conf. Advanced Visual Interfaces*, article no. 10, pp. 1–9. ACM, 10 2020. doi: 10.1145/3399715.3399875 3

[22] R. Cutura, S. Holzer, M. Aupetit, and M. Sedlmair. VisCoDeR: A Tool for Visually Comparing Dimensionality Reduction Algorithms. In *ESANN*, pp. 105–110, 2018. 1, 3

[23] R. Cutura, C. Kralj, and M. Sedlmair. DRUID_{JS}—A JavaScript Library for Dimensionality Reduction. In *Visualization Conf.*, pp. 111–115. IEEE, 10 2020. doi: 10.1109/VIS47514.2020.00029 4

[24] K. Darwish, P. Stefanov, M. Aupetit, and P. Nakov. Unsupervised User Stance Detection on Twitter. In M. D. Choudhury, R. Chunara, A. Culotta, and B. F. Welles, eds., *Int. Conf. on Web and Social Media*, pp. 141–152. AAAI, 5 2020. doi: 10.1609/icwsm.v14i1.7286 2

[25] C. Demattei, N. Molinari, and J.-P. Daurès. Arbitrarily Shaped Multiple Spatial Cluster Detection for Case Event Data. *Computational Statistics & Data Analysis*, 51(8):3931–3945, 5 2007. doi: 10.1016/j.csda.2006.03.011 3, 4

[26] H. Doraiswamy, J. Tierny, P. J. Silva, L. G. Nonato, and C. Silva. Topomap: A 0-dimensional homology preserving projection of high-dimensional data. *Trans. Visualization and Computer Graphics*, 27(2):561–571, 10 2020. doi: 10.1109/TVCG.2020.3030441 3, 5, 9

[27] M. Espadoto, R. M. Martins, A. Kerren, N. S. T. Hirata, and A. C. Telea. Toward a Quantitative Survey of Dimension Reduction Techniques. *Trans. Visualization and Computer Graphics*, 27(3):2153–2173, 3 2021. doi: 10.1109/TVCG.2019.2944182 1, 9

[28] C. Faloutsos and K.-I. D. Lin. FastMap: A Fast Algorithm for Indexing, Data-Mining and Visualization of Traditional and Multimedia Datasets. In *Conf. Management of Data*, pp. 163–174. ACM, 5 1995. doi: 10.1145/223784.223812 1

[29] R. A. Fisher. The Use of Multiple Measurements in Taxonomic Problems. *Annals of Eugenics*, 7(2):179–188, 9 1936. doi: 10.1111/j.1469-1809.1936.tb02137.x 4, 8, 13, 14

[30] M. Gardner. Mathematical games—in which "Monster" Curves force Redefinition of the Word "Curve". *Scientific American*, 235(6):124–133, 12 1976. 2

[31] J. C. Gower and G. J. Ross. Minimum Spanning Trees and Single Linkage Cluster Analysis. *J. Royal Statistical Society: Series C (Applied Statistics)*, 18(1):54–64, 1969. doi: 10.2307/2346439 3, 5, 9

[32] M. M. Hamza, E. Ullah, A. Baggag, H. Bensmail, M. Sedlmair, and M. Aupetit. ClustML: A Measure of Cluster Pattern Complexity in Scatterplots learnt from Human-labeled Groupings. *Int. Conf. Information Visualisation*, 23(2):105–122, 1 2024. doi: 10.1177/14738716231220536 9

[33] J. A. Hartigan. Printer Graphics for Clustering. *J. Statistical Computation and Simulation*, 4(3):187–213, 5 1975. doi: 10.1080/00949657508810123 1, 2, 3

[34] D. Hilbert. Über die stetige Abbildung einer Linie auf ein Flächenstück. In *Dritter Band: Analysis· Grundlagen der Mathematik· Physik Verschiedenes*, pp. 1–2. Springer Nature, 1 1935. doi: 10.1007/978-3-662-38452-7_1 2, 3, 4

[35] P. Horton and K. Nakai. A Probabilistic Classification System for Predicting the Cellular Localization Sites of Proteins. pp. 109—115. AAAI, 1996. 4, 5, 13

[36] Á. Ipkovich and J. Abonyi. Neighborhood Ranking-Based Feature Selection. *Access*, 12:20152–20168, 2024. doi: 10.1109/ACCESS.2024.3362677 4

[37] H. Jeon, M. Aupetit, S. Lee, K. Ko, Y. Kim, G. J. Quadri, and J. Seo. Distortion-aware Brushing for Reliable Cluster Analysis in Multidimensional Projections. *Trans. Visualization and Computer Graphics*, pp. 1–18, 9 2025. doi: 10.1109/TVCG.2025.3615314 3, 9

[38] H. Jeon, M. Aupetit, D. Shin, A. Cho, S. Park, and J. Seo. Measuring the Validity of Clustering Validation Datasets. *Trans. Pattern Analysis & Machine Intelligence*, 47:5045–5058, 6 2025. doi: 10.1109/TPAMI.2025.3548011 2

[39] H. Jeon, Y.-H. Kuo, M. Aupetit, K.-L. Ma, and J. Seo. Classes are Not Clusters: Improving Label-Based Evaluation of Dimensionality Reduction. *Trans. Visualization and Computer Graphics*, 30(1):781–791, 11 2024. doi: 10.1109/TVCG.2023.3327187 3, 4, 9

[40] E. Kreyszig. *Introductory Functional Analysis with Applications*. Theorem 3.4-6, p. 157. Wiley, 1989. 1, 3

[41] J. B. Kruskal. Multidimensional Scaling by Optimizing Goodness of Fit to A Nonmetric Hypothesis. *Psychometrika*, 29(1):1–27, 7 1964. doi: 10.

1007/BF02289565 2, 4

[42] J. A. Lee and M. Verleysen. Quality Assessment of Dimensionality Reduction: Rank-based Criteria. *Neurocomputing*, 72(7):1431–1443, 3 2009. Advances in Machine Learning and Computational Intelligence. doi: 10.1016/j.neucom.2008.12.017 2

[43] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. *Procedia Computer Science*, 4:538–547, 2011. Int. Conf. Computational Science. doi: 10.1016/j.procs.2011.04.056 2, 9

[44] J. A. Lee and M. Verleysen. Two Key Properties of Dimensionality Reduction Methods. In *Symp. Computational Intelligence and Data Mining*, pp. 163–170. IEEE, 12 2014. doi: 10.1109/CIDM.2014.7008663 2

[45] S. Lespinats and M. Aupetit. CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings. In *Computer Graphics Forum*, vol. 30, pp. 113–125. Wiley Online Library, John Wiley and Sons, 2 2011. doi: 10.1111/j.1467-8659.2010.01835.x 1

[46] S. Lespinats, B. Fertil, P. Villemain, and J. Hérault. RankVisu: Mapping from the Neighborhood Network. *Neurocomputing*, 72(13):2964–2978, 7 2009. Hybrid Learning Machines (HAIS 2007) / Recent Developments in Natural Computation (ICNC 2007). doi: 10.1016/j.neucom.2009.04.008 2

[47] T. Li, C. Meng, H. Xu, and J. Yu. Hilbert Curve Projection Distance for Distribution Comparison. *Trans. Pattern Analysis & Machine Intelligence*, 46(07):4993–5007, 7 2024. doi: 10.1109/TPAMI.2024.3363780 3

[48] W. Li, Y. Zhang, Y. Sun, W. Wang, M. Li, W. Zhang, and X. Lin. Approximate Nearest Neighbor Search on High Dimensional Data — Experiments, Analyses, and Improvement. *Trans. on Knowledge & Data Engineering*, 32(8):1475–1488, 2020. doi: 10.1109/TKDE.2019.2909204 4

[49] R. M. Martins, D. B. Coimbra, R. Minghim, and A. C. Telea. Visual Analysis of Dimensionality Reduction Quality for Parameterized Projections. *Computer Graphics Forum*, 41:26–42, 6 2014. doi: 10.1016/j.cag.2014.01.006 1

[50] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction, 9 2020. doi: 10.48550/arXiv.1802.03426 1, 9

[51] B. Moon, H. Jagadish, C. Faloutsos, and J. Saltz. Analysis of the Clustering Properties of the Hilbert Space-filling Curve. *Trans. Knowledge and Data Engineering*, 13(1):124–141, 1 2001. doi: 10.1109/69.908985 3

[52] Q. Q. Ngo and L. Linsen. Interactive Generation of 1D Embeddings from 2D Multi-dimensional Data Projections. In J. Krüger, M. Niessner, and J. Stückler, eds., *Vision, Modeling, and Visualization*. Eurographics Association, 10 2020. doi: 10.2312/vmv.20201190 2

[53] L. G. Nonato and M. Aupetit. Multidimensional Projection for Visual Analytics: Linking Techniques with Distortions, Tasks, and Layout Enrichment. *Trans. Visualization and Computer Graphics*, 25(8):2650–2673, 8 2019. doi: 10.1109/TVCG.2018.2846735 1, 2, 3, 4, 9

[54] K. Nybo. dredviz: dimensionality reduction for information visualization. http://cis.legacy.ics.tkk.fi/projects/mi/software/dredviz/. Accessed: 2025-12-10. 4

[55] F. V. Paulovich, A. Arleo, and S. van den Elzen. When Dimensionality Reduction Meets Graph (Drawing) Theory: Introducing a Common Framework, Challenges and Opportunities. *Computer Graphics Forum*, 44, 2025. doi: 10.1111/cgf.70105 3

[56] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11):559–572, 6 1901. doi: 10.1080/14786440109462720 1, 2, 3, 13

[57] D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and Evaluation of Similarity Measures for Pairs of Clusterings. *Knowledge and Information Systems*, 19(3):361–394, 7 2009. doi: 10.1007/s10115-008-0150-6 9

[58] M. Raj and R. T. Whitaker. Visualizing Multidimensional Data with Order Statistics. *Computer Graphics Forum*, 37(3):277–287, 7 2018. doi: 10.1111/cgf.13419 2

[59] P. E. Rauber, A. X. Falcão, and A. C. Telea. Visualizing Time-dependent Data using Dynamic t-SNE. In *Proceedings of the Eurographics / IEEE VGTC Conference on Visualization: Short Papers*, EuroVis '16, 5 pages, pp. 73—77. Eurographics Association, 6 2016. doi: 10.5555/3058878.3058894 2

[60] D. Sacha, H. Senaratne, B. C. Kwon, G. Ellis, and D. A. Keim. The Role of Uncertainty, Awareness, and Trust in Visual Analytics. *Trans. Visualization and Computer Graphics*, 22(1):240–249, 2016. doi: 10.1109/TVCG.2015.2467591 1, 9

[61] G. Sanderson. Hilbert's Curve: Is infinite math useful? https://www.3blue1brown.com/?v=hilbert-curve, July 21, 2017. Accessed: 2025-12-10. 9

[62] L. J. P. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *J. Machine Learning Research*, 9:2579–2605, 11 2008. 1, 2

[63] J. Venna and S. Kaski. Local Multidimensional Scaling. *Neural Networks*, 19(6–7):889–899, 7 2006. doi: 10.1016/j.neunet.2006.05.014 2, 4

[64] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information Retrieval Perspective to Nonlinear Dimensionality Reduction for Data Visualization. *J. Machine Learning Research*, 11(2), 2 2010. 1, 2, 9

[65] J. Wang, J. Zhong, G. Chen, M. Li, F.-x. Wu, and Y. Pan. ClusterViz: A Cytoscape APP for Cluster Analysis of Biological Network. *IEEE/ACM Trans. Computational Biology and Bioinformatics*, 12(04):815–822, 7 2015. doi: 10.1109/TCBB.2014.2361348 2

[66] M. Wattenberg. A Note on Space-filling Visualizations and Space-filling Curves. In *Information Visualization Symp.*, pp. 181–186. IEEE, 11 2005. doi: 10.1109/INFVIS.2005.1532145 3

[67] J. Wenskovitch, I. Crandell, N. Ramakrishnan, L. House, S. Leman, and C. North. Towards a Systematic Combination of Dimension Reduction and Clustering in Visual Analytics. *Trans. Visualization and Computer Graphics*, 24(1):131–141, 8 2018. doi: 10.1109/TVCG.2017.2745258 2

[68] J. Wulms, J. Buchmuller, W. Meulemans, K. Verbeek, and B. Speckmann. Stable Visual Summaries for Trajectory Collections. In *Pacific Visualization Symp.*, pp. 61–70. IEEE, Los Alamitos, CA, USA, 4 2021. doi: 10.1109/PacificVis52677.2021.00016 2

[69] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting Dimensionality Reduction Techniques for Visual Cluster Analysis: An Empirical Study. *Trans. Visualization and Computer Graphics*, 28(1):529–539, 1 2022. doi: 10.1109/TVCG.2021.3114694 2, 9

[70] L. Zhou, C. R. Johnson, and D. Weiskopf. Data-Driven Space-Filling Curves. *Trans. Visualization and Computer Graphics*, 27(2):1591–1600, 2 2021. doi: 10.1109/TVCG.2020.3030473 2, 9

Supplementary Document for **ISilDR**: **I**sometric **S**eriation-based **D**imensionality **R**eduction for Visual Cluster Analysis

A Proofs for the Theorems 1 and 2

In this section, we provide the proofs for Theorems 1 and 2 in the main paper.

A.1 Isometric Seriation-based Dimensionality Reduction

We want to prove Theorem 2 for the two variants of mD ISilDR.

First, we demonstrate that when we use ISilDR to project data into 1D axis, distances can never decrease (Theorem S1). We then extend this proof to the mD case based on the full space (Theorem S2), then to the mD case based on the disjoint subspaces (Theorem S3). With those theorems (Theorems S1 to S3) proven, all options for Theorem 2 are proven.

A.1.1 ISilDR: the 1D Case

The Euclidean distance is $\mathbf{Y}_{a,b} = |Y_a - Y_b|$ in 1D.

Theorem S1 (No decreasing distances in a 1D ISilDR)**.** A projection $Y \in \mathbb{R}^N$ of the input data $X \in \mathbb{R}^{N\times M}$ created with ISilDR Def. 1 using any permutation $\pi : I \to I : i \neq j \implies \pi(i) \neq \pi(j)$, then

$$\forall i \neq j, \left|Y_i - Y_j\right| \geq \mathbf{X}_{i,j}.$$

Proof. Assume $i > j > 0$, $\pi(0) := \pi(1)$, $Y_{\pi(1)} = 0$. The 1D ISilDR projection Y of points X is defined as:

$$Y_{\pi(i)} = \sum_{k=1}^{i} \mathbf{X}_{\pi(k-1),\pi(k)}.$$

Therefore,

$$\begin{aligned}\left|Y_{\pi(i)} - Y_{\pi(j)}\right| &= \left|\sum_{k=1}^{i} \mathbf{X}_{\pi(k-1),\pi(k)} - \sum_{k=1}^{j} \mathbf{X}_{\pi(k-1),\pi(k)}\right| \\ &= \left|\sum_{k=j+1}^{i} \mathbf{X}_{\pi(k-1),\pi(k)}\right| \\ &= \sum_{k=j+1}^{i} \mathbf{X}_{\pi(k-1),\pi(k)}\end{aligned}$$

If $i = j+1$, then $\left|Y_{\pi(i)} - Y_{\pi(j)}\right| \triangleq \mathbf{X}_{\pi(i),\pi(j)}$.
If $i > j+1$, then

$$\left|Y_{\pi(i)} - Y_{\pi(j)}\right| = \sum_{k=j+1}^{i} \mathbf{X}_{\pi(k-1),\pi(k)} \geq \mathbf{X}_{\pi(i),\pi(j)}$$

due to the triangle inequality of the Euclidean norm.

Hence, $\forall i \neq j, \left|Y_{\pi(i)} - Y_{\pi(j)}\right| \geq \mathbf{X}_{\pi(i),\pi(j)}$ and as π is a permutation of I, we finally get:

$$\forall i \neq j, \left|Y_i - Y_j\right| \geq \mathbf{X}_{i,j} \tag{5}$$

□

A.1.2 ISilDR: the mD Case

Theorem S2 (No decreasing distances in an mD ISilDR using distances from the full space)**.** Let $X \in \mathbb{R}^{N\times M}$ be the input data, and $Y \in \mathbb{R}^{N\times m}$ be the output with $m > 1$. Let $\pi_k : I \to I$ ($k \in \{1,\ldots,m\}$) be permutations of I ($X_{\pi_k(1)}$ is the first point of the ordering π_k).

When constructing Y according to Def. 2, then pairs of points in Y verify:

$$\forall i, j \in I : \mathbf{Y}_{i,j} \geq \mathbf{X}_{i,j}.$$

Proof. Creating $\widetilde{Y}$ by concatenating m 1D projections using Def. 1, $\widetilde{Y} = \left[Y_{\pi_1}, \quad \ldots, \quad Y_{\pi_m}\right]$
From Theorem S1

$$\forall k \in \{1,\ldots,m\} : \forall i \neq j : \left|\widetilde{Y}_{i,k} - \widetilde{Y}_{j,k}\right| \geq \mathbf{X}_{i,j}$$

We have:

$$\begin{aligned}\forall i, j \in I, \widetilde{\mathbf{Y}}^2_{i,j} &= \sum_{k=1}^{m} \left(\widetilde{y}_{i,k} - \widetilde{y}_{j,k}\right)^2 \\ &\geq \sum_{k=1}^{m} \mathbf{X}^2_{i,j} = m \cdot (\mathbf{X}_{i,j})^2.\end{aligned}$$

Hence,

$$\widetilde{\mathbf{Y}}_{i,j} \geq m^{-1/2} \cdot \mathbf{X}_{i,j}.$$

When we scale $Y = m^{1/2} \cdot \widetilde{Y}$ we get

$$\mathbf{Y}_{i,j} = m^{1/2} \widetilde{\mathbf{Y}}_{i,j} \geq \mathbf{X}_{i,j}.$$

□

Theorem S3 (No decreasing distances in a mD ISilDR using distances from disjoint orthogonal subspaces)**.** Let $X \in \mathbb{R}^{N\times M}$ be the input data, and $Y \in \mathbb{R}^{N\times m}$ be the output with $m > 1$. Further, let $\pi_k : I \to I$ be permutations of I ($k \in \{1,\ldots,m\}$) and $F_k \subset \{1,\ldots,M\}$ be the feature sets of the respective column subsets X_{F_k}, where $\bigcup_{k=1}^{m} F_k = \{1,\ldots,M\}$, $\forall i : F_i \neq \emptyset$ and $\forall i, j : F_i \cap F_j = \emptyset$. When constructing Y, according to Def. 3, by concatenating m individual 1D projections $Y^{F_k}_{\pi_k}$ for each subspace X^{F_k}, the pairwise distances verify:

$$\forall i, j \in \{1,\ldots,N\} : \mathbf{Y}_{i,j} \geq \mathbf{X}_{i,j}.$$

Proof. From Theorem S1

$$\forall k \in \{1,\ldots,m\} : \forall i \neq j : \left|y^{F_k}_{i,k} - y^{F_k}_{j,k}\right| \geq \mathbf{X}^{F_k}_{i,j}$$

$$\delta\left(y_i, y_j\right)^2 = \sum_{k=1}^{m} \left(y^{F_k}_{i,k} - y^{F_k}_{j,k}\right)^2 \geq \sum_{k=1}^{m} (\mathbf{X}^{F_k}_{i,j})^2.$$

And as $\bigcup_{k=1}^{m} F_k = \{1,\ldots,M\}$ and $\bigcap_{k=1}^{m} F_k = \emptyset$, we get:

$$\sum_{k=1}^{m} (\mathbf{X}^{F_k}_{i,j})^2 = \sum_{k=1}^{m} \sum_{q\in F_k} (x_{i,q} - x_{j,q})^2 = (\mathbf{X}_{i,j})^2$$

Hence,

$$\mathbf{Y}_{i,j} \geq \mathbf{X}_{i,j}$$

□

A.2 Orthogonal Linear Projection

Now, we prove Theorem 1. Let $X \in \mathbb{R}^{N\times M}$ the input data, and $V \in \mathbb{R}^{M\times M}$ an orthogonal matrix. If the matrix $Y \in \mathbb{R}^{N\times M}$ is the dot product $Y = X \cdot V$, as V acts as an isometry, we get $\mathbf{X}_{i,j} = \delta(x_i \cdot V, x_j \cdot V) = \mathbf{Y}_{i,j}$.

An OLP transforms the input data X with a truncated orthogonal matrix $V \in \mathbb{R}^{M\times m}$. For instance, PCA [56] computes the column eigenvectors v_i ($i \in \{1,\dots,M\}$) of the correlation matrix of X and put them together in a matrix $\widetilde{V} = \begin{bmatrix} v_1 & \cdots & v_M \end{bmatrix}$. When using PCA for dimensionality reduction, this matrix gets truncated to the m eigenvectors with the m largest eigenvalues $V = \begin{bmatrix} v_1 & \cdots & v_m \end{bmatrix}$. A PCA projection is computed by the dot product of the input matrix with the truncated orthogonal matrix: $X \cdot V = Y \in \mathbb{R}^{N\times m}$.

Proof of Theorem 1

We need to prove that the distances in an OLP projection $Y \in \mathbb{R}^{N\times m}$ of the input data $X \in \mathbb{R}^{N\times M}$ never increase.

Proof. The Euclidean distances of any pair of points of the input data X, are

$$\forall i,j \in \{1,\dots,N\} : \mathbf{X}_{i,j} = \sqrt{\textstyle\sum_{k=1}^{M} \left(x_{i,k} - x_{j,k}\right)^2}$$

Let $\widetilde{V}$ be an orthogonal matrix $\widetilde{V} \in \mathbb{R}^{M\times M}$. Distances between pairs of points in $X \cdot V$ are the same as in X as $\widetilde{V}$ being orthogonal, it acts as an isometry. So, the Euclidean distances between any pair of points of the orthogonally transformed input $\widetilde{Y} = X \cdot \widetilde{V}$ verify:

$$\begin{aligned}\forall i,j \in I : \delta\left(\widetilde{Y}_i, \widetilde{Y}_j\right)^2 &= \delta\left(X\widetilde{V}_i, X\widetilde{V}_j\right)^2 \\ &= \sum_{k=1}^{M} \left(X\widetilde{V}_{i,k} - X\widetilde{V}_{j,k}\right)^2 = \delta\left(X_i, X_j\right)^2.\end{aligned}$$

Let $V_i = \begin{bmatrix} \widetilde{V}_{i,1} & \cdots & \widetilde{V}_{i,m} \end{bmatrix}$ be the first m columns of $\widetilde{V}_i$. and let Y be the m-dimensional orthogonal projection of X: $Y = X \cdot V$.

The Euclidean distance of any pair of points in the projection Y is $\delta\left(Y_i, Y_j\right)$ and verifies:

$$\begin{aligned}\forall i,j \in I : \delta\left(Y_i, Y_j\right)^2 &= \delta\left(XV_i, XV_j\right)^2 \\ &= \sum_{k=1}^{m} \left(X\widetilde{V}_{i,k} - X\widetilde{V}_{j,k}\right)^2 = \delta\left(X_i, X_j\right)^2.\end{aligned}$$

We have:

$$\delta\left(\widetilde{Y}_i, \widetilde{Y}_j\right)^2 - \delta\left(Y_i, Y_j\right)^2 = \sum_{k=m+1}^{M} (X\widetilde{V}_{i,k} - X\widetilde{V}_{j,k})^2 \geq 0$$

hence:

$$\forall i,j \in I : \delta\left(Y_i, Y_j\right) \leq \delta\left(X_i, X_j\right).$$

□

B Examples

The four 1D IsilDR projections with their individual cluster coloring of the Iris dataset from Fig. 10 are shown in Fig. S1.

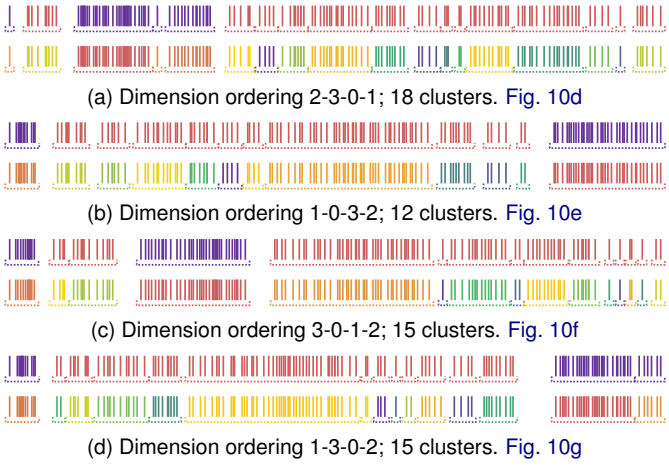
We provide the empirical analysis for the Iris [29] dataset here in Fig. S2.

The scatterplot matrices for the two datasets, Iris and Ecoli [35], are provided in Fig. S3.

C False and Missing Neighbor alternative definitions

Definition 7 (Rank-Wise False and Missing Neighbors)**.** Let $X \in \mathbb{R}^{N\times M}$ be the input data, and $Y \in \mathbb{R}^{N\times m}$ be a projection of X. Let $U_k^Z(\cdot)$ be a function that maps to the indices of the k-nearest neighbors of a given point of a dataset Z.

Definition 7.1. For $i,j \in \{1,\dots,N\}$, if $U_k^Y(i) \ni j \notin U_k^X(i)$, then y_j is a rank-wise false neighbor of y_i.



(a) Dimension ordering 2-3-0-1; 18 clusters. Fig. 10d

(b) Dimension ordering 1-0-3-2; 12 clusters. Fig. 10e

(c) Dimension ordering 3-0-1-2; 15 clusters. Fig. 10f

(d) Dimension ordering 1-3-0-2; 15 clusters. Fig. 10g

Fig. S1: Projections of the Iris [29] dataset. We use an ε-graph for the clustering (Sec. 3.5.2) with $\varepsilon = 1$. The first row of each subfigure shows the coloring based on the masterview (PCA) used in Fig. 10, the second row shows each ε-cluster defined in the 1D **HilDR** projection space in a separate color.

Definition 7.2. For $i,j \in \{1,\dots,N\}$, if $U_k^Y(i) \not\ni j \in U_k^X(i)$, then y_j is a rank-wise missing neighbor of y_i.

Definition 8 ($\varepsilon\mu$-Wise False and Missing Neighbors)**.** Let $X \in \mathbb{R}^{N\times M}$ be the input data, and $Y \in \mathbb{R}^{N\times m}$ be a projection of X. Let $U_\varepsilon^{\mathbf{Z}}(\cdot)$ be a function that maps to the indices of those points in Z which are closer than ε. Let $\varepsilon, \mu \in \mathbb{R}^+$ be two independent radii.

Definition 8.1. For $i,j \in \{1,\dots,N\}$, if $j \in U_\varepsilon^Y(i)$ but $j \notin U_\mu^X(i)$, then y_i is a $\varepsilon\mu$-wise false neighbor to y_j.

Definition 8.2. For $i,j \in \{1,\dots,N\}$, if $j \notin U_\varepsilon^Y(i)$ but $j \in U_\mu^X(i)$, then y_i is a $\varepsilon\mu$-wise missing neighbor to y_j.

D OLP and ISilDR distortions based on alternative neighborhoods definitions

Even if an ISilDR does not have false ε-neighbors according to Def. 4.4, it can have rank-wise false neighbors (see Fig. S4), as well as $\varepsilon\mu$-wise false neighbors (as ε and μ can be chosen independently, imagine Fig. S4 with ε the radius of U_2^X and μ the radius of U_2^Y).

Similarly, an OLP does not have missing ε-neighbors according to Def. 4.3, but it can have rank-wise missing neighbors, as well as $\varepsilon\mu$-wise missing neighbors.

Finding neighborhoods other than ε-neighbors, making ISilDR FN-free and OLP MN-free, remains an open problem.

E Screenshots of the interactive system

Fig. S5 shows screenshots of our interactive system.

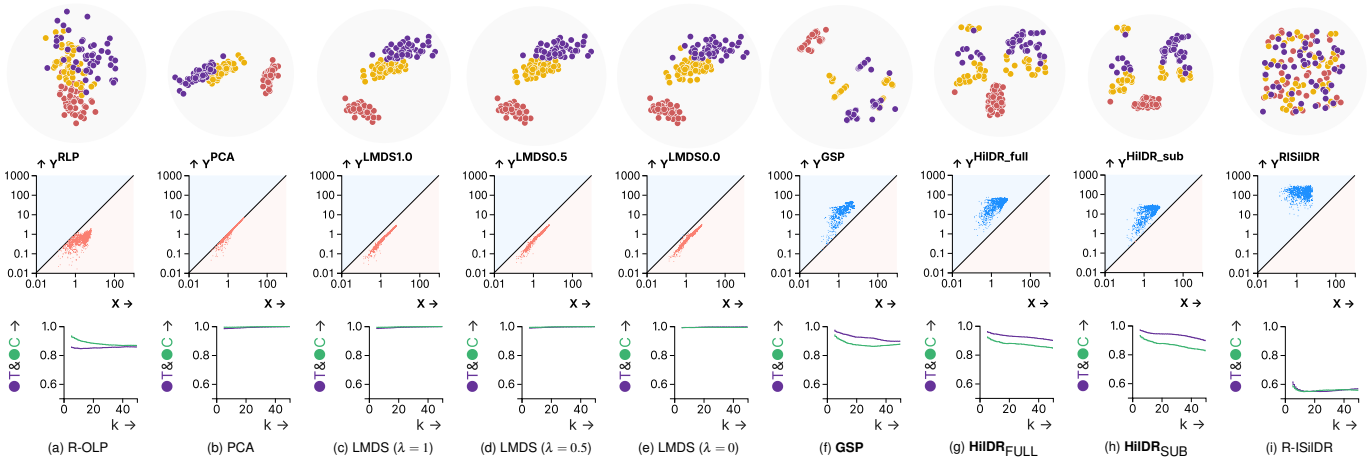
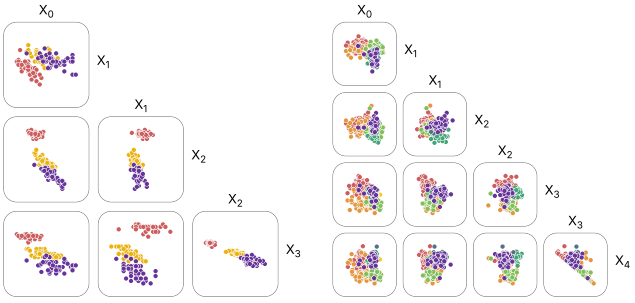


Fig. S2: Projections of Iris [29] dataset, colored by their labels. Points on the diagonal (black line) of the Shepard diagram represent unchanged distances, red points mean the distance in LD is smaller than in MD, blue points mean the distance in LD is bigger than in MD.The Trustworthiness and Continuity measures of the projection with increasing parameter k as a line chart.



(a) Iris dataset (b) Ecoli dataset

Fig. S3: SPLOMs of Iris and Ecoli datasets.

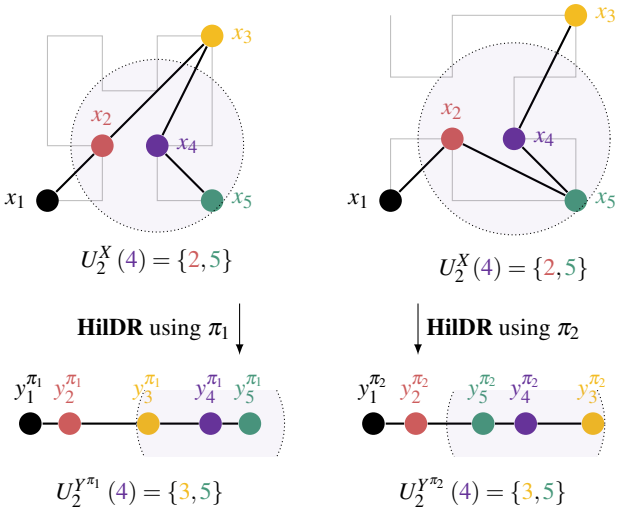
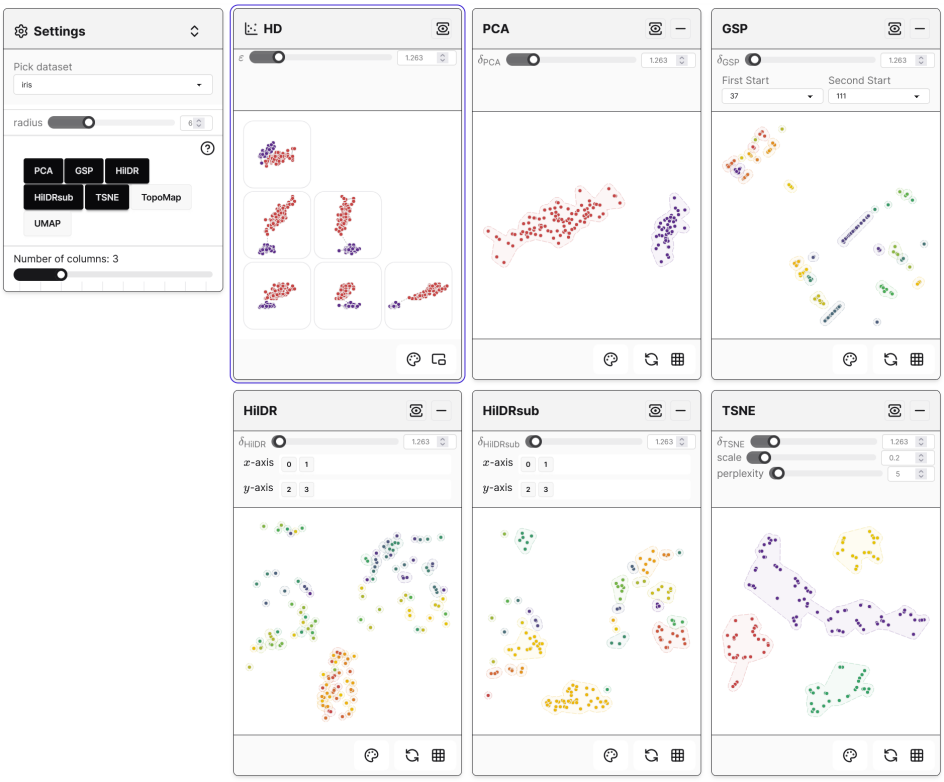


Fig. S4: **ISilDR can have rank-wise or $\varepsilon\mu$-wise false neighbors.** Two projections of X using **HilDR** with permutations π_1 and π_2. The two nearest neighbors of X_4 are $U_2^X(4) = \{2,5\}$, and for Y^{π_1} and Y^{π_2} they are $U_2^{Y^{\pi_1}}(4) = U_2^{Y^{\pi_2}}(4) = \{3,5\}$. According to Def. 7.1 $y_3^{\pi_1}$ (and $y_3^{\pi_2}$) are rank-wise false neighbors to $y_4^{\pi_1}$ (or $y_4^{\pi_2}$). Also, according to Def. 7.2 $y_2^{\pi_1}$ (and $y_2^{\pi_2}$) are rank-wise missing neighbors to $y_4^{\pi_1}$ (or $y_4^{\pi_2}$). Considering μ, the radius of U_2^Y and $\varepsilon \neq \mu$, the radius of U_2^X, μ and ε determine the same neighborhoods as the two-nearest neighbors, showing ISilDR can also have $\varepsilon\mu$-wise false neighbors.



(a) The 2D projection views

(b) In the first row, the HD block shows a PCP plot of the MD data. In the second row, the first two blocks show the 2D scatterplots of PCA and tSNE, the right two blocks show the Rugplots of **HiIDR** and **GSP**.

Fig. S5: Screenshots of the interactive system implemented in Javascript.