Neurocomputing 112 (2013) 109–123

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Visualizing the quality of dimensionality reduction

Bassam Mokbel [a,*], Wouter Lueks [b,c], Andrej Gisbrecht [a], Barbara Hammer [a]

[a] Bielefeld University—CITEC Centre of Excellence, Germany
[b] University of Groningen—Faculty of Mathematics and Natural Sciences, The Netherlands
[c] University of Nijmegen—Faculty of Science, The Netherlands

ARTICLE INFO

Available online 6 March 2013

Keywords:
Nonlinear dimensionality reduction
Data visualization
Quality assessment
Co-ranking matrix

ABSTRACT

The growing number of dimensionality reduction methods available for data visualization has recently inspired the development of formal measures to evaluate the resulting low-dimensional representation independently from the methods' inherent criteria. Many evaluation measures can be summarized based on the co-ranking matrix. In this work, we analyze the characteristics of the co-ranking framework, focusing on interpretability and controllability in evaluation scenarios where a fine-grained assessment of a given visualization is desired. We extend the framework in two ways: (i) we propose how to link the evaluation to point-wise quality measures which can be used directly to augment the evaluated visualization and highlight erroneous regions; (ii) we improve the parameterization of the quality measure to offer more direct control over the evaluation's focus, and thus help the user to investigate more specific characteristics of the visualization.

© 2013 Elsevier B.V. All rights reserved.

1. Introduction

In recent years, many dimensionality reduction (DR) techniques have been developed primarily for data visualization, see e.g. [1–4] for overviews. A DR method is used to map high-dimensional data to two- or three-dimensional vectors, in order to display them in a scatter plot or point cloud. This allows field experts to investigate the structure and distribution of large amounts of data in a user-friendly and easily accessible way. It is crucial that the visualization adequately resembles the original high-dimensional data structure and distribution. From a theoretical point of view, DR constitutes an ill-posed problem: not all the structure and relations that exist in intrinsically high-dimensional data can be faithfully represented in the lower-dimensional space, and it is not clear which relations should be preserved. The application task dictates which concessions to make. Here, and in the following, we assume that the intrinsic dimensionality of the original data is higher than the one of the embedding space. Therefore, the user faces the problem of choosing an appropriate DR technique and an adequate configuration of its parameters, along with other methodic choices that affect the DR procedure, e.g. preprocessing steps.

Dimensionality reduction methods: The variety of possible strategies has resulted in the development of many different DR techniques. Often they optimize an objective function which formalizes the preservation goal. For an overview of established DR methods, see e.g. [1–4]. As a simple example, linear dimensionality reduction such as principal component analysis (PCA) minimizes the loss of information as measured by the sum squared error. Due to its simplicity and well-understood behavior, it is often used for data visualization, albeit its linearity severely restricts its applicability. Modern DR methods usually use nonlinear projections of the data into low dimensions to appropriately visualize, rather than preprocess, complex data sets. The methods differ in their objectives which guide the projection. Several approaches have been proposed which can be interpreted as nonlinear extensions of PCA, such as kernel PCA [5], auto-encoding neural networks [6], or projections based on principal curves which pass through the 'center' of the high-dimensional data [7]. Some visualization techniques aim to preserve distances, such as multidimensional scaling (MDS) [8] and its variants like Sammon's map [9], or Isomap [10], the latter referring to geodesic distances in the data. Others take into account local neighborhood structures and project these accordingly, examples are the classical self-organizing map (SOM) [11], the generative topographic mapping (GTM) [12] which use a fixed topological structure, locally linear embedding (LLE) [13] which measures local structure in terms of local linear relationships, and maximum variance unfolding (MVU) [14] which unfolds data while preserving direct neighbors. Laplacian eigenmaps [15] are also based on the local neighborhood graph, but they take a more principled approach by referring to the spectral properties of the resulting dissimilarity matrix. Another possibility is to glue local linear projections such

* Corresponding author. Tel.: +49 52 110612135.
E-mail address:
bmokbel@techfak.uni-bielefeld.de (B. Mokbel).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.11.046

as tangent spaces together using appropriate transforms as in tangent space alignment or manifold charting [16]. Stochastic neighbor embedding (SNE) [17], t-distributed SNE (t-SNE) [3], and the neighbor retrieval visualizer (NeRV) [18] try to match the probability distributions induced by the pairwise data dissimilarities in the original space and the projection space, respectively. There exist many more techniques and variants of dimensionality reducing visualization, partially related to the methods mentioned above, such as Isotop, XOM, local MDS, factor analysis, curvilinear component analysis, etc., see [19–23].

The different approaches result in qualitatively very different visualizations for a given data set. Therefore, it is not clear a priori, which DR technique is best suited for the task at hand. In addition, virtually all recent techniques have parameters to control (in some way) the preservation strategy for the embedding. Hence, depending on the chosen parameters, even a single DR method can lead to vastly diverse results. Moreover, many nonlinear DR techniques do not arrive at a unique solution due to random aspects of the algorithm. Instead, they can produce different outputs in every run, corresponding to different local optima of the objective. Therefore, it is possible that qualitatively different solutions can be obtained by a single method with a single set of model parameters.

Quality measures: Usually it is not clear whether differences in the dimensionality reduction (between different methods, parameters or runs) represent different relevant aspects in the data or signify unsuitability of a method. Further, it can happen that suboptimal results are obtained simply because of numerical problems, such as (bad) local optima. At the same time, it is very hard for humans to judge the quality of a given embedding by visual inspection. The user cannot compare it against a ground truth, as this data is inaccessible due to its high dimensionality. Therefore, we need formal measures which judge the quality of a given data embedding. Such formal measures should evaluate, in an automated and objective way, in how far the structure of the original data is preserved in the low-dimensional representation. Apart from their importance for practical applications, quality measures are generally relevant to automatically evaluate and compare DR techniques for research. As reported in [18], a high percentage of publications on data visualization evaluates results in terms of visual impression only—in [18], about 40% of the 69 referenced papers did not use any quantitative evaluation criterion. Even if formal evaluation criteria are used, these differ from one application to the next, referring e.g. to a local misclassification error for labeled data [4,18], the reconstruction error provided an inverse mapping is possible [24], or local preservation of neighborhoods [18]. Further, many popular benchmark data sets in the literature are artificially generated and thus are of limited use for a realistic evaluation of DR methods [4]. Although a few real life data sets are currently available (see e.g. [3]), there does not exist a large variety of data encompassing different characteristics together with suitable evaluation criteria.

In this paper, we will first give a short overview over existing approaches of quality assessment for dimensionality reduction in Section 2, and will discuss some fundamental features which distinguish their strategies, concluding with the general motivation for our own approach. Thereafter, we will focus on the co-ranking matrix from [24], which serves as a unifying framework to represent several other measures. In Section 3, we will briefly describe the co-ranking matrix itself, and, in Section 4, propose to augment data visualizations by point-wise quality contributions based on the co-ranking framework. Section 5 discusses how a fairly simple parameterization in established quality measures causes problems regarding the interpretability of the evaluation results. We propose a new parameterization which allows for more fine-grained control over the evaluation focus, and which facilitates a more specific analysis of the given visualization. After we demonstrate the benefits of our approach on several artificial examples, we show, in Section 6, how it performs in real-world visualization scenarios in comparison with the former model. In Section 7, we summarize our findings, and close with an outlook over future research.

2. Principles of quality assessment for DR

Several quality criteria to evaluate DR have been proposed in recent years, see [24] for an overview of the more prominent measures. However, the problem to define formal evaluation criteria suffers from the ill-posedness of DR itself: it is not clear a priori which structural aspects of the data should be preserved in a given task. Generally, the existing quality measures evaluate in how far the original data relations agree with the ones produced by the embedding. A similar notion forms the basis for most objective functions in DR methods, but the specifics and priorities of the agreement calculation are a matter of ongoing research and debate. By formalizing an objective, every DR strategy gives rise to a perfect mapping in terms of the global optimum of this objective, and thus it incorporates a quality measure in itself. However, the goal of DR is to produce a representative embedding of the data and, thus, algorithmic aspects such as easy optimization are vital, while a quality measure is used to gain insight into the properties of the embedding. Therefore, the quality measure can and should be general as well as understandable, so the user can easily interpret its results, and thus judge the trustworthiness of the visualization.

Regarding application scenarios, we believe that a formal quality evaluation can assist the user in two ways, to which we will refer in the further discussion:

(a) Given a data set, formal measures help to compare different DR methods along with their parameter settings, as demonstrated in extensive experiments, e.g. in [18,24,25]. Therefore, by iterative comparison, a DR method's parameter configuration could be optimized interactively. A relatively coarse-grained, overall quality assessment seems sufficient for this purpose.
(b) Given a single visualization, formal measures can provide the user with information about the qualitative characteristics of the observed embedding. Since the user wants to gain knowledge about the original data, it is beneficial to analyze in detail the compromised representation that is displayed. For this task, fine-grained evidence is necessary, see [26,27].

In the following, we will briefly discuss existing strategies for quality assessment in the literature and point out some basic distinguishing features. We assume the following basic setting: the DR method maps high dimensional data $\Xi = \{\xi_1, \ldots, \xi_N\} \subset \mathbb{R}^{\mathrm{H}}$ to low-dimensional points $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^{\mathrm{L}}$, with $\mathrm{L} < \mathrm{H}$ and $\mathrm{L} = 2$ or $\mathrm{L} = 3$ for the purpose of visualization. Data is either characterized directly as vectors or via dissimilarities (which might be non-Euclidean), depending on the chosen DR technique. DR evaluation compares characteristics derived from the data $\{\xi_1, \ldots, \xi_N\}$ to corresponding characteristics derived from the projections $\{x_1, \ldots, x_N\}$. Since a formal mathematical characterization together with a unifying framework of many DR evaluation techniques has already been developed [24], we do not aim at a formal definition of the particular methods. Instead, we highlight distinguishing characteristics of the evaluation methods.

Distances vs. ranks: Most DR evaluation techniques relate to pairwise distances of data in some way. One fundamental distinguishing aspect is whether the pairwise distances in the

high-dimensional data are compared directly with the low-dimensional setting, or if only their order, i.e. ranks, is considered. All evaluation measures mentioned in [24,28] use ranks, whereas in [18] the criteria *precision* and *recall* may be evaluated for any form of proximity measure, including ranks of distances, or distances itself. From the measures presented in [7], the *quality of point neighborhood preservation* and *quality of group compactness* are both based on K nearest neighbors, i.e. they consider ranks; while the *quality of distance mapping* can be evaluated for both, distances and ranks alike. While absolute distance information is lost when only the order of distances is considered, it has the benefit that any notion of pairwise proximity in the original data (e.g., distances, dissimilarities, similarities, neighborhood probabilities) is comparable with the Euclidean distances of the embedded data points, since ordering the neighbors of a point is possible in all these cases. Further, ranks are invariant to monotonic transformations of the distances.

Neighborhood scales: Many quality measures aim to give an overview of the visualization's characteristics on different scales, by considering the agreement rates over varying neighborhood sizes (usually averaged over all data points). This facilitates to some extent the fine-grained analysis mentioned in our introductory statement (b). The neighborhoods are either defined via hyperspheres of a radius ϵ centered at each point, or, alternatively, as the K nearest neighbors of each point. All measures discussed in [24,29,28] use K-neighborhoods, while in [18] ϵ-hyperspheres are considered. Note, that the latter case is more general, since an ϵ-radius can serve as a boundary for any kind of proximity, including ranks.[1]

Agreement evaluation: Based on these neighborhoods for some fixed K or ϵ, there are different possibilities to evaluate the agreement between the characteristics in the high-dimensional data and its counterpart in the embedding. Some quality measures simply calculate the ratio of agreed points within these regions, see e.g. [24,29]. Others consider a weighted combination of the agreement rate inside and outside of the neighborhoods, like the *mean relative rank errors* from [1]. A recent criterion known as *local strict rank order preservation* [28] counts strictly preserved ranks. Instead of counting the number of agreed neighbors, the *quality of distance mapping* [7] uses the correlation of pairwise distances between the original and the embedded data. For the *precision* and *recall* for DR, as defined in [18], one can use ranks instead of pairwise distances between the data, which leads to defining regions of ϵ nearest neighbors. Then, precision and recall are both equal to calculating the average number of agreeing neighbors, which coincides exactly with the *quality* Q_{NX} from [24], the *quality of point neighborhood preservation* in [7], as well as the *agreement rate* from [29]; criteria which were all proposed independently. Supplemental to the *quality* from [24], the *behavior* indicator gives insight about the types of errors which occur in the visualization: either points become closer in the embedding, or points are farther apart than in the original, called *intrusive* or *extrusive* behavior, indicated by values below or above zero respectively. While most of these concepts aim at our scenario described in (a), the behavior indicator reveals more details about the embedding's characteristics.

Aggregation of pointwise contributions: Point-wise agreement rates (i.e. independently regarding every point's neighborhood) are usually aggregated to fewer nominal values, in order to deliver a compact evaluation result, like a curve over growing ϵ or K. For the aggregation, a simple average is often used. While this is beneficial when comparing several DR techniques for the same data set, as addressed in statement (a), the aggregation hides the local quality characteristics of the embedding, which would be beneficial regarding our statement (b).

Scale-independent criteria: To obtain even fewer nominal values which subsume the quality on all (or some important) neighborhood scales, different possibilities can be found in the literature. In [25] averaging the quality curve $Q_{\mathrm{NX}}(K)$ over certain ranges of K has been proposed. A splitting point $K_{\max}$ is defined as the first maximum of the curve with respect to its baseline. Then, the mean quality for all $k \le K_{\max}$ represents the local quality, whereas the mean quality over all $K > K_{\max}$ defines the global quality. Other measures to judge the overall topology at once are, for example, the *quality of distance mapping* proposed in [7] which calculates the Pearson or Spearman correlation coefficients between all pairwise distances in the high-dimensional versus the low-dimensional setting, yielding a single value. For efficiency, the authors propose to calculate the correlation only on a representative subset of pairwise distances, selected by the *natural PCA* procedure, see [7]. For the topographic mapping with *self-organizing maps* (SOMs) [11], the *topographic product* has been proposed, which yields a single value to assess the topographic disturbances in the map, see [30]. It basically considers the distances between pairs of nearest neighbors, hence it is easily possible to generalize the topographic product to any high- and low-dimensional point configurations (without the fixed lattice of a SOM).

A single nominal quality rating greatly benefits overall comparison of DR methods as stated in (a), while a fine-grained analysis of a given visualization is not supported.

Supervised evaluation: There are also measures which take a class labeling of the data into account, like e.g. the *quality of group compactness* in [7], the K nearest neighbor error of the projections in [18], or which even introduce a local labeling in the original data space to judge the preservation of local neighborhoods via this labeling [4]. We will not discuss these further, since we focus strictly on an unsupervised evaluation scenario.

Concluding remarks: From the existing literature, we see that many quality measures are suitable for a situation where the user wants an overall comparison of a number of different embeddings of the same data, e.g. originating from different DR methods or different parameter settings, as described in statement (a) in Section 2. However, there are only few approaches which aim at a more fine-grained analysis of a single visualization, mentioned in (b). Some measures are useful for compromises between (a) and (b), by evaluating the quality over all neighborhood scales, e.g. in [18,24]. However, only the works [26,27] aim fully towards the scenario (b), by integrating visual cues about local reliability directly into the embedding. Their idea is to provide the user with sufficient information to compensate for the distortions in the observed visualization, when reasoning about the original data. These approaches are, however, not directly linked to any of the referenced formal quality measures. Therefore, our goal is to extend the formal evaluation based on the well-established *co-ranking matrix* [24] toward a more fine-grained analysis.

After introducing the co-ranking matrix in the next section, we will utilize a decomposition into point-wise quality contributions in Section 4. In Section 5 we will point out certain disadvantages of the quality framework with regard to our purpose of fine-grained analysis and control, and propose to circumvent these disadvantages with a different parameterization.

3. Evaluating DR based on the co-ranking matrix

Referring to the high-dimensional data set $\Xi = \{\xi_1, \ldots, \xi_N\} \subset \mathbb{R}^{\mathrm{H}}$ and the low-dimensional dataset $X = \{x_1, \ldots, x_N\} \subset \mathbb{R}^{\mathrm{L}}$, let δ_{ij} be the distance from ξ_i to ξ_j in $\mathbb{R}^{\mathrm{H}}$ and d_{ij} the distance from x_i to x_j in $\mathbb{R}^{\mathrm{L}}$.

[1] When distances are replaced by their respective ranks, limiting to a radius of size ϵ for a point means choosing its ϵ nearest neighbors.

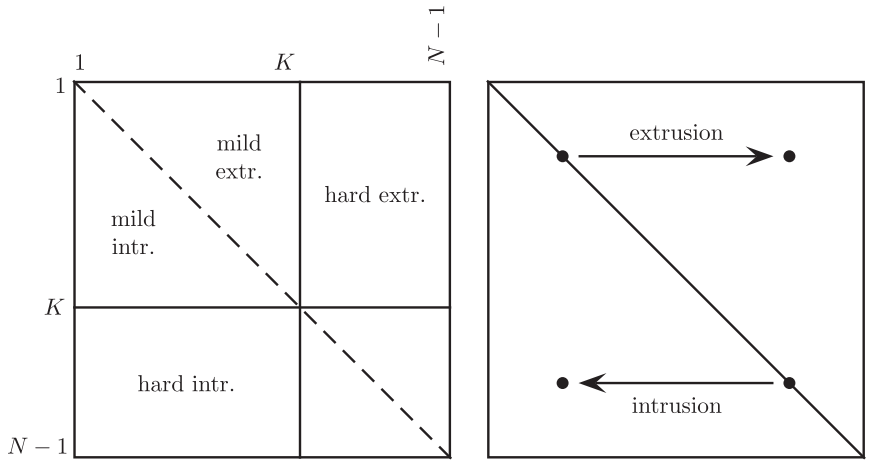


Fig. 1. Large-scale structure of the co-ranking matrix. On the left, the matrix is split into blocks to show different types of intrusions and extrusions. In a perfect mapping, the co-ranking matrix will be a diagonal matrix. The image on the right shows how rank differences will alter the matrix. If a neighbor moves further away in the embedding (an extrusion) it will appear to the right of the diagonal. Similarly, intrusions appear to the left of the diagonal.

The rank of ξ_j with respect to ξ_i in $\mathbb{R}^H$ is given by

$$\rho_{ij} = \left|\{k \mid \delta_{ik} < \delta_{ij} \quad \text{or} \quad (\delta_{ik} = \delta_{ij} \quad \text{and} \quad 1 \le k < j \le N)\}\right|.$$

Analogously, the rank of x_j with respect to x_i in the low-dimensional space is

$$r_{ij} = \left|\{k \mid d_{ik} < d_{ij} \quad \text{or} \ (d_{ik} = d_{ij} \quad \text{and} \quad 1 \le k < j \le N)\}\right|.$$

The differences $R_{ij} = r_{ij} - \rho_{ij}$ are the *rank errors*. The co-ranking matrix $\mathbf{C}$ [24] can be seen as a histogram of all rank errors, and is defined by

$$\mathbf{C}_{kl} = \left|\{(i,j) \mid \rho_{ij} = k \quad \text{and} \quad r_{ij} = l\}\right|.$$

Pairs of points which change their rank between the original data and its projection are considered errors of the DR procedure. They result in non-zero off-diagonal entries in the co-ranking matrix. A point x_j with $\rho_{ij} > r_{ij}$ is called an *intrusion*, with $\rho_{ij} < r_{ij}$ it is an *extrusion*. Usually, a DR method cannot embed all relationships of data faithfully. Often, the focus is on the preservation of local relationships. The co-ranking matrix offers a framework, in which several existing evaluation measures can be expressed, as pointed out in [24]: *local continuity meta-criterion* (LCMC) [31], *trustworthiness* & *continuity* (T&C) [32], and *mean relative rank errors* (MRRE) [1]. Essentially, these quality measures correspond to weighted sums of entries $\mathbf{C}_{kl}$ of the co-ranking matrix for regions $k \le K$ and/or $l \le K$, with a fixed neighborhood range K.

In [24], a comprehensible (unweighted) sum has been proposed, the *quality* Q_{NX}

$$Q_{\mathrm{NX}}(K) = \frac{1}{KN} \sum_{k=1}^{K} \sum_{l=1}^{K} \mathbf{C}_{kl} = \frac{1}{KN} \sum_{i=1}^{N} |A_{\xi_i} \cap B_{x_i}|. \tag{1}$$

where $A_{\xi_i} = \{\xi_j \mid \rho_{ij} \le K\}$ and $B_{x_i} = \{x_j \mid r_{ij} \le K\}$ are the sets of K nearest neighbors of point ξ_i in the high-dimensional data, and, respectively, x_i in the embedding. Therefore, the meaning of this sum is simply the average ratio of K nearest neighbors coinciding in the original and the embedded data. Therefore, it summarizes all 'benevolent' points which maintain a rank below K, which are also called *mild* in- and extrusions. Fig. 1 shows a schematic picture of how the co-ranking matrix is partitioned via K, and how intrusions and extrusions appear in the matrix.

To display the quality, usually a curve of $Q_{\mathrm{NX}}(K)$ is plotted for a range of different settings of K. An example is given in Fig. 2 for the classical *swiss roll* data set, which has often been used for illustration purposes in the DR literature, see e.g. [1]. In our case, the original three-dimensional data consists of 1000 points sampled from the curled two-dimensional manifold, see Fig. 2a. It was reduced to two dimensions using the well-known DR method t-SNE [3] with the perplexity parameter set to 50, see Fig. 2c. The 2D embedding produced a piecewise 'unrolled' view of the original spiral strip, where some continuous regions were separated and the data is depicted as three distinct patches.[2] In the embedding, most local neighbors stay in the proximity, corresponding to a quality close to 1, while not all neighbors are preserved in larger neighborhood sizes, see Fig. 2b. Here, the nonlinear structure of the swiss roll comes into the play: while points on different ends of the swiss roll are relatively close as measured using the Euclidean distances in 3D, these are far away in the 2D unfolding of the spiral strip. The expected quality of a random mapping serves as a baseline for $Q_{\mathrm{NX}}(K)$, see [24,25,29] for a formal derivation. It is displayed as the dotted line in Fig. 2b. Therefore, with K approaching the total number of points, quality values of 1 are reached slowly, a necessity corresponding to the baseline.

4. Point-wise quality measure

We argue that it is important to provide the user with information about the reliability of the displayed embedding, as mentioned in (b). Integrated visual cues indicating the reliability of every single mapped point can help the user reason about the original data based on the embedding. Ideally, the user is not only able to identify erroneous regions in the visualized data, but can also get an intuition about the structure of the original data. In a real visualization task, the expert user would have semantical knowledge about the data that might imply certain structural assumptions or expectations. Often, additional information is available, like the membership to semantically meaningful classes in the data. When the expert combines such semantic knowledge with the given visualization, the augmented display might help to distinguish whether the observed local errors are artifacts of the DR procedure, or structures which are in fact contradictory in

[2] Note that we did not use pairwise geodesic distances to represent the original data, despite our knowledge of the underlying manifold. Instead, we calculated Euclidean distances in the original three-dimensional space to reveal more clearly the effects of the quality evaluation. Moreover, the t-SNE method is generally more suited to embed data which is arranged in clusters, as opposed to data lying continuously on a manifold. We deliberately chose the method for this example to demonstrate the quality evaluation with the typical effects of separating or condensing neighboring points due to the method's inherent assumption of an underlying cluster structure.

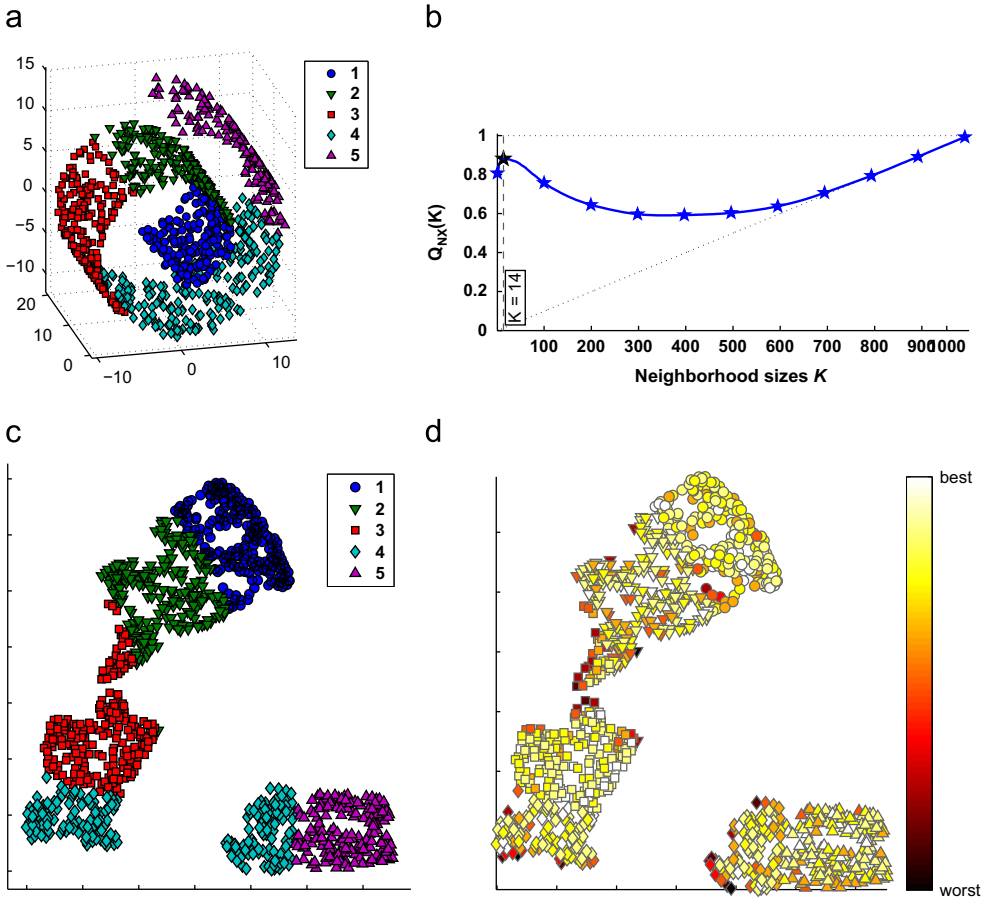


Fig. 2. An example of the qualitative evaluation for an embedding of the well-known artificial *swiss roll* data. On the upper left, the original 3D data is shown, and on the lower left is the 2D embedding obtained by the t-SNE method (with a perplexity of 50). The different symbols serve as a reference to the original positions on the spiral-shaped manifold. The upper right shows the classical evaluation via the quality graph over $Q_{NX}(K)$. The lower right shows the embedding, colored by the proposed point-wise qualities $Q_{NX}^{x_i}(14)$. While the DR method mostly 'unrolled' the original manifold rather truthfully, the strip is torn into several pieces, and the locations of the tears are clearly indicated by the coloring. (a) Original 3D data, (b) $Q_{NX}K$, (c) 2D t-SNE embedding and (d) Point-wise quality of t-SNE embedding. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

low-dimensions. While the approaches presented in [26,27] provide very effective heuristics to do so, surprisingly, none of the formal evaluation measures mentioned so far have been directly integrated into the visualization display. As mentioned, many of the measures explained in Section 2 are aggregated values consisting of point-wise quality contributions (or error rates analogously), and thus yield the possibility to be extended in such a way.

Pointwise co-ranking matrices: In the following we will derive the point-wise quality contributions which are aggregated in the measure $Q_{NX}(K)$. A co-ranking matrix can be seen as the joint histogram of ranks in the high- and low-dimensional data, as stated in [24]. For every single point, it contains the ranks of all its $N-1$ neighbors. Every co-ranking matrix $\mathbf{C}$ can therefore be decomposed into per-point permutation matrices $\mathbf{C}^{x_i}$ for every point $x_i \in X$ with $\mathbf{C} = \sum_{i=1}^{N} \mathbf{C}^{x_i}$ where

$$\mathbf{C}_{kl}^{x_i} = |\{j | \rho_{ij} = k \quad \text{and} \quad r_{ij} = l\}|.$$

Hence, the point-wise contributions of the quality Q_{NX} directly follow as

$$Q_{NX}^{x_i}(K) = \sum_{k \le K} \sum_{l \le K} \mathbf{C}_{kl}^{i} / K = |A_{\xi_i} \cap B_{x_i}| / K$$

which, averaged over all points, again yields the quality measure: $Q_{NX}(K) = \sum_{i=1}^{N} Q_{NX}^{x_i}(K)/N$.

Thus, every mapped point can be colored based on its quality $Q_{NX}^{x_i}(K)$ for relevant K. The parameter K is either chosen according to relevant structural criteria such as a local extremum of the curve Q_{NX}, or determined interactively according to the user's needs.

Fig. 2d shows an example for the swiss roll data set, where the points are colored by $Q_{NX}^{x_i}(14)$ with $K=14$ chosen according to the first local optimum of the quality curve. While it indicates small errors for almost all the points in the inner parts of the patches, it also reveals the positions of stronger topological mismatches on the borders of the visualized patches. These errors are caused by the unrolling and tearing of the original manifold, and are clearly revealed by the coloring via point-wise quality. Hence this augmentation of the DR according to local quality highlights those regions where the user cannot rely on the visualization.

5. Parameterization of the quality measure

Even in very simple settings, however, it is difficult to interpret the shape of the curve $Q_{NX}(K)$ and the related local quality measure $Q_{NX}^{x_i}(K)$, in particular the parameter K is a fairly simple, but sometimes unintuitive control mechanism. To demonstrate this problem, we will consider a few simple examples of two-dimensional toy data, where we performed no reduction of the data dimensionality, but used deliberately defined artificial 'mappings' to map the original data to a different configuration of points in the plane. Although we have direct access to the original data structure as well as the specific characteristics of the

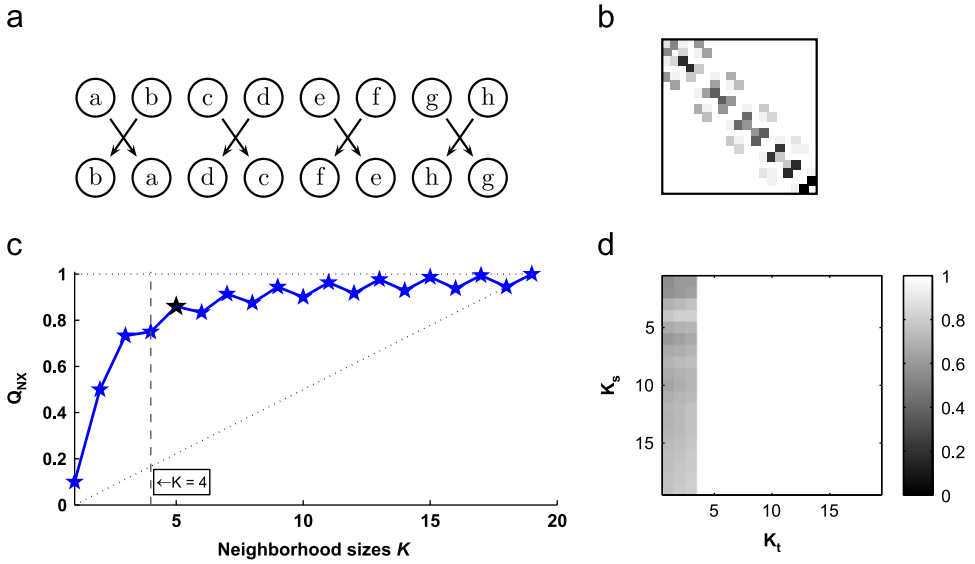


Fig. 3. The upper left shows the artificial mapping, which is a simple switching scheme of a row of one-dimensional points. Obviously, rank errors are at most four (in case of tie breaks) in this setting. This is mirrored by the shape of a co-ranking matrix for the same setting with 20 points (upper right) for which four off-diagonals are non-vanishing. However, the established measure $Q_{NX}(K)$ is below 1 for almost all K (on the bottom left), which is hard to link back to the mapping's characteristics. For our proposed measure (on the bottom right), $Q_{ND}(K_s,K_t)=1$ for all $K \geq 4$. (a) Switching of points, (b) $\mathbf{C}$, (c) $Q_{NX}(K)$ and (d) $Q_{ND}(K_s,K_t)$.

mapping in these cases, we found that from the results of the quality measure Q_{NX} the types of the deliberately implanted errors are hard to recognize.

First, we consider a very simple scenario: a row of equidistant points is mapped to a row where the points are swapped in pairs, as depicted in Fig. 3a. Any even number of points could be chosen arbitrarily. When examining this scenario, we find that the maximum absolute rank error between the original and the switched points is 4 for the entire data set, and independent from the total number of points (for example, when point d moves left, and its right neighbor e moves right).[3] Intuitively, if we consider rank error sizes up to 4 as acceptable, this mapping is perfect. This is, however, not indicated by $Q_{NX}(4)$ in the graph in Fig. 3c (which displays the quality for a row of 20 points which are swapped in this manner): the quality is below one for most $K \geq 4$. It is hardly possible to gain insight about the characteristics of the errors based on the observation of $Q_{NX}(K)$ and the mapped points alone, although the errors in this scenario can be fully characterized by local pairwise swapping.

The problem arises, because small rank errors can have an effect over larger ranges of K: regarding some reference point x_i, let us consider a faraway neighbor x_j with the original rank $\rho_{ij}=K$ For the quality $Q_{NX}(K)$, this point is considered benign (i.e. it adds to the quality) as long as its rank stays at K or intrudes to some lower rank $1 \leq r_{ij} < K$, whereas this neighbor would be regarded as erroneous immediately with just a slightly higher rank of $K+1$ for instance. On the other hand, a close neighbor, e.g. with rank 1, is allowed to extrude up to a rank of K and still adds to the quality rating, although the rank difference can be rather large. This seems to be an unbalanced characteristic of the quality measure in general.

A look at the co-ranking matrix in Fig. 3b reveals the distribution of rank changes for this simple example. Since the rank error is always smaller than 5, only 4 off-diagonals of the co-ranking matrix are not equal to 0, since the ith off-diagonal corresponds to rank errors of size i. However, the quality Q_{NX} is a sum over a square block of the co-ranking matrix, like many other DR evaluation measures described in [24]. This observation also suggests how the quality measure can be altered to achieve a more appropriate parameterization: rather than considering a rectangular sub-matrix, it should focus on a limited number of off-diagonals corresponding to the size of the rank deviation which is considered to be acceptable.

Looking at the rather comprehensible and straightforward definition of Q_{NX}, we find that the parameter K serves two different purposes: on the one hand, K identifies a region of interest by determining the size of the neighborhood of every point in the original data, namely $\rho_{ij} \leq K$. On the other hand, it determines the size and shape of errors which are tolerated for points in the region of interest: every $r_{ij} \leq K$ is acceptable and adds to the overall quality. This parameterization has the effect that small rank errors can contribute to the shape of the curve $Q_{NX}(K)$ on every scale of K. While there is no immediate drawback when merely comparing several DR methods, i.e. the usage scenario described in (a) in Section 2, the effect can be problematic when a fine-grained analysis of a visualization is desired, like in case (b) in Section 2. As stated in Section 2, the quality measure Q_{NX} is similar (or equal) to several other evaluation criteria which rely on the same part $\mathbf{C}_{kl},k,l \leq K$ of the co-ranking matrix. Hence, this problem is present in all these evaluation measures.

To circumvent the described problem, we propose a different, more fine-grained assessment of quality based on the co-ranking matrix, which (i) identifies benign points by their amount of deviation from the original rank, rather than their absolute rank in the embedding, and (ii) allows for separate control over the region of interest and the size of the tolerated errors. We therefore replace the single parameter K by the pair (K_s,K_t), where K_s determines the region of interest (alias the significant ranks) and K_t is the size of tolerated rank errors. Further, rather than tolerating errors within a certain region of the projection, we explicitly consider a limit on the absolute rank errors. The new measure is defined as

$$Q_{ND}(K_s,K_t) = \frac{1}{K_s N} \sum_{i \leq K_s} \sum_{j:|i-j| \leq K_t} \mathbf{C}_{ij}.$$

Since the second sum is limited to entries $j : |i-j| \leq K_t$, i.e. rank errors $|i-j|$ smaller than K_t, we now sum over a part of the co-ranking

[3] Note that for these equidistant points, ties in the pairwise distances need to be broken to arrive at proper ranks. In case of a tie, we define that the point with the lower alphabetical letter gets assigned the lower rank.

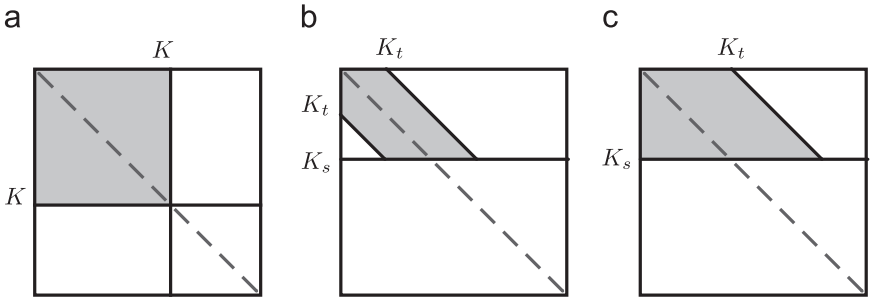


Fig. 4. Change of the summation area of the co-ranking matrix for precise control over the region of interest and the tolerated rank errors. (a) Original $Q_{NX}(K)$, (b) $Q_{ND}(K_s,K_t)$, K_t limits w.r.t. diagonal and (c) $Q_{ND}(K_s,K_t)$ for $K_s=K_t$.

matrix which is oriented according to the diagonal, see the schematic in Fig. 4b. By controlling the two parameters of the quality measure, the user can assess the compliance with specific requirements for the embedding. For example, common tasks would be to assess:

(I) the preservation of local relationships (chosen by small K_s and small K_t);
(II) the amount of errors originating in fairly local neighborhoods, but are deviating largely from the original rank (small K_s, large K_t);
(III) the preservation of global relationships in the data (large K_s, smaller K_t).

To get a rich impression of a visualization's qualitative characteristics, the quality $Q_{ND}(K_s,K_t)$ is now parameterized by two values. Hence, rather than in a single curve, the results are now represented by a surface. The full quality surface can easily be displayed as a colored matrix, where the position (K_s,K_t) is assigned a color value according to $Q_{ND}(K_s,K_t)$, see Fig. 3d for an example. The matrix in Fig. 3d shows the results for our example of 20 swapped points. It clearly reveals that all entries for $K_t \geq 4$ yield the maximum quality, which is the expected behavior.

In the following artificial example, we will further demonstrate the more directly controllable characteristics of our approach. We consider three simple scenarios, mapped from two-dimensional points to a new point distribution in 2D. The original data consist of three well-separated Gaussian clusters, containing 100 points each, see Fig. 5a. As a 'mapping', we consider the points obtained by (i) a random permutation of the points within every cluster, see Fig. 5b, (ii) a switch of the two leftmost clusters, see Fig. 5c, and (iii) the middle and leftmost cluster stacked on top of each other, see Fig. 5d. These artificial mappings represent typical behavior of DR embeddings since they capture (i) local distortions, (ii) a tearing of regions, and (iii) an overlay of regions, which are common effects due to the low dimensionality of the projection space.

The resulting curves for Q_{NX} are depicted in Fig. 6a. Although we know the exact behavior of the mapping in this case, it is not easy to link the entire shape of the curves to the characteristics of the respective mapping. In setting (i), the random permutation of points within the clusters causes a vast number of local errors, which is clearly indicated by the low quality for $K < 100$. Farther neighbors change their rank as well, because of the permutation within the neighboring clusters. However, the absolute size of all rank errors in the mapping is strictly below 100, when considering only a single cluster, which cannot be inferred on the basis of the quality curve. The quality matrix for the new measure Q_{ND} in Fig. 6b clearly shows the errors which are present rather steadily over all scales of K_s, whereas the quality is perfect for all pairs $(K_t > 100, K_s < 100)$, which implies that the absolute size of rank

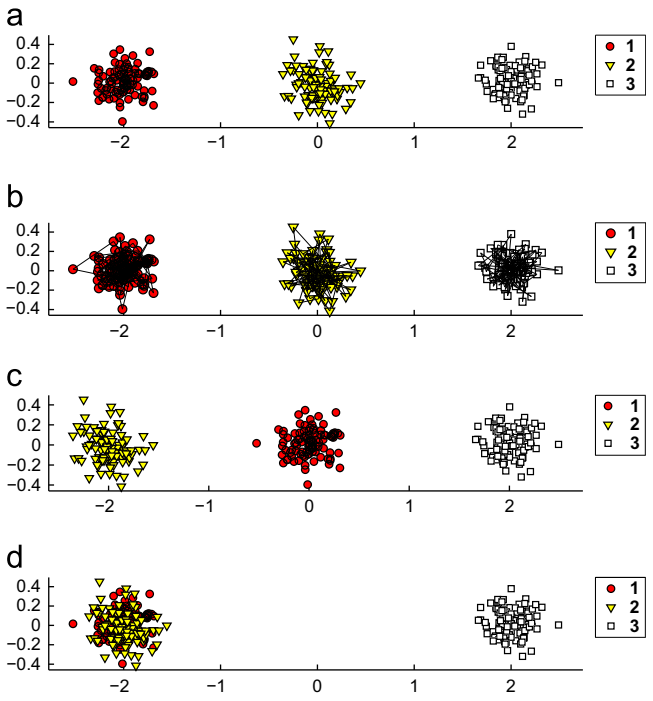


Fig. 5. These scatter plots show a simple example of deliberately designed artificial mappings, which resemble typical effects of DR procedures and serve as a demonstration benchmark for our quality evaluation. The upper plot depicts the original data consisting of 3 Gaussian clusters in 2D. The plot below shows how the data was randomly shuffled within each cluster, where black lines are drawn from every point to its original position, to demonstrate the permutation. The last two plots show, respectively, how two clusters are switched, and stacked on top of each other. (a) Original clusters, (b) Mapping (i): shuffled within clusters, (c) Mapping (ii): left two clusters switched and (d) Mapping (iii): left two clusters merged.

errors caused by the mapping for a single cluster is below this range. When considering large neighborhoods of interest with $K_s > 100$, the quality is very good for $100 < K_t < 150$, and perfect for all $K_t > 150$. The type of errors that appear here are more rare, the extreme case would be, that a point on the very right of a cluster moves to the very left of its cluster, and a neighbor originally on the left, moves to the very right of its cluster. The absolute rank error for this type cannot exceed half of the total number of points, as indicated by $Q_{ND}(K_s,K_t) = 1$ for all $K_t > 150$. This mapping refers to the evaluation tasks (I) and (III) as described in Section 5, i.e. the upper left part of the quality surface for (I), and the lower left/middle part for (III).

For mapping (ii), the curve of Q_{NX} in Fig. 6a reveals that there are no errors on a small neighborhood scale (below the cluster size of 100), whereas the quality drops severely beyond this scale. The corresponding matrix of Q_{ND} in Fig. 6c gives us the same

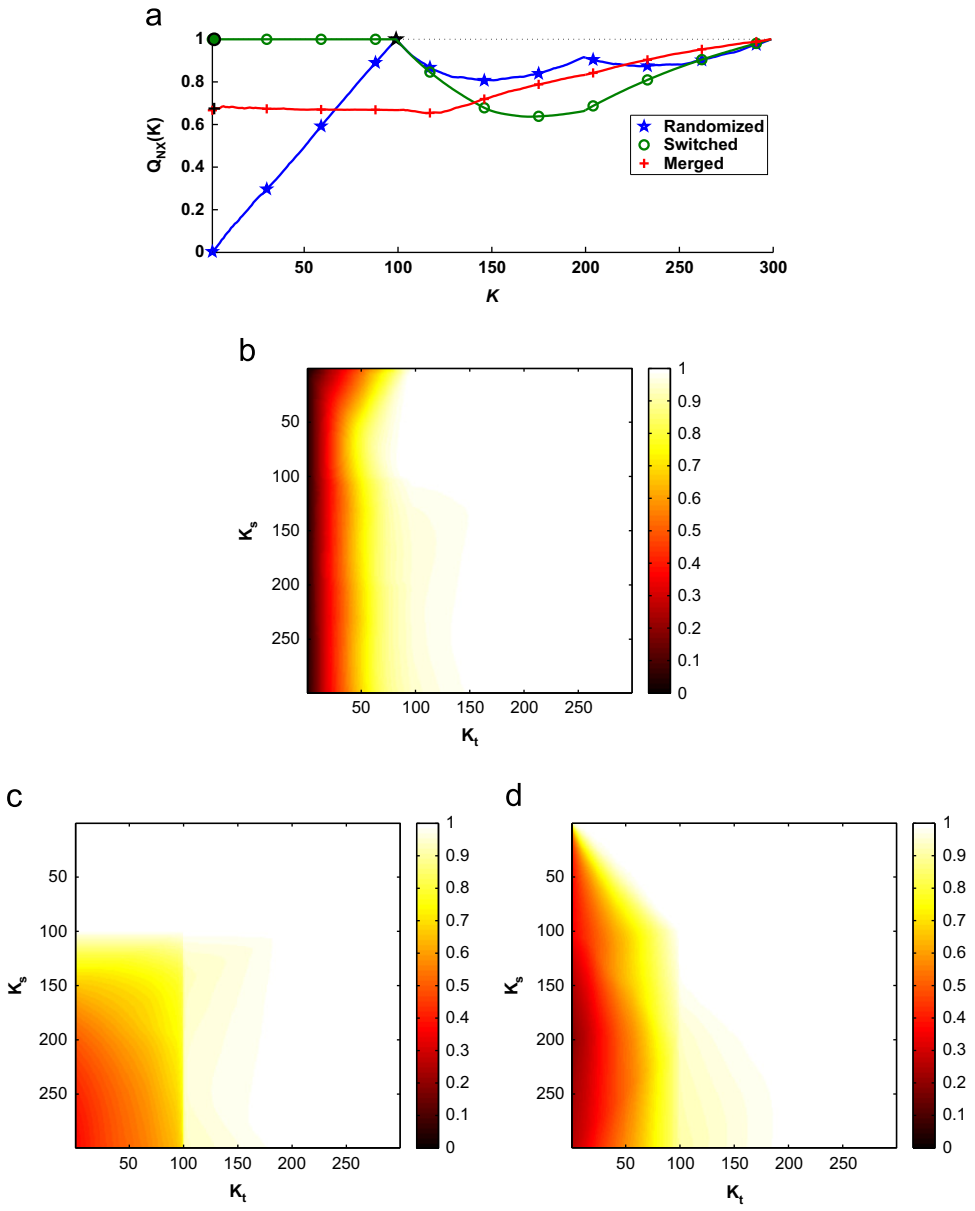


Fig. 6. The figures show the evaluation for the artificial mappings from Fig. 5, first with curves obtained by the classical quality measure Q_{NX} (the top figure), and with the quality surfaces resulting from the newly proposed Q_{ND} measure (in the three remaining figures), each being a counterpart to one of the curves above. (a) $Q_{NX}(K)$, (b) $Q_{ND}(K_s,K_t)$, shuffled within clusters, (c) $Q_{ND}(K_s,K_t)$, left two clusters switched and (d) $Q_{ND}(K_s,K_t)$, left two clusters merged.

information, but also reveals that the absolute size of rank errors is below 200 by showing a perfect quality for all $K \geq 200$. This is expected, since there are only two clusters involved in errors. We also see a sharp rise in quality at $K_t = 100$, because for the points of the rightmost cluster, two-thirds of all neighbors (the points of the other clusters) change their ranks by exactly 100 due to the switching. From the perspective of the leftmost cluster, there are also some rank errors of size 100–200, which is indicated by the slight coloring in the region $100 \leq K_t < 200$. Mapping (ii) refers to the evaluation task (III) in Section 5, i.e. the lower left to middle part of the surface.

In the evaluation for mapping (iii), the curve of Q_{NX} shows a steadily diminished quality until $K \approx 125$, a scale which cannot be linked to the structural knowledge about the data. Thereafter, the curve steadily rises to the maximum. From this, we can gather that there are relatively little global errors, however, the matrix for Q_{ND} in Fig. 6d gives more insight: we can see that the errors originating in small regions (small K_s) are rather small, i.e. there are errors only for $K_t < K_s$ approximately. On larger scales, the number of errors increases along with the tolerance K_t, which implies that the absolute size of rank errors increases. This is expected, since the stacking of the clusters causes small deviations from the original ranks when considering small neighborhoods, as well as large errors when considering large neighborhoods. However, the quality is perfect for $K_t > 200$ which, again, suggests that there are only two clusters involved in the occurring errors. This mapping is linked to the evaluation tasks (I) and (II).

If the computational cost to calculate $Q_{ND}(K_s,K_t)$ for all pairs of $(K_s,K_t) \in \{1 \ldots N-1\}^2$ should be reduced in a practical visualization scenario, it is reasonable to calculate only the quality values for the following three curves instead of the full surface:

- For $Q_{ND}(K_s,K_t)$ with $K_s = K_t \in \{1 \ldots N-1\}$, which resembles the original curve from $Q_{NX}(K)$ over growing neighborhood sizes,

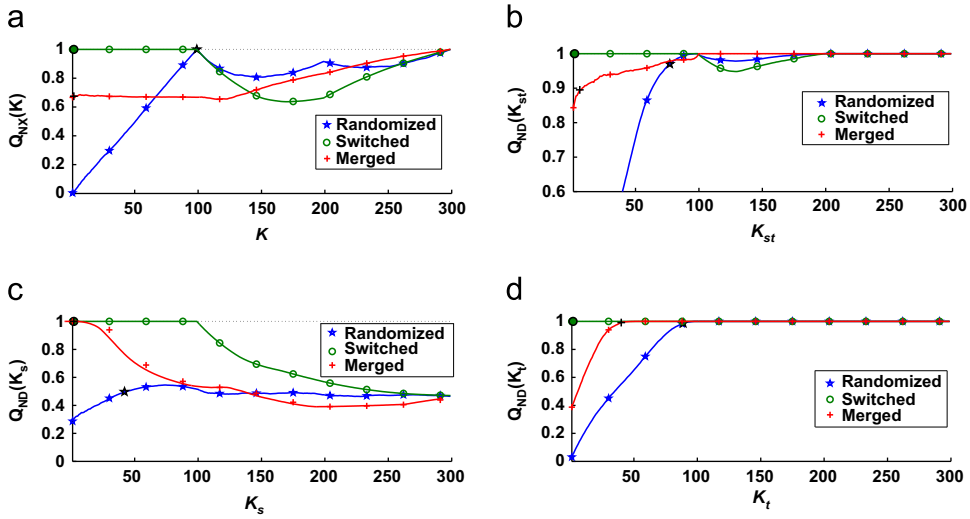


Fig. 7. This figure shows a different representation of the qualitative evaluation presented in Fig. 6 for the artificial mappings of three clusters shown in Fig. 5. In the upper left are the curves for $Q_{NX}(K)$ (the same as in Fig. 6a). The other three plots show an alternative, lean representation of the Q_{ND} measure, which needs less computational effort. The upper right shows $Q_{ND}(K_{st})$ for all $K_{st} \coloneqq K_s = K_t \in \{1 \dots N-1\}$, meaning that the measure tolerates all absolute rank errors which are smaller than the current neighborhood of interest. (Here, the y-axis is scaled differently to highlight the details.) The graphs in the bottom left capture the mapping's characteristics under a fixed assumption about the failure tolerance, as only the region of interest grows, in this case for an error tolerance of 30. The bottom right shows the graphs for $Q_{ND}(30,K_t)$ for all $K_t \in \{1 \dots N-1\}$, i.e. the average quality of the 30 nearest neighbors, as the failure tolerance increases. These graphs capture the essential content of the surfaces shown in Fig. 6, requiring far less computational effort. (a) $Q_{NX}(K)$, (b) $Q_{ND}(K_{st})$, (c) $Q_{ND}(K_s,30)$ and (d) $Q_{ND}(30,K_t)$.

but with a different area of summation, taking the error size into account; the corresponding part of the co-ranking matrix is depicted in the schematic in Fig. 4c. This means that the size of tolerated errors is growing as the considered region of interest gets larger. We then denote the measure by $Q_{ND}(K_{st})$ with $K_{st} \coloneqq K_s = K_t$.

- For the mapping's characteristics under a fixed assumption about the failure tolerance, as only the region of interest grows, i.e. $Q_{ND}(K_s,K_t)$ for all $K_s \in \{1 \dots N-1\}$ and fixed K_t.
- For a fixed neighborhood size of interest, as the failure tolerance increases, i.e. $Q_{ND}(K_s,K_t)$ for all $K_t \in \{1 \dots N-1\}$ with a constant K_s.

In the latter two cases, the respective fixed parameters can be selected according to the user's prospect, e.g. on which scale the visualization is required to be trustworthy. Fig. 7 shows how the combination of these curves for Q_{ND} offers an adequate approximation of the full surfaces from Fig. 6.

The evaluation in these artificial cases is simplified by the assumption of equal cluster sizes, which yield a natural threshold for the parameters. While this was helpful to clarify the proposed parameterization, the benefits of the two parameters become apparent when the aggregated overview of the mapping quality is combined with a point-wise evaluation, which is introduced in the following section.

Controllable point-wise quality: For the new measure Q_{ND}, the definition of the point-wise quality is analogous to the one from Section 4:

$$Q_{ND}^{x_i}(K_s,K_t) = \sum_{k \le K_s l:|k-l| \le K_t} \mathbf{C}^i_{kl}/K_s$$

where $Q_{ND}(K_s,K_t) = \sum_{i=1}^{N} Q_{ND}^{x_i}(K_s,K_t)/N$. Here, the benefits of the new parameterization are particularly noticeable, since the user is able to tune the parameters to make specific types of embedding errors directly visible. We consider the example from Fig. 2, and show how the previous point-wise quality compares to the new definition in Fig. 8. The problem of Q_{NX} described above becomes apparent when looking at $Q_{NX}^{x_i}$: at a scale K, the measure considers very different types of errors at once. In this case, we see small rank errors caused by fairly local permutations of points within the unfolded pieces of the strip, which exhibit a lighter color. Also, we observe a small number of strongly colored points on the edges where tearing occurred and lead to larger rank errors. In contrast, the new measure $Q_{ND}^{x_i}$ exclusively singles out the tearing, since in this case the small rank errors within the unfolded patches are below the tolerance threshold of $K_t=14$, while larger errors which originate in small regions of 14 nearest neighbors (K_s) are caused by the tearing only, and diminish the quality in this parameter configuration.

6. Experiments with real-world data

In this section, we demonstrate the quality evaluation framework on two real-world data sets, and showcase the augmented visualization along with the classical evaluation by the quality curve Q_{NX}. For the dimensionality reduction of the data, we applied the standard linear technique PCA projecting the data on the first two principal components, as well as the well-known modern nonlinear method t-SNE (Figs. 9 and 10).

Runner data: The first data set is a motion capture sequence, freely available from the *Open Motion Data Project* at the Ohio State University.[4] It contains the three-dimensional positions of 34 tracking markers over 217 time steps. The sequence shows a person, who begins to run from a forward-leaning position, and takes about five strides during which the inclination of the body becomes upright. Fig. 11 shows 2D-embeddings of the original 102-dimensional points. To clarify the sequential relation of the visualized data, we connected points from consecutive frames by a line. We deliberately chose this data set, because here the user has additional knowledge about the original underlying manifold, and can directly inspect where a DR technique did not represent the manifold truthfully. Therefore, the visual augmentation to detect tearing and overlapping of the manifold is superfluous.

[4] sequence *Figure Run 1* from http://accad.osu.edu/research/mocap/mocap_data.htm.

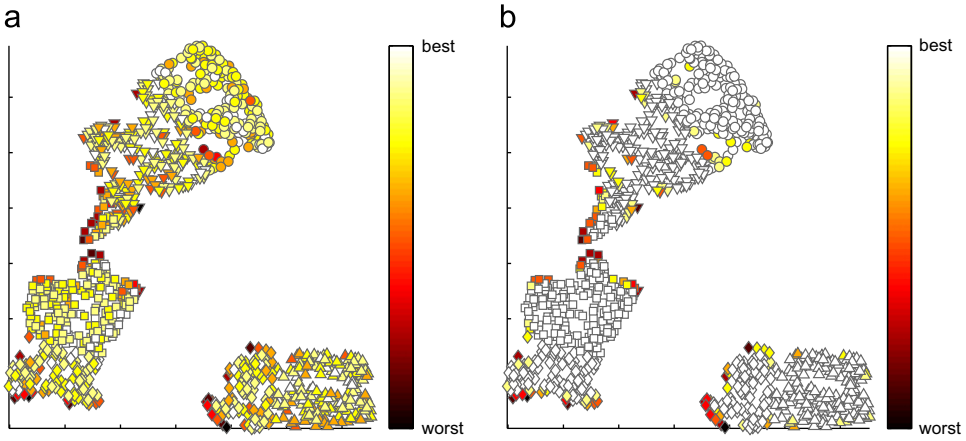


Fig. 8. This figure compares the two proposed point-wise quality measures for the same embedding of the *swiss roll* data by t-SNE, as introduced in Fig. 2. On the left, we show the points colored according to $Q_{NX}^{x_i}(14)$ (the same as in Fig. 2d). On the right, the points' color coding is obtained by $Q_{ND}^{x_i}(14,14)$. Both measures highlight the tearing of the original manifold, but $Q_{ND}^{x_i}$ shows only the torn regions and almost no local errors within the unrolled patches, since absolute rank errors below 14 are explicitly tolerated. (The sequence of class labels from the inside to the outside of the original spiral-shaped manifold is: ○▽□⋄△, see Fig. 2a.). (a) $Q_{NX}^{x_i}(14)$ and (b) $Q_{ND}^{x_i}(14,14)$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

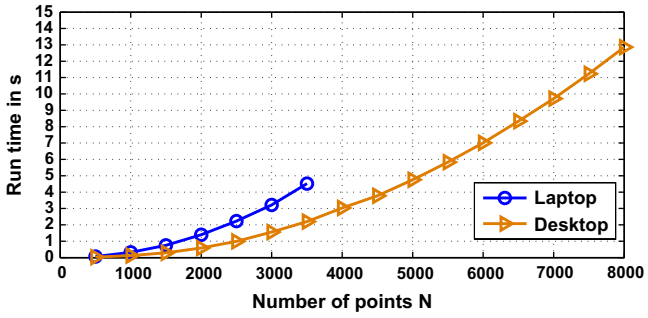


Fig. 9. The figure shows a small survey about the computing time to calculate the co-ranking matrix from given ranks ρ_{ij},r_{ij}. The ranks were calculated from uniformly random points in 10 dimensions which were randomly mapped to points in two dimensions. We used various data set sizes $N \in \{500,\ldots,8000\}$ and tracked the run time of a standard Matlab implementation. One curve shows the computation times on a standard laptop machine with a 2.0 GHz dual core processor and 2 Gigabytes of RAM, where the memory limitation allowed a maximum set size of 3500 points. The other curve represents the same experiment on a modern desktop computer using 4 CPU cores with 2.5 GHz each, and 6 Gigabytes of memory.

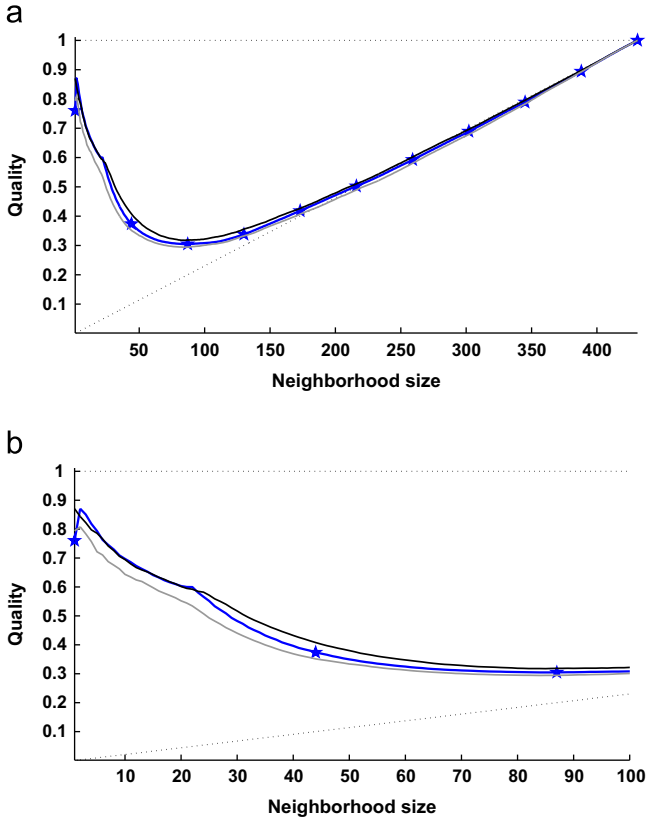


Fig. 10. The figure shows the quality Q_{NX} for random subsamples of the t-SNE embedding from Fig. 15. As a reference, the line marked by pentagrams is the original quality from the full data set, as given in 14a. We sampled 20 times a random subset of 30% of the total points, and calculated the Q_{NX} based on this subset only. From the 20 iterations, the gray line shows the minimal outcoming value for the respective neighborhood size, while the black line shows the maximum. The neighborhood sizes of the original curve were aligned in relation to the respective value in the sampled case. The upper figure displays the graphs for all possible neighborhood sizes, and the lower figure shows a zoomed view, focusing on the neighborhoods up to 100 points only. The deviation from the original curve is fairly small, although the co-ranking matrix is based only on the subsample. (a) Subsampling of Q_{NX} and (b) Subsampling of Q_{NX}.

However, since data lying on an (unknown) underlying manifold structure are common in general practical applications, this showcases the insights we can gain from the point-wise quality evaluation.

The embeddings of both, PCA and t-SNE, show a similar shape of a 'tail' which leads into a spiral structure. This can be explained by the sequence starting from a leaning posture (the tail) and progressing to several strides of upright running (the spiral). In case of PCA, the sequence of points is overlapping at several positions, while the t-SNE method splits some of the consecutive points apart but shows less overlap in general. The t-SNE embedding also produced some crowds and zig–zag shapes along the point sequence. We report the corresponding quality curves in Fig. 12.

In case of the PCA, both measures $Q_{NX}^{x_i}$ and $Q_{ND}^{x_i}$ show that some of the overlapping regions have a reduced quality. However, the measure Q_{ND} identifies less regions to be severely erroneous, due to the tolerance of $K_t=20$, compare, for example, the overlap on the left of the scatter plots in Fig. 11a and b. This indicates that the rank error for these points must be below 20, while the highlighted regions contain larger errors. Depending on the practical purpose, the user may want to be aware of the severe mismatches, neglecting tolerable errors. Similar characteristics can be observed in the Fig. 11c and d for t-SNE. The tearing of the sequence is distinctly highlighted as erroneous in the coloring by

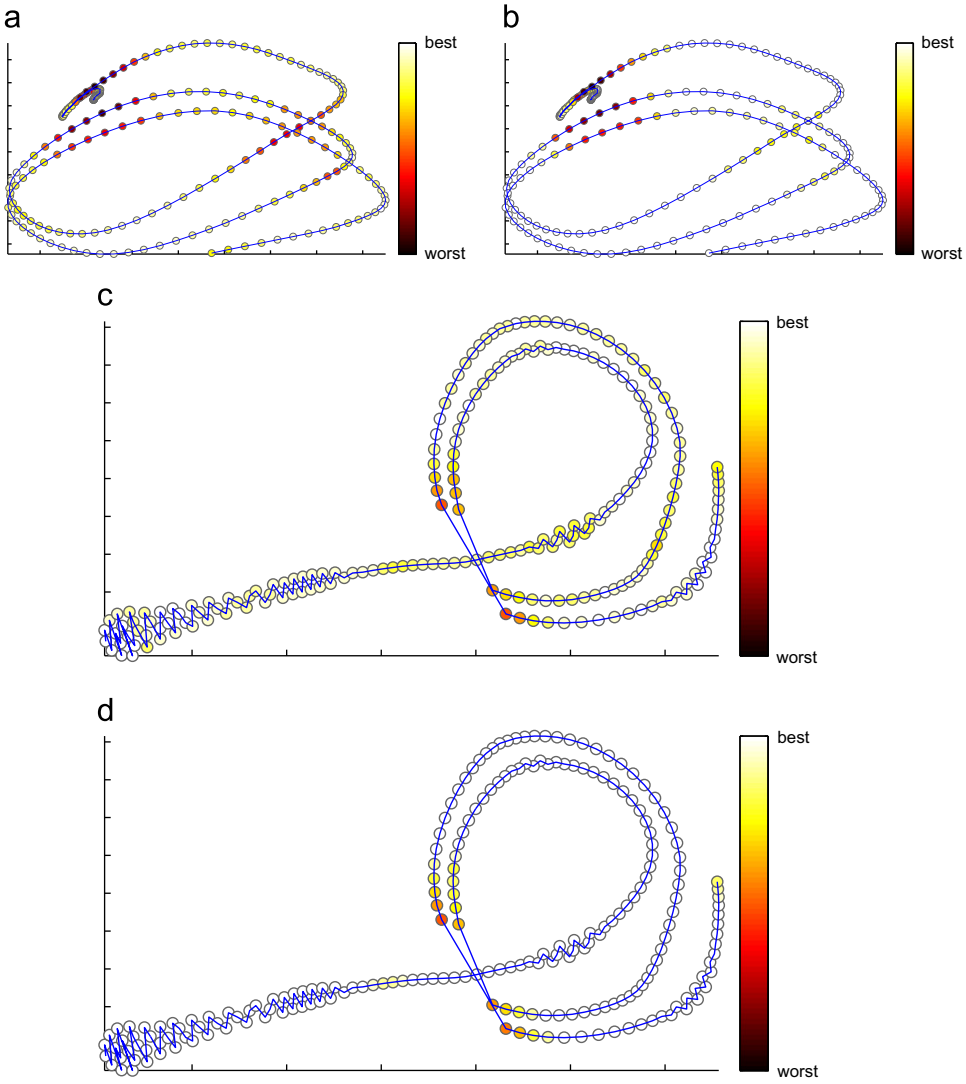


Fig. 11. The figure shows a PCA and a t-SNE embedding (with perplexity 30) of the *runner* data, each colored by the pointwise quality $Q_{NX}^{x_i}(20)$ and $Q_{ND}^{x_i}(20,20)$, respectively. Points from consecutive motion capture frames are connected by a line. For $Q_{NX}^{x_i}(20)$, various types of errors arising in neighborhoods of 20 points are highlighted all at once, while $Q_{ND}^{x_i}(20,20)$ is able to identify where the neighbors deviate from their original rank by more than 20. This gives a clearer indication of where the underlying manifold is not truthfully represented, e.g. torn apart in case of t-SNE. (a) PCA, $Q_{NX}^{x_i}(20)$, (b) PCA, $Q_{ND}^{x_i}(20,20)$, (c) t-SNE, $Q_{NX}^{x_i}(20)$ and (d) t-SNE, $Q_{ND}^{x_i}(20,20)$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

$Q_{ND}^{x_i}$. Here, the advantage of the parameterization becomes particularly apparent: the user may choose via K_t not to highlight the tolerable local errors which are caused by the crowding and zig–zag patterns.

COIL-20 data: The second data set, the *Columbia University Image Library* from [33], consists of 1440 gray-value images in a resolution of 128×128, which show 20 small objects, photographed from 72 consecutive rotation angles. Each class in the data corresponds to photos of one particular object. Because of the consecutive angles, we can assume that the original data is clustered by their class membership, and that within a class, the data are situated on a ring-shaped manifold.

In Figs. 13 and 15, we show an embedding by PCA and t-SNE, respectively, with points colored by their quality $Q_{ND}^{x_i}(10,10)$. Since we chose $K_s = K_t = 10$, the measure highlights absolute rank errors larger than 10, originally situated among the 10 nearest neighbors of a point. Fig. 14 shows the corresponding quality curves. The PCA embedding is generally of a low quality, as indicated by the coloring in Fig. 13b, and the curves in Fig. 14. For the t-SNE embedding, the coloring in Fig. 15b reveals distinct defects in the clusters, seemingly either caused by the tearing or contracting of the original manifold within a cluster, or by overlaying separated clusters.

MNIST data: The third data set *MNIST* from [34] consists of 60,000 gray-value images of handwritten digits[5] from 0 to 9. Each image comes at a resolution of 28×28 and is therefore represented as a vector of 784 dimensions. Applying t-SNE on the full data set of 60,000 images is not feasible in terms of memory demand and computational effort. We therefore used a random sample of 10,000 points for our experiments. For this data set, we have no prior assumption about an underlying manifold structure, but we can assume that there are clusters according to the ten digits.

Fig. 16 shows a visualization with t-SNE. We now omitted the corresponding PCA embedding since it shows a considerably inferior quality, similar to the case of the *COIL-20* data. In Fig. 18, the embedding is colored by the point-wise quality $Q_{ND}^{x_i}(300,300)$, and Fig. 17 shows the quality curves. The embedding shows the expected cluster structure according to the digits;

[5] For further information, see http://yann.lecun.com/exdb/mnist/.

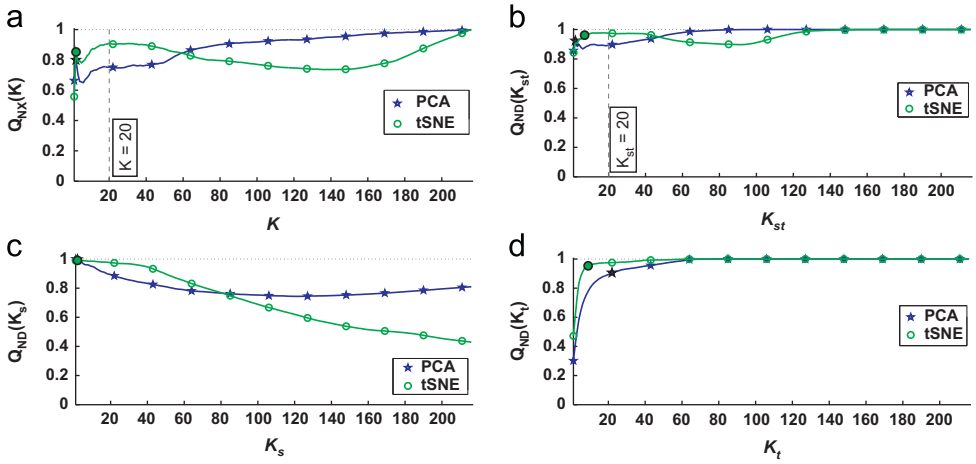


Fig. 12. The quality curves of Q_{NX} (upper left figure) and Q_{ND} (the remaining figures) for the PCA and the t-SNE embedding of the *runner* data, see Fig. 11. The curves show that the t-SNE embedding is more truthful in displaying the local relationships, whereas PCA preserves more of the ranks in the larger neighborhoods. The curves for $Q_{ND}(K_s,20)$ in the lower left (for a fixed error tolerance of 20) reveal that there are many errors in the t-SNE embedding when considering larger neighborhoods. On the other hand, the curves over $Q_{ND}(20,K_t)$ in the lower right show that the errors in the t-SNE embedding are only of very small magnitude when considering 20 nearest neighbors. (a) $Q_{NX}(K)$, (b) $Q_{ND}(K_{st})$, (c) $Q_{ND}(K_s,20)$ and (d) $Q_{ND}(20,K_t)$.

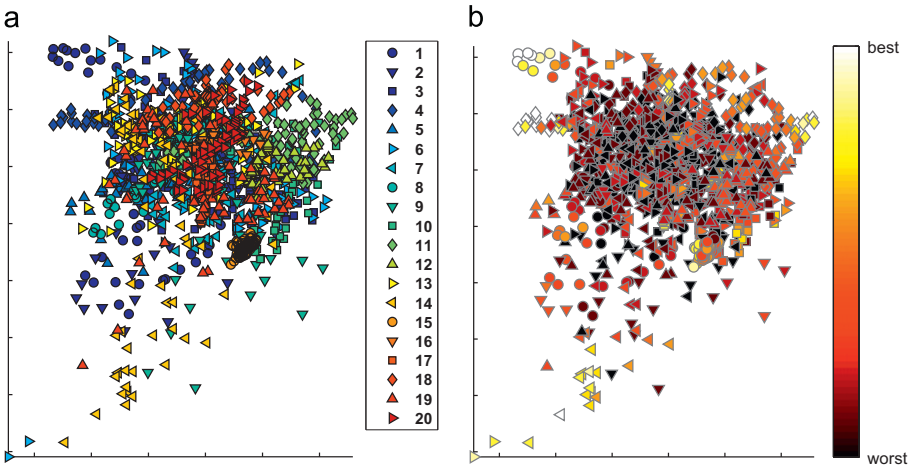


Fig. 13. On the left, we show a PCA embedding of the *COIL-20* data where each point's class membership is represented by a distinct combination of a marker symbol and color. On the right, points are colored by their quality $Q^{x_i}_{ND}(10,10)$. The embedding exhibits a low quality at almost every position, since there are many rank errors >10 occurring in the 10 nearest neighbors. (a) PCA embedding and (b) $Q^{x_i}_{ND}(10,10)$ for PCA embedding. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

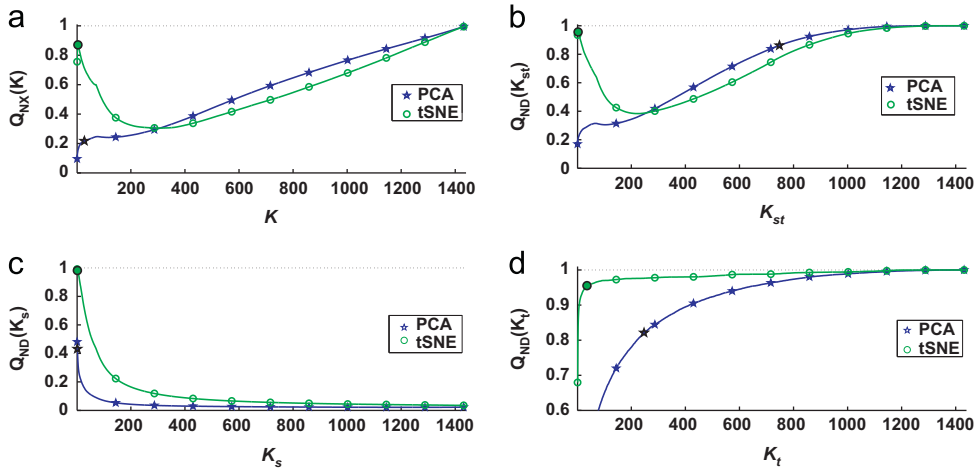


Fig. 14. This figure shows qualitative evaluations for the PCA and t-SNE embeddings of the *COIL-20* data (see Figs. 13a and 15a). The curves in the upper left figure are the quality Q_{NX}, while the other three figures show the quality curves resulting from the Q_{ND} measure. Both, Q_{NX} and Q_{ND} clearly identify that the PCA embedding fails to reliably represent the local relationships, while the t-SNE embedding sacrifices some of the global relationships but is generally depicting the smaller neighborhoods rather truthfully. (a) $Q_{NX}(K)$, (b) $Q_{ND}(K_{st})$, (c) $Q_{ND}(K_s,10)$ and (d) $Q_{ND}(10,K_t)$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

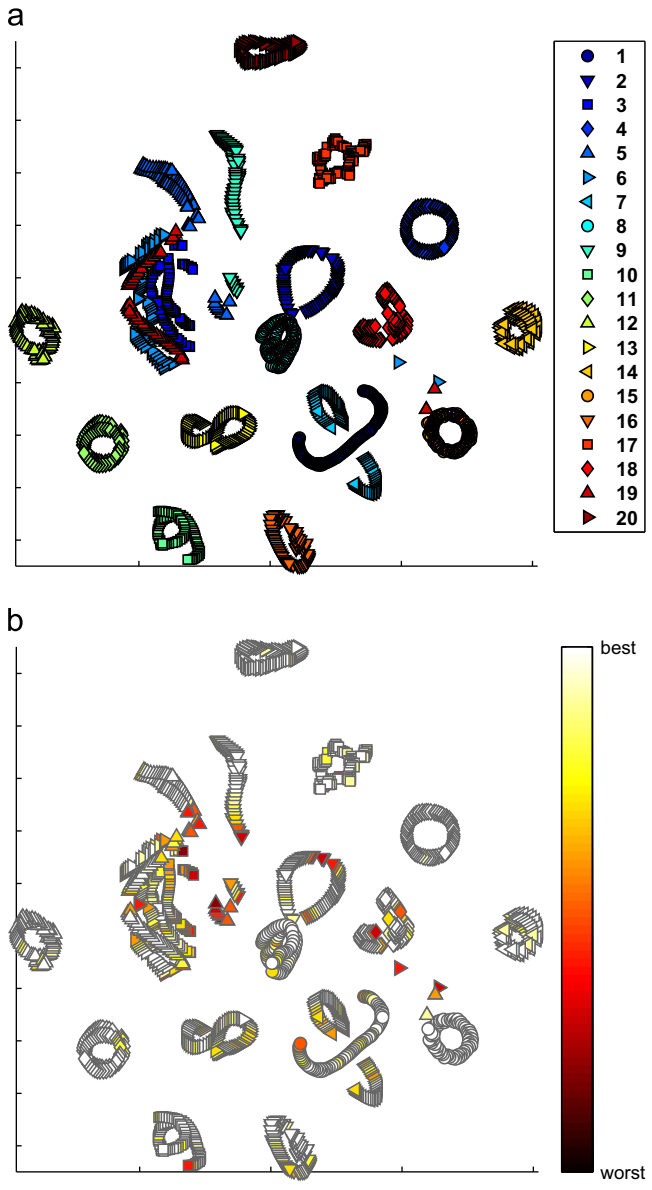


Fig. 15. The figure shows a t-SNE embedding of the *COIL*-20 data, the perplexity parameter was set to 15. In the upper visualization, points are marked according to their classes by combinations of marker shape and color. The points in the lower picture are colored by their quality $Q_{ND}^{\kappa_i}(10,10)$. Since the region of interest K_s as well as the failure tolerance K_t were both set to 10, we can see where some of the clusters, which are assumed to be on a ring-shaped manifold originally, have been torn or contracted severely. (a) t-SNE embedding and (b) $Q_{ND}^{\kappa_i}(10,10)$ for t-SNE embedding. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

however, the classes are only weakly separated and they are partially overlapping. Although the overall quality is diminished by these effects, we can see in the point-wise evaluation that the errors are less pronounced for the digits 0, 1, 6, and 7. The other digits show a lower quality, especially in the border regions, presumably caused by the stronger overlaps.

Computational effort and speedup: In real world data sets, such as *MNIST*, sizes in the order of several thousand data points become more and more common. Since the computational demand for the discussed quality evaluation is rather high, we address this topic shortly. If ranks have been calculated, assembling the pointwise co-ranking matrices requires a lookup operation for every pair of points, therefore the time complexity is $\mathcal{O}(N^2)$. To give an impression of the practical computational effort,

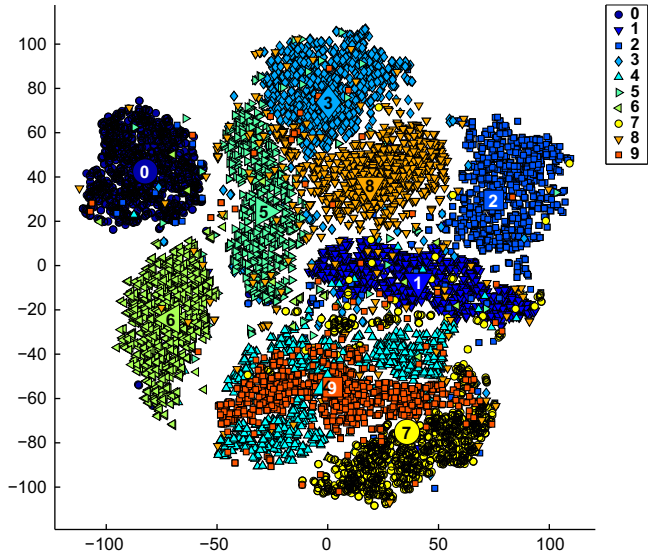


Fig. 16. This figure shows a t-SNE embedding of the *MNIST* data consisting of 10,000 data points, where the perplexity parameter was set to 30. Each point's class membership is represented by a distinct combination of a marker symbol and color. Additionally we highlighted the corresponding digit in each cluster center. We see that the data are arranged in clusters according to the classes, but are generally close together with some significant overlaps between classes. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

we tracked the run time to calculate the classical co-ranking matrix for random mappings of sizes between 500 and 8000 points on a standard laptop, as well as a modern desktop computer, see Fig. 9.

To tackle this practical issue, we investigated in how far a random subsampling of the points affects the outcome of the quality. In a small experiment, we performed a subsampling of the t-SNE embedding of the *COIL*-20 data from Fig. 15, where we randomly sampled 30% of the points, i.e. 432 out of 1440 (using the same subset of the original as well as the embedded points). We repeated this procedure 20 times, evaluating the quality Q_{NX} every time. In Fig. 10 we show the respective maximum and minimum of the resulting curves together with the original quality curve as given in 14a. The figure shows that the deviation from the original curve is relatively small, from which we can conclude that subsampling seems to be a valid possibility to approximate the quality evaluation using less computational effort. While this shows only Q_{NX} exemplarily, we observed a similar effect for the Q_{ND} measure.

Subsampling could open the way towards an interactive graphical user interface, where the user can observe a given visualization augmented by the point-wise quality, and directly try different parameter settings for K_s, K_t. Since the computational effort and memory demands can be limited by sampling a fixed number of points, the interface can be updated instantly and the user can quickly browse various combinations. Together with techniques to accelerate the DR process itself, see e.g. [35], mapping and evaluation would become feasible even for very large data sets.

7. Discussion and future work

We have discussed existing quality measures for dimensionality reduction with regard to their suitability for practical usage scenarios. While there are several evaluation methods which are suited for an overall assessment and comparison of different

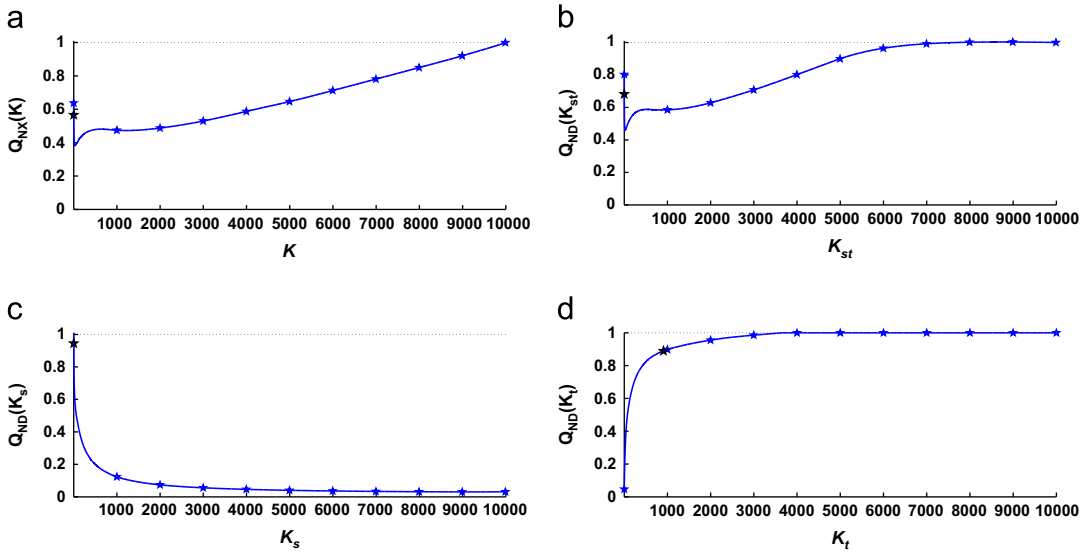


Fig. 17. This figure shows the qualitative evaluations for the t-SNE embedding of the *MNIST* data (see Fig. 16). The curves in the upper left figure are the quality Q_{NX}, while the other three figures show the quality curves resulting from the Q_{ND} measure. The curves indicate that the distortions in the mapping are generally quite large. Even in a range of 70 neighbors, there are many errors, and the rank errors have a size of up to 3000, see the curve for $Q_{ND}(70,K_t)$. Only the very small neighborhood ranges are depicted rather truthfully, as seen on the very left of the curves for $Q_{NX}(K)$ and $Q_{ND}(K_{st})$. (a) $Q_{NX}(K)$, (b) $Q_{ND}(K_{st})$, (c) $Q_{ND}(K_s,70)$ and (d) $Q_{ND}(70,K_t)$. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

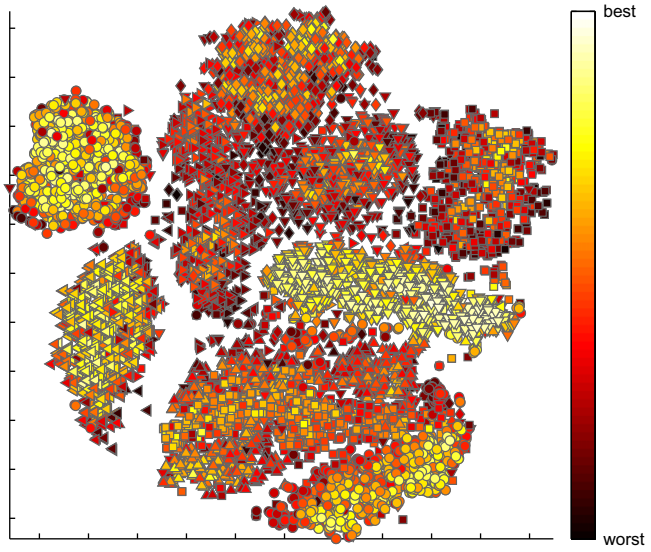


Fig. 18. The figure shows the point-wise quality $Q_{ND}^{x_i}(300,300)$ for the t-SNE embedding of the *MNIST* data from Fig. 16. As expected from the curves of Fig. 17, we generally see many errors in the visualization. Furthermore, we can observe that the overlaps of the classes cause stronger errors. This is less pronounced for the digits 0, 1, 6, and 7. The classes 2, 3, 5, and 8 show many disturbances in the defined range of 300 neighbors, deviating from the original rank by more than 300. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

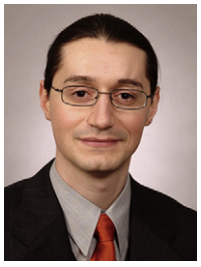
visualizations, we observed a lack of formally established techniques that facilitate a more detailed evaluation of a single visualization. Therefore, we proposed to extend general formal quality criteria to yield a richer impression of the embedding's local characteristics. We based our approach on the established co-ranking framework which unifies several quality criteria, and presented a point-wise quality measure following directly from individual co-ranking matrices. These local quality ratings can be used to augment the given data embedding by meaningful color values which highlight distortions in the visualization for user-specified neighborhood scales. We further suggested to improve the parameterization of the established quality measure to enable more control over the evaluation's focus. In several artificial and real-world experiments we demonstrated the benefits of our evaluation framework, and discussed possibilities for speed-up with an interactive user interface in mind.

In future research, we will investigate which parameterization of the quality measure leads to better interpretable results in user studies. Since DR is usually applied to represent very high-dimensional data and our evaluation is based on the agreement of ranks, further ongoing research will focus on certain phenomena of high-dimensional data distributions and their influence on our evaluation scheme. In [1] the meaninglessness of distances is discussed as part of the so-called curse of dimensionality, and a related issue known as *hubness* is pointed out in [36], which leads to some points appearing very often among the K nearest neighbors of other points in the data. It seems important to investigate the influence of such effects on the evaluation of rank agreement. Additionally, we plan to investigate further how truthfully we can approximate the quality evaluation by using a small subsample from the original and the embedded data points.

References

[1] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.
[2] K. Bunte, M. Biehl, B. Hammer, A general framework for dimensionality-reducing data visualization mapping, Neural Comput. 24 (3) (2012) 771–804.
[3] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[4] L. van der Maaten, E. Postma, H. van den Herik, Dimensionality Reduction: A Comparative Review, Technical Report, Tilburg University Technical Report, TiCC-TR 2009-005, 2009.
[5] B. Schölkopf, A.J. Smola, K.R. Müller, Kernel principal component analysis, Adv. Kernel Methods: Support Vector Learn. (1999) 327–352.
[6] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (2006) 504–507.
[7] A. Gorban, A. Zinovyev, Principal manifolds and graphs in practice: from molecular biology to dynamical systems, Int. J. Neural Syst. 20 (3) (2010) 219–232.
[8] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer-Verlag, New York, 1997.
[9] J.W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 18 (5) (1969) 401–409.
[10] J.B. Tenenbaum, V. De Silva, J.C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, Science 290 (5500) (2000) 2319–2323.

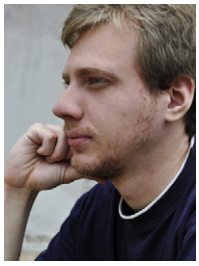
[11] T. Kohonen, Self-Organizing Maps, Springer, 1995.
[12] C.M. Bishop, M. Svensén, C.K.I. Williams, Gtm: the generative topographic mapping, Neural Comput. 10 (1998) 215–234.
[13] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[14] K. Weinberger, L.K. Saul, An introduction to nonlinear dimensionality reduction by maximum variance unfolding, in: Proceedings of the National Conference on Artificial Intelligence, Boston, MA, 2006, pp. 1683–1686.
[15] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in Neural Information Processing Systems, vol. 14, MIT Press, 2001, pp. 585–591.
[16] Z. Zhang, H. Zha, Principal manifolds and nonlinear dimension reduction via local tangent space alignment, SIAM J. Sci. Comput. 26 (2002) 313–338.
[17] G. Hinton, S. Roweis, Stochastic neighbor embedding, Adv. Neural Inf. Process. Syst. 15 (2003) 833–840.
[18] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, J. Mach. Learn. Res. 11 (2010) 451–490.
[19] J.A. Lee, M. Verleysen, Nonlinear projection with the isotop method, in: J.R. Dorronsoro (Ed.), ICANN, Lecture Notes in Computer Science, vol. 2415, Springer, 2002, pp. 933–938.
[20] K. Bunte, B. Hammer, T. Villmann, M. Biehl, A. Wismüller, Neighbor embedding xom for dimension reduction and visualization, Neurocomputing 74 (9) (2011) 1340–1350.
[21] J. Venna, S. Kaski, Local multidimensional scaling, Neural Networks 19 (6) (2006) 889–899.
[22] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, 1979.
[23] P. Demartines, J. Herault, Curvilinear Component Analysis: A Self-Organizing Neural Network for Nonlinear Mapping of Data Sets, 1997.
[24] J. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72 (7–9) (2009) 1431–1443.
[25] J. Lee, M. Verleysen, Scale-independent quality criteria for dimensionality reduction, Pattern Recognition Lett. 31 (2010) 2248–2257.
[26] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques, Neurocomputing 70 (7–9) (2007) 1304–1330.
[27] A. Ultsch, H. Siemon, Kohonen's self organizing feature maps for exploratory data analysis, in: Proceedings of the INNC'90, Kluwer, 1990, pp. 305–308.
[28] J. Sun, C. Fyfe, M. Crowe, Extending sammon mapping with Bregman divergences, Inf. Sci. 187 (2012) 72–92.
[29] S. France, D. Carroll, Development of an agreement metric based upon the rand index for the evaluation of dimensionality reduction techniques, with applications to mapping customer data, in: Proceedings of MLDM '07, Springer-Verlag, Berlin, Heidelberg, 2007, pp. 499–517.
[30] H.-U. Bauer, K. Pawelzik, T. Geisel, A topographic product for the optimization of self-organizing feature maps, in: NIPS, 1991, pp. 1141–1147.
[31] L. Chen, A. Buja, Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis, J. Am. Stat. Assoc. 104 (485) (2009) 209–219.
[32] J. Venna, S. Kaski, Local multidimensional scaling, Neural Networks 19 (2006) 889–899.
[33] S.A. Nene, S.K. Nayar, H. Murase, Columbia Object Image Library (COIL-20), Technical Report, February 1996.
[34] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[35] A. Gisbrecht, B. Mokbel, B. Hammer, Linear basis-function t-SNE for fast nonlinear dimensionality reduction, in: IJCNN, IEEE, 2012, pp. 1–8.
[36] M. Radovanović, A. Nanopoulos, M. Ivanović, Hubs in space: popular nearest neighbors in high-dimensional data, J. Mach. Learn. Res. 11 (2010) 2487–2531.



Bassam Mokbel is currently a PhD-student at Bielefeld University, Germany, in the research group for theoretical computer science within the Center of Excellence for Cognitive Interaction Technology (CITEC). He studied computer science at Clausthal University of Technology in Germany, where he received his Diploma degree in 2009, and later became a research assistant in the state-funded collaborative research program "IT Ecosystems" about complex and heterogeneous distributed computer systems.

He joined the CITEC in April 2010, and is currently working in the context of the DFG priority programme 1527 for "Autonomous Learning".

Wouter Lueks is currently a PhD student in the Digital Security group of the Radboud University Nijmegen. He studied both Mathematics and Computing Science at the University of Groningen, where he received his master degrees in 2011. As part of these studies he visisted the Cognitive Interaction Technology Center of Excellence at Bielefeld University, Germany.



Andrej Gisbrecht received his Diploma in Computer Science in 2009 from the Clausthal University of Technology, Germany, and continued there as a PhD-student. Since early 2010 he is a PhD-student at the Cognitive Interaction Technology Center of Excellence at Bielefeld University, Germany.

After completing her diploma studies in pure mathematics in 1995, Barbara Hammer received her Ph.D. in Computer Science in 1995 and her venia legendi in Computer Science in 2003, both from the University of Osnabrueck, Germany. From 2000–2004, she was leading the junior research group 'Learning with Neural Methods in Structured Domains' which was funded within the frame of an innovation initiative of Lower Saxony, before accepting an offer as professor for Theoretical Computer Science at Clausthal University of Technology in 2004. Starting from April 2010, she moved to the CITEC center of excellence at Bielefeld University as professor for Theoretical Computer Science for Cognitive Systems.