2015-02

Graph-based measures to assist user assessment of multidimensional projections

Neurocomputing, Amsterdam, v. 150, p. 583-598, Feb. 2015
http://www.producao.usp.br/handle/BDPI/50995

Neurocomputing 150 (2015) 583–598

Contents lists available at ScienceDirect

Neurocomputing

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

Graph-based measures to assist user assessment of multidimensional projections

Robson Motta *, Rosane Minghim, Alneu de Andrade Lopes, Maria Cristina F. Oliveira

University of São Paulo - Institute of Mathematics and Computer Science (ICMC), São Carlos, Brazil

ARTICLE INFO

Article history:
Received 1 December 2013
Received in revised form
24 August 2014
Accepted 11 September 2014
Available online 6 November 2014

Keywords:
Quantitative evaluation
Visual analysis models
Metrics and benchmarks
Multidimensional data
Dimension reduction evaluation

ABSTRACT

Multidimensional projections are valuable tools to generate visualizations that support exploratory analysis of a wide variety of complex high-dimensional data. However, projection mappings obtained from different techniques vary considerably, and users exploring the mappings or selecting between projection techniques still have limited assistance in their task. Current methods to assess projection quality fail to capture properties that are paramount to user interpretation, such as the capability of conveying class information, or the preservation of groups and neighborhoods from the original space. In this paper we propose a unifying framework to derive objective measures of the local behavior of projection mappings that support interpreting the mappings and comparing solutions regarding several properties. A quality value is computed for each data point, from which a single global value may be also assigned to the projection. Measures are computed from a recently introduced data graph model known as *Extended Minimum Spanning Tree* (EMST). Measurements of the topology of EMST graphs, built relative to the original and projected data representations, are scale independent and afford evaluation of multiple properties. We introduce measures of visual properties and of preservation of properties from the original space. They are targeted at (i) depicting class segregation capability; (ii) quantifying 'neighborhood purity' regarding classes; (iii) evaluating neighborhood preservation; and finally (iv) evaluating group preservation. We introduce the measures and illustrate how they can inform users about the local and global behavior of projection techniques considering multiple mappings of artificial and real data sets.

1. Introduction

Multidimensional projection techniques map high dimensional data points into low dimensional(2D or 3D) spaces striving to minimize the information loss. Such projections can be generated by dimension reduction strategies when their target space is typically two- or three-dimensional. Projection techniques have been deserving great attention, with several recent solutions introduced focusing on improvements such as achieving higher precision, giving users additional control over the outcome, or improving scalability [1–5]. Information loss may be measured in different ways, depending on the goal of the projection, e.g., to preserve relative distances, or to preserve neighborhoods or neighborhood ranks, to highlight outliers or yet to favor class or group segregation.

Projections are usually displayed as 2D scatter plots and support visual analysis in a wide variety of applications, typically requiring users to identify similarity of individuals in a class or cluster, as well as outliers based on proximity [6–9]. Nonetheless, their usage poses many practical issues, e.g., which technique would handle the data better, which is the best parameterization, how reliable is the data mapping, or yet how to interpret a perceived pattern. Indeed, projection interpretation is highly subjective and analysts need help to grasp the real meaning of a point layout.

Quantitative and qualitative measures strategies are available to assess and compare projection mappings. Some graphical representations often employed for qualitative assessments are illustrated in Fig. 1, namely a distance histogram, exemplified in Fig. 1(a); a similarity matrix, in Fig. 1(b), which can convey whether data classes are highly separable; and a Shepard diagram in Fig. 1(c), which plots the pairwise distances computed in the projected versus in the original data space. Albeit useful, such representations convey limited information on the faithfulness of a mapping to the original data space. Some quantitative measures that support objective assessments of the global or local behavior of projection mappings regarding distinct properties are reviewed and discussed in Section 2.

The outcome of many quality measures is affected both by the choice of projection and by parameterization. Additionally, good parameter values may depend on data characteristics unknown to the analyst, for example, users are often required to set distance thresholds or neighborhood sizes. Best choices vary across data sets and even within a data set, e.g., a data point located in a very

* Corresponding author.
E-mail address: rmotta@icmc.usp.br (R. Motta).

http://dx.doi.org/10.1016/j.neucom.2014.09.063
0925-2312/

sparse region may have its closest neighbor far away, whereas one in a dense region can have many very close neighbors. Using cluster models as reference is inadequate, as an arbitrarily chosen clustering does not necessarily reflect the data distribution. Additionally, the value of directly comparing distances in sparse high-dimensional spaces with distances in the reduced space is limited. Thus, there is need for investigation of new evaluation approaches that capture quality variations due to choice of projection techniques or parameter settings.

The measures introduced in this paper rely on connectivity patterns identified on a graph model that does not require setting neighborhood sizes or other parameters. In building the graph, connection patterns are established considering the density distribution across the data space, so that neighborhood relations are mapped to vertex connectivity in order to reflect local data properties. Clusters may then be identified on the graph regardless of shape or density. The ability to build graph models from data in original and in projected spaces enables assessing data property preservation by comparing graph topologies, rather than neighborhood distances or ranks. Our graph-based framework thus avoids several of the shortcomings mentioned above and introduces a novel approach to measuring projection quality that is potentially useful both to enhance their interpretation and to assess their reliability. It is important to stress that the measures proposed are meant to evaluate 2D mappings generated by multidimensional projections for visualization purposes, and not as an approach to evaluate general dimension reduction techniques, aimed at obtaining lower-dimensional embeddings that characterize a non-linear manifold in which the data would lie.

This paper is organized as follows. In Section 2 we discuss related work on numerical measures for assessing quality of multidimensional projections. In Section 3 we briefly discuss some similarity-based graph models and justify our choice of a particular similarity graph as the underlying model to derive numerical measures of projection quality. In Section 4 we introduce and describe five measures to assess specific projection properties. In Section 5 we illustrate their application to projections of two data sets, created with four techniques, and also compare their behavior with that of some existing quality measures. We summarize the contribution and present the conclusions in Section 6.

2. Related work

Quantitative measures can tell to which extent a projection conveys or preserves properties such as class distribution, dissimilarities or neighborhoods. Marghescu [6], for example, employs cluster validity measures in this context: data is clustered in both spaces and the resulting models compared to quantify how the original clusters have been preserved. This approach is limited in that a clustering model extracted by a particular algorithm is not necessarily faithful to the data, and moreover intra-cluster and inter-cluster relationships are largely ignored. Indeed, assessing projections solely as a cluster solution is quite restrictive, as they can reveal much more about the data.

Stress functions are a classical measure of the distance preservation capability of projections, e.g., Morrison and Chalmers [11] use stress to compare projections obtained with various techniques and varying parameters. Nonetheless, very cluttered projections may have excellent stress values [7] and similar stress functions may lead to different perceptions of quality.

These inherent limitations motivated additional graphical representations based on definitions of neighborhood. The *Neighborhood Preservation* [12], for example, computes for each data point how many of its K-nearest neighbors in the projection are also in its original K-neighborhood (for a particular K), averaging the numbers over all points to yield a single value. A curve is obtained varying K, as illustrated in Fig. 2(a) for three projections of the *Optidigts* data (see Section 5). Naturally, as K increases so does precision. Therefore, this measure is suitable to compare projections rather than as a stand-alone measure of quality. Venna and Kaski [13] compute neighborhood preservation in a similar manner, but they explicitly consider the non-coincident nearest neighbors in the original and in the projected spaces, ranking these disparate elements based on how misplaced they are. Again, the output depends on the choice of K.

Lespinats and Aupetit [14] introduce *CheckViz* as a method to qualify projection mappings generated by nonlinear techniques. They define a perceptually uniform 2D color coding of the projection area that allows observers to detect the presence of geometrical mapping distortions such as tears – when neighboring data instances are mapped far apart – and false neighborhoods – when distant ones are mapped close. Their solution affords detecting such distortions at a glance, preventing mistaken interpretations of the mappings. Earlier work by Aupetit [15] introduced measures of outliers and other projection artifacts, as well as an approach to visualize projection distortions by overlaying quality measures on a Voronoi cell decomposition of the 2D projection mapping. Other approaches also employ visual representations of the quality of neighborhoods coloring and connecting missing and false neighbors through appropriate value averaging and graph bundling [16].

Other measures verify projection properties relative to given data class information. One example is the *Neighborhood Hit* curve to quantify class segregation capability, illustrated in Fig. 2(b). It computes, for a data point with class label l, the proportion of its K-nearest neighbors also labeled l, and averages over all points. It reveals to which extent the groups observable in the projection

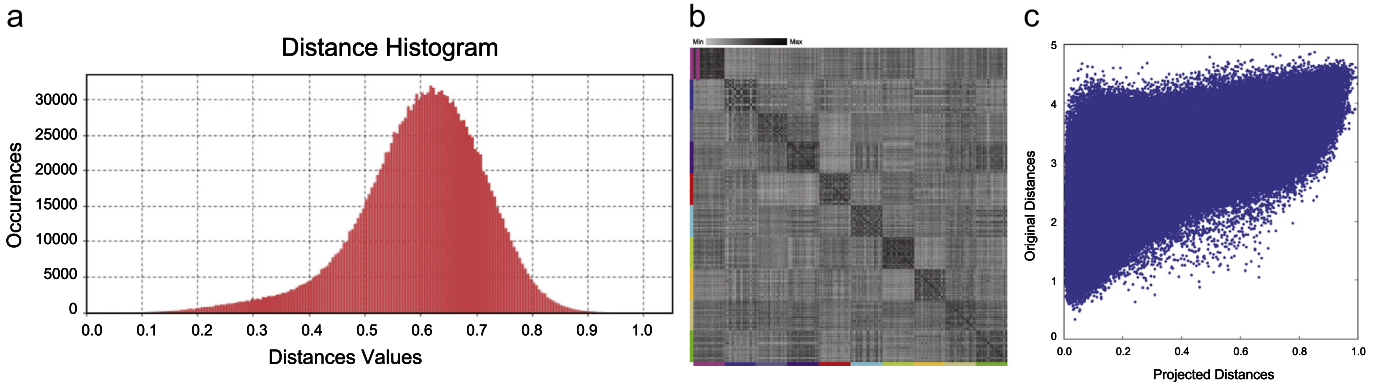


Fig. 1. Visual representations for assessing projection quality based on distance preservation: (a) distance histogram; (b) distance similarity matrix; and (c) Shepard diagram. Projection of the *Optidigts* data set (see Section 5) with Sammon's mapping [10] and distances computed with the Euclidean metric.

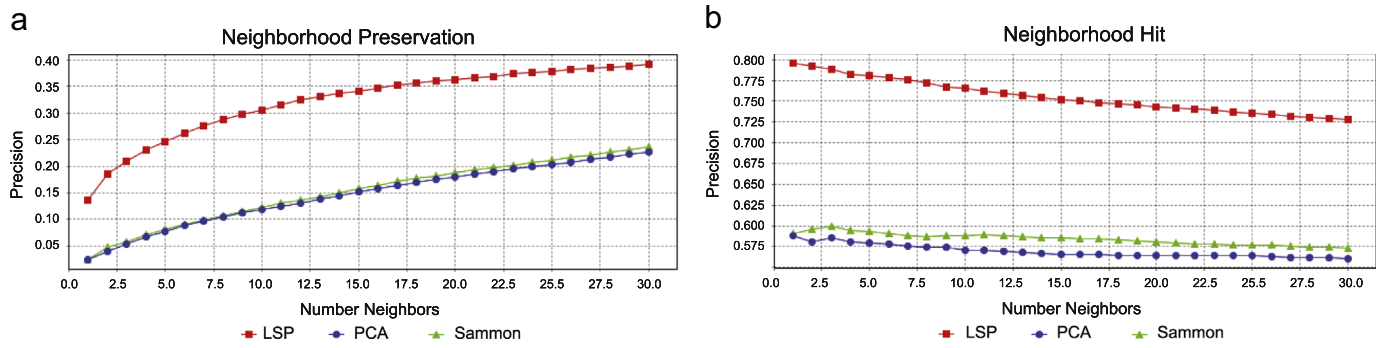


Fig. 2. Projection evaluation curves: (a) *Neighborhood Preservation* and (b) *Neighborhood Hit* curves for $K = 1, \ldots, 30$, relative to projections of *Optdigits* obtained with three techniques.

reflect the given class structure; it is, nonetheless, highly affected by class boundaries: data points near class borders are likely closer to points from another classes than to their own-class nearest-neighbors. Such points will have low *Neighborhood Hit* values even if properly placed in the projection, as far as similarity preservation is concerned.

Sips et al. [8] propose two quantitative measures of class consistency. The *distance consistency* verifies class separation: for each data point it checks whether it is closer to the centroid of its own class than to the centroid of any other class. The *distribution consistency* measure generalizes the previous idea, considering the entropy of the spatial distribution of the classes in the projection. Regions containing points from a single class achieve maximum consistency values, which decrease in regions containing points from distinct classes.

The *class density measure* [9] ranks a set of candidate projections based on their class segregation capability. The data points in each class are considered separately, generating a continuous representation with a smooth density function based on local neighborhoods. The resulting continuous representations are inspected to identify class overlap after estimating individual pixel densities.

Previously mentioned measures do express relevant aspects of projections, afford to some extent comparing solutions and consider properties relevant to their interpretation, such as grouping capability. Nonetheless, the question remains as to how they relate to the users´ perception of quality. The work by Sedlmair et al. [17] tackles this problem, introducing a taxonomy of visual cluster separation factors in projection scatter plots, which is given as a guide to evaluate cluster separability measures. The authors also conducted a systematic qualitative study comparing human observations with numerical measures of quality, and found out that measures failed to match human perception of quality in nearly 50% of the cases. They evaluated *distance consistency*, *distribution consistency* and *class density measure* (referred to as centroid-based and grid-based measures) on projections showing groups with varying visual properties, e.g., single group, group with multiple sub-groups, groups of different sizes, shapes, densities, etc. The work concludes that grouping measures cannot capture all such variations.

Several strategies proposed to evaluate the quality of dimensionality reduction (DR) techniques in general are relevant to projection evaluation for visualization purposes. Lee and Verleysen [18] define a co-ranking matrix, which is a joint histogram of point ranks. The rank of a data point p_i with respect to p_j is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } k < j)\}|$. A similar rank can be computed in the reduced space, i.e., $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } k < j)\}|$. A co-ranking matrix which conveys the preservation of point rankings in both spaces is defined as

$$Q = [q_{kl}]_{1 \le k,l \le N-1} \text{ with } q_{k,l} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}| \quad (1)$$

An element q_{ii} yields how many points ranked i in the data space remain ranked i in the reduced space. Off-diagonal elements q_{ij} inform how many elements originally ranked as i have been re-ranked as j after the dimension reduction. The rank error is defined as the difference $\rho_{ij} - r_{ij}$, with the event of a positive/negative rank error being called an *intrusion*/*extrusion*, respectively. Notice that such events occur with respect to a neighborhood size K: for values of K such that $r_{ij} \le K \le \rho_{ij}$ point p_j is an intruder in the K-ary neighborhood of the reduced space, with respect to the original neighborhood, i.e., the mapping approximated the points and introduced 'false' neighbors in the reduced space. An extruder signals that points have been separated, incurring in missing 'real' neighbors.

The co-ranking matrix Q contains complete information about how ranks have been preserved in a given low-dimensional representation, and as such it provides a framework for several assessments of the neighborhood preservation. Various types of intrusions and extrusions can be associated with different blocks of Q and neighborhood sizes K. The error is thus measured relative to a specific neighborhood size K. In a follow up work, the authors propose criteria to identify the most appropriate choice for K [19], which depends upon the desired compromise between preserving local (small) and global (large) neighborhoods. This is done by computing two overall measures, Q_{local} and Q_{global}, from the co-ranking matrix.

Mokbel et al. [20] extend the rank-based framework by introducing pointwise measures that follow directly from individual co-ranking matrices. These can be mapped directly to the point cloud visualization, enriching the information and enhancing interpretability. They further suggest an improved parameterization of the quality measures Q_{local} and Q_{global} to ensure better control of the evaluation focus and a more fine-grained analysis. Their solution identifies benign points by their relative deviation from the original rank, rather than their absolute rank in the original space, and allows for separate control of the region of interest and the size of the tolerated errors. Parameter K is replaced by a pair (K_s, K_t), where the first determines a region of interest and the second determines the size of the tolerated rank errors. Users can vary both parameters to prioritize e.g., the preservation of local or global relationships in a quality assessment.

Measures derived from the co-ranking framework suffer from the same limitations of other approaches based on K-neighborhoods computed either from distances or ranks: by fixing neighborhood sizes *a priori* one ignores that the number of relevant neighbors is likely to vary across the data set and K should vary accordingly to capture these local properties.

Venna et al. [21] frame the specific visualization task of projecting data as an information retrieval task. Given a data set $\{x_i\}_{i=1}^{N}$, and its projection mapping $\{y_i\}_{i=1}^{N}$, let P_i and Q_i denote, respectively, defined neighborhoods of an element x_i and of its

projection y_i, with $|P_i| = r_i$ and $|Q_i| = k_i$. They define three variables: (i) the number of elements on both sets, or the *true positives* for x_i, denoted as $N_{TP,i}$; the number of elements in Q_i but not in P_i, the *false positives*, denoted $N_{FP,i}$, and the number of elements in P_i that are not in Q_i, or the misses, $N_{MISS,i}$. These variables are mapped into measures of precision and recall, analogous to those traditionally employed in information retrieval, as described in Eqs. (2) and 3. These measures can be weighted into a single measure for a combined quantitative evaluation.

$$precision(i) = \frac{N_{TP,i}}{k_i} = 1 - \frac{N_{FP,i}}{k_i} \tag{2}$$

$$recall(i) = \frac{N_{TP,i}}{r_i} = 1 - \frac{N_{MISS,i}}{r_i} \tag{3}$$

Authors highlight that considering fixed-size neighborhoods causes all neighborhood violations to be equally penalized. In order to overcome this problem, they propose replacing the fixed neighborhoods by one that considers a probability distribution for all points, adopting a control variable to ensure that the probability decays as the distance increases. They then introduce novel measures of *precision* and *recall* based on comparing the probability distributions in the original and in the reduced spaces. As interpreting these measures are not straightforward, they introduce a modified version that replaces the point probability distributions by point rankings, yielding measures that are similar in nature to those obtained with the co-ranking framework [18].

Evaluation measurements presented thus far offer interesting sources of information on various properties of projections to data analysts and designers of projection techniques. Many solutions consider visual coding properties or the preservation of class information (for labeled data). Methods that address property preservation as compared to the original space typically focus on comparing neighborhoods in terms of distances or rankings, e.g., those based on the co-ranking matrix. However, users face many practical difficulties. In order to investigate multiple characteristics of a projection mapping they will have to resort to several unrelated approaches. Even the co-ranking framework, that yields several measurements focuses on the particular property of neighborhood preservation – then other properties of the mappings relevant to visualization tasks are ignored.

We introduce a graph-based framework as an integrated approach that consistently reflects multiple properties of multidimensional projections. The framework allows deriving multiple measurements that code both the visual configuration of mappings as well as their relationship with point distribution in the original space. The goals are: (i) to enhance interpretability, by helping analysts to understand what a particular projection is coding and how reliable it is; (ii) to support evaluating multiple quality properties of a particular projection mapping; and (iii) to enable comparing distinct projection mappings in terms of their observable properties. The graph-based solution prevents neighborhood scale issues from interfering in the assessment.

Using measurements derived from graph models to evaluate visual representations is not novel: 'scagnostics' [22,23], from 'scatter plot diagnostics' is based on extracting geometric features that describe scatter plots. Wilkinson et al. [22] extend and generalize the original idea by deriving features from graphs built from geometric representations. Similarly to us, they employ Minimum Spanning Trees, but their focus is on highlighting anomalous and interesting patterns in large scatter plot matrices of attribute data, rather than on evaluating multidimensional projections, that is, no relationship between the original and projected spaces can be drawn from the original formulation of their work, a fundamental task in evaluating projections. In this respect, our approach extends scagnostics to include aspects of relationships between the two data representation spaces involved.

3. Modeling the data as a similarity graph

A similarity-based data graph model is at the core of our proposed strategy to assess projections. Graphs that capture the relevant (dis)similarity properties of a tabular data set may be built by taking data instances and their pairwise distances as vertices and potential weighted edges. Such graphs are extensively employed in data mining tasks such as similarity search or clustering. Some relevant examples include the *Minimum Spanning Tree* (*MST*) built from the complete weighted graph; various *K-nearest neighbors* (*KNN*) graphs [24–26], and some recent structural-based graphs [27–29]. Building *KNN* models requires defining a neighborhood size *K*. This is critical, as in most cases capturing the 'relevant' connection patterns would require different settings across regions of the data space. The structural-based models define which edges to include considering certain properties such as clustering [27] or neighborhood density [28]. The Shortest-Path Graphs, as proposed by Berg et al. [29], adopt a criterion whereby a data a point can end up more connected to distant points than to closer ones. Thus, they are not a proper choice for tasks that require connectivity to reflect distances.

We argue that similarity graphs afford various useful analysis of projection-based visualizations, which are known to support exploration tasks driven by (dis)similarity as conveyed by spatial proximity. However, for such purposes highly parameterized models are not appropriate, i.e., deriving the graph should not require parameter tuning by users. Moreover, one must favor models capable of correctly capturing the local and global neighborhood relations into connectivity patterns.

From those mentioned above, the *MST* (the minimum-weight connected tree built from the complete graph) and its derived *EMST - Extended MST* graph [28] are non-parametric. Additionally, the *EMST* is by construction very effective in capturing the local variations in neighborhood relations and mapping them into connectivity patterns. The graph is built in two stages: after obtaining the *MST*, the *EMST* construction starts with an empty edge set to which edges are added according to connection patterns identified in the *MST*. Two values are used to compute the connectivity threshold of each vertex: a *local distance information* and a *global distance distribution*. Two vertices will be connected only if their pairwise distance is bellow the threshold. The idea is formalized in the definition below. The edge set A includes the *MST* edges plus the additional incident edges to each vertex in the complete graph G_C that satisfy the *limit* threshold. The whole process is detailed in the *EMST* Algorithm that follows, including the computation of *limit*. Table 1 introduces the notation employed in the algorithm and in describing the proposed measures.

$$\begin{aligned} &Define: EMST(V, A)| \\ &\quad V = \{v_1, \ldots, v_N\} \\ &\quad A = \{(v_i, v_j) | (v_i, v_j) \in MST\} \cup \{(v_i, v_j) \in G_C | \delta(x_i, x_j) \leq limit(v_i)\} \end{aligned}$$

Algorithm 1 initially computes the complete graph G_C and its corresponding $MST(V, A')$. The *EMST* edges will be gradually added to an initially empty edge set A. The *limit* threshold is computed for each vertex v_i as shown in line 14. Q_{dd} is the global distance distribution factor, computed for G_C according to Eq. (4), where D stands for the set of all (normalized) pairwise distances, and the standard deviation sd, $0 \leq sd \leq 0.5$, is taken as an indicator of distance distribution variability, with $max(sd) = 0.5$ adopted as a normalization factor. Small values of Q_{dd} indicate little variability

in pairwise distances, and capturing neighborhoods thus requires a strict connectivity criterion, i.e., data points must be very close in order to be connected. Q_{dd} values close to one indicate high variability in the distances, and thus the connectivity criterion must be relaxed. Thus, *limit* is computed by adding to Q_{dd} (weighted by $\langle w^{A'} \rangle$) the greatest value between the vertex' minimum edge weight ($w_{ij}^{G_C}$) and the graph's average weight $\langle w^{A'} \rangle$, which accounts for the local distance information factor. The final graph is $G_{EMST}(V, A)$, where $A \supset A'$. The algorithm may be similarly applied to the reduced space with distance function d. Computational complexity is determined by the cost of building the MST, which is $O(N \log(N))$, and by the iterations in lines 12–16, clearly $O(N^2)$. Thus, building the EMST has cost $O(N^2)$.

$$Q_{dd}(D) = sd(D)/\max(sd) \tag{4}$$

Algorithm 1. Extended minimum spanning tree graph model (*EMST*).

1: **Input**:
2: Set of data instances: $X = \{x_1, \ldots, x_n\}$ // or $Y = \{y_1, \ldots, y_n\}$
3: Distance function: $\delta(x_i, x_j)$ // or $d(y_i, y_j)$
4: **Output**:
5: *EMST* graph: (*V*,*A*)
6:
7: Vertices $V \leftarrow X$
8: Distance matrix D (pairwise dissimilarities $\leftarrow \delta(X)$ // normalized
9: $G_C \leftarrow$ CompleteGraph(V, D) // complete weighted (by distances) graph
10: Edges $A' \leftarrow MST(G_C)$ // edges from *MST* (G_C)
11: Edges $A \leftarrow \emptyset$
12: For each vertex v_i in V
13: $v_j \leftarrow$ closest vertice to v_i in G_C
14: $limit \leftarrow max(w_{ij}^{G_C}, \langle w^{A'} \rangle) + (\langle w^{A'} \rangle \cdot Q_{dd}(D))$
15: For each vertex v_k in V
16: $A \leftarrow A \cup \{a_{ik} | a_{ik} \in G_C \text{ and } w_{ik}^{G_C} \leq limit\}$
17:
18: **Returns** (*V*,*A*)

Table 1
Notation.

N	Number of data points
m	Original data dimensionality
p_i	ith data point (x_i in R^m or y_i in R^2)
L	Number of classes
l	A particular class label
$l(p_i)$	Class label of the point p_i
$\|S\|$	Number of elements of a set S
$d(y_i, y_j)$	Pairwise distance function in R^2
$\delta(x_i, x_j)$	Pairwise distance function in R^m
$G(V, A, W)$	Weighted graph model
v_i	Vertex representing p_i in a graph model
(v_i, v_j, w_{ij})	Weighted edge connecting vertices v_i and v_j
$N_i^G = \{v_j \| (v_i, v_j) \in A\}$	Neighborhood of p_i in $EMST(V, A)$
$N_{i,l}^G$	The set of neighbors of p_i (in G) labeled l
$EMST^{R^2}$	EMST graph built from the data projected in R^2
$EMST^{R^m}$	EMST graph built from the data in R^m
$\langle w^A \rangle$	Average weight for graph G(V,A)
w_{ij}^G	Edge weight between v_i and v_j in graph G
$C_i^{R^2}$	Two-dimensional cluster so that $p_i \in C_i^{R^2}$
$C_i^{R^m}$	m-dimensional cluster so that $p_i \in C_i^{R^m}$

Fig. 3 illustrates the MST and EMST models of a synthetic 2D data set. By construction, the connectivity patterns in the *EMST* effectively capture data clusters and outliers and faithfully reflect the original spatial data distribution, as the graph maps data points in dense regions into highly connected vertices and points in sparse regions into vertices with few or no connections [28]. The ability to preserve local and global neighborhoods render it suitable as an underlying model to assess projections, by comparing graph topology in the original and projected spaces. A set of measures and their computation are explained next.

4. Measuring projection quality

We introduce two measures that code the visual configuration of a mapping regarding class information – i.e., the relationship between observed groups and given labels – and three measures that reflect the mapping's consistency with the original space – i.e., the preservation of relevant data properties, as summarized in Table 2. They are computed pointwise to account for the local behavior of techniques, and averaging the point values yields a global assessment of the projection (or of any arbitrary selection of points, for that matter). All measures vary in the range [0, 1], with higher values indicating better scenarios. They are initially illustrated on the well-known Swiss Roll and Iris data sets. The Swiss

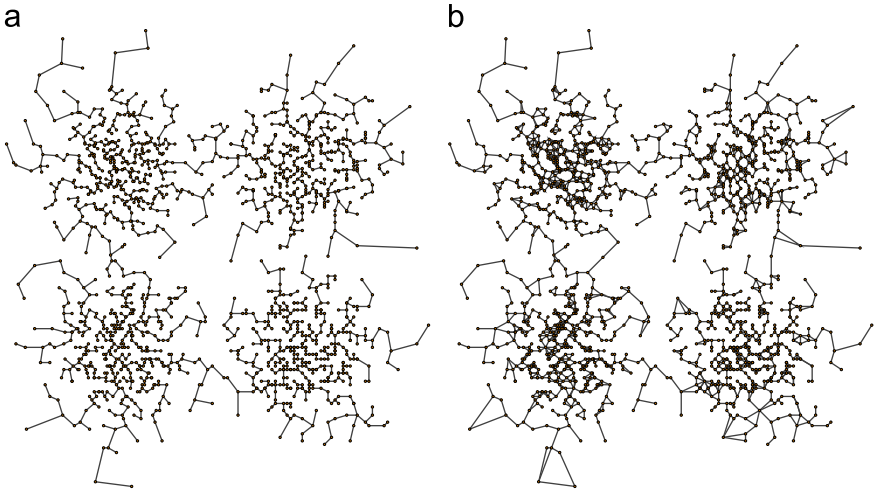


Fig. 3. *MST* and *EMST* graphs of the 2D Swiss Roll with 1600 data points. Dissimilarity measured with the Euclidean distance. (a) MST (1599 edges) and (b) EMST (3180 edges).

Table 2
Summary description of the proposed measures.

Measure	Requires labels	Considers original space	Short description
Class separation	Yes	No	Measures the class 'purity' in the neighborhood of the reference point, i.e., whether classes are visually well segregated
Class aggregation	Yes	No	Measures the visual proximity (aggregation) of points in a particular class (that of the reference point)
Class separation validation	Yes	Yes	Measures the class 'purity' in the neighborhood of the reference point in the projected space as compared to the original space
Neighborhood validation	No	Yes	Measures to which extent the neighborhood of the reference point has been preserved relative to the original space
Group validation	No	Yes	Measures whether groups observed in the projection are indeed formed by closer points in the original space, in average

Roll[1] is formed by points generated from a 2D distribution with four Gaussians randomly sampled with different centers, representing four classes. A 3D distribution is obtained with the mapping $(x,y)\rightarrow(xcosx,y,xsinx)$. Fig. 6(a) and (b) illustrates both distributions with 1600 points colored by class (400 points per class). Iris is formed by 150 data instances describing 3 varieties of the flower (50 from each variety).[2] Each instance is represented by 4 attributes from which *sepal length* and *sepal width* are employed in the 2D mappings shown.

4.1. Measures of class distribution

Two measures, defined as *Class Separation* and *Class Aggregation*, enable comparing multiple mappings regarding their ability to convey the given data class distribution. As such, they are computed from the $EMST^{R^2}$ graph(denoted simply as *EMST* in their definition). Both measures attempt to quantify how effectively a projection conveys the class distribution in terms of purity and segregation, by assessing the *EMST*-neighborhood of a data point p_i with class label l.

We first define functions σ and γ, as follows: given $v_j \in N_i^{EMST}$, then $\sigma(v_i,v_j)=1$ if both p_i and p_j have the same label, otherwise $\sigma(v_i,v_j)=0$; $\gamma(v_i,v_j)=1$ if v_j is reachable from v_i traversing only edges v_a,v_b for which $\sigma(v_a,v_b)=1$, otherwise $\gamma(v_i,v_j)=0$.

Class Separation of a point p_i is defined, see Eq. (5), as the percentage of its neighbors (*EMST*-adjacents) also labeled l. Higher values indicate that p_i is surrounded mostly by points from the same class, lower values indicate a neighborhood with class mixture.

$$class\_separation(p_i)=\frac{1}{|N_i^{EMST}|}\sum_{j}^{N_i^{EMST}}\sigma(v_i,v_j) \tag{5}$$

Class Aggregation, computed with Eq. (6), measures the visual aggregation of data points from a particular class. Higher values indicate stronger class grouping, whereas lower values indicate spatial spreading of the class by the projection.

$$class\_aggregation(p_i)=\frac{\sum_j^N\sigma(v_i,v_j)\cdot\gamma(v_i,v_j)}{\sum_j^N\sigma(v_i,v_j)} \tag{6}$$

Fig. 4 shows in (a) the classes in the 2D Swiss Roll, and the corresponding *Class Separation* (b) and *Class Aggregation* (c) point measures. In this and the following figures we adopt the *Heated Objects* color scale [30], where darker colors indicate higher values; the numbers shown refer to the global value of the corresponding measure. As class separability is good, both measures display a similar behavior, with lower values mostly near class boundaries. Still, some borderline points have high *Class Separation* and low *Class Aggregation*: a point has mostly neighbors from its own class, but it is not well grouped with the other points in its class.

The Iris mappings of *Class Separation* and *Class Aggregation* are shown in Fig. 5. Notice that in both cases points from the separable class (pink) have higher values. The 2D view indeed seggregates this class from the other two. Class Aggregation values are visibly lower for the gray and green classes, which are not well grouped.

4.2. Measures of data property preservation

These measures assess how a projection preserves certain data properties of interest, i.e., to which extent it conveys a faithful image of the data in R^m regarding a target property. Properties of neighborhoods, class and clustering structure are computed from the $EMST^{R^2}$ and $EMST^{R^m}$ graphs, and compared to capture whether (i) a property observed in the projection is consistent with the original space, called *precision*, and (ii) properties that hold in R^m have been mapped in the projection, called *recall*. Usage of these terms is consistent with that by Venna et al. [21], who employed the same concepts in evaluating neighborhood preservation.

Precision and *recall* are defined respectively as $TP/(TP+FP)$ and $TP/(TP+FN)$, where *TP* stands for the number of true positives, *FP* for the number false positives and *FN* the number of false negatives. The specific meanings of TP, FP and FN depend on the property under assessment. *Precision* and *recall* may be taken directly as standalone measures or unified into a single *F-measure*, as in Eq. (7). Although they are equally weighted in Eq. (7) other choices could be justified.

$$F\text{-}measure=\frac{2\cdot precision\cdot recall}{precision+recall} \tag{7}$$

We now describe how *precision* and *recall* are defined to obtain point measures of *Class Separation Validation* (for labeled data), *Neighborhood Validation* and *Group Validation*. Measures are again illustrated on the Swiss Roll data, interpreting the 2D spatial distribution as a 'projection' of the 3D one.

4.2.1. Class separation validation

This measure quantifies 'neighborhood purity' regarding class, i.e., the class composition of data point neighborhoods. The goal is to verify whether the composition of class neighborhoods observed in the projection is consistent with that in R^m, with *precision* and *recall* computed as in Eq. (8).

In computing *precision*, the denominator $TP+FP$ is given by the fraction of points in the p_i-neighborhood in the projection that have the same label as p_i. *TP* refers to how many are indeed in the p_i-neighborhood in the original space. Analogously, in computing *recall*, $TP+FN$ is computed as the fraction of points in the p_i-

[1] http://people.cs.uchicago.edu/~dinoj/manifold/swissroll.html.
[2] http://archive.ics.uci.edu/ml/datasets/Iris.

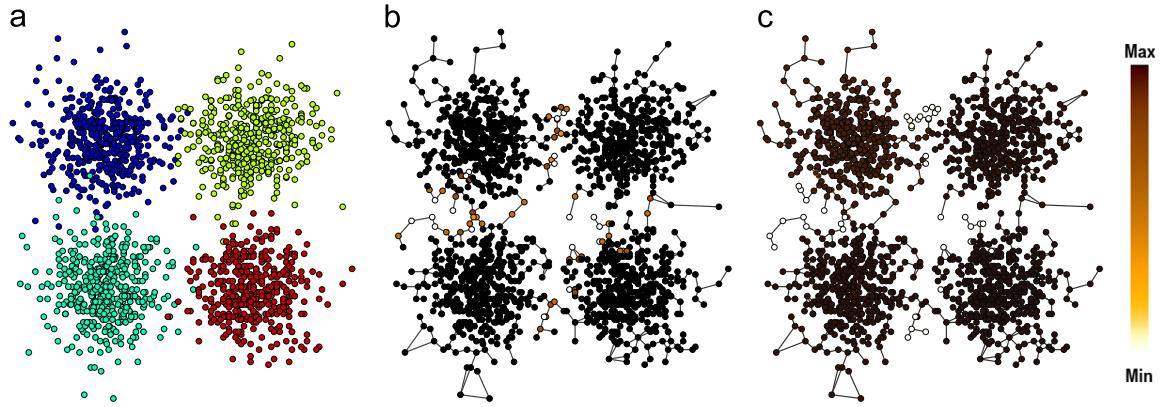


Fig. 4. 2D Swiss Roll: (a) four data classes; (b) class separation; and (c) class aggregation mapped with the *Heated Objects* color scale (darker is better).

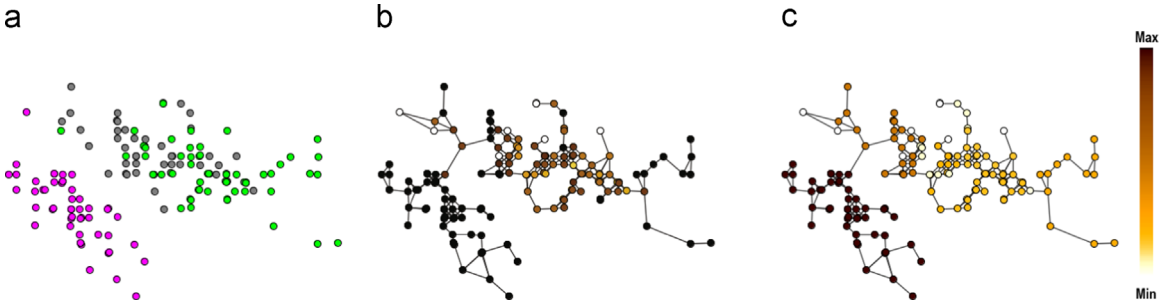


Fig. 5. 2D Iris (a) three data classes; (b) class separation and (c) class aggregation mapped with the *Heated Objects* color scale (darker is better).

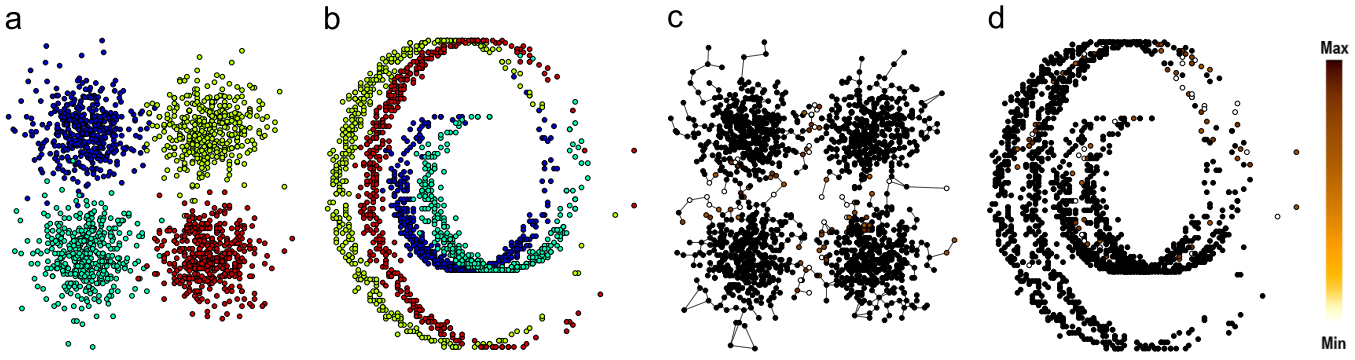


Fig. 6. Swiss Roll: (a) 2D and (b) 3D distributions with color mapping the 4 classes. Color mappings of *Class Separation Validation* in (c) 2D and (d) 3D distributions. (*Heated Objects* scale, darker is better.)

neighborhood in the original space that are also labeled as p_i.

$$class\_separation\_validation(p_i)\begin{cases} precision = \dfrac{min\left(\dfrac{|N_{i,l(p_i)}^{EMST^{R^2}}|}{|N_i^{EMST^{R^2}}|}, \dfrac{|N_{i,l(p_i)}^{EMST^{R^m}}|}{|N_i^{EMST^{R^m}}|}\right)}{\dfrac{|N_{i,l(p_i)}^{EMST^{R^2}}|}{|N_i^{EMST^{R^2}}|}} \\ recall = \dfrac{min\left(\dfrac{|N_{i,l(p_i)}^{EMST^{R^2}}|}{|N_i^{EMST^{R^2}}|}, \dfrac{|N_{i,l(p_i)}^{EMST^{R^m}}|}{|N_i^{EMST^{R^m}}|}\right)}{\dfrac{|N_{i,l(p_i)}^{EMST^{R^m}}|}{|N_i^{EMST^{R^m}}|}} \end{cases} \tag{8}$$

Fig. 6 illustrates the pointwise *Class Separation Validation* mapped to both the 2D and 3D Swiss Roll distributions for comparison. The mapping in the 2D distribution indicates that class neighborhoods are mostly very well-preserved. Preservation is poor for a few points near class boundaries, but other borderline points do have mixed-class neighborhoods observable in 3D that have been preserved in the 2D distribution, and as such they have high *Class Separation Validation* values.

4.2.2. Neighborhood validation

Following the same rationale, the measure of *Neighborhood Validation* attempts to quantify how the projection preserves the original p_i-neighborhood, regardless of class. *Precision* and *Recall* are computed with Eq. (9) and combined into an *F-Measure* of *Neighborhood Validation*. A data point has high *precision* when its neighbors in the projection mostly coincide with those from the original space. *Recall* is high if most neighbors from the original space are preserved in the projection.

Thus, *TP* for *Neighborhood Validation* stands for how many points are in the p_i-neighborhood in both the projected and

original spaces. $TP+FP$ is the size of the p_i-neighborhood in the projected space, and $TP+FN$ its size in the original space.

$$neighborhood\_validation(p_i)\begin{cases} precision = \dfrac{|N_i^{EMST^{R^2}} \cap N_i^{EMST^{R^m}}|}{|N_i^{EMST^{R^2}}|} \\ recall = \dfrac{|N_i^{EMST^{R^2}} \cap N_i^{EMST^{R^m}}|}{|N_i^{EMST^{R^m}}|} \end{cases} \quad (9)$$

The *Neighborhood Validation* values for the Swiss Roll are depicted in Fig. 7, where (a) and (b) show color mappings in the 2D and 3D distributions, respectively. Notice that lower values occur in the class boundaries and in the extremes of the Gaussian distributions. Figure (c) left shows the 2D distribution with two points highlighted (red circled) that have *Neighborhood Validation* equal to 0 (white) and to 1 (black). The same points and their EMST-adjacent are highlighted in the 3D distribution view to the right, with their corresponding neighbors shown in brown. Notice that the two neighbors of the black point are indeed neighbors in the 3D distribution, unlike those of the white point.

4.2.3. Group validation

The outcome of a particular clustering should not be taken as a reference to assess projection properties other than the clustering itself. Still, a validation measure of group formation is potentially valuable, considering that detecting and inspecting groups of similar/dissimilar elements is at the heart of many analysis tasks conducted on projections. This is true both in scenarios where no class structure is given, as when the goal is to verify/validate a given class structure. Rather than taking a cluster model as ground truth, we use clusters extracted in the original and reduced spaces as relative references against which to check the projection's capability of retaining and conveying groups of similar elements in R^m.

We extract clusters from the $EMST^{R^m}$ and $EMST^{R^2}$ graphs with the relational *Adaptive Clustering* (AC) algorithm [31]. In order to determine the 'ideal' number of clusters, it searches for a solution that maximizes intra-cluster edges and minimizes edges to vertices external to the cluster. Relational clustering algorithms are not biased by shape or density when identifying clusters. Previous results suggest AC as a robust choice for our purposes: Motta et al. [28] conducted an empirical comparison of cluster models extracted with several relational and non-relational (agglomerative, divisive and particional) clustering algorithms, considering 23 numerical and 10 textual data sets and three state-of-the-art external cluster evaluation measures (Rand Index, Adjusted Rand Index and F-Score). The relational clustering models performed better, in general, and authors conjecture that this may be due to their ability to handle data sets with varying topological features. In particular, the AC algorithm applied on EMST models of the data produced the best results, in general, as compared to other relational and non-relational solutions.

The rationale for computing *Group Validation* is to verify (i) whether the average pairwise distance computed in R^m for points in a cluster $C_i^{R^2}$ is indeed lower than the average pairwise distance of the points external to it; and (ii) whether the average pairwise distance computed in R^2 of points in a cluster $C_i^{R^m}$ is indeed lower than the average pairwise distance of the points outside the cluster.

Precision of a point $p_i \in C_i^{R^2}$ is computed as described in Eq. (10). The value will be high if the clusters $C_i^{R^2}$ are indeed formed by data points that are closer in R^m, i.e., more similar, as compared to the

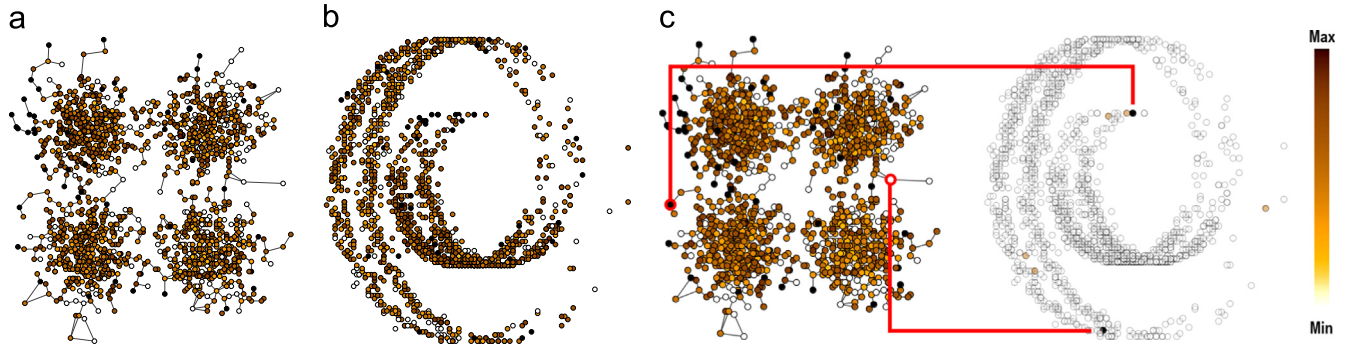


Fig. 7. Swiss Roll: (a) 2D and (b) 3D distributions with the *Neighborhood Validation* measure mapped to color. (c) shows (left) two points highlighted in the 2D mapping, one with maximum (black) and another with minimum (white) preservation, and (right) both points in the 3D view and their corresponding neighbors. (*Heated Objects* color scale, darker is better.).

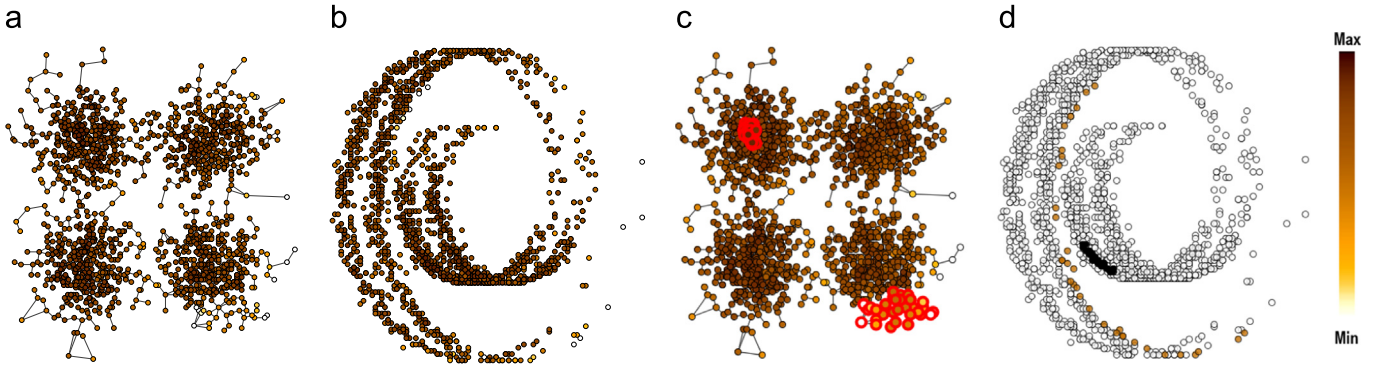


Fig. 8. Swiss Roll: (a) 2D, and (b) 3D distributions, mapping *Group Validation*; (c) two groups highlighted, with good (darker) and poor preservation (lighter); (d) both groups in 3D (black points are from the group at the top and brown points from the one at the bottom). (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this article.)

Table 3
Data sets.

Name	Content	Classes	#Items	#Attributes	Dissimilarity
Optdigits	Handwritten digits (test set)	10	1797	64	Euclidean
News2011	RSS news feeds	23	1771	834	Cosine
Optdigits class distribution (%):					
10.2, 10.1, 10.1, 10.1, 10.1, 10.0, 10.0, 9.9, 9.8, 9.7					
News2011 class distribution (%):					
19.0, 14.7, 9.4, 6.2, 6.2, 5.8, 5.3, 5.1, 5.0, 3.5, 2.5, 2.0, 2.0, 2.0, 1.9, 1.6, 1.6, 1.5, 1.2, 1.0, 0.9, 0.9, 0.7					

rest. *Recall* is computed similarly. High recall indicates good recovery of the original group structure, i.e., points within a cluster $C_i^{R^m}$ have been projected spatially closer than those outside the cluster.

$$group\_validation(p_i)\begin{cases} precision = \dfrac{\frac{1}{|C_i^{R^2}|}\sum_{p_j}^{C_i^{R^2}}\delta(p_i,p_j)}{\frac{1}{N}\sum_{p_k}^{N}\delta(p_j,p_k)} = \dfrac{N\sum_{p_j}^{C_i^{R^2}}\delta(p_i,p_j)}{|C_i^{R^2}|\sum_{p_k}^{N}\delta(p_j,p_k)} \\ recall = \dfrac{\frac{1}{|C_i^{R^m}|}\sum_{p_j}^{C_i^{R^m}}d(p_i,p_j)}{\frac{1}{N}\sum_{p_k}^{N}d(p_j,p_k)} = \dfrac{N\sum_{p_j}^{C_i^{R^m}}d(p_i,p_j)}{|C_i^{R^m}|\sum_{p_k}^{N}d(p_j,p_k)} \end{cases} \tag{10}$$

Group Validation values for Swiss Roll are shown in Fig. 8 (a) and (b). Figure (c) shows (a) with two clusters highlighted, and (d) shows them in the 3D view. Values are high in the top group, shown in brown in (d). The bottom group has points with lower values and is shown in black in (d). Notice that the first (good preservation) is indeed very cohesive in 3D, whereas the second (poor preservation) is not really a group in 3D.

5. Results and discussion

In this section we illustrate the applicability of the proposed measures to assess and compare projections in Section 5.1, and compare our measures of neighborhood preservation with those obtained with the co-ranking matrix framework in Section 5.2.

5.1. Applying the measures on real data sets

We illustrate the measures and discuss their applicability on multiple projections of two labeled data sets summarized in Table 3, namely the Optical Recognition of Handwritten Digits (*Optdigits*) from UCI,[3] which contains handwritten occurrences of the 0–9 digits; and *News2011*, a collection of RSS feeds collected from various news providers (AP, CNN, Reuters and BBC) during 4 weeks in June and July 2011 and manually labeled according to their topic [32].

The *News2011* corpus is described by a Vector Space Model with 834 terms and is a highly unbalanced data set, as detailed in Table 3. A *K*NN classifier (with $K=3$) with 10-fold cross validation achieved very good classification precision (91.8%), suggesting that classes are well-formed, a property that should be reflected in good projection mappings. Digits in *Optidigits* are described by 8x8 bitmaps, and the data is class balanced (see Table 3). Table 4 shows the confusion matrix resulting from running a *K*NN classifier with $K=3$, which achieved 86.7% classification precision: notice that digits 8 and 2 have often been misclassified as 1, both 8 and 9 are also sometimes misclassified as 3. A pairwise similarity matrix of distances in R^m is shown to the left of Table 4.

We pick a reduced subset of four projection techniques, two classic and two recent ones, characterized by adopting very distinct mapping approaches and sufficient for our purposes of illustrating the potential usefulness of the proposed measures. PCA [33] is a standard statistical dimension reduction method, Sammon's mapping [10] is a typical distance preservation MDS strategy, the *Least Square Projection* (LSP) [7] has been designed to favor preservation of local neighborhoods over preservation of global distances, and t-SNE [1] is a novel dimension reduction method that retains probability distributions rather than distances. We used locally available Java implementations of PCA, LSP and Sammon's Mapping, and the t-SNE implementation provided by the authors.[4] For PCA we take the two first principal components as the projected dimensions. Sammon's was run with the default settings of 1797 iterations and magic factor 0.3. For t-SNE we also considered the defaults, setting *perplexity* to 30 and taking the first 30 PCA principal components for the dimension reduction prior to the t-SNE mapping. The LSP parameters *number of control points* and *neighborhood size* were set to the suggested defaults, namely 10 % of the points (179 for *Optdigits* and 177 for *News2011*) and a neighborhood size of 15 points [7].

Table 5 depicts the projections, with points colored by the given class in the first column, whereas columns 2–6 show them color mapped according to each measure introduced in Section 4. Each cell also shows the summary measure for the projection (and recall and precision values in parentheses, when applicable). Next we interpret these measurements, replicating some of the figures for clarity and convenience.

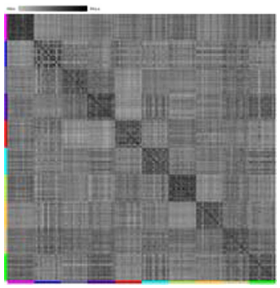
Let us start inspecting the behavior of *Class Separation* and *Class Aggregation* on the *News2011* data, highlighting the LSP and PCA solutions already shown in Table 5 (Fig. 9). One observes that in LSP most points have very high class separation, the exceptions being those placed in-between visual agglomerates. The PCA solution, on the other hand, mixes the classes in the central region, even though high *Class Separation* is observed in certain areas. PCA is a dimension reduction approach, rather than a projection: its poor performance here is a direct consequence of the impossibility of capturing the variability of the 23 data classes with just two principal components.

Points in classes that have not been split into multiple regions will have higher *Class Aggregation*. LSP again has higher values as compared to PCA, but still few classes form single agglomerates, as reflected by the generally lower point values of *Class Aggregation* as compared to *Class Separation*. Notice the yellow class (split into two regions separated by points in the orange class), the blue/salmon/rosée classes close to the central region, as well as the elongated reddish class further down. Most points with lower values are near group boundaries or in-between groups. Some classes do aggregate well in PCA (e.g. blue and olive green) and

[3] UCI KDD Archive, http://www.archive.ics.uci.edu/ml/datasets/Optical+Recognition+of+Handwritten+Digits.

[4] http://www.homepage.tudelft.nl/19j49/t-SNE.html.

Table 4
Similarity matrix and confusion matrix for *Optdigits* data.



0	1	2	3	4	5	6	7	8	9	← classified as
173	1	0	0	1	1	2	0	0	0	class 0
0	182	0	0	0	0	0	0	0	0	class 1
0	22	152	1	0	0	0	1	1	0	class 2
0	10	3	159	0	4	0	3	0	4	class 3
0	11	0	0	168	0	0	2	0	0	class 4
2	7	0	5	2	162	1	1	0	2	class 5
1	4	0	0	1	0	175	0	0	0	class 6
0	5	1	0	1	0	0	172	0	0	class 7
0	66	2	20	1	5	3	3	71	3	class 8
1	9	0	18	0	4	0	2	2	144	class 9

their points have higher values; mid-range values also occur in the yellow, salmon, rosée and green classes. Still, other classes split and overlap in the central region, resulting in a low global value of *Class Aggregation* for PCA. Referring back to Table 5 one observes that t-SNE outperforms the other solutions on both data sets, with excellent *Class Separation* and *Class Aggregation*. In the *Optdigits* projection most points have very high *Class Aggregation*, with the exception of the blue class (digit 1), which has been split into in three sub-groups.

The *Class Separation Validation* helps interpreting whether regions with high 'class purity' in the projection also exist in the original space. Inspecting the LSP and t-SNE projections replicated in Fig. 10, it is noticeable that higher values (dark brown and black) are mostly associated with points in classes well-segregated, such as those of digits 0, 6, 2, 7 and 3 in the LSP and all groups in t-SNE. In LSP, points from class 8 are spread and have neighbors from several other classes (e.g., 5, 3, 2, 1) (reminding that the *K*-NN classifier misclassified many samples from class 8). Again, all classes are well separated in t-SNE. The overall good values of *Class Separation Validation* are more due to precision than to recall. Values of *precision* are quite high (average above 0.75) on all projections considered (see Table 5) whereas values of *recall* are, in general, lower (still, above 0.59). We conclude that the projections actually improved class segregation relative to the original space, which is good news for users who rely on projections to validate data classes. All techniques performed well on both data sets, but again the t-SNE mappings delivered the best results.

Fig. 11 shows *Neighborhood Validation* for the *Optdigits* projections. Values are higher on points in visual agglomerates, and otherwise generally low. Groups are observable, however, only in LSP and t-SNE. Unlike Sammon's, which optimizes a global error function of the original versus projected distances, or PCA, which does not operate on distances or neighborhoods, LSP has been designed to preserve local neighborhoods and t-SNE mappings are known to reflect well the similarities between high-dimensional data points. Their better results on neighborhood preservation are expected, with the best overall performance being again by t-SNE.

Column 5 in Table 5 indicates that most points have poor *Neighborhood Validation* on all solutions, with the exception of t-SNE on *Optdigits*, confirming the difficulty in preserving neighborhoods, as indicated by the overall low values of *recall*. *Precision* performs slightly higher, but still hinting loss of the original neighborhoods. t-SNE and LSP show the best overall performance in terms of both recall and precision.

On data with well-defined classes one expects good correlation between the classes and *m*-dimensional clusters extracted by an effective clustering procedure. We do know that classes in *Optdigits* are well-defined, apart from certain digits with higher writing variability, and thus some class mixture is expected. The AC clustering algorithm identified 15 clusters in the $EMST^{R^m}$ graph (and 41 and 35 clusters, respectively, in the $EMST^{R^2}$ graphs from LSP and t-SNE). Table 6 shows the class distribution of the clusters in R^m. As anticipated, 10 out of the 15 include mostly points from a single class (above 98% purity), and class purity is under 90% only for classes 8 and 11. Classes 1, 3, 7 and 9 have each been split into two or three sub-clusters, indicating greater writing variability.

As only LSP and t-SNE favor the perception of grouping structures, we analyze further their mappings of *Optidigits* regarding *Group Validation* (F-measure, precision and recall), replicated in Fig. 12. Notice that *recall* is typically higher, particularly in regions grouping points from a single class. High recall indicates that the projection placement does reflect groupings from the original space. We observe in LSP that the higher values occur mostly in clusters with high purity and low variability, such as classes 0 and 6. The t-SNE mapping outperforms LSP regarding *Group Validation* both in recall and precision.

Both mappings have lower values of precision, hinting that clusters identified by the AC algorithm in the projection do not necessarily correspond to points more grouped in R^m. This is an expected effect of the space 'folding' incurred in the dimension reduction – remember the AC algorithm identified 41 and 35 clusters in the LSP and t-SNE mappings, respectively, against only 15 in the original space.

Column 6 in Table 5 shows that, similarly to *Neighborhood Validation*, values of *Group Validation* are typically low. However, unlike the neighborhood measure, recall values are often better than those of precision. Higher recall indicates that a projection is doing a good job of spatially approximating points grouped in R^m; lower precision indicates that clusters identified in the mapping do not necessarily reflect groups in R^m.

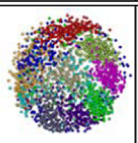
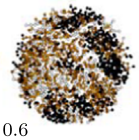
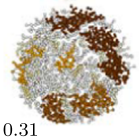
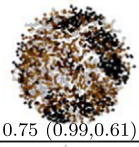
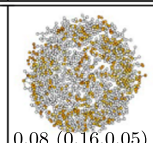
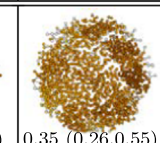
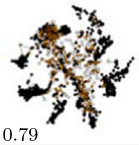
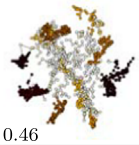
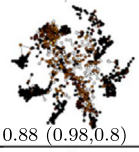
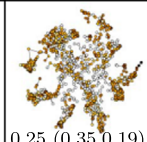
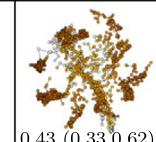
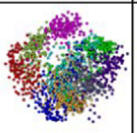
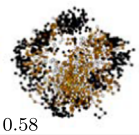
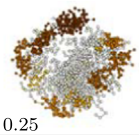
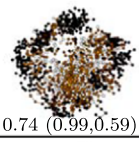
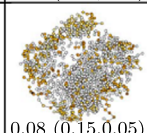
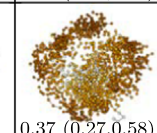
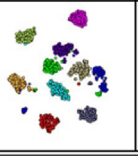
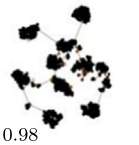
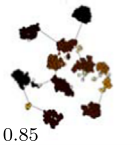
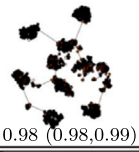
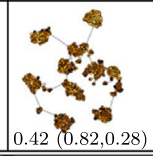
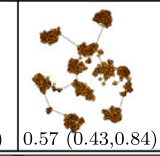
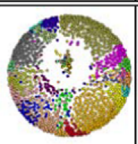
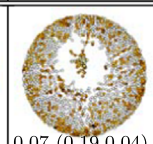
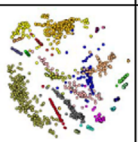
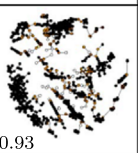
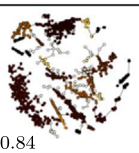
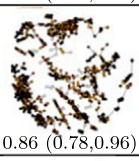
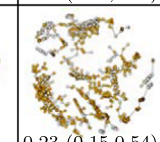
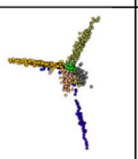
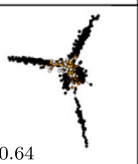
5.2. Comparing with other measures of neighborhood preservation

We compare assessments of neighborhood preservation obtained with measures based on the co-ranking matrix and with our *Neighborhood Validation*, taking as example t-SNE projections of the *Coil-20* data set[5] (with *perplexity* set to 15, as in [20]). This consists of 1440 images (128×128 bitmaps) of 20 objects: each object characterizes a class, and has been photographed at 72 distinct rotations, at 5 degree increments. Rotationally symmetric objects are thus described by highly similar images, unlike non-symmetric ones (see Fig. 13).

Fig. 14 (a) shows the data classes, Fig. 14(b)–(d) shows, respectively, the mappings of global *Neighborhood Validation* and corresponding precision and recall; (e)–(g) show measures derived from the co-ranking matrix: (e) refers to the one proposed by Lee and Verleysen [19], with $K_{max} = 4$ (the optimal value as obtained with the strategy introduced by the authors) and

[5] http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.php.

Table 5
Column 1: projections of *Optdigits* and *News2011* (points colored by class); Columns 2–6: corresponding measures of visual properties and property preservation. Point values mapped using the Heated Objects color scale, where darker is better. Each cell shows the average value relative to the projection and in parenthesis the *precision* and *recall* values of the preservation measures.

		classes	class separation	class aggregation	class separation validation	neighborhood validation	cluster validation
Optdigits	Sammon's		 0.6	 0.31	 0.75 (0.99,0.61)	 0.08 (0.16,0.05)	 0.35 (0.26,0.55)
Optdigits	LSP		 0.79	 0.46	 0.88 (0.98,0.8)	 0.25 (0.35,0.19)	 0.43 (0.33,0.62)
Optdigits	PCA		 0.58	 0.25	 0.74 (0.99,0.59)	 0.08 (0.15,0.05)	 0.37 (0.27,0.58)
Optdigits	t-SNE		 0.98	 0.85	 0.98 (0.98,0.99)	 0.42 (0.82,0.28)	 0.57 (0.43,0.84)
News2011	Sammon's		 0.77	 0.60	 0.81 (0.81,0.82)	 0.07 (0.19,0.04)	 0.15 (0.09,0.46)
News2011	LSP		 0.93	 0.84	 0.86 (0.78,0.96)	 0.15 (0.28,0.1)	 0.23 (0.15,0.54)
News2011	PCA		 0.64	 0.5	 0.78 (0.84,0.72)	0.05 (0.14,0.03)	0.17 (0.1,0.57)
News2011	t-SNE		0.98	0.93	0.87 (0.77,0.99)	0.12 (0.39,0.07)	0.27 (0.18,0.58)

mapped pointwise; (f) shows the measure by Mokbel et al.[20] with $K_t = K_s = K_{max} = 4$, and finally the mapping in (g) shows the same now considering $K_s = 4$, $K_t = 10$ (an arbitrary choice for K_t).

Overall, the color mappings are consistent and indicate that neighborhood preservation is mostly very high, according to all measures. They all show that some points from a few classes have lower values, e.g., 6, 12, 14, 15, 16, 17, 19, signaling some discrepancy in the original and projected neighborhoods. We notice, however, that preservation as measured by *Neighborhood Validation (NV)* is worse in classes 12, 15, 16 and 17 than measured by the other methods. Fig. 13(e)–(h) shows the objects described by each (at rotations 0, 120 and 240 degrees): unlike all the others in the database, these four are rotationally symmetric objects with highly similar description images that correspond to data points very close in the high-dimensional space. Ideally, a projection should group these corresponding points very strongly. Measure NV is indicating that the mapping is not reflecting this: precision is good ('real' neighbors are preserved), but the values of NV *recall* are the lowest from all classes (see Table 7).

In Table 8 we present a comparative overview of our proposed measures and others described in the literature, regarding their overall behavior and how they contemplate certain properties. Measures have been grouped in three categories: those strictly concerned with how classes are conveyed in the visual display, those based on assessing how the mapping affects neighborhoods, and those based on assessing the quality of groups formed.

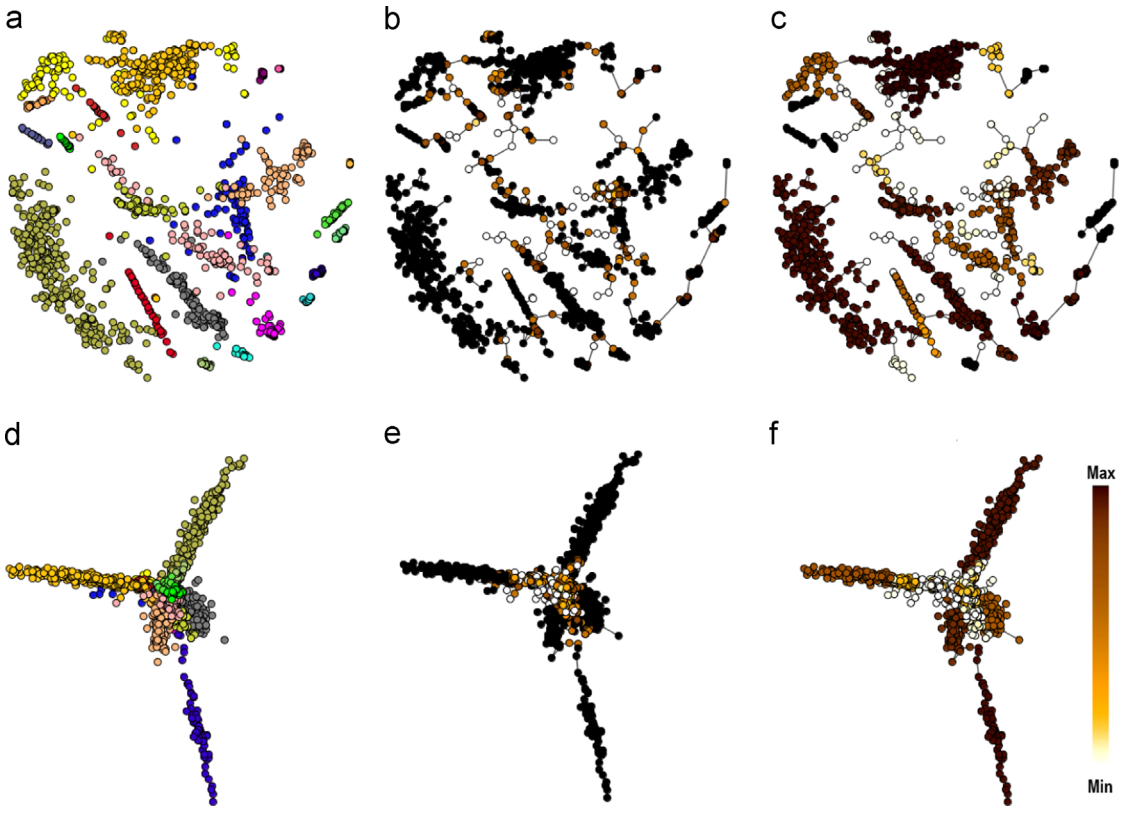


Fig. 9. LSP and PCA projections of *News2011*, and corresponding *Class Separation* and *Class Aggregation* values mapped to color (darker is better). Numbers refer to the average measure for the projection. (a) LSP classes, (b) class separation: 0.93, (c) class aggregation: 0.84, (d) PCA: classes, (e) class separation: 0.64, (f) class aggregation: 0.5.

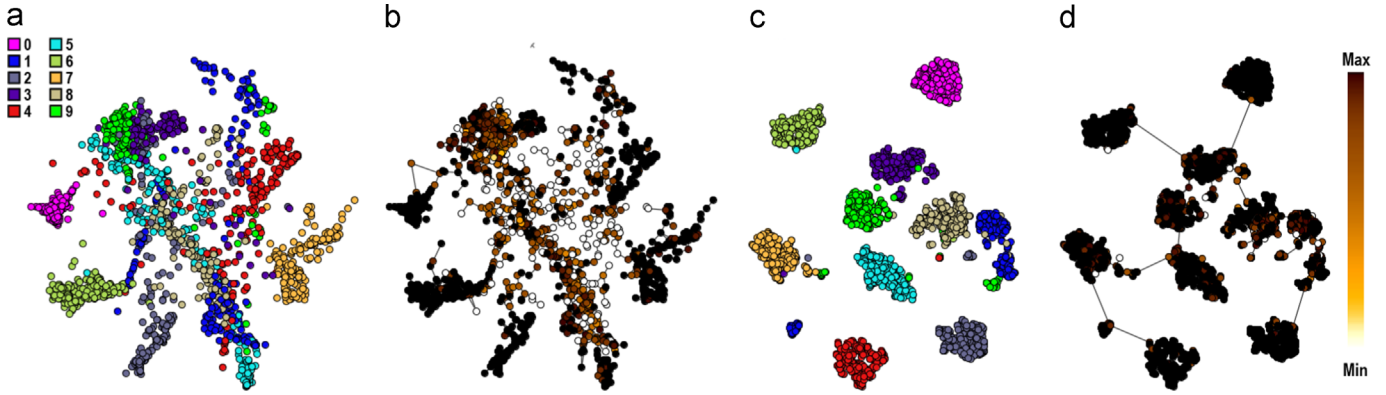


Fig. 10. LSP and t-SNE projections of *Optdigits*: (a) LSP with classes; (b) LSP mapping *Class Separation Validation*; (c) t-SNE with classes; (d) t-SNE mapping *Class Separation Validation* (darker is better). Summary measure for the projection is shown in parentheses.

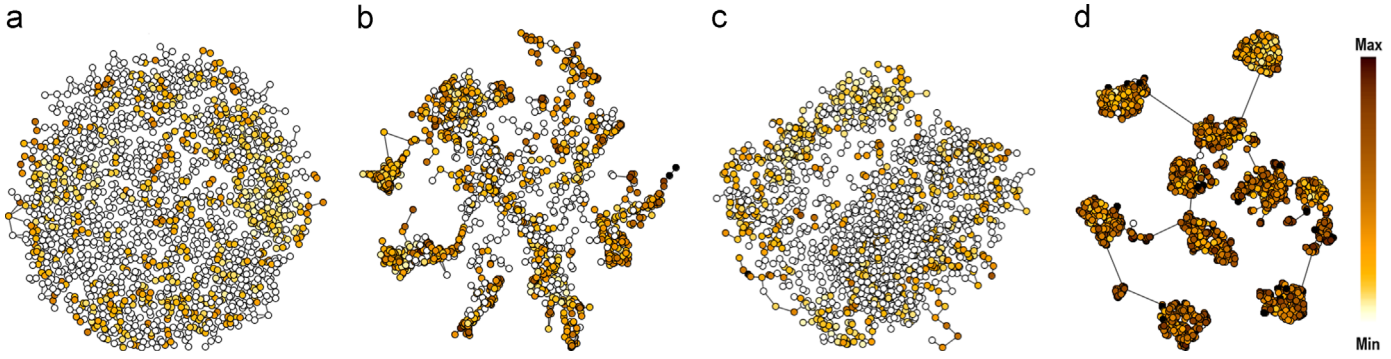


Fig. 11. Sammon's, LSP, PCA and t-SNE projections of *Optdigits*, with *Neighborhood Validation* mapped to color (darker is better). (a) Sammon's (0.08), (b) LSP (0.25), (c) PCA (0.08), and (d) t-SNE (0.42).

Table 6
Class composition of the 15 *m*-dimensional clusters extracted from *Optidigits* by the AC algorithm.

Cluster id	Class	Purity (%)	Points per class									
			0	1	2	3	4	5	6	7	8	9
Cluster 0	0	100	178	0	0	0	0	0	0	0	0	0
Cluster 1	1	98	0	99	0	0	0	0	0	0	2	0
Cluster 2	1	96.6	0	56	0	0	0	0	0	0	1	1
Cluster 3	1	100	0	27	0	0	0	0	0	0	0	0
Cluster 4	2	100	0	0	176	0	0	0	0	0	0	0
Cluster 5	3	92.3	0	0	0	12	0	0	0	0	0	1
Cluster 6	3	98.2	0	0	1	164	0	0	0	0	1	1
Cluster 7	4	100	0	0	0	0	175	0	0	0	0	0
Cluster 8	5	98.9	0	0	0	0	0	173	0	0	0	2
Cluster 9	6	98.9	0	0	0	0	0	1	180	0	1	0
Cluster 10	7	97.1	0	0	0	2	3	0	0	166	0	0
Cluster 11	7	56.5	0	0	0	0	0	0	0	13	0	10
Cluster 12	8	89.9	0	0	0	5	3	7	1	0	169	3
Cluster 13	9	100	0	0	0	0	0	0	0	0	0	19
Cluster 14	9	99.3	0	0	0	0	0	1	0	0	0	143

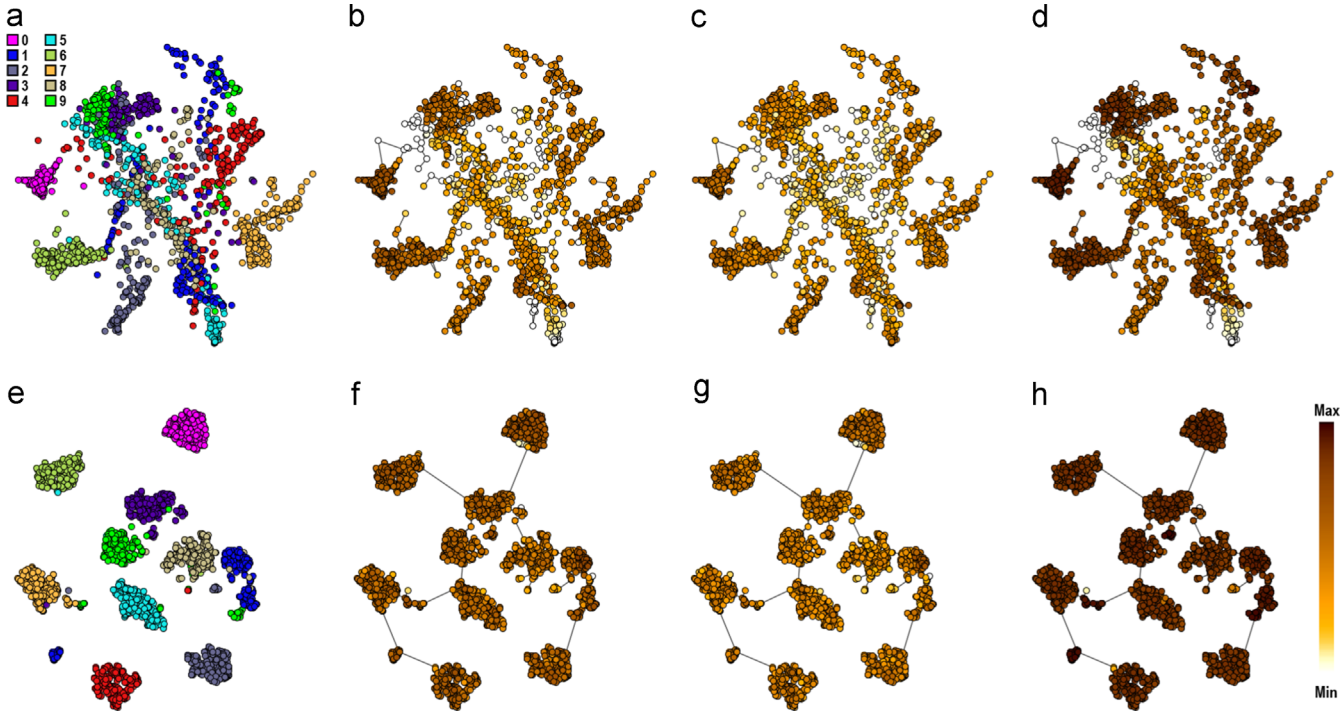


Fig. 12. LSP and t-SNE projections of *Optdigits*, with color mappings of the *Group Validation* measure (b) and (f); *Group Validation precision* (c) and (g); and *Group Validation recall* (d) and (h). (a) LSP classes, (b) group validation (0.43), (c) group validation (precision: 0.33), (d) group validation (recall: 0.62), (e) t-SNE classes, (f) group validation (0.57), (g) group validation (precision: 0.43), and (h) group validation (recall: 0.84).

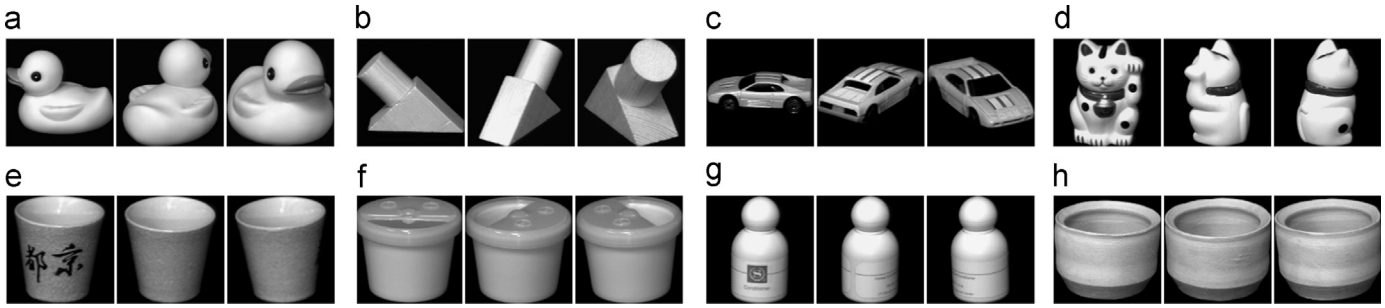


Fig. 13. Samples of eight objects from *Coil-20* viewed at three distinct rotations. (a)–(d) illustrate classes with high *Neighborhood Validation*, (e)–(h) illustrate the classes with worst *Neighborhood Validation* in Fig. 14.

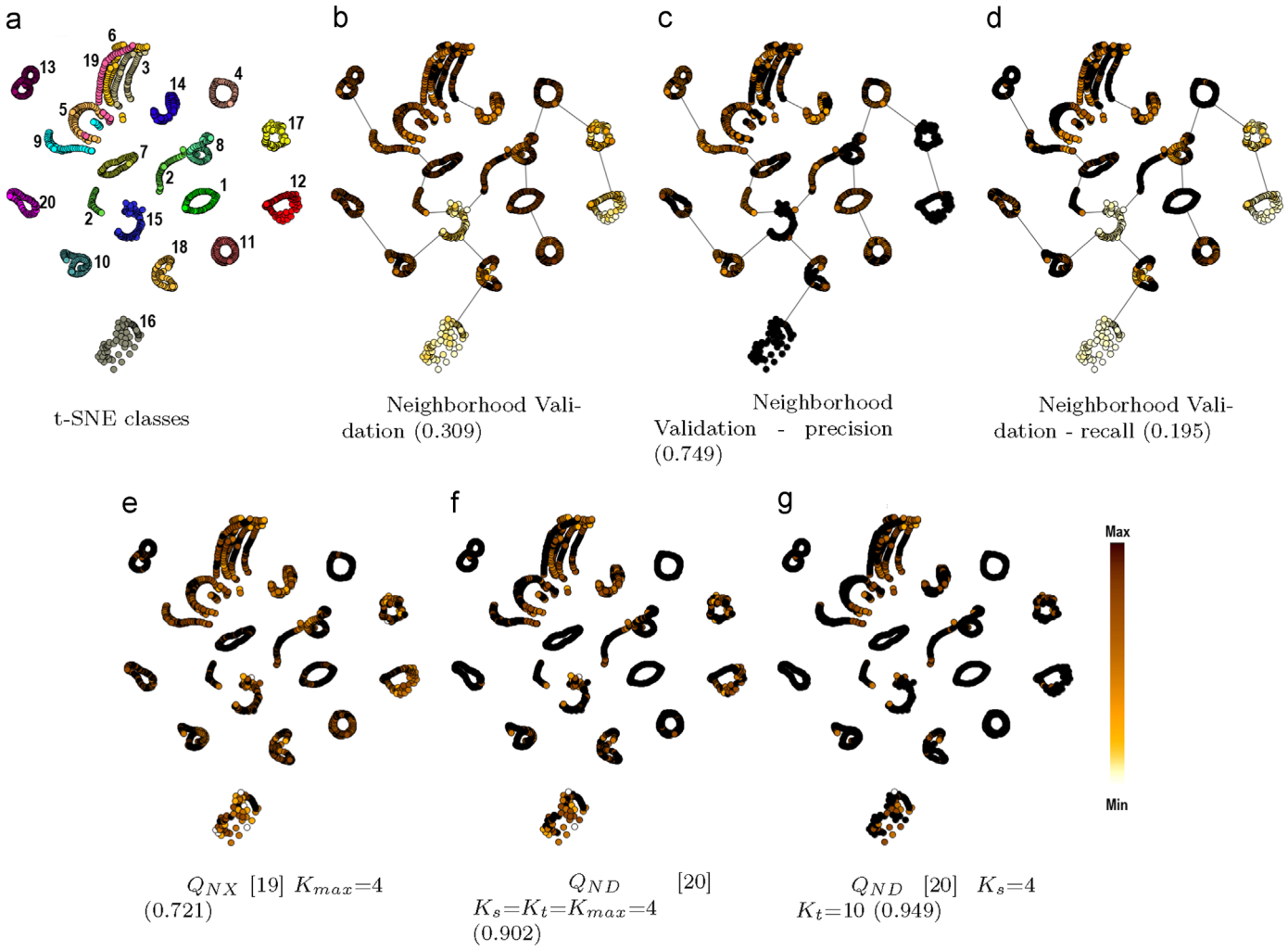


Fig. 14. t-SNE projections of *Coil-20*. (a) shows the classes, and (b)–(g) show different neighborhood preservation measures mapped to color (darker is better).

6. Conclusion

We addressed the problem of giving users objective quantitative measures to interpret, assess and compare projections of high-dimensional data regarding visual properties of class segregation, as well as of preservation of neighborhoods and groupings as compared to the original space. We derive such measures using the *EMST* similarity graph, which translates relevant data neighborhood patterns into graph connections with no need to specify neighborhood sizes or other parameters. The *EMST* provides a single unifying framework from which several topological measures can be derived that capture multiple projection properties. This capability was exemplified by two measures that assess how a visual mapping distinguishes different classes and aggregates points in a single class, and by three measures that verify property preservation relative to the original space, one quantifying 'neighborhood purity' regarding class (if given), and others quantifying neighborhood preservation and grouping capability. The proposed measures are meant to reflect properties that are both understandable by analysts and important to interpret and evaluate projections. Previous evaluation efforts do not convey, either explicitly or effectively, such a broad set of properties. Our measures can be computed and displayed relative to individual points, to arbitrary groups of points (e.g., a class, or a selection), or to the projection as a whole. As such, they serve both to infer local properties of the mappings and to detect local artifacts, and also as summary indicators of global behavior when comparing alternative mappings of the same data. Several examples were presented to illustrate how the measures can assist interpretation of mappings and assess their reliability regarding those properties. Our experiments

Table 7
Neighborhood Validation precision and *recall* for the classes of the *Coil-20* data set.

Class	Average precision	Average recall
1	0.660	0.995
2	0.592	0.968
3	0.824	0.922
4	0.675	0.990
5	0.680	0.817
6	0.713	0.793
7	0.852	0.938
8	0.941	0.622
9	0.561	0.883
10	0.609	0.904
11	0.733	0.808
12	0.997	**0.073**
13	0.607	0.984
14	0.531	0.666
15	0.984	**0.060**
16	0.997	**0.049**
17	0.994	**0.119**
18	0.841	0.714
19	0.739	0.612
20	0.859	0.796

Table 8
Comparative overview.

A. Visual measures (class related)	[A1]	[A2]	[A3]	[A4]	[A5]	[A6]	[A7]	[A8]	[A9]
Distance consistency [8]	X	X	X	X			X		
Distribution consistency [8]	X	X		X			X		X
Class density measure [9]	X	X			X		X		
Class separation	X	X	X			X	X		X
Class aggregation	X	X	X			X		X	X
B. Neighborhood-based measures	[B1]	[B2]	[B3]	[B4]	[B5]	[B6]	[B7]	[B8]	[B9]
Neighborhood preservation [12]	X	X		X		X			
Trustworthiness [13]	X	X		X		X	X		X
Lee and Verleysen [18]	X			X		X			X
Mokbel et al. [20]	X	X		X		X			X
Class separation validation	X	X	X		X	X	X	X	
Neighborhood validation	X	X	X		X	X	X		
C. Cluster-based measures	[C1]	[C2]	[C3]	[C4]	[C5]	[C6]	[C7]		
Marghescu [6]	X		X	X			X		
Group validation	X	X	X	X	X	X			

Features

[A1]	[B1]	[C1]	: Global computation (whole projection)
[A2]	[B2]	[C2]	: Local computation (points, regions)
[A3]	[B3]	[C3]	: User parameterization not required
[A4]			: Neighborhood definition not required
[A5]	[B4]		: Single neighborhood size defined for all points
[A6]	[B5]		: Neighborhood sizes tailored to individual points
[A7]			: Evaluates class purity of regions
[A8]			: Detects class integrity in the projection (classes well aggregated)
[A9]			: Robust to variations in data spatial distribution (shape) and density
	[B6]	[C4]	: Evaluates projection consistency (precision) with the original space
	[B7]	[C5]	: Evaluates projection recovery (recall) of the original space
	[B8]		: Evaluates class separation as compared to the original space
	[B9]		: Measures the distortion with regards to the original neighborhoods
		[C6]	: Robust to choice of clusters (prevents bias due to choice of technique)
		[C7]	: Enables evaluation of distances among clusters

involved different projection techniques run with their default values. However, the methodology would be equally applicable to study the parameterization of a target projection technique.

Other measures can be derived from the *EMST*, e.g., it is straightforward to identify highly connected data points (hubs), or outliers. A projection-based visualization system might offer users a *toolkit* of quality measures, from which s/he can select those relevant to a specific problem or task. They can be mapped visually not only to the point cloud, but to other complementary views, such as distribution histograms of their values. Such multiple views can be coordinated and coupled with functionalities for searching and filtering, e.g., to highlight points with extreme values of one or multiple target properties. We are working on such a system (available at http://vicg.icmc.usp.br/infovis2/EMSTProjectionEvaluation), to be made available in its current and future versions. Future work includes investigating the measures on additional data sets and their usability from an end-user perspective. Another line for further work is to investigate centrality and other graph measures as feature vectors descriptive of projections, that could be useful to identify potentially relevant mappings from a very large set of alternatives.

Acknowledgments

The authors acknowledge the financial support of FAPESP (The State of São Paulo Research Funding Agency) (Grant nos. 2009/03306-8 and 2011/22749-8) and CNPq (the Brazilian Federal Research Funding Agency).

References

[1] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.
[2] S. Ingram, T. Munzner, M. Olano, Glimmer: multilevel MDS on the GPU, IEEE Trans. Vis. Comput. Graph. 15 (2009) 249–261.
[3] F.V. Paulovich, C.T. Silva, L.G. Nonato, Two-phase mapping for projecting massive data sets, IEEE Trans. Vis. Comput. Graph. 16 (2010) 1281–1290.
[4] J. Poco, R. Etemadpour, F. Paulovich, T. Long, P. Rosenthal, M. Oliveira, L. Linsen, R. Minghim, A framework for exploring multidimensional data with 3D projections, Comput. Graph. Forum 30 (2011) 1111–1120.
[5] P. Joia, F.V. Paulovich, D. Coimbra, J.A. Cuminato, L.G. Nonato, Local affine multidimensional projection, IEEE Trans. Vis. Comput. Graph. 17 (2011) 2563–2571.
[6] D. Marghescu, Evaluating the effectiveness of projection techniques in visual data mining, in: Proceedings of the 6th International Conference on Visualization, Imaging, and Image Processing, ACTA Press, 2006, pp. 94–103.
[7] F.V. Paulovich, L.G. Nonato, R. Minghim, H. Levkowitz, Least square projection: a fast high precision multidimensional projection technique and its application to document mapping, IEEE Trans. Vis. Comput. Graph. 14 (2008) 564–575.
[8] M. Sips, B. Neubert, J.P. Lewis, P. Hanrahan, Selecting good views of high-dimensional data using class consistency, Comput. Graph. Forum 28 (2009) 831–838.
[9] A. Tatu, G. Albuquerque, M. Eisemann, J. Schneidewind, H. Theisel, M. Magnor, D. A. Keim, Combining automated analysis and visualization techniques for effective exploration of high-dimensional data, in: Proceedings of the IEEE Symposium on Visual Analytics Science and Technology, IEEE CS Press, 2009, pp. 59–66.
[10] J.W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. 18 (1969) 401–409.
[11] A. Morrison, M. Chalmers, A pivot-based routine for improved parent-finding in hybrid MDS, Inf. Vis. 3 (2004) 109–122.
[12] F.V. Paulovich, R. Minghim, Text map explorer: a tool to create and explore document maps, in: Proceedings of the 10th International Conference on Information Visualisation, IEEE CS Press, 2006, pp. 245–251.
[13] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: an experimental study, in: Proceedings of the International Conference on Artificial Neural Networks, Springer-Verlag, 2001, pp. 485–491.
[14] S. Lespinats, M. Aupetit, Checkviz: sanity check and topological clues for linear and non-linear mappings, Comput. Graph. Forum 30 (2011) 113–125.
[15] M. Aupetit, Visualizing distortions and recovering topology in continuous projection techniques, Neurocomputing 70 (2007) 1304–1330.
[16] R.M. Martins, D.B. Coimbra, R. Minghim, A. Telea, Visual analysis of dimensionality reduction quality for parameterized projections, Comput. Graph. 41 (2014) 26–42.
[17] M. Sedlmair, A. Tatu, T. Munzner, M. Tory, A taxonomy of visual cluster separation factors, Comput. Graph. Forum 31 (2012) 1335–1344.

[18] J.A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72 (2009) 1431–1443.
[19] J.A. Lee, M. Verleysen, Scale-independent quality criteria for dimensionality reduction, Pattern Recognit. Lett. 31 (2010) 2248–2257.
[20] B. Mokbel, W. Lueks, A. Gisbrecht, B. Hammer, Visualizing the quality of dimensionality reduction, Neurocomputing 112 (2013) 109–123.
[21] 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.
[22] L. Wilkinson, A. Anand, R. Grossman, Graph-theoretic scagnostics, in: Proceedings of the 2005 IEEE Symposium on Information Visualization, IEEE CS Press, 2005, pp. 21.
[23] L. Wilkinson, G. Wills, Scagnostics distributions, J. Comput. Graph. Stat. 17 (2008) 473–491.
[24] X. Zhu, Semi-Supervised Learning Literature Survey, Technical Report, Computer Sciences, University of Wisconsin-Madison, 2005.
[25] K. Aoyama, K. Saito, H. Sawada, N. Ueda, Fast approximate similarity search based on degree-reduced neighborhood graphs, in: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '11, ACM, 2011, pp. 1055–1063.
[26] A.E. Bayá, P.M. Granitto, Clustering gene expression data with a penalized graph-based metric, BMC Bioinform. 12 (2011) 2.
[27] R. Motta, A.A. Lopes, M.C.F. Oliveira, Centrality measures from complex networks in active learning, in: Proceedings of the 12th International Conference on Discovery Science, Springer-Verlag, 2009, pp. 184–196.
[28] R. Motta, B.M. Nogueira, A.M. Jorge, A. de Andrade Lopes, S.O. Rezende, M.C.F. de Oliveira, Comparing relational and non-relational algorithms for clustering propositional data, in: Proceedings of the 28th Annual ACM Symposium on Applied Computing, SAC '13, ACM, 2013, pp. 150–155.
[29] M. de Berg, W. Meulemans, B. Speckmann, Delineating imprecise regions via shortest-path graphs, in: Proceedings 19th International Conference on Advances in Geographic Information Systems, GIS '11, ACM, 2011, pp. 271–280.
[30] H. Levkowitz, Color Theory and Modeling for Computer Graphics, Visualization, and Multimedia Applications, Kluwer Academic Publishers, Norwell, MA, USA, 1997.
[31] Z. Ye, S. Hu, J. Yu, Adaptive clustering algorithm for community detection in complex networks, Phys. Rev. E 78 (2008) 046115.
[32] F.S. Roman, R.D. de Pinho, R. Minghim, M.C.F. Oliveira, A study on the role of similarity measures in visual text analytics, in: Proceedings of the 6th International Conference on Information Visualization Theory and Applications, 2013, pp. 429–438.
[33] I.T. Jolliffe, Principal Component Analysis, second ed., Springer, New York, 2002.

Robson Motta received his PhD from the Universidade de São Paulo in São Carlos, Brazil. He is interested in information visualization, data mining, and big data analytics.

Rosane Minghim is an associate professor at Universidade de São Paulo in São Carlos, Brazil. She is interested in all aspects of visualization, information visualization, visual analytics and a wide variety of applications.

Alneu de Andrade Lopes is an assistant professor at University of São Paulo in São Carlos. He is member of the Machine Learning Group (Biocom). His research interests lie in the fields of Machine Learning and Data Mining, in particular in Graph-Based Relational Learning.

Maria Cristina F. Oliveira is currently a professor at the Computer Science Department of the Instituto de Ciências Matemáticas e de Computação, at the University of São Paulo. Her research interests are in visual analytics and visual data mining techniques and applications.