A methodology to compare Dimensionality Reduction algorithms in terms of loss of quality

Antonio Gracia , Santiago González, Víctor Robles, Ernestina Menasalvas

A B S T R A C T

Dimensionality Reduction (DR) is attracting more attention these days as a result of the increasing need to handle huge amounts of data effectively. DR methods allow the number of initial features to be reduced considerably until a set of them is found that allows the original properties of the data to be kept. However, their use entails an inherent loss of quality that is likely to affect the understanding of the data, in terms of data analysis. This loss of quality could be determinant when selecting a DR method, because of the nature of each method.

In this paper, we propose a methodology that allows different DR methods to be analyzed and compared as regards the loss of quality produced by them. This methodology makes use of the concept of preservation of geometry (quality assessment criteria) to assess the loss of quality. Experiments have been carried out by using the most well-known DR algorithms and quality assessment criteria, based on the literature. These experiments have been applied on 12 real-world datasets.

Results obtained so far show that it is possible to establish a method to select the most appropriate DR method, in terms of minimum loss of quality. Experiments have also highlighted some interesting relationships between the quality assessment criteria. Finally, the methodology allows the appropriate choice of dimensionality for reducing data to be established, whilst giving rise to a minimum loss of quality.

1. Introduction

The use of Dimensionality Reduction (DR) in recent decades has been motivated by the difficulties in analyzing very high dimensional data. Historically, the main DR applications have been, amongst others, the elimination of data redundancy and noise, the reduction in the number of features for minimizing the computational cost in data pre-processing, the identification of the most discriminative features and the reduction of features for visualization tasks.

However, the use of DR entails an inherent loss of quality that is likely to affect the understanding of the data, in terms of data mining. That is, patterns discovered and extracted from a dimensionally reduced data will probably be a small part of the patterns extracted from the original data. Furthermore, the meaning of these patterns may be altered by this reduction.

On the other hand, each DR algorithm has been created to achieve a specific aim, which defines its specific nature. It is also true that, depending on its specification, a DR algorithm can give rise to more or less loss of quality at the time of reducing the data.

Different comparative studies comparing the different DR algorithms are currently being addressed in the literature [74,108,66]. Specifically, a set of quality assessment criteria, based on geometry-preservation concepts, have been used in several comparative research studies [77,37,119]. However, these studies are not sufficiently complete because of the lack of quality criteria and datasets used, as well as the fact that an exhaustive analysis of the geometry preservation is not carried out throughout the entire DR process (instead, it is carried out on a particular dimensionality, usually 2).

In this paper we propose a methodology for comparing DR algorithms based on the concept of loss of quality. Thus, the loss of quality could be strongly linked to the preservation of geometry. That is, the greater the loss of quality, the less the preservation of geometry. Hence, this methodology uses 11 quality assessment criteria to make a comparative analysis. Furthermore, this new approach attempts to address some of the shortcomings of the aforementioned studies.

The rest of this paper is structured as follows: Section 2 explains the basic concepts of a DR process and classification of DR algorithms. Quality assessment measures to calculate the preservation of geometry of data, used in the proposed methodology, are presented in Section 3. Previous comparative studies on DR, presented as related work, are detailed in Section 4. The proposed methodology for the comparison of DR methods is presented in Section 5. In Section 6 the environment for carrying out the experiments is described. The experimental results are also presented. Finally, Section 7 draws the main conclusions of the paper.

2. Dimensionality reduction methods

2.1. Basis

Based on the nomenclature stated in Table 1, Dimensionality Reduction (DR) can be defined as follows: X is made up of n datavectors $x_i (i \in 1, 2, \ldots, n)$ with dimensionality D. The DR techniques transform X with dimensionality D into a new dataset Y with a *target* dimensionality d' (where $d' < D$, often $d' \ll D$), while retaining the original geometric structure of high-dimensional data as much as possible [113]. The fundamental assumption that justifies the DR is that the original data actually lies, at least approximately, on a manifold (often nonlinear) of lower dimension than the original data space. The aim of DR is to find a representation of that manifold (a coordinate system) that will allow X to be projected on it and obtain Y, that is a low-dimensional and compact representation of the data.

Let d be the intrinsic dimensionality of the dataset. The intrinsic dimensionality of data is the minimum number of parameters needed to account for the observed properties of the data [29,62]. Ideally, the reduced representation Y should have a dimensionality that corresponds to the intrinsic dimensionality of the data.

There are currently two canonical ways of dealing with data when carrying out a DR process. The first one does so in a linear way (Linear Dimensionality Reduction or LDR), while the second one is in a nonlinear way (Nonlinear Dimensionality Reduction or NLDR). LDR handles data containing linear dependencies. However, they are not powerful enough to deal with complex data. NLDR methods are assumed to be more powerful than linear ones, since the procedure to connect the latent variables (aka intrinsic dimensionality) to the observed ones (the dimensionality of the original space) may be much more

Table 1
Main nomenclature.

Notation	Description
D	Dimensionality of the high-dimensional data
d	Intrinsic dimensionality of the high-dimensional data
n	Total number of datapoints
M	Topological manifold
$\Re^D$	D-Dimensional Euclidean space where high-dimensional datapoints lie
$\Re^d$	d-Dimensional Euclidean space (low-dimensional space using d dimensionality)
x_i	the ith datapoint in $\Re^D$
y_i	the ith datapoint in $\Re^d$
X	Original dataset in $\Re^D$ $(X = x_1, x_2, \ldots, x_n)$.
Y	Reduced dataset in $\Re^d$ $(Y = y_1, y_2, \ldots, y_n)$.
Dg	Pairwise geodesic distance matrix in $\Re^D$
δ	Pairwise euclidean distance matrix in $\Re^D$
ζ	Pairwise euclidean distance matrix in $\Re^d$
Dg_{ij}	Pairwise geodesic distance between x_i and x_j
δ_{ij}	Pairwise euclidean distance between x_i and x_j
ζ_{ij}	Pairwise euclidean distance between y_i and y_j
k	Number of neighbors of a datapoint
Xi_k	Set of k nearest neighbors of x_i
Yi_k	Set of k nearest neighbors of y_i

complex than a simple matrix multiplication operation. Furthermore, the behavior of many data, such as a DNA Microarray, cannot be explained by means of LDR because it may contain essential multiple nonlinear relationships between attributes that cannot simply be interpreted by using linear models. This suggests the design of other techniques (NLDR methods) in order to highlight the true underlying structure of the data. These methods assume that data are generated in accordance with a nonlinear model [62].

2.2. Classification in DR

Different taxonomies or classification of DR techniques, in terms of Feature Extraction (FE), have been proposed. van der Maaten et al. [74] carried out a thorough comparative review of the most important linear DR techniques, and twelve front-ranked NLDR techniques. They divided the DR techniques into two criteria (Fig. 1).

First of all, they took into consideration the convex and non-convex intrinsic nature of the techniques. Convex techniques optimize an objective function that does not contain any local optima (i.e., the solution space is convex [12]), whereas non-convex techniques optimize objective functions that do contain local optima. The second division criterion is related to full or sparse spectral techniques. The first one carries out an eigendecomposition of a full matrix that captures the covariance between dimensions or the pairwise similarities between datapoints. The other case solves a sparse eigenproblem.

John A. Lee et al. proposed a different taxonomy of DR-FE techniques [62] in accordance with procedures that reduce the features or dimensionality of the data by preserving the overall shape of the geometry, or by preserving the local properties and neighborhood information of the data. Thus, there is a possibility of distinguishing both the local and global quality. [18].

2.2.1. Distance preservation

Historically, distance preservation (DP) has been the first criterion used to achieve a DR in a nonlinear way. From the point of view of an ideal case, the preservation of the *pairwise distances* measured in a dataset ensures that the low-dimensional embedding inherits the main geometric properties of the data, such as the overall shape. However, in nonlinear cases distances cannot be perfectly preserved. To explain this, it is necessary to define a *manifold*. A topological manifold M is a topological space that is locally Euclidean, meaning that around every point of M there is a neighborhood that is topologically the same as the open unit ball in $\Re^d$ [67].

DP methods can be divided into three groups as considered by Lee et al.[62]:

Spatial distance algorithms as Multidimensional Scaling (MDS) [17], Sammon Mapping or Curvilinear Component Analysis.

Geodesic distances and, specifically graph distances, were conceived to deal with some of the shortcomings in the spatial metrics (Fig. 3). The geodesic distance between two points is defined as the distance along the mathematical manifold where the data points are embedded. It can be partially approximated by constructing a neighborhood graph, and considering the distances between the points as paths in the graph (Fig. 2). Examples of algorithms using this distance are Isomap, Geodesic Nonlinear Mapping (GNLM) [23,61,59] and Curvilinear Distance Analysis (CDA) [58,59].

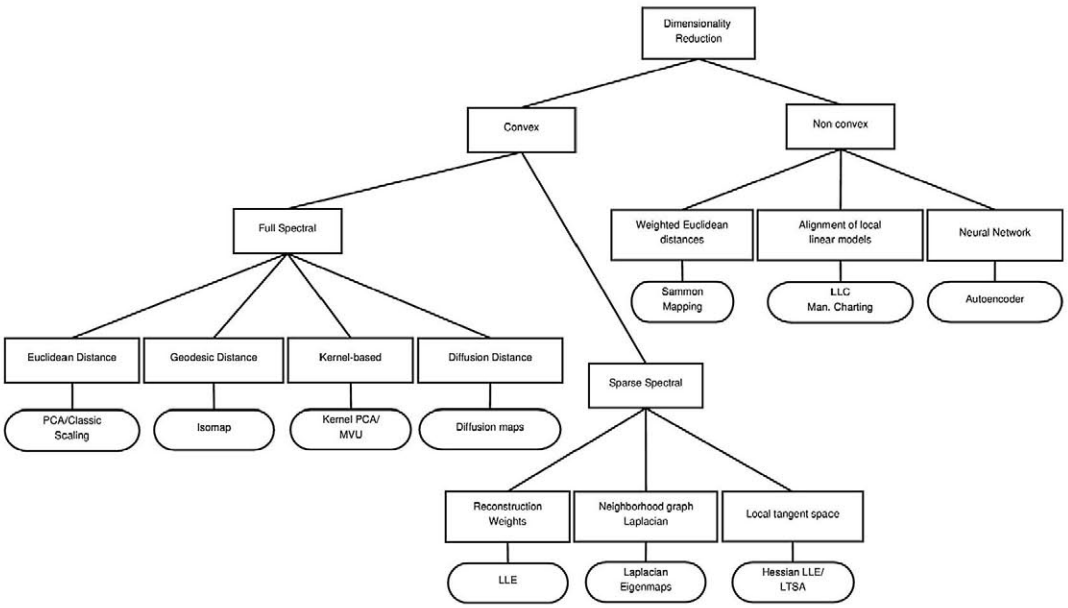

Fig. 1. van der Maaten's Taxonomy (taken from [74]).

Other distances There are also NLDR methods that rely on less geometrically intuitive ideas. These techniques are characterized by the use of other distances. For instance, Kernel PCA [88], which is closely related to the spectral methods.

2.2.2. Topology preservation

Techniques that reduce the dimensionality of the data by preserving their *topology* (TP) rather than their pairwise distances are also called local preservation approaches. These techniques help to overcome the drawback of using the DP principle: the manifold could be constrained with distance conditions and, in many situations, the embedding of a manifold requires some flexibility because some sub-regions must be locally stretched or shrunk to embed them into other dimensional spaces.

Most of these techniques work with a discrete mapping model, and the topology is also defined in a discrete way. This discrete representation of the topology is called a lattice [3], i.e., a set of points regularly and homogeneously spaced on a graph. Topology preservation (TP) techniques can be divided into two types according to the kind of topology they use. The first one deals with methods relying on a *predefined lattice*, i.e., the lattice is fixed in advance and cannot change after the DR process has begun. Self-Organizing Maps (SOM's) [48] and Generative Topographic Mapping (GTM) [9] are well-known as predefined lattice methods. The second group contains methods working with a *data-driven lattice*. This concept means that the shape of the lattice can be modified or entirely built while the methods are running. Locally linear embedding, Laplacian eigenmaps and Isotop [57] are in this category. As we will see in future sections, maybe working with ranks is the best and most reliable criterion.

2.3. Methods

Once the possible classifications of DR algorithms are presented, it is interesting to highlight those algorithms that are most used in the literature. Table 2 presents each one, with its references in the literature as well as its preservation criterion (DP, TP or other). These are also the DR algorithms used in our experiments (described in Section 6).

3. Quality assessment measures for DR

There are many different quality assessment measures for evaluating the performance of the DR algorithms. Historically, most of the approaches have focused on evaluating the local-neighborhood-preservation and the overall-structure-holding performance of the DR methods. In this section the most used measures in the literature are classified (using global or local preservation criteria) and described (see Table 3). Firstly, local-based approaches are presented. Secondly, global-based approaches are explained and finally, several approaches based on different criteria are described.

Before explaining the different approaches for quality assessment, it is very important to highlight a basic concept to better understand the following measures.

Multidimensional scaling. Multidimensional scaling (MDS) is a statistical method for fitting a set of points in a space so that the distances between points correspond as closely as possible to a given set of dissimilarities between a set of objects. Developed primarily by psychometricians and statisticians, MDS is widely used in a variety of disciplines for visualization and DR. The literature on MDS includes books [10,17,25] and book chapters ([24, chapter 5]; [53, chapter 5]; [76, chapter 14]; and [91, Section 5.5]). The method devised by Torgerson [104] and Gower [36], called classic MDS and principal coordinate analysis, could be formulated as an optimization problem with an objective function whose minimum value is called the *stress* criterion.

Later, Kruskal [51,52] defined MDS in terms of the minimization of this *stress* criterion, which is simply a measure of the lack of fit between dissimilarities δ and fitted distances ζ. In the simplest case, *stress* is a residual sum of the squares:

$$Stress_D(y_1,\ldots,y_n) = \left(\sum_{i\neq j=1\ldots n} (\delta_{ij} - \|\zeta_{ij}\|)^2\right)^{\frac{1}{2}} \tag{1}$$

where the outer square root provides greater spread to small values. For a given δ, MDS minimizes *Stress* over all different configurations $(y_1,\ldots,y_n)^T$, thought of as $n \times D$-dimensional hypervectors of unknown parameters. The minimization is carried out by gradient descent applied to $Stress_D$, viewed as a function on $\Re^{nD}$.

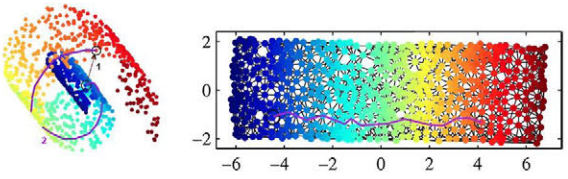

Fig. 2. This dataset consists of a list of 3-dimensional points. It is, a two-dimensional manifold embedded into a three-dimensional space (taken from [62]).

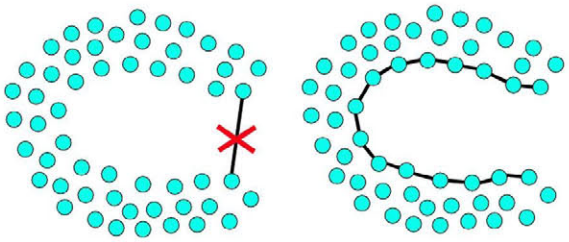

Fig. 3. Left: when performing an unfolding process, the appearance of short circuit induced by the Euclidean distance is likely. Right: the benefits of the geodesic distance. The two points are not neighbors as they are far away in accordance with the geodesic distance.

Table 2
Most used DR algorithms in the literature, listed chronologically.

Year	DR algorithm	Reference	Criterion
1901	Principal Component Analysis (PCA)	[44,41]	Other
1969	Sammon Mapping (SM)	[86]	DP
1997	Curvilinear Component Analysis (CCA)	[20]	DP
1998	Kernel PCA (KPCA)	[88,89]	DP
2000	Isomap	[103,18]	DP
2000	Locally Linear Embedding (LLE)	[85,87]	TP
2001	Linear Discriminant Analysis (LDA)	[22,38,29]	Other
2001	Laplacian Eigenmaps (LE)	[5,6]	TP
2004	Maximum Variance Unfolding (MVU)	[114,117,116]	DP
2006	Diffusion Maps (DM)	[80,56]	TP
2008	*t*-Stochastic Neighbor Embedding (*t*-SNE)	[72]	TP

Table 3
Summary of methods for evaluating the quality of DR algorithms, listed chronologically.

Year	Name of the measure	Criterion	Reference
1962	Sheppard Diagram (SD)	Global	[93,94]
1964	Kruskal Stress Measure (S)	Global	[51,52]
1969	Sammon Stress (S_S)	Global	[86]
1988	Spearman's Rho (S_R)	Local	[98]
1992	Topological Product (T_{Pr})	Local	[4]
1997	Topological Function (T_F)	Local	[111]
2000	Residual Variance (R_V)	Global	[103]
2000	König's Measure (K_M)	Local	[49]
2001	Trustworthiness & Continuity (T&C)	Local	[107]
2003	Classification error rate	classification error	[87,114,106]
2006	Local Continuity Meta-Criterion (Q_k)	Local	[13,14]
2006	Agreement Rate (A_R)/Corrected Agreement Rate (CA_R)	Local	[2]
2007	Mean Relative Rank Errors (MRRE)	Local	[62,65,64]
2009	Procrustes Measure (P_M)/Modified Procrustes Measure (P_{MC})	Local	[31]
2009	Co-ranking Matrix (Q)	Local	[65,63]
2011	Global Measure (Q_Y)	Local and global	[77]
2011	The Relative Error (R_E)	Global	[37]
2012	Normalization independent embedding quality assessment (NIEQA)	Local/global/local&global	[119]

3.1. Local-neighborhood-preservation approaches

Spearman's Rho Siegel and Castellan presented one of the first measures to estimate the *topology preservation* (TP) after a DR process, Spearman's rho (S_R) [98]. This measure estimates the correlation of rank order data. That is, it tries to assess how well the corresponding projection preserves the order of pairwise distances between data-points in high-dimensional space. In order to compute the S_R, the following equation is used

$$S_R = 1 - \frac{6\sum_{i=1}^{T}(z(i) - \hat{z}(i))^2}{T^3 - T} \tag{2}$$

where $z(i), i = 1, T$ is the different rank (order numbers) of pairwise distances in the original space, sorted in ascending order. $\hat{z}(i), i = 1, T$ is the same for the output space. T is the total number of distances to be compared ($T = n(n-1)/2$). The interval is $S_R \in [-1, 1]$, where 1 means a perfect preservation. S_R is often used for estimating the TP with a view to reducing dimen-

sionality [8,34,50,7]. Karbauskaite and Dzemyda [45] successfully demonstrated that S_R can be used to analize the TP when visualizing the data through the embeddings generated by the LLE algorithm.

In contrast to MDS, that only focuses on fitting the distances from δ to ζ (the order of these distances do not matter), the S_R measure also takes into account the rank of pairwise distances in δ and ζ for the quality assessment.

Topological Product. The following attempts are found in the particular case of SOM. In this sense, Bauer and Pawelzik proposed the topographic product (T_{Pr}) [4]. T_{Pr} is one of the oldest measures that quantifies the TP features of the SOM, and it is a measure for the preservation of distances within the local neighborhoods. Let $Q_1(i,j)$ be the distance between point i in $\Re^D$ and its jth nearest neighbor as measured by distance orderings of their images in $\Re^d$, divided by the distance between point i in $\Re^D$ and its jth nearest neighbor as measured by distance orderings in $\Re^D$. $Q_2(i,j)$ gives analogous information where i and j are points in $\Re^d$. The Qs are then combined to yield a single number T_{Pr}, the *topological product*, which defines the quality of the mapping:

$$T_{Pr} = \frac{1}{n(n-1)} \sum_{g=1}^{n} \sum_{f=1}^{n-1} log\left(\Pi_{p=1}^{f} Q_1(g,p) Q_2(g,p)\right)^{\frac{1}{2f}} \tag{3}$$

The result of the T_{Pr} indicates whether the size of the map is appropriate to fit into the dataset. $T_{Pr} = 0$ means a perfectly order-preserving map.

Topological Function. Five years later Villmann et al. presented the topological function (T_F, 1997) [110]. The T_F was one of the simplest TP measures in SOM. T_F is based in the Delaunay triangulation graph D of the weight vectors. These vectors w_i and w_j were defined as being adjacent on the manifold V, if their receptive fields are adjacent. Thus, the adjacency of these receptive fields can be approximated by computing C (the connectivity matrix) of the induced Delaunay triangulation graph D:

1. Given a data sample, find its first best matching unit i and second best matching unit j.
2. Create a synaptic link between neurons i and j, i.e. set $C_{ij} = 1$.
3. Go back to step 1 and repeat for all datasamples.

If the number of weight vectors is *"dense"* enough on the manifold V, then D represents a perfect TP mapping of V that also preserves the paths on V. Villmann et al. demonstrated that the T_F presents reliable results only for almost linear datasets [111].

König's Measure. König. A developed a TP measure, the König's measure (K_M) [49]. K_M was used to estimate the local preservation of the maps, obtained when using self-organizing neural networks. It is also based on the analysis of rank order in the input and output spaces. The K_M is calculated as follows:

$$K_M = \frac{1}{3k_1 n} \sum_{i=1}^{n} \sum_{j=1}^{k_1} KM_{ij} \tag{4}$$

$K_M \in [0,1]$, where 1 means a perfect preservation. KM_{ij} represents the TP between point i and j, and k_1 is the neighborhood value.

Trustworthiness & Continuity. Venna and Kaski proposed a method which assesses two different concepts, trustworthiness and continuity (T&C) [107]. It is based on the exchange of indices of neighboring samples in D and d (by using the pairwise Euclidean distances), respectively. The T&C criterion involves two evaluations, the trustworthiness and the continuity measure, defined, respectively, as:

$$M_T = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in U_k(i) \notin V_k(i)} (r(i,j) - k) \tag{5}$$

$$M_C = 1 - \frac{2}{nk(2n-3k-1)} \sum_{i=1}^{n} \sum_{j \in V_k(i) \notin U_k(i)} (\hat{r}(i,j) - k) \tag{6}$$

where k is the size of the neighborhood, $r(i,j)$ and $\hat{r}(i,j)$ are the rank of x_j and y_j in the ordering according to the distance from $x_i(y_i)$ in the original (representational) space. $U_k(i)$ and $V_k(i)$ are the set of those data samples that are in k of $x_i(y_i)$ in the representational (original) space. As regards the meaning of M_T and M_C, the former measures the degree of trustworthiness that data points which were originally farther away enter the neighborhood of a sample in the embeddings. The latter evaluates the degree of continuity that data points that are originally in the neighborhood are pushed farther away in data representations. Therefore, the T&C measure is defined as:

$$Q_T = \alpha M_T + (1-\alpha) M_C \tag{7}$$

where $\alpha \in [0,1]$ is the compromise parameter. The trade-off between the two terms, tunable by a parameter α, governs the trade-off between trustworthiness and continuity. A properly selected α value, can reflect the consistency between the local

neighborhoods of the original data and the corresponding ones in the embeddings calculated by the NLDR method. The interval of $Q_T \in [0,1]$ which are the higher values means a good preservation of trustworthiness and continuity.

Local Continuity Meta-Criterion. There are also several methods that assess the performance of the DR algorithms by checking the degree of overlap between the neighboring sets of a data sample and of their corresponding embedding. This is the case of the Local Continuity Meta-Criterion (Q_k) [13,14], presented by Chen and Buja. The Q_k can be defined as:

$$Q_k = 1 - \frac{1}{nk}\sum_{i=1}^{n}\left|\Psi_k^x(i) \bigcap \Psi_k^y(i)\right| - \frac{k^2}{n-1} \tag{8}$$

where k is the pre-specified size of the neighborhood, $\Psi_k^x(i)$ is the index set of $x_i's$ k points and $\Psi_k^y(i)$ is the index set of $y_i's$ k points. If the overlap between two k neighboring sets of the original and representational sets is computing, the Q_k gives a general measurement for the local faithfulness of the computed embeddings. The interval of $Q_k \in [0,1]$, whose values next to 1 mean a high neighborhood overlap between the two dimensional spaces, and next to 0 values the opposite.

In contrast to T_{Pr}, which attempts to measure the distance preservation (DP) between the local neighbors, the Q_k measure focuses on comparing the identities of these local neighbors.

Agreement Rate/Corrected Agreement Rate. The agreement rate (A_R, originally called 'rate of agreement in local structure') technique was presented by Akkucuk and Carroll [2]. This method is very similar to Q_k, and RAND or corrected RAND index [84,42]. A_R was originally developed for comparing embeddings of sets of objects in [1]. It works as follows: it takes two configuration of points X and Y. For each embedding, A_R calculates the distances between each pair of datapoints, this give us δ and ζ. For each datapoint, it calculates its neighborhood in both configurations, producing Xi_k and Yi_k. Finally, it attempts to compute the percentage of overlapping datapoints in the neighborhood of each point, for X and Y. Here, the order is not important. Let us consider u_i as the number of overlapping points in both Xi_k and Yi_k, for datapoint i. Therefore, the A_R is

$$A_R = \frac{1}{kn}\sum_{i=1}^{n} u_i \tag{9}$$

where an A_R value equal to 1 means a perfect preservation. The authors also suggested another quality criterion called the corrected agreement rate (CA_R). This method computes an A_R, by randomly rearranging the indices of datapoints in Y. France and Carroll also proposed a method in [27], where they combined the use of the A_R and RAND index in order to assess DR methods.

Mean Relative Rank Errors Lee and Verleysen developed a quality assessment measure, the mean relative rank errors (MRRE) [62,65,64]. It is based on ranks of pairwise Euclidean distances within local neighborhoods. In 2009, Karbauskaite et al. analyzed the efficiency of MRRE when reducing the dimensionality using LLE [46]. The MRRE criterion is based on a very similar principle to that of the T&C, but it includes two elements defined as

$$W_T = 1 - \frac{1}{H_k}\sum_{i=1}^{n}\sum_{j\in U_k(i)}\frac{\left|r(i,j) - \widehat{r}(i,j)\right|}{r(i,j)} \tag{10}$$

$$W_C = 1 - \frac{1}{H_k}\sum_{i=1}^{n}\sum_{j\in V_k(i)}\frac{\left|r(i,j) - \widehat{r}(i,j)\right|}{\widehat{r}(i,j)} \tag{11}$$

where k is the size of the neighborhood and Eq. 12 is the normalizing factor. The MRRE criterion is Eq. 13 where $\beta \in [0,1]$ is the compromise parameter. The main difference between the MRRE and the T&C is that the first one considers all of the k samples in the representational (original) space, and the latter focuses on the k of the samples in the representational (original) space but not in the original (representational) space. Although we are talking about subtle differences between them, they are significant enough to be considered. H_k is a normalizing factor. The interval of $Q_M \in [0,1]$, whose values near to 0 will indicate a small rank error in the final embedding, are result of the error-based nature of MRRE.

$$H_k = n\sum_{i=1}^{k}\frac{|n-2i+1|}{i} \tag{12}$$

$$Q_M = \beta W_T + (1-\beta)W_C \tag{13}$$

In contrast to S_R, which focuses on assesing how well the corresponding low-dimensional projection preserves the order of pairwise distances between the high-dimensional data points converted to ranks, the Q_M measure evaluates (using an error value) that the order of Xi_k and Yi_k is the same.

Procrustes Measure/Modified Procrustes Measure. The Procrustes analysis [96,97,90] has been widely used for the study of the distribution of a set of shapes. Based on this concept, Goldberg and Ritov [31] developed the Procrustes measure (P_M). P_M allows the isometric embeddings to be compared. The method can be described as follows: using the procrustes analysis, the aim is to find a rigid motion (a translation and a rotation), after which Xi_k best when it coincides with Yi_k (for $i = 1$ to n). Once the transformation has been computed, the local similarity for the i-th element is calculated as

$$L_{similarity}(Xi_k, Yi_k) = \sum_{j=1}^{k} \left\| Xi_j - \partial Yi_j - \Im \right\|_2^2 \tag{14}$$

∂ is the selected rotation matrix and $\Im$ the translation vector ($\|\ldots\|_2$ indicates the L_2 norm for a vector). To finish, the P_M value is obtained by

$$P_M = 1/n \sum_{i=1}^{n} L_{similarity}(Xi_k, Yi_k) / \|Xi_k B_k\|_F^2 \tag{15}$$

$$B_k = I_k - \frac{1}{k} q_k q_k^T \tag{16}$$

where I_k represents the identity matrix of size $k \times k$, q_k a k dimensional column vector of ones, and $\|\ldots\|_F$ indicates the Frobenius norm for a matrix. A P_M close to zero means a perfect preservation. At this point, it is important to highlight that P_M was originally devised for assessing the quality of isometric embeddings, such as Isomap or MDS. Nevertheless, P_M will fail when assessing normalized embeddings (such as LLE, [32]), as they are known to distort the local neighborhood. To overcome this, in [31] the authors also suggested a modified version (P_{MC}) that addresses this drawback. This version eliminates the global scaling factor in each neighborhood, so it is appropriate for conformal embeddings. To summarize, the main difference between the two versions of the measure is that P_M takes into account the stretch/shrink factor, and P_{MC} does not. In the particular case of different scaling of coordinates in low dimensional embeddings, neither P_M nor P_{MC} solves the problem.

Co-ranking Matrix. Many different concepts and quality criteria for DR can be summarized using the Co-ranking framework (Q), presented by Lee and Verleysen [65,63]. Several of the aforementioned methods (based on distance ranking in local neighborhoods: Q_k, MRRE, T&C), are easily unified into an overall framework. Q works as follows: let $\rho_{i,j}$ be the rank of x_j respect to x_i in $\Re^D$,

$$\rho_{i,j} = \left| \{ k | \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leqslant k < j \leqslant n) \} \right| \tag{17}$$

and $\tau_{i,j}$ is the rank of y_j in respect to y_i in $\Re^d$,

$$\tau_{i,j} = \left| \{ k | \zeta_{ik} < \zeta_{ij} \text{ or } (\zeta_{ik} = \zeta_{ij} \text{ and } 1 \leqslant k < j \leqslant n) \} \right| \tag{18}$$

Therefore, Q can be defined as

$$Q_{kl} = \left| \left\{ (i,j) | \rho_{i,j} = k \text{ and } \tau_{i,j} = l \right\} \right| \tag{19}$$

The errors after the DR process are reflected in the non-diagonal entries of Q. So, an *intrusion* can be defined as a point j where $\rho_{i,j} > \tau_{i,j}$ (i.e. points entering a neighborhood erroneously). If $\rho_{i,j} < \tau_{i,j}$ it is called an *extrusion* (points leaving a neighborhood erroneously). Q provides a framework, in which several existing evaluation measures can be expressed in an intuitive method for visualizing the differences between Q_k, MRRE and T&C. Basically, these quality criteria correspond to weighted sums of entries Q_{kl} of Q for different regions as $k, l \leqslant K$ and a fixed neighborhood range K (Fig. 4).

Lee and Verleysen also proposed a new criterion in [65], Q_{NX}. Q_{NX} is the criterion that summarizes Q in the very simplest way, without arbitrary choices (weighting schemes, coefficient, scale preference, etc). It is defined as

$$Q_{NX}(K) = \frac{1}{Kn} \sum_{k=1}^{K} \sum_{l=1}^{K} Q_{kl} \tag{20}$$

$Q_{NX}(K)$ is the same as Q_k without the subtraction of the 'random' baseline. Note that $Q_{NX}(K)$, Q_k and A_R basically represent the same. Here, the range is $Q_{NX}(K) \in [0,1]$, where 1 means a perfect embedding. There are two other quality criteria, $B_{NX}(K)$ and $R_{NX}(K)$. The first one subtracts elements of Q that are above or below the main diagonal: it indicates whether a given embedding tends to favor intrusions or extrusions. The range is $B_{NX}(K) \in [-1,1]$. The sign depends on the dominating type of errors ($B_{NX}(K) > 0$ represents intrusions, and $B_{NX}(K) < 0$ are extrusions). Zero means an equal number of intrusions and extrusions.

The last one, $R_{NX}(K)$ [60], can be considered a *renormalized* Q_k, allowing us to compare values at different scales. $R_{NX}(K)$ is based on Q_k with a baseline subtraction and a normalization: it indicates the relative improvement in a random embedding. Thus, the main advantage of $R_{NX}(K)$ is straightforward, as two different embeddings can be compared on different values of K. This is very difficult to achieve and less interpretable with other criteria. The range is $R_{NX}(K) \in [0,1]$, where 1 represents a perfect embedding.

In [66], Lee and Verleysen studied and proposed several solutions to solve the issue of overall scale dependency.

3.2. Global-structure-holding approaches

Shepard Diagram and Kruskal Stress Measure. Shepard presented in [93,94] the Shepard Diagram (SD). The SD is known to be one of the oldest DP methods. The SD can be formally considered as the diagram obtained by plotting the $n(n-1)/2$ distances of the original configuration δ against the approximated distances ζ (Fig. 5). The SD visualizes the goodness-of-fit of all

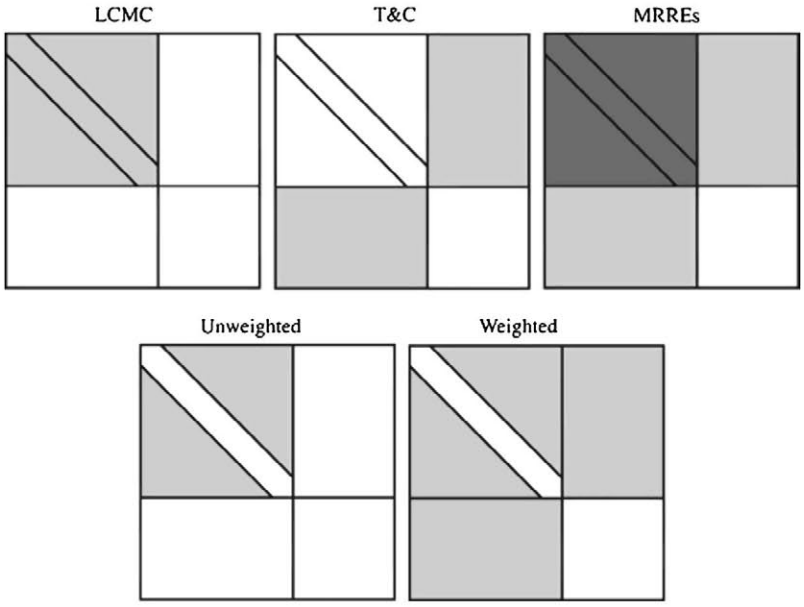

Fig. 4. Co-ranking matrix (reproduced with permission from [65]).

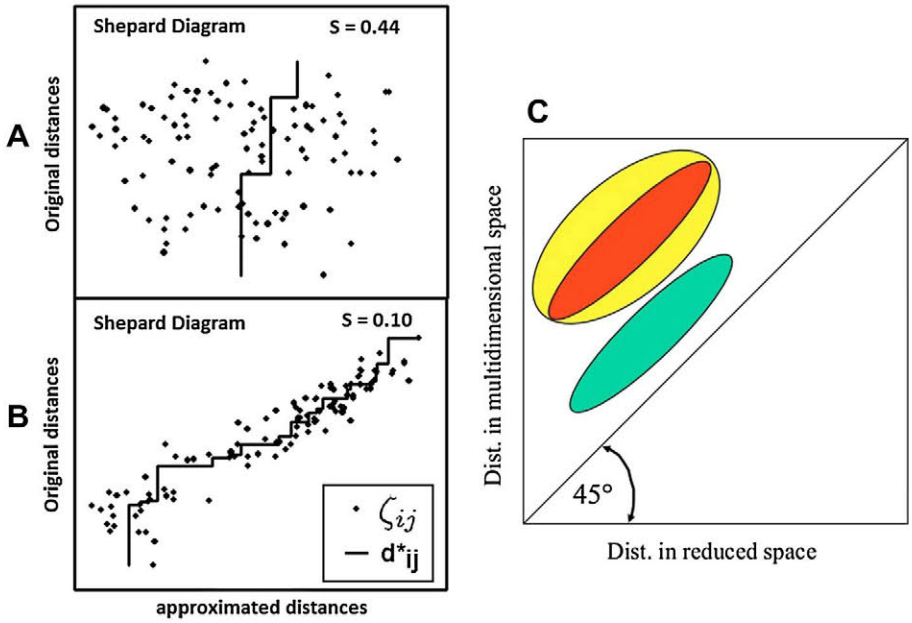

Fig. 5. Shepard Diagram example. A and B: different types of diagrams (the ideal case is when all the points lie in the diagonal line. It means that all the distances in the reduced space match the original distances, so the representation in B is better than in A). C: intuitive explanation of the SD diagrams; Original distances on a vertical axis, embedded distances on a horizontal axis. Green represents projection in a reduced space accounting for a high fraction of variance (relative positions of points are similar). Red represents projection accounting for a small fraction of variance (relative projections of objects are similar). Yellow represents projection accounting for a small fraction of variance (but the relative projection of objects differ in the two spaces). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sets of distances. It can be useful to detect anisotropic distortions in the representation. Thus, Kruskal [51,52] proposed a measure for the deviation from monotonicity between δ and ζ, called the stress function (S):

$$S = \sqrt{\frac{\sum_{ij}(\zeta_{ij} - d^*_{ij})^2}{\sum_{ij}\zeta^2_{ij}}} \tag{21}$$

Note that δ does not appear in this equation. Instead, the discrepancy between ζ_{ij} and the target distances d^*_{ij} are measured. d^*_{ij} can be computed by finding the monotonic regression [51,52]. In the SD, instead of showing individual points for d^*_{ij}, they are connected by a solid line. The target distances d^*_{ij} represent the distances that lead to a perfect monotonic relationship to δ that minimizes S for the given ζ_{ij}. S is a 'lack of fit' measure: if S equals 0, there is a perfect monotonic relationship between ζ_{ij} and δ_{ij}.

Note that, the Shepard Diagram and the Kruskal Stress Measure are based on a concept very similar to MDS, since in fact they originate from it.

Sammon Stress. The Sammon stress (S_S, Eq. 23) [86] measure is also used in order to compare the DR algorithms, in terms of DP. Examples of other error measures frequently used for structure preservation are S stress [100] (Eq. 22) and Quadratic error [11] (Eq. 24).

$$S\ stress = \sqrt{\frac{1}{n}\frac{\sum_{i<j}(\delta_{ij}^2 - \zeta_{ij}^2)^2}{\sum_{i<j}\delta_{ij}^4}} \tag{22}$$

$$Sammon\ stress(S_s) = \frac{1}{\sum_{i<j}\delta_{ij}}\sum_{i<j}\frac{(\delta_{ij} - \zeta_{ij})^2}{\delta_{ij}} \tag{23}$$

$$Quadratic\ Loss = \sum_{i<j}(\delta_{ij} - \zeta_{ij})^2 \tag{24}$$

The range is $S_S \in [0, +\infty)$, where 0 represents a perfect DP, and the quality decreases as the DP increases in value. S_S must be minimized by carrying out a gradient descent, or by other means, usually involving iterative methods.

Residual Variance. Tenenbaum et al. [103] used the residual variance (RV) for assessing the overall quality of an embedding. RV is computed by $RV = 1 - R^2(G_X, \zeta)$, where $R(G_X, \zeta)$ represents the standard linear correlation coefficients over all entries of G_X and ζ. The term G_X is the graph distance matrix [103]. The range is $RV \in [0, 1]$, where 0 value represents a perfect quality of the embedding. An RV quality criterion has been also successfully applied to choose the embedding dimensionality for Isomap [103].

The Relative Error. Handa [37] introduced the Relative error (R_E) to be used with another quality criteria, such as MRRE and Q_k, for evaluating the quality of DR methods. The R_E is calculated as

$$R_E = \sum_{i}^{n}\sum_{j=i+1}^{n}\frac{|Dg_{ij} - \zeta_{ij}|}{Dg_{ij}n(n+1)/2} \tag{25}$$

3.3. Others

There are different quality criteria approaches that do not merely focus on evaluating the TP or DP, for example: Classification error rate, Global Measure and NIEQA.

Classification Error Rate. Another approach mentioned in the literature consists of using an indirect accuracy index, such as a *classification error*. See [87,114] and other references in [106]. It can be defined as:

$$C_e = Acc_{\Re^D} - Acc_{\Re^d} \tag{26}$$

where $Acc_{\Re^D}$ is the classification accuracy in $\Re^D$, and $Acc_{\Re^d}$ represents the same in $\Re^d$. Logically, the classification error can be used only with labeled data.

The last two quality measures have recently appeared, and they share a particular feature: they combine both local and global quality measure approaches. Here, the main aim is to provide an overall or 'mixture' value that assesses the TP and DP capabilities of a DR algorithm.

Global Measure. Meng et al. [77] proposed a new quality criteria (Q_Y) that evaluates the neighborhood-preserving and global-structure performances when performing manifold learning tasks. To compute Q_Y, the shortest path tree (SPT) is generated from the k neighborhood graph. After this, the global-structure assessment is calculated using the Spearman rank order correlation, defined in the rankings of branch lengths (Q_{GB}). So, the overall assessment (Q_Y) can be defined as a linear combination of the global assessment, Q_{GB}, and a local assessment, such as Q_k (Q_M or Q_T could also be used). Then, $Q_Y = \mu Q_{GB} + (1 - \mu)Q_k$, where $\mu \in [0, 1]$ and represents a parameter to balance Q_{GB} and Q_k in quality assessment. Q_Y is valued between 0 and 1, where 1 represents a perfect global-structure-preserving.

In contrast to methods, such as S_R and Q_M, Q_Y provides a more sophisticate and complete approach, since it assesses both local and global quality. However, there is a certain similarity to S_R measure, as it uses the Spearman rank order correlation on the main branches of the SPT in order to evaluate the DP.

Normalization independent embedding quality assessment. Zhang et al. [119] presented a normalization independent embedding quality criterion, for manifold learning purposes (NIEQA). In the paper, they first developed a measure called the anisotropic scaling independent measure (ASIM), which compares the similarity between two configurations under rigid

motion and anisotropic coordinate scaling. NIEQA is based on ASIM, and consists of three assessments, a local one, a global one and a linear combination of the two. In our review we use the local one, so we merely focus on it. The local measure evaluates how well local neighborhood information is preserved under anisotropic coordinate scaling and rigid motion. That is, the local assessment is defined as:

$$NIEQA_{LOCAL}(X,Y) = \frac{1}{n}\sum_{i=1}^{n} M_{asim}(X_i, Y_i) \tag{27}$$

where $M_{asim}(Xi_k, Yi_k)$ is the ASIM value for index i. NIEQA is valued between 0 and 1, where 0 represents a perfect preservation. NIEQA has three characteristics to be highlighted: it can be applied to both normalized and isometric embeddings, it can provide both local and global assessments, and it can serve as a natural tool for model selection and evaluation tasks.

4. Related work

Different comparative studies amongst the different DR algorithms are currently being carried out as reported in the literature. In this section the most complete studies will be described, in chronological order.

Pölzlbauer [82] presented a comparative study in which he described some of the major SOM quality measuring methods. The aim was to test empirically how well the measures are suited for different map sizes. Finally, the author highlighted several advantages and disadvantages for each method. In the same year, Fukumizu et al. [28] proposed a novel DR kernel-based approach, KDR, for supervised learning problems. KDR provides data visualization capabilities, it can also identify and select important explanatory variables in regression and it can yield a better classification performance than the performance achieved with the full-dimensional covariate space.

Vinay et al. worked [112] on a comparison of the DR techniques for text retrieval. Basically, they compared four different DR techniques and assessed their performance in the context of text retrieval. They concluded that ICA (Independent component analysis) and PCA offered the best improvements. In the field of text clustering, Tang et al. presented in [101] a study of the comparison and the combination of DR techniques for efficient text clustering. Thus, they compared the performance of six DR algorithms when applied to text clustering. DR algorithms consisted of three DR-FE algorithms: ICA, Latent Semantic Indexing (LSI), Random Projection (RP); and three DR-FS algorithms based on Document Frequency (DF), mean TF-IDF (TI) and Term Frequency Variance (TfV). They observed that for DR-FE, the ranking (considering classification accuracy and stability) was: ICA > LSI > RP. However, in the case of DR-FS methods, DF was inferior compared to TI and TfV.

Chikhi et al. [16] carried out a comparative review of DR techniques for web structure mining. They used several DR algorithms (PCA, Non-negative Matrix Factorization – NMF, ICA and RP) in order to extract the implicit structures hidden in the web hyperlink connectivity. The conclusions were that NMF outperforms PCA and ICA in terms of stability and interpretability of the discovered structures. In the same year [81], Ohbuchi et al. experimentally compared six DR algorithms for their efficacy in the context of shape-based 3D model retrieval. They discovered that nonlinear manifold learning algorithms (KPCA, Locality Preserving Projections - LLP, LLE, LE, Isomap) performed better than the linear one (PCA). Specifically, LE and LLE algorithms produced significant gains in retrieval performance for different shape features. France and Carroll introduced in [27] a new metric (A_R) for evaluating the performance of DR techniques. Furthermore, they proposed three potential uses for the measure: comparing DR techniques, tuning parameters, and selecting solutions in techniques with local optima.

Lacoste-Julien et al. [55] presented a new method, DiscLDA, based on a variation of the LDA algorithm for DR and classification tasks. DiscLDA retains the ability of the LDA approach to find useful low-dimensional representations of documents, and also to make use of discriminative side information (labels) in forming these representations. Tsang et al. [105] focused on the attributes reduction with fuzzy rough sets. They developed an algorithm using a discernibility matrix to compute all the attribute reductions.

van der Maaten et al. [74] carried out one of the most extensive and complete comparative studies in the field of DR. They investigated the performances of the NLDR techniques in artificial and natural tasks. To do so, the authors carried out a comparison between several DR algorithms, by using the T&C quality criteria on artificial and natural datasets. They concluded that NLDR methods performed well in artificial tasks, but that this does not necessarily extend to real-world tasks. They also suggested how the performance of the NLDR techniques may be improved. Karbauskaite and Dzemyda tested the efficacy of several TP measures in [46]. Specifically, they used the K_M, MRRE and S_R criteria for estimating the TP of a manifold after unfolding it in a low-dimensional space. The authors pointed out that K_M and MRRE produced better results than S_R in all the cases. In the same year, Ji and Ye [43] studied the role of DR in multi-label classification. They proposed a new iterative algorithm and showed that when the least squares loss is used in classification, the joint learning decouples into two separate components.

Venna et al. [108] presented a new DR algorithm, Neighborhood Retrieval Visualizer (NeRV), as well as new measures of visualization quality (*mean smoothed precision* and *mean smoothed recall* methods). The performance of NeRV was compared with 11 unsupervised DR algorithms: PCA, MDS, LLE, LE, Hessian-based locally linear embedding (HLLE), Isomap, CCA, CDA, MVU, Landmark MVU (LMVU), and local MDS (LMDS). Two NeRV approaches were developed: one supervised and another unsupervised. To compare the methods, the authors used five pairs of quality measures: mean smoothed precision-mean smoothed recall, mean precision-mean recall curves, mean rank-based smoothed precision-mean rank-based smoothed recall, T&C criteria, and the classification error. The tests showed that NeRV outperformed existing DR methods. Lee and

Verleysen [66] suggested a way of summarizing the quality criteria that are based on ranks and neighborhoods into a single scalar value. This allows the user to compare DR methods in a straightforward way. Qian and Davidson [83] studied a novel joint learning framework which carries out optimization for DR and multi-label inference in semi-supervised setting. The experimental results validated the performance of their approach, and demonstrated the effectiveness of connecting DR and learning tasks.

With the aim of validating their new quality assessment criterion, Meng et al. [77] compared it to four quality criteria: T&C, MRRE, S_R and Q_k, when reducing the dimensionality through different DR algorithms. In particular, the authors used PCA, MDS, ICA, Isomap, LLE, LE, HLLE, Local Tangent Space Alignment (LTSA), MVU, Locally Linear Coordination (LLC), Neighborhood Preserving Embedding (NPE), and Linearity Preserving Projection (LPP) for the experiments. Handa [37] also analyzed the effect of DR through manifold learning for evolutionary learning. He proposed a method for reducing the difficulty in designing the allocation of sensors. To achieve this, he used Isomap and LLE for DR tasks and compared them by using R_E, MRRE and Q_k measures. In the same year, Lespinats and Aupetit [68] proposed the CheckViz method to evaluate the mapping quality at one single glance. Particularly, they defined a two-dimensional perceptually uniform color coding which allows tears and false neighborhoods to be visualized, the two elementary and complementary types of geometrical mapping distortions, straight onto the map at the location where they occur.

Recently, Zhang et al. developed a new quality assessment method for manifold learning tasks [119]. In the paper, they conducted an exhaustive comparison with other quality criteria (P_M, P_{MC}, R_V and Q_k) in order to test the efficacy of the new method. Empirical tests on synthetic and real data demonstrated the effectiveness of the proposed method. Chen and Lin [15] presented a novel approach, to Label Space DR (LSDR, is a paradigm for multi-label classification with many classes) that considers both the label and the feature parts. The approach is based on minimizing an upper bound of the popular Hamming loss. They demonstrated that their approach is more effective than existing ones to LSDR across many real-world datasets. Gan et al. [30] proposed a filter-dominating hybrid SFFS method, aiming at high efficiency and insignificant accuracy sacrifice for high-dimensional feature subset selection.

Very recently, Mokbel et al. [78] proposed a way of linking the evaluation to point-wise quality measures which can be used directly to augment the evaluated visualization and highlight erroneous regions. Furthermore, they improved 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. Finally, Musa [79] carried out a comparison of $\ell 1$-regularized logistic regression, PCA, KPCA and ICA for feature selection in classification tasks. To do so, he assessed the performance of these methods using different statistical measures, e.g.: accuracy, sensitivity, specificity, precision, the area under receiver operating characteristic curve and the receiver operating characteristic analysis.

5. Proposed methodology

The goal of this methodology is to compare different DR methods in terms of loss of quality. To achieve this, the loss of quality is quantified when reducing the dimensionality of the data over a pre-specified dimensional range. The loss of quality concept is defined as:

$$Quality\ Loss = (1 - quality\ value) \tag{28}$$

where 1 represents a perfect preservation of geometry, and the *quality value* is the value obtained by a particular quality measure. The domain for *quality value* is $[0,1]$, where 0 means the worst preservation of geometry and 1 is the best possible result (this is explained better in Section 5.2). The *loss of quality* is the achieved *quality value* subtracted from 1. Therefore, the smaller *loss of quality value*, the better preservation of geometry.

The *loss of quality* concept could be seen as a simple way of referring to the process of losing the original data geometry associated with a reduction in the data dimensionality, when using a DR algorithm. The rationale for using this concept is that we wanted the methodology presented here to emphasise the loss of quality that occurs in a DR process, rather than the value itself obtained by a quality measure.

The methodology is based on the following steps (Fig. 6): dimensional thresholding computation, quality loss quantifier curves (a.k.a. *QLQC*, explained below in Section 5.2) obtaining, increasing/decreasing stability function and quantification analysis of loss of quality.

In the first step, a dimensionality interval (by using the minor and major thresholds) is defined in order to quantify the loss of quality over the dimensionality reduction process. After this, the quality curves associated to each assessment measure are obtained. The increasing/decreasing stability function deals with the selection of those curves that meet a set of constraints. Finally, an analysis of the loss of quality on the selected curves is carried out.

5.1. Dimensional thresholding computation

In order to quantify the loss of quality in a DR process, it is necessary to define a major (N') and minor (n') dimensionality threshold. The minor threshold n' is considered a fixed value independent of the data and the DR algorithms. This value is usually the lowest possible dimensionality to be reduced (2 dimensions).

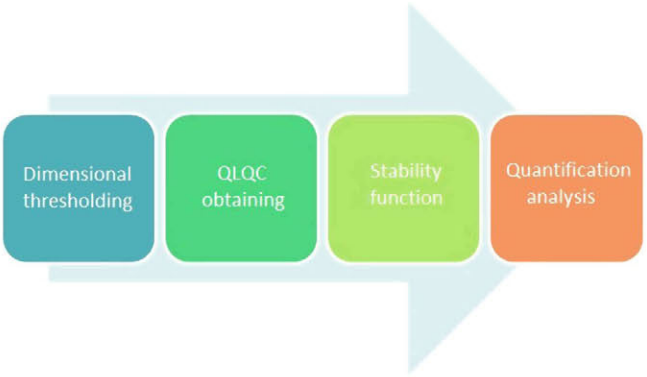

Fig. 6. Proposed methodology.

The major threshold N' is limited by the selected DR algorithms. Theoretically, the methodology presented here proposes that N' could correspond to the dimensionality value of the original dataset in order to carry out a more extensive study of the loss of quality, but in fact there are some cases in which this is not always possible due to technical issues. That is, there are several DR algorithms that do not allow us to select a target dimensionality greater than the number of individuals of the data analyzed in the study. There is a simple theoretical justification for this: according to linear algebra and vector spans, the intrinsic dimensionality of a given set of points can never be higher than the number of points. Therefore, some DR methods explicitly exclude the option to reduce the data dimensionality to any number larger than the number of points.

In this study, the major threshold N' is usually limited to the number of individuals (instances) of the data.

5.2. Quality Loss Quantifier Curves (QLQC) obtaining

In order to quantify the loss of quality when performing a DR task, 11 quality assessment measures have been selected from Section 3 (see Table 3). The selection criterion for these measures is closely related to the number of times they have been cited in the literature, particularly through studies with similar characteristics (see Section 4). This fact reinforces the importance of using them. So, for achieving real and significant values in the loss of quality estimation, the use of widely referenced methods in the literature was absolutely necessary.

As regards the inclusion of recently developed measures, such as Q_Y and $NIEQA_{LOCAL}$, they are considered as an interesting source of analysis. They provide a fresh approach, and have also demonstrated some desirable properties which the oldest ones lack.

The codification of the S_S, Q_M, M_T, M_C and Q_k measures were implemented by us. The P_M and P_{MC} methods were implemented thanks to the code kindly provided by the original authors (Goldberg and Ritov). The Co-ranking matrix code belongs to Lee and Verleysen. The Q_Y measure was implemented thanks to the code provided by the authors (Meng et al). Finally, to implement the NIEQA measure, the original code (Zhang et al.) was used.

Note that, all the *quality values* produced by a measure have been represented in the same range [0,1], where 0 is the worst value and 1 is the best possible result (perfect preservation of geometry). In the case of S_S, Q_M, P_M, P_{MC} and $NIEQA_{LOCAL}$ measures, these values were modified from the original measure ($1 - measurement$).

Our methodology computes a set of *QLQC* as the result of evaluating the loss of quality by using the 11 quality measures in all the range of dimensions from N' to n' (Fig. 7). The quality values provided by each measure can be considered as a single *QLQ* curve in which the X axis represents the range for dimensionalities where the data will be embedded, and Y axis the quality value of the measurement.

It is worth noting that, local measures such as $Q_{NX}, R_{NX}, Q_k, M_C, M_T$ and Q_M are usually evaluated on increasing k values. Therefore each measure yields a curve formed by the quality values obtained using different values of k. The methodology described here is not intended to carry out a study of the neighborhood, this is out of the scope of this paper. As a first approach, we were interested in studying the loss of quality over a wide interval of dimensions, by using a prefixed k value. Specifically, in the experiments we used $k = 7$.

The rationale for selecting this value is related to k parameter in some of the DR algorithms we have used. Isomap, LE, LLE and MVU also use a k local (also $k = 7$) parameter for evaluating the neighborhood before reducing the data dimensionality. Thus, we were looking for a high uniformity between different methods, in terms of parameter settings.

5.3. Increasing/decreasing stability function

One of the main challenges is related to selecting those curves of the plot that could be useful and provide valid information when quantifying results and drawing conclusions. That is, we select those curves in which the quality values are

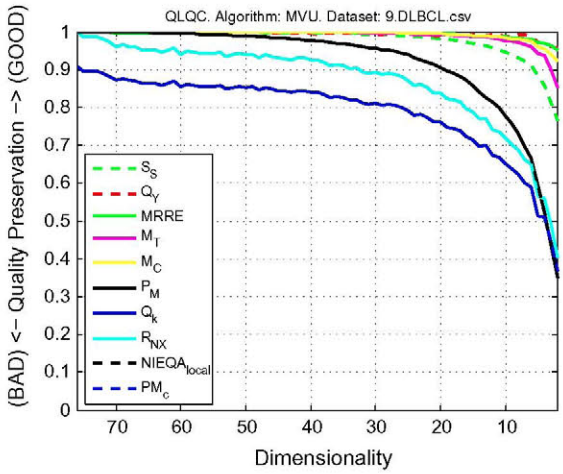

Fig. 7. Example of QLQC plot for a particular dataset, by using a DR algorithm (MVU).

gradual, stable and decrease (analogously, the loss of quality increases) during the DR process, just as we start from a N' dimensionality and progressively reduce it until n'. In other words, a DR algorithm is used on N'-dimensional data, first to yield an N'-1-dimensional embedding, then a second time to yield N'-2-dimensional data, and so on until the n'-dimension. The input data for the DR algorithm are always the original data. Thus, what we reduce is the target dimensionality in which the data will be embedded (from N'-1 to n').

The increasing/decreasing stability function ($S_{I/D}$) arises, firstly, in order to select those curves we consider suitable in order to study the loss of quality. After obtaining the *QLQC*, some of the curves showed a strange, unstable and erratic behavior. This behavior largely depends on the DR algorithm used. This irregular behavior makes the analysis of the loss of quality over a dimensionality interval difficult. By observing experimentally the behavior of many of these curves, we realized that a large proportion of them tended to decrease as the dimension decreased from N' to n'.

This fact should be considered a *natural* and *intuitive* concept, since for dimensionalities close to n' the quality values should be considerably smaller than for dimensionalities close to N'. Therefore, we decided that we should select those curves that showed this trend, since selecting other curves would make the process of extracting patterns or carrying out a clustering difficult, due to their unexpected and irregular behavior (see Fig. 8).

After this, we considered how to select those curves that meet this natural and intuitive constraint of progresive loss of quality. We wanted to use a statistical method or technique that exists in the literature. However, nowadays there is no statistical method in the literature that considers the concepts of the stability or growth of a curve. For this reason, it has been

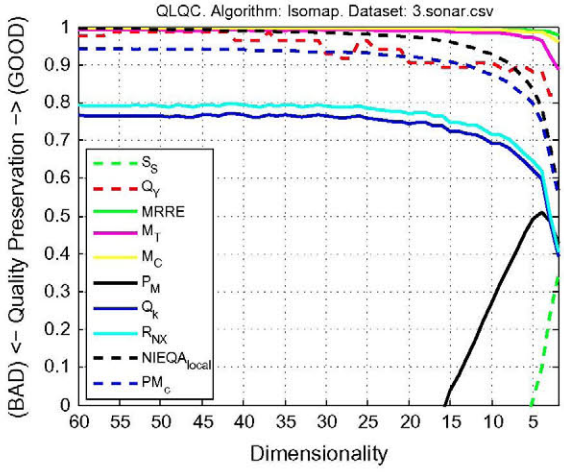

Fig. 8. QLQC containing curves that violate the increasing/decreasing stability criterion. The red and green dashed lines (that is, the quality curves generated by the Q_Y and S_S measures) and black line (P_M) violate the increasing/decreasing stability criterion. These curves do not reach the minimum threshold to be considered suitable to analyze. The blue and light blue lines (Q_k and R_{NX} measures) present low values of increasing/decreasing stability, and the rest present high values of increasing/decreasing stability since they are smooth and have a decreasing behavior. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

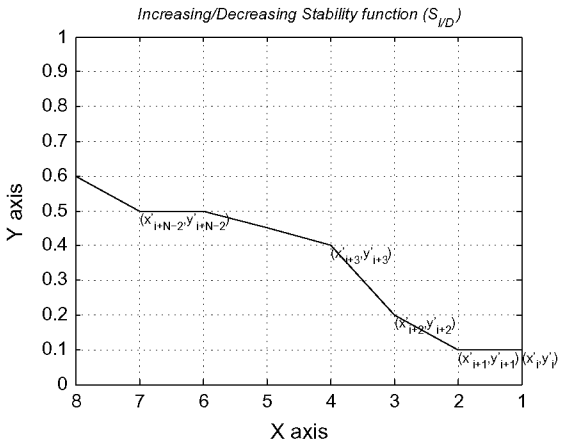

Fig. 9. Increasing/decreasing stability function.

designed and developed in order to carry out an analysis of the increasing/decreasing stability of a curve. This technique should be able to model the curve, that is, to provide us a value detailing the stability of the curve (without peaks). It should also provide information on its increasing/decreasing behavior. We have considered the selection of other techniques in the literature, statistical or otherwise, but as we did not find any method that achieves either of our aims or both of them, we decided to implement it.

So, the increasing/decreasing stability function ($S_{I/D}$) carries out an analysis of the behavior of the curve in terms of positive or negative growth in a stable way. Thus, $S_{I/D}$ represents how and to what extent a curve shows an increasing/decreasing behavior and, at the same time, how is the stability of the curve during the process. A curve can be considered stable by a full analysis of the oscillatory and fluctuating motion or, for that matter, by checking the existence of peaks in opposite directions. The bigger the oscillations, the smaller the $S_{I/D}$ final value.

Fig. 9 illustrates the behavior of the function. Let $\Delta x'_i = x'_{i+1} - x'_i$ and $\Delta y'_i = y'_{i+1} - y'_i$ be the increments in X and Y axes, respectively. The mean slope M_m is computed as in (29) and represents the mean of the slopes in the different sections of the curve. As M_m is not normalized, we do it in order to make future computations easier, through the following Eq. 24

$$M_m = \frac{1}{N-1}\sum_{i=1}^{N-1}\frac{\Delta y'_i}{\Delta x'_i} \tag{29}$$

$$M_{mn} = \frac{2\arctan(M_m)}{\pi} \tag{30}$$

where the interval of $M_{mn} \in [-1, 1]$, 1 represents a positive mean slope of 90°, 0 is a 0° mean slope, and -1 represents a negative mean slope of $-90°$ in respect to the X axis line. As we saw above, this value penalizes the 0 value slope sections of the curve. Finally, $S_{I/D}$ is computed as

$$S_{I/D} = \frac{\sum_{i=1}^{N-1} C_p}{\sum_{i=1}^{N-1}\sqrt{\Delta y'^2_i + \Delta x'^2_i}} \tag{31}$$

where the denominator calculates the total length of the curve, and the numerator calculates the partial contributions (C_p) in each section of the curve. Thus, C_p is computed by the following conditions

$$C_p = \begin{cases} \sqrt{\Delta y'^2_i + \Delta x'^2_i}, & \text{if } \Delta y'_i > 0, \\ -\sqrt{\Delta y'^2_i + \Delta x'^2_i}, & \text{if } \Delta y'_i < 0, \\ 0, & \text{if } \Delta y'_i = 0 \text{ and } M_{mn} = 0, \\ M_{mn}\sqrt{\Delta y'^2_i + \Delta x'^2_i}, & \text{if } \Delta y'_i = 0 \text{ and } M_{mn} \neq 0, \end{cases} \tag{32}$$

$S_{I/D} \in [-1, 1]$, where 1 represents a perfect increasing stability, 0 absence of increasing/decreasing stability, and -1 perfect decreasing stability. Basically, $S_{I/D}$ computes the total length of the curve and carries out an analysis of the contribution of each section of the curve to the total length, according to its positive, null or negative growth. In the case of 0 slope sections, the total mean slope is analyzed to penalize the value of $S_{I/D}$ in a way proportional to the curve, that is, according to the general trend of the curve.

It is important to discard those curves from the plot that show a high instability ($S_{I/D}$ values close to 0), or excessively low quality values. For selecting a proper threshold for $S_{I/D}$ values, the approach analyzes the boxplot of the absolute values of $S_{I/D}$ obtained for all the curves, and selects a particular minimum threshold. Thus, those curves in which the $S_{I/D}$ value is less than this threshold will be automatically discarded (see Section 6.1 to see the rationale for the selection of these values).

5.4. Quantification analysis of loss of quality

Once the stabilized curves have been selected, the methodology proposes a quantification analysis of them. These analyses could be from a simple analysis of the loss of quality in a certain interesting dimensionality to a more complex data analysis. In this way, three different kinds of analysis are proposed as a starting point:

Clustering of methods according to the loss of quality throughout the entire DR process. In order to detect similar behaviors when reducing the dimensionality of the data, in terms of loss of quality, a clustering process of the DR algorithms has been carried out.
Relationship between different preservation of geometry measures. The Pearson correlation indicates whether two different curves are linearly correlated or not. Nevertheless, it cannot detect differences in correlation when curves having the same proportion but different magnitudes. In this sense, for analyzing the similarity we should take into account both proportions and magnitudes (loss of quality values). Therefore, a new modified version of the Pearson correlation of two different curves i and j is proposed.

$$Corr_{ij} = |P_{ij}| - \left(1 - \frac{cv_i}{cv_j}\right) \tag{33}$$

where P_{ij} is the Pearson correlation between the curves i and j, and cv_i and cv_j are the variation coefficients (cv = standard deviation/mean) of curves i and j, respectively. Taking into account that the variation coefficient determines the possible variability in relation to the mean of the population [39], in this case this determines the possible stability of a curve in relation to the mean of its values.
Note that $cv_j > cv_i$ for all cases, so the denominator must always be the greatest of the two values. The equation part $(cv_i/cv_j) \in [0,1]$ represents how similar both curves are in terms of variability, 0 being the representation of different curves and 1 when the variability of both curves is the same. Thus, in this way, $(1 - cv_i/cv_j)$ penalizes the Pearson correlation when the proportions and magnitudes of the variabilities of both curves are different, even if these are correlated. The interval of $Corr_{ij} \in [-1,1]$, where 1 represents a perfect correlation and -1 indicates the absence of it. This equation evaluates the correlation between two distributions of data, considering the coefficient of variance of both distributions.
Loss of quality trend analysis from M into B dimension. Here the methodology represents the differences in loss of quality trend when the data is reduced from N' into M and from N' into B dimensions, B being lower than M. With this analysis we can conclude that any DR algorithm is stable or not (its trend is always the same) in the different dimensionality reductions in terms of loss of quality.

6. Experiments

In this paper, as an example of the use of the proposed methodology with real data, several experiments have been proposed. The implementation has been completely carried out in Matlab software. The environmental setting is made up of:

12 DR algorithms (2 linear, 9 nonlinear) (Table 2), also presented in Section 2. Four main packages have been used for encoding the different algorithms: The Matlab Toolbox for Dimensionality Reduction [71], Matlab package for Isomap (MIT) [102], Matlab package for MVU [115] and SOM Toolbox 2.0 [109]. As regards the input parameter settings of the methods, in most of the cases the default values (proposed by the authors) have been used. Generally, these default values have been previously verified empirically to be suitable for the different experiments (see Table 4).
In this methodology, we decided to use a parameter selection criterion based on default settings, as experimentally recommended by most of the authors of the DR algorithms. We only changed the k value of the DR algorithms, in order to make them coincide with the number of nearest neighbors in the quality measures. The rest, such as *perplexity* (in t-SNE); *epochs* (in CCA); t and σ (in DM); and some kernel parameters were set as default.
We must clarify that, in this first approach, the aim of the methodology is not to experiment with these parameters, but to provide a methodology able to produce reliable results. We were interested in analyzing the results derived from a default configuration of the DR algorithms. However, we must leave open the possibility of experimenting with different initial configurations.
12 real-world datasets, where eight of them have been selected from the UCI Machine Learning Repository (Table 5). As regards their nature, 3 of the selected datasets are exclusively of DNA microarray origin (Leukemia, DLBCL and SRBCT's), 5 of them belong to other medical nature (Breast Cancer Wisconsin, SPECTF Heart, Prostate, Parkinsons and neurons), and other fields (Connectionist Bench, Glass Identification and Libras Movement).

Table 4
DR algorithms and parameter settings for the experiments.

Method	Package	Parameter settings	Reference
PCA	The Matlab Toolbox for Dimensionality Reduction (2012)	None (default)	[71]
LDA	The Matlab Toolbox for Dimensionality Reduction	None (default)	[71]
Isomap	Matlab package for Isomap (MIT, 2000)	$K = 7$	[102]
$KPCA_{gaussian}$	The Matlab Toolbox for Dimensionality Reduction	$\kappa(x_i, x_j) = e^{\frac{-\lvert x_i - x_j \rvert^2}{\sigma^2}}$	[71]
$KPCA_{polynomial}$	The Matlab Toolbox for Dimensionality Reduction	$\kappa(x_i, x_j) = (x_i \cdot x_j)^2$	[71]
LE	The Matlab Toolbox for Dimensionality Reduction	$K = 7$, $\sigma = 1.0$ (default)	[71]
LLE	The Matlab Toolbox for Dimensionality Reduction	$K = 7$	[71]
DM	The Matlab Toolbox for Dimensionality Reduction	$t = 1.0$ (default), $\sigma = 1.0$ (default)	[71]
t-SNE	The Matlab Toolbox for Dimensionality Reduction	*perplexity* = 30 (default)	[71]
SM	The Matlab Toolbox for Dimensionality Reduction	None (default)	[71]
MVU	Matlab package for MVU (2012)	$K = 7$	[115]
CCA	SOM Toolbox 2.0 (2005)	*epochs* = 10 (default)	[109]

Table 5
Real-world datasets used in the experiments.

Dataset	Instances	Features	Reference	Intrinsic dimensionality (d)
Breast Cancer Wisconsin (Diagnostic, 1995)	569	30	[75]	6
Connectionist Bench (Sonar, Mines versus Rocks, 1988)	208	60	[35]	9
SPECTF Heart (2001)	267	44	[54]	11
Breast Cancer Wisconsin (Prognostic, 1995)	198	33	[99,75,118]	5
Prostate (2008)	380	9	[92]	6
Glass Identification (1988)	107	9	[26]	5
Parkinsons (2007)	195	22	[70,69]	3
Leukemia (1999)	72	5147	[33]	18
Diffuse large B-cell lymphomas (DLBCL, 2002)	77	7070	[95]	15
Gardener Classificator (neurons, 2013)	241	368	[19]	12
Small Round Blue Cell Tumors (SRBCT's, 2001)	83	2308	[47]	10
Libras Movement (2009)	330	90	[21]	6

Note that, in order to obtain the intrinsic dimensionality for each dataset, Maximum likelihood (MLE) and Eigenvalue-based estimators [73] were used, by calculating the integer mean value of both estimators.

6.1. Applying the methodology

The first step in the methodology is the dimensional thresholding calculation. The minor threshold n' of all the experiments has been fixed at 2, that is the lowest dimensionality possible. On the other hand, the major thresholds N' have been calculated depending on the original number of dimensions and instances of the data considered (as stated in Section 5.1). So, the major thresholds N' are: 30 in the Breast Cancer Wisconsin Diagnostic, 60 in the Connectionist Bench, 44 in SPECTF Heart, 33 in the Breast Cancer Wisconsin Prognostic, 9 in Prostate, 9 in Glass Identification, 22 in Parkinsons, 72 in Leukemia, 77 in DLBCL, 100 in neurons, 83 in SRBCTs and 90 in Libras. Note that, for DNA microarray data (Leukemia, DLBCL and SRBCTs), N' is constrained to the number of individuals of each dataset due to the technical limitations of the DR algorithms. In the neurons dataset, N' is set to 100 since it has been observed that greater values do not give rise to loss of qualities, thus these cases are of no interest to the study. For the rest of the datasets, the N' value is the original dimensionality of the data.

In order to obtain the *QLQC* plots, all the curves must be calculated. Based on the 12 DR algorithms and the 12 datasets, the method calculates $11_{measures} \times 12_{algorithms} \times 12_{datasets} = 1,584$ curves for studying the loss of quality resulting from a DR process.

For each curve, the $S_{l/D}$ value is calculated. In order to select the sufficiently stable curves, a minimum threshold is necessary. To select this threshold, a boxplot of the absolute values of $S_{l/D}$ obtained throughout the 1,584 curves was carried out. When analyzing the distribution of the boxplot, it could make sense to discard those curves whose stability value is less than the second quartil of the boxplot, that is 0.3005.

Rationale for the $S_{l/D}$ minimum threshold The main values obtained when representing by using the boxplot technique were: *quartile 1* (0.08, $Q1$), *quartile 2* (0.3005, median or $Q2$) and *quartile 3* (0.8, $Q3$). At first glance, selecting a threshold value from which a curve meets the decreasing stability constraints was not easy, thus an empirical study of the behavior of the curves using $Q1, Q2$ and $Q3$ was carried out. To this end, we selected the curves with $S_{l/D}$ values equal to or greater than the selected quartile, and they were plotted using the 2D scatterplot technique. When using $Q1$ (0.08) as the threshold, we realized that almost all the selected curves behaved in a highly unstable and erratic manner and they did not meet the decreasing stability constraint. Therefore, $Q1$ was discounted and $Q2$ was studied. When $Q2$ (0.3005) was used as a candidate threshold, almost the 100% of the curves exceeding this threshold showed a strong decreasing stability behavior. Finally,

Q2 was selected as the threshold for $S_{I/D}$ values, and roughly one third (474) of the total curves (1584) were selected for further analysis. It is important to note that: firstly, when testing Q3 as the threshold value, many fewer curves were selected, so we decided to work with Q2. Secondly, we decided to select for the $S_{I/D}$ threshold, the first value (and it should be statistically justified) that allowed us to achieve curves that meet the decreasing stability constraints, and that is why we did not select an intermediate value between quartiles.

Moreover, the curves whose quality values obtained in 2, 3 and intrinsic dimension were outside a specific interval were also discounted.

Rationale for the selection of the quality value interval Solely based on normalization criteria, we were interested in selecting those curves whose quality values in 2, 3 and intrinsic dimensions were in the interval $[0, 1]$, and discarding the rest. By definition, all the quality measures, except S_S sometimes, provide quality values within that range. In its original definition [86], $S_S \in [0, \infty)$. However, as we normalized all the measures so that 1 is the best quality value (see Section 5.2), the new range for this measure was $S_S \in [1, -\infty)$ (where $-\infty$ is the worst value) and therefore there were still a few curves with quality values of less than 0 (even after filtering by the $S_{I/D}$ threshold). As has already been said, we wanted to study quality values in the interval $[0, 1]$, therefore this constraint discounted the rest of the curves that did not meet this condition. After discounting the curves outside the $[0, 1]$ interval, we also realized that all the curves, in fact, curiously presented quality values greater than 0.198.

Thus, using two ways of filtering the curves, we ensured that we were selecting decreasing and stable curves (as the target dimensionality is reduced from N' to n'), as well as quality values in the $[0, 1]$ interval (see Fig. 10). The aim is to be able to quantify the loss of quality in a DR process.

It is worth mentioning that, after applying the two constraints (quality and stability) imposed on the selected curves, no DR algorithm fails uniformly. That is, to a greater or lesser degree, all the DR algorithms yield *QTQC* enough to accurately quantify the loss of quality through these curves. Specifically, the distribution of the selected curves for each of the DR algorithms is as follows: PCA (67 curves), MVU (66), $KPCA_{poly}$ (57), Isomap (57), LE (47), $KPCA_{gauss}$ (42), SM (37), LDA (26), DM (21),

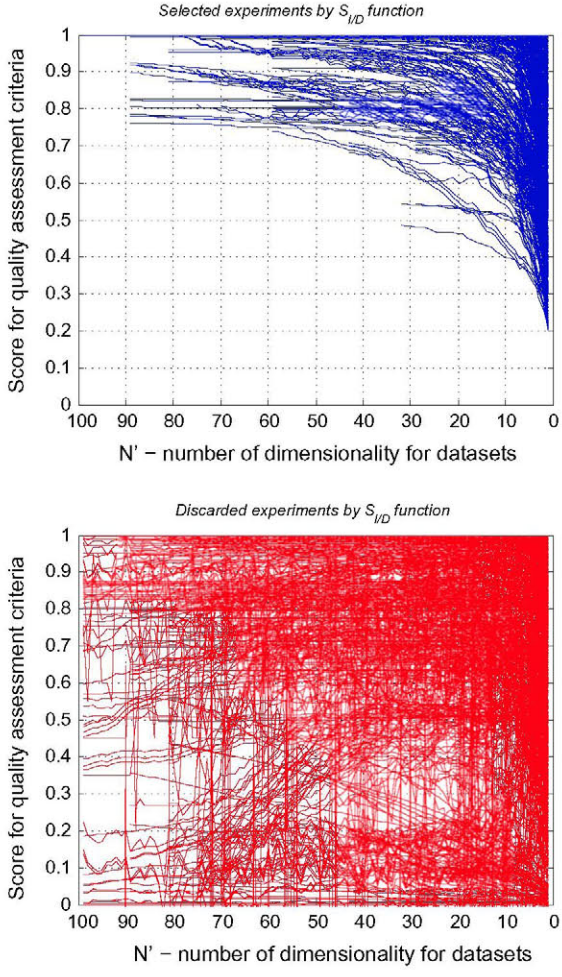

Fig. 10. Selected experiments (top) versus discarded experiments (bottom).

t-SNE (21), CCA (18) and LLE (15). From this distribution we conclude that, from a point of view based on stability and quality criteria, the DR algorithms that produce more suitable curves for studying the loss of quality are PCA, MVU, $KPCA_{poly}$ and Isomap, whilst LLE and CCA performed the worst.

6.1.1. The relationship between different preservation of geometry measures

One of the quantification analyses made using this methodology is the relationship between the quality criteria during the DR process, when using the different DR algorithms.

So, firstly the proposed correlation (Section 5.4, Eq. 33) for each pair of measures in the different datasets was calculated. After analyzing all the values through a boxplot, it was decided to analyze only those pairs of measures whose correlation was greater than the third quartile (0.612), in order to see the possible real relationships between the measures.

Fig. 11A represents those relationships between pairs of measures that are greater than 0.612 as opposed to the total number of relationships in all the datasets (as a percentage). So, for example, P_{MC} versus $NIEQA_{local}$ has a correlation greater than 0.612 in 14.2% of cases in all the datasets, while P_{MC} versus S_S only has a correlation greater than the threshold in 2.059% of the cases.

After that, it is also interesting to analyze the correlation values of these pairs of measures when they are greater than 0.612. So, Fig. 11B presents the statistic values (mean, median and standard deviation) calculated from the correlations. For example, the P_{MC} versus $NIEQA_{local}$ correlation mean is 0.821 with a median of 0.829 and a std. deviation of 0.107.

Several conclusions can be extracted from the previous figure. Firstly, the pairs of measures which are correlated the greatest number of times (presented on the left-hand side of the figure) are those that have the highest values in correlation (mean values greater than 0.78). However, when the pairs are correlated fewer times (right-hand side of the figure), the mean values decrease. This makes sense because if a pair of measures are really correlated, this event will be repeated several times with a high value, although the nature of the data has changed.

It is worth highlighting the strong correlation between measures with a similar nature, such as Q_{NX}, R_{NX} and Qk, since all of them are based on the ranking of the nearest neighboring concepts. Furthermore, P_M, P_{MC}, and $NIEQA_{LOCAL}$ are closely correlated to each other because they work using the procrustes analysis methods. It is also observed that there is a high correlation between these two groups. Although they work in different ways, both were originally devised to assess the local TP after a DR process.

There is also another group of measures that present a high correlation between themselves but only in a few of cases (presented in Table 6). Within this group is, for example, S_S and $NIEQA_{LOCAL}$, where the first one evaluates the global preservation, whilst the other one calculates the local preservation. The same happens with S_S and P_{MC}.

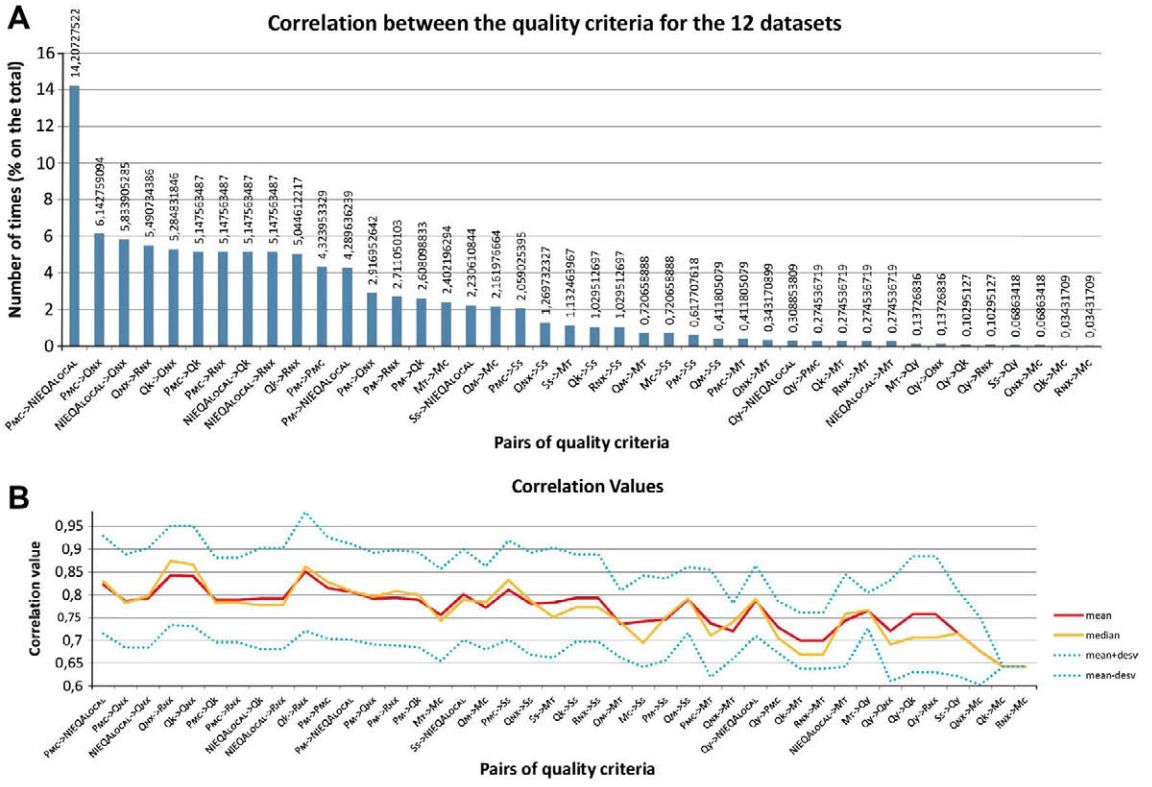

Fig. 11. (A) Correlations between pairs of quality measures in all datasets greater than 0.612. (B) Statistical values of correlation for each pair of measures.

Table 6
High correlated pairs of measures for all datasets.

Pair of measures	Times correlated (%)	Mean	Median	Mean + desv	Mean-desv
$P_M; Q_{NX}$	2.916	0.791	0.796	0.892	0.691
$P_M; R_{NX}$	2.711	0.793	0.807	0.898	0.688
P_M;Qk	2.608	0.788	0.800	0.892	0.684
$M_T; M_C$	2.402	0.755	0.742	0.857	0.654
$S_S; NIEQA_{LOCAL}$	2.230	0.800	0.789	0.900	0.701
$Q_M; M_C$	2.161	0.771	0.783	0.863	0.680
$P_{MC}; S_S$	2.059	0.810	0.832	0.919	0.702
...;...	...	...	...	...	...

Finally, the $Q_k - M_C$ and $R_{NX} - M_C$ correlations have the same value and both without deviation. This is because the three measures come from the same idea of preservation of geometry and there is only one case of study that has the correlation greater than 0.612.

At this point, it could be interesting to analyze the correlation between the measures for each dataset. This may help contrast the previous results and also give details of relationship between measures depending on the nature of the data.

To do this, firstly 22 figures (two per dataset), included in Supplementary Material, were obtained. Only in one of the dataset were there no figures, as there were no stable curves with correlations outperforming the threshold. Like the previous results (the averaged way with all the datasets), there is a very strong correlation between the following groups of measures: the first group consists of 3 measures, Q_{NX}, R_{NX}, and Qk. It is noted that, for absolutely all the datasets, the same pattern is repeated and these 3 measures are highly correlated. Furthermore, the second group showing a very high degree of correlation is made up of P_M, P_{MC} and $NIEQA_{LOCAL}$. This is, by far, the most correlated group of all datasets, and it is also strongly correlated with the first group. For each one of the datasets, there is always a large number of correlation cases between the members of the first group, the members of the second group, and between these two groups. This coincides entirely with the conclusions drawn in the previous section for all the datasets.

High correlations are often reported between S_S and P_{MC} and $NIEQA_{LOCAL}$ measures. Furthermore, M_T, M_C and Q_M also present a large number of correlation cases between themselves. However, the Q_Y criterion lacks any direct correlation with other criteria (it does not appear in the figures, or with a low degree) because, as we pointed out earlier, of its peculiar nature.

The strong correlation between these measures is confirmed when their mean correlation values are observed (see the Supplementary Material). Note that, correlation values in the [0.75, 1.0] range can be considered as very high, since the original Pearson correlation function was modified in order to be stricter.

To sum up, the conclusions presented by separating per dataset confirm the high degree of correlation between the different groups of quality measures.

6.1.2. Comparative study and clustering of DR methods

The analysis of the loss of quality during the DR process from N'D to 2D obtained by each DR algorithm are presented here. The aim is to compare these results, in order to highlight the 'quality preservation' skills of the DR algorithms. Then, we show which type of DR algorithms usually carry out DR tasks while producing minimum losses of quality. To achieve this, the quality values obtained by the different DR algorithms are compared in a set of key dimensions, as 2D, 3D, ID (intrinsic dimensionality, d) and N'D.

The Mann–Whitney Wilcoxon signed-rank test is a non-parametric statistical hypothesis test used when comparing two samples. This can matematically demonstrate whether two samples came from the same population, or if the distribution of one sample is stochastically greater than the other.

In this study, the Wilcox test was used to compare each pair of DR algorithms (if one algorithm is better than other, in terms of loss of quality), based on its mean loss of quality values for each DR algorithm. A p-value (probability) less than or equal than 0.05 affirms the asumption of improvement from one over the other algorithm. Fig. 12 shows the mean loss of quality values achieved by the different algorithms.

First of all, note that, in fact, the real input data for the statistical test consist of all the values that produce these mean values. It can be clearly seen that, for dimensionalities close to N', the loss of quality produced by the DR algorithms is significantly less than for dimensionalities close to 2. The Wilcox statistical test provides us information as to which DR algorithm produces a lower loss of quality, as regards other algorithms. Thus, the Wilcox test is carried out on all the different possible pairs of algorithms. Fig. 13 shows the $p - values$ for each pair.

If the number of times a DR algorithm presents lower loss of quality values than the rest is counted, a preliminary classification of the DR algorithms is obtained, as regards the loss of quality produced (Fig. 13).

A greater number means that a DR algorithm produces fewer loss of quality values than other DR algorithms more often. This is always positive. The worst results are obtained by $KPCA_{gauss}$ and LDA (0), that is, they generate the greatest loss of quality values, so they never outperform the remaining algorithms. In the case of LDA, this fact can be explained as follows: LDA is characterized by reducing the data dimensionality for improving the classification accuracy, therefore this impacts negatively on the loss of quality, i.e. an improvement in classification tasks affects the efficiency when preserving the original

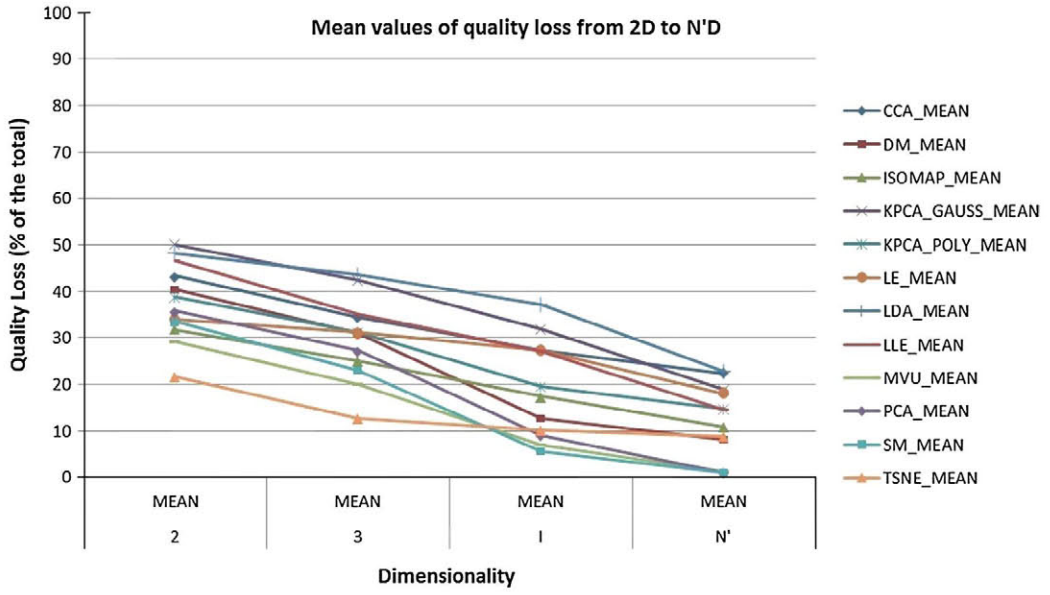

Fig. 12. Mean values of loss of quality from N'D to 2D, for each DR algorithm. A set of key dimensions, as 2D, 3D, ID and N'D have been selected for the study.

	CCA	DM	ISOMAP	KPCA_GAUSS	KPCA_POLY	LE	LDA	LLE	MVU	PCA	SM	t-SNE
CCA	1,000	0,813	0,993	0,018	0,830	0,724	0,158	0,631	0,999	0,927	0,985	1,000
DM	0,192	1,000	0,900	0,011	0,439	0,458	0,047	0,271	0,926	0,697	0,882	0,997
ISOMAP	0,008	0,103	1,000	0,000	0,076	0,039	0,002	0,035	0,655	0,264	0,497	0,950
KPCA_GAUSS	0,983	0,989	1,000	1,000	0,991	0,994	0,655	0,948	1,000	0,994	1,000	1,000
KPCA_POLY	0,174	0,568	0,926	0,009	1,000	0,484	0,039	0,304	0,947	0,725	0,911	0,998
LE	0,282	0,549	0,962	0,006	0,523	1,000	0,047	0,287	0,970	0,736	0,926	0,999
LDA	0,846	0,955	0,998	0,351	0,962	0,955	1,000	0,969	1,000	0,993	0,999	1,000
LLE	0,376	0,734	0,966	0,053	0,701	0,718	0,032	1,000	0,977	0,814	0,968	0,999
MVU	0,002	0,076	0,351	0,000	0,055	0,031	0,000	0,024	1,000	0,157	0,445	0,888
PCA	0,076	0,309	0,741	0,006	0,280	0,269	0,007	0,190	0,847	1,000	0,827	0,988
SM	0,015	0,122	0,510	0,000	0,092	0,076	0,001	0,034	0,561	0,177	1,000	0,882
t-SNE	0,000	0,003	0,052	0,000	0,002	0,001	0,000	0,001	0,115	0,013	0,122	1,000

Fig. 13. Results of the Wilcox statistical test, comparing each pair of DR algorithms. The p-values are shown. Green means that, a particular DR algorithm produces a lower loss of quality than another algorithm. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

geometry of the data. In addition, LDA is known to have a linear nature and behavior, thus confirming the difficulty in preserving the quality when using linear approaches. LLE and CCA also indicate high loss of quality values (see [32] for LLE), compared to the rest of the algorithms (in addition to being very unstable, as mentioned at the end of Section 6.1).

However, the best results are achieved by t-SNE, MVU and ISOMAP. These sophisticated algorithms base their nature on data embedding by: computation of the conditional probability distributions that represent similarities in both dimensional spaces (t-SNE), preservation of the distances between the k-nearest neighbors by means of a neighborhood graph G (ISOMAP and MVU, the former uses geodesic distances and the latter euclidean distances).

Finally, a density-based clustering algorithm is performed, the Farthest First algorithm [40]. The aim is to detect different groups of curves in Fig. 12 in order to highlight common behaviors for the DR algorithms, during a DR process. By studying the results of the mean loss of quality values for each DR algorithm, the behavior of the curves can be described, as can the clusters. The results of the clustering algorithm are shown in Fig. 14.

Three different clusters and one outlier is observed. The detected outlier is the t-SNE algorithm (in orange). This clearly coincides with the results of the Wilcox test (ranked as the best algorithm) as, in addition to obtaining the lowest loss of quality values, the shape of the curve during the DR process is quite different from the remaining algorithms. The greatest leap in loss of quality is produced in 2D in respect to that produced in 3D, since from 3D to N'D the loss is lower. However, it is quickly realized that, in the 3 clusters, there is a strong similarity between the curves inside a particular cluster. Curves grouped in the same cluster show very similar transitions during the loss of quality process (from N'D to 2D).

The first cluster, which groups DR algorithms that give rise to the lowest loss of quality, is made up of by PCA, SM and MVU algorithms. All of them show moderate loss of qualities from N'D to ID. However, from there to 2D there is a huge leap in loss of quality. The results obtained by SM and MVU coincide with the Wilcox test, since both algorithms give rise to mild loss of qualities with respect to the rest and occupy a front-ranked position.

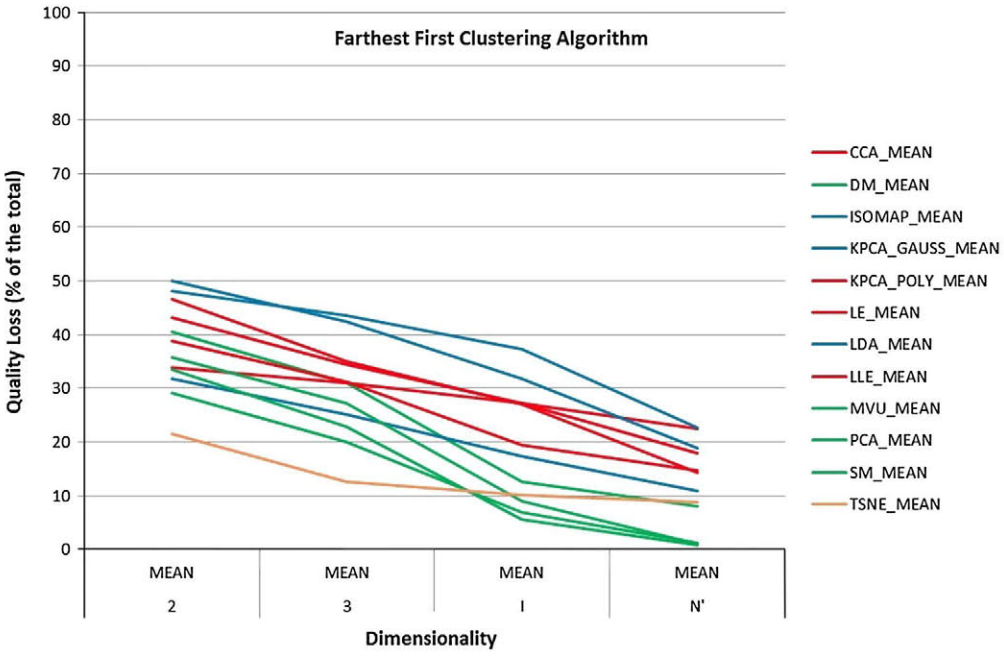

Fig. 14. Farthest First clustering algorithm. Green, blue and red represent the three clusters, while the orange indicates the outlier. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

The second cluster groups most of the algorithms together, it is made up of LLE, LE, CCA and $KPCA_{poly}$. The behavior of this group is slightly different as, unlike other clusters, there is a clear linearity in the loss of quality.

The last cluster groups the LDA, $KPCA_{gauss}$ and ISOMAP algorithms together. This is the group that presents higher loss of qualities. The following observations can be made: the first is that the results of LDA and $KPCA_{gauss}$ coincide with the Wilcox test, since they appear in the lowest positions in the ranking. The second one is that the bad results of Isomap in the clustering algorithm, does not coincide at all with the Wilcox test. This can be explained by the fact that the input data for the Wilcox test and the clustering algorithm are different, in value and quantity. Thereby, the results are different because of variability in the data. The clusters have been defined according to the average of the data. Thus, any curve varying in [−3*standard.dev, 3*standard.dev] range could be grouped into different clusters.

6.1.3. Loss of quality trend analysis from 3 into 2 dimensions

Like the first analysis made using this methodology, a comparative study based on the loss of quality produced when reducing multivariate data to two and three dimensions was proposed. This particular case has been separated from the present paper, since a more detailed study was necessary.

This study demonstrates that, only when switching from 3D to 2D, does it reach maximum values of 48.62% and mean values of 30.483% of the total loss of quality for many case studies with the presented datasets. These values can be considered noticeably high and suggest the suitability of the third dimension for reducing and visualizing data. Furthermore, the theoretical results have been reinforced by a set of visual tests carried out by a group of 40 individuals. The aim of these tests was to test whether, by using the visualization, the theoretical results obtained by the methodology coincide with the expertise and perception of the users when they work with 2D and 3D spaces.

6.2. Computation times

Table 7 describes the computation times per dataset (in hours), as well as the % of the CPU time used by each quality measure, as regards the total CPU time (in hours).

A single column has been used for each of the most computationally high measures, while the rest of the measures have been grouped together in the *Rest of measures* column due to their insignificant computation time as regards the total time.

The quality measures that have required the longest computing times have been $NIEQA_{LOCAL}$, and Q_Y. This makes sense, as $NIEQA_{LOCAL}$ carries out a Procrustes analysis, and this is computationally very demanding. The motivation behind $NIEQA_{LOCAL}$ is geometric matching, that is, assessing how similar two sets of observations are under rigid motion and coordinate scaling (by using operations with matrices). If they match each other well, then $NIEQA_{LOCAL}$ converges quickly. However, if they do not match, the convergence would be slow, since the iteration process gets stuck in finding an optimal transformation which will never match them well. For Q_Y, the bottlecneck is in the shortest path tree constructed from the k neighborhood graph.

Table 7
Computation times (in hours) per dataset. Columns $P_M, P_{MC}, NIEQA_{LOCAL}, Q_Y$, *Rest of measures* and *DR methods* show the % of the CPU time used, as regards the *Total CPU time (hours)*. The largest values are printed in bold.

Dataset	Instances	Features	%of the total CPU time (by quality measures and DR methods)						Total CPU time (h)
			P_M	P_{MC}	$NIEQA_{LOCAL}$	Q_Y	Rest of measures	DR methods	
1. Breast Cancer (Diagnostic)	569	30	0.1	0.09	**55.82**	**43.94**	0.03	0.02	41.3
2. Connectionist Bench	208	60	0.124	0.124	**96.96**	**2.78**	0.01	0.002	**79.79**
3. SPECTF	267	44	0.15	0.14	**91.48**	**8.13**	0.08	0.02	43.06
4. Breast Cancer (Prognostic)	198	33	0.17	0.18	**91.23**	**8.4**	0.01	0.01	10.52
5. Prostate	380	9	0.36	0.36	**16.39**	**82.75**	0.13	0.01	1.66
6. Glass Identification	107	9	0.83	0.83	**65.21**	**33.11**	0.002	0.018	0.10
7. Parkinsons	195	22	0.32	0.34	**84.22**	**15.04**	0.05	0.03	3.84
8. Leukemia	72	5147	0.19	0.21	**74.25**	**25.32**	0.02	0.01	**85.03**
9. DLBCL	77	7070	0.15	0.17	**82.98**	**16.68**	0.01	0.01	**102.22**
10. Neurons	241	368	0.2	0.19	**75.44**	**24.14**	0.025	0.005	**123.01**
11. SRBCT's	83	2308	0.17	0.17	**89.08**	**10.57**	0.007	0.003	**95.26**
12. Libras Movement	330	90	0.15	0.15	**84.98**	**14.7**	0.015	0.005	**107.61**
									Total = 693.41

Table 8
Computation times (in seconds) per dataset and DR method. The largest values are printed in bold.

Dataset	Total CPU time (by DR methods)											Total CPU time (s)
	PCA	LDA	Isomap	$KPCA_{gauss}$	$KPCA_{poly}$	LE	LLE	DM	*t*-SNE	SM	MVU	CCA	
1. Breast Cancer (Diagnostic)	0.02	0.16	71.36	3.74	19.44	2.78	6.42	6.27	**681.95**	**903.97**	**1225.12**	49.55	2970.78
2. Connectionist Bench	0.004	0.03	11.96	0.58	3.53	0.46	1.16	1.03	**123.66**	**164.06**	**222.48**	9.18	538.134
3. SPECTF	0.03	0.17	74.59	4.9	20.27	2.3	6.69	6.54	**719.8**	**942.27**	**1271**	54.7	3103.26
4. Breast Cancer (Prognostic)	0.004	0.02	10.07	0.47	2.37	0.34	0.81	0.89	**86.56**	**114.91**	**165.63**	6.42	388.494
5. Prostate	0	0.003	1.35	0.07	0.37	0.04	0.12	0.12	**13.19**	**16.51**	**23.73**	0.77	56.273
6. Glass Identification	3.85E−05	0	0.08	0.003	0.02	0.003	0.006	0.007	**0.82**	**1.19**	**1.48**	0.26	3.869
7. Parkinsons	0.004	0.023	9.93	0.52	2.7	0.37	0.89	0.87	**94.8**	**125.85**	**170.56**	7.03	413.547
8. Leukemia	0.03	0.17	72.44	3.75	20.01	2.76	6.3	6.45	**701.74**	**933.24**	**1260.72**	51.02	3058.63
9. DLBCL	0.03	0.2	87.3	4.63	24.06	3.32	7.94	7.76	**842.53**	**1118.47**	**1515.83**	62.54	3674.61
10. Neurons	0.02	0.12	53.23	2.78	14.67	2	4.78	4.67	**502**	**673**	**912.11**	37.61	2206.99
11. SRBCT's	0.01	0.02	24.32	1.29	6.71	0.97	2.21	2.06	**234.95**	**312.9**	**422.71**	17.24	1025.39
12. Libras Movement	0.02	0.11	46.41	2.44	12.36	1.75	4.08	4.08	**444.52**	**584.78**	**797.96**	32.72	1931.23
Total CPU time (s)	0.17	1.02	463.04	25.17	126.51	17.093	41.406	40.74	**4446.52**	**5891.15**	**7989.33**	329.04	**Total = 19371.2**

The estimated total time to complete all the experiments has been approximately 693 h, equivalent to 28.8 days of sequential running. For the experiments, 5 desktop computers were used. Therefore, each computer took roughly 5.76 days to complete its tasks. The features of each computer were as follows: Intel i5 2.8 Ghz, 8 GB RAM.

Table 8 also shows the computation time of the DR process for each of the DR algorithms. Although the computation times for the DR methods are very small compared to the total CPU time for the experiments, two different facts can be highlighted: PCA, LDA, LE and $KPCA_{gaussian}$ methods are computationally very light; on the other hand, SM, *t*-SNE and MVU methods are, by far, computationally very demanding.

Finally, Fig. 15 shows a relationship between the number of features for each dataset versus the CPU time taken to complete the experiments.

As can be seen in the figure, the computation time for each dataset behaves approximately linearly with respect to the number of attributes that have been considered for each dataset. This clearly indicates that the bottleneck for quality measures could be strongly related to the number of attributes, as well as the number of instances to be evaluated.

7. Discussion

This paper proposes a new methodology that allows the analysis and comparison of different DR methods as regards the loss of quality they give rise to when carrying out a DR process. As a step prior to the presentation of the methodology, an exhaustive, chronological and comparative review of the quality assessment criteria in DR for measuring the loss of quality is also provided. By using this methodology, it is possible to analyse the curve generated by the loss of quality produced when

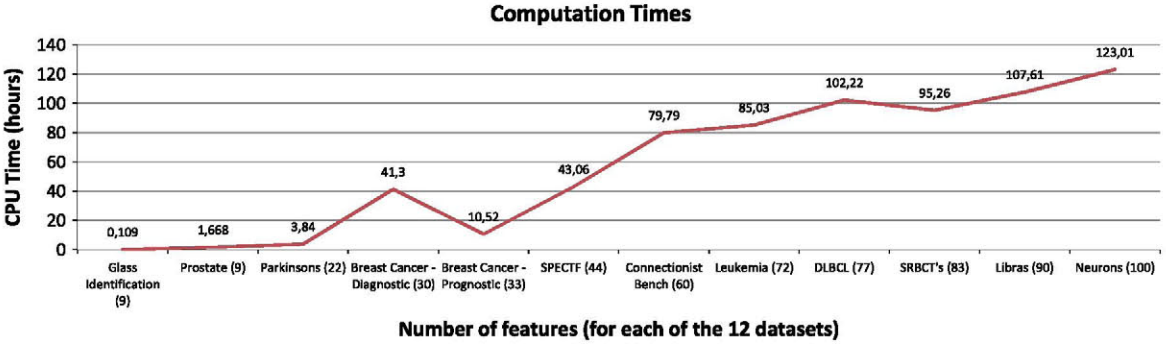

Fig. 15. Number of features versus CPU time.

reducing the dimensionality of a dataset from its original space, to a lower space. This also gives rise a study focusing on interesting dimensionalities, such as 2, 3, or intrinsic dimensionality. Particularly, for scatterplot visualization techniques (usually using 2 and 3 dimensions), it could be very useful to know the behavior, in terms of loss of quality, of a DR algorithm on a dataset. High loss of quality values could indicate the suitability of using another DR algorithm or dimensionality space for embedding the data. Other studies in the literature quantify, in a very superficial way, some loss of quality values for a particular case. However, the lack of a methodology for analyzing the entire loss of quality process in DR tasks has brought about this study.

It is also worth highlighting that all the selected DR algorithms for our study are unsupervised except LDA (supervised). The reason for including both supervised and unsupervised DR algorithms is simple. On the one hand, the aim was to study a wide range of quality assessment measures after a DR process. This included the large majority that are unsupervised, but we also considered it necessary to mention C_e, that is a supervised quality indicator. On the other hand, the rationale for including a supervised DR algorithm such as LDA in our methodology is as follows: to be able to demonstrate by using quality indicators that LDA was originally devised to reduce the dimensionality of the data in order to improve the classification accuracy. This fact, indeed, impoverished the results in terms of quality preservation (as confirmed in Section 6.1.2).

In order to test our methodology, three different kinds of analysis are proposed. The first one is a new way of classifying the current DR algorithms, according to their natural preservation of geometry skills on real-world datasets. *t*-SNE, MVU and Isomap algorithms have been demonstrated in these cases to preserve the original quality contained in the data, in a more effective way than the remaining algorithms. However, $KPCA_{gauss}$ and LDA performed worst in the experiments. To select the experiments for studying properly the loss of quality, the increasing/decreasing stability function ($S_{I/D}$) is also proposed.

For detecting similar behaviors when reducing the dimensionality of the data, in terms of loss of quality, a clustering process of the DR algorithms has been carried out. This second analysis reports results that indicate 4 different groups of algorithms. *t*-SNE indicated a differentiated behavior of the remaining algorithms when performing DR tasks, and thus achieved the best results. PCA, SM and MVU algorithms reduced the dimensionality of the data in a very similar way. Conversely, LDA, $KPCA_{gauss}$ and ISOMAP algorithms also showed common features when reducing the dimensionality of the data.

A final analysis is also presented, as regards the correlation between the different quality criteria when assessing the DR process. There is a very strong correlation between several criteria. For a more accurate measurement of the correlations, a modification of the original Pearson correlation coefficient is presented. Therefore, P_M, P_{MC} and $NIEQA_{LOCAL}$ proved to be strongly correlated. Q_{NX}, R_{NX}, and Qk showed high correlation values. All these criteria and many other have shown strong correlations, independently of the nature of the dataset. However, the Q_Y criterion lacks direct correlation with other criteria because of its peculiar nature.

The study presented here opens up a wide range of possibilities for carrying out a deeper comparative study of the DR algorithms according their geometry preservation skills, as well as the inclusion of other metrics or quality criteria for enriching the loss of quality evaluation. Furthermore, the results obtained through this methodology could be extended through the implementation of a more complex kind of data analysis for studying the behavior of the loss of quality in DR tasks. Finally, in the near future, this methodology will be included as a part of a tool, that carries out a Feature Subset Selection (FSS) on DNA microarray data. Its aim is to identify biomarkers for subsequent 2D and 3D linear scatterplot visualization techniques, as well as supervised learning tasks for new samples. Part of this process involves the use of our methodology for reporting on the amount of loss of quality produced during the entire process.

Acknowledgements

The authors thankfully acknowledge the computer resources, technical expertise and assistance provided by the Centro de Supercomputación y Visualización de Madrid (CeSViMa), the Spanish Supercomputing Network and the Graphics-Interactive Systems Group at TU Darmstadt, Germany. The study was supported by the Spanish Ministry of Economy and Competitiveness (Grant TIN2010-21289-C02-02 and the Cajal Blue Brain Project, Spanish partner of the Blue Brain Project initiative from EPFL).

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.ins.2014.02.068.

References

[1] U. Akkucuk, Nonlinear Mapping: Approaches Based on Optimizing an Index of Continuity and Applying Classical Metric MDS on Revised Distances, Rutgers University, 2004.
[2] U. Akkucuk, J.D. Carroll, PARAMAP vs. Isomap: a comparison of two nonlinear mapping algorithms, J. Classif. 23 (2006) 221–254.
[3] W.W.R. Ball, H.S.M. Coxeter, Mathematical Recreations and Essays, 13th ed., Dover, New York, 1987.
[4] H.U. Bauer, K. Pawelzik, Quantifying the neighborhood preservation of self-organizing feature maps, IEEE Trans. Neural Networks 3 (1992) 570–579.
[5] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: Advances in Neural Information Processing Systems, vol. 14, 2001, pp. 585–591.
[6] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, Neural Comput. 15 (2003) 1373–1396.
[7] J. Bernataviciene, G. Dzemyda, O. Kurasova, V. Marcinkevicius, Optimal decisions in combining the SOM with nonlinear projection methods, Eur. J. Oper. Res. 173 (2006) 729–745.
[8] J.C. Bezdek, N.R. Pal, An index of topological preservation for feature extraction, Pattern Recogn. 28 (1995) 381–391.
[9] C.M. Bishop, C.K.I. Williams, GTM: the generative topographic mapping, Neural Comput. 10 (1998) 215–234.
[10] I. Borg, P. Groenen, Modern Multidimensional Scaling: Theory and Applications, Springer, New York, 1997.
[11] I. Borg, J. Lingoes, Multidimensional Similarity Structure Analysis, Springer Verlag, New York, 1987.
[12] S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, Vandenberghe, 2004.
[13] L. Chen, Local Multidimensional Scaling for Nonlinear Dimension Reduction, Graph Layout and Proximity Analysis, Ph.D. Thesis, University of Pennsylvania, 2006.
[14] L. Chen, A. Buja, Local multidimensional scaling for nonlinear dimension reduction, and proximity analysis, J. Am. Stat. Assoc. 104 (2009) 209–219.
[15] Y.N. Chen, H.T. Lin, Feature-aware label space dimension reduction for multi-label classification, in: Advances in Neural Information Processing Systems, vol. 25, 2012, pp. 1538–1546.
[16] N. Chikhi, B. Rothenburger, N. Aussenac-Gilles, A comparison of dimensionality reduction techniques for web structure mining, in: IEEE/WIC/ACM International Conference on Web, Intelligence, 2007, pp. 116–119.
[17] T.F. Cox, M.A.A. Cox, Multidimensional Scaling, Chapman & Hall, London, 1994.
[18] V. De Silva, J.B. Tenenbaum, Global versus local methods in nonlinear dimensionality reduction, in: Advances in Neural Information Processing Systems 15, vol. 15, 2003, pp. 705–712.
[19] J. DeFelipe, P. Lopez-Cruz, R. Benavides-Piccione, C. Bielza, P. Larrañaga, et al, New insights into the classification and nomenclature of cortical GABAergic interneurons, Nat. Rev. Neurosci. 14 (2013) 202–216.
[20] P. Demartines, J. Herault, Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Netw. 8 (1997) 148–154.
[21] D.B. Dias, R.C.B. Madeo, T. Rocha, H.H. Bíscaro, S.M. Peres, Hand movement recognition for brazilian sign language: a study using distance-based neural networks, in: Proceedings of the 2009 international joint conference on Neural Networks, IJCNN'09, IEEE Press, 2009, pp. 2355–2362.
[22] R. Duda, P. Hart, D. Stork, Pattern classification, Pattern Classification and Scene Analysis: Pattern Classification, Wiley, 2001.
[23] P.A. Estévez, A.M. Chong, Geodesic Nonlinear Mapping Using the Neural Gas Network, in: IJCNN, 2006, pp. 3287–3294.
[24] B. Everitt, G.G. Dunn, B. Everitt, Applied Multivariate Data Analysis, Edward Arnold, London, 1991.
[25] B. Everitt, S. Rabe-Hesketh, The Analysis of Proximity Data, Kendall's Library of Statistics, No. 4, Edward Arnold, London, 1997.
[26] I.W. Evett, E.J. Spiehler, Knowledge Based Systems, Halsted Press, New York, NY, USA, 1988. pp. 152–160.
[27] 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: P. Perner (Ed.), Machine Learning and Data Mining in Pattern Recognition, Lecture Notes in Computer Science, vol. 4571, Springer, Berlin, Heidelberg, 2007, pp. 499–517.
[28] K. Fukumizu, F.R. Bach, M.I. Jordan, Dimensionality Reduction for Supervised Learning with Reproducing Kernel Hilbert Spaces, J. Mach. Learn. Res. 5 (2004) 73–99.
[29] K. Fukunaga, Introduction to Statistical Pattern Recognition, 2nd ed., Academic Press Professional Inc., San Diego, CA, USA, 1990.
[30] J. Gan, B. Awwad, Shiekh Hasan, C. Tsui, A filter-dominating hybrid sequential forward floating search method for feature subset selection in high-dimensional space, Int. J. Mach. Learning Cybern. (2012) 1–11.
[31] Y. Goldberg, Y. Ritov, Local procrustes for manifold embedding: a measure of embedding quality and embedding algorithms, Mach. Learning 77 (2009) 1–25.
[32] Y. Goldberg, A. Zakai, D. Kushnir, Y. Ritov, Manifold learning: the price of normalization, J. Mach. Learn. Res. 9 (2008) 1909–1939.
[33] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, E.S. Lander, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring, Science 286 (1999) 531–537.
[34] G. Goodhill, S. Finch, T. Sejnowski, Quantifying neighbourhood preservation in topographic mappings, in: Proceedings of the 3rd Joint Symposium on Neural Computation, 1996.
[35] P.R. Gorman, T.J. Sejnowski, Analysis of hidden units in a layered network trained to classify sonar targets, Neural Networks 1 (1988) 75–89.
[36] J.C. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.
[37] H. Handa, On the effect of dimensionality reduction by Manifold Learning for Evolutionary Learning, Evolving Syst. 2 (2011) 235–247.
[38] T. Hastie, R. Tibshirani, J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Darinka Springer & Janez, Springer Series in Statistics Series, 2001.
[39] W.A. Hendricks, K.W. Robey, The sampling distribution of the coefficient of variation, Ann. Math. Statist. 7 (1936) 129–132.
[40] D.S. Hochbaum, D.B. Shmoys, A best possible heuristic for the k-center problem, Math. Oper. Res. 10 (1985) 180–184.
[41] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psych. 24 (1933).
[42] L. Hubert, P. Arabie, Comparing partitions, J. Classif. 2 (1985) 193–218.
[43] S. Ji, J. Ye, Linear dimensionality reduction for multi-label classification, in: Proceedings of the 21st International Jont Conference on Artificial Intelligence, IJCAI'09, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2009, pp. 1077–1082.
[44] I.T. Jolliffe, Principal Component Analysis, Springer-Verlag, 1986.
[45] R. Karbauskaitė, O. Kurasova, G. Dzemyda, Selection of the number of neighbours of each data point for the locally linear embedding algorithm, Inf. Technol. Control 36 (2007) 359–364.
[46] R. Karbauskaite, G. Dzemyda, Topology preservation measures in the visualization of manifold-type multidimensional data, Inf. Lith. Acad. Sci. 20 (2009) 235–254.
[47] J. Khan, J.S. Wei, M. Ringner, L.H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C.R. Antonescu, C. Peterson, P.S. Meltzer, Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks, Nat. Med. 7 (2001) 673–679.
[48] T. Kohonen, M.R. Schroeder, T.S. Huang, Self-Organizing Maps, 3rd edition., Springer-Verlag New York Inc., Secaucus, NJ, USA, 2001.

[49] A. König, Interactive visualization and analysis of hierarchical neural projections for data mining, IEEE Trans. Neural Netw. Learning Syst. 11 (2000) 615–624.
[50] O. Kouropteva, O. Okun, M. PietikSinen, Incremental locally linear embedding algorithm, in: SCIA, Lecture Notes in Computer Science, vol. 3540, Springer, 2005, pp. 521–530.
[51] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–27.
[52] J. Kruskal, Nonmetric multidimensional scaling: a numerical method, Psychometrika 29 (1964) 115–129.
[53] W.J. Krzanowski, F.H.C. Marriott, Multivariate Analysis, Part 1: Distributions, Ordination and Inference, Edward Arnold, London, 1994.
[54] L.A. Kurgan, K.J. Cios, R. Tadeusiewicz, M.R. Ogiela, L.S. Goodenday, Knowledge discovery approach to automated cardiac SPECT diagnosis, Artif. Intell. Med. 23 (2001) 149.
[55] S. Lacoste-Julien, F. Sha, M.I. Jordan, DiscLDA: discriminative learning for dimensionality reduction and classification, in: NIPS, Curran Associates Inc., 2008, pp. 897–904.
[56] S. Lafon, A.B. Lee, Diffusion maps and coarse-graining: a unified framework for dimensionality reduction, graph partitioning, and data set parameterization, IEEE Trans. Pattern Anal. Mach. Intell. 28 (2006) 1393–1403.
[57] J.A. Lee, C. Archambeau, M. Verleysen, Locally linear embedding versus isotop, in: ESANN, 2003, pp. 527–534.
[58] J.A. Lee, A. Lendasse, N. Donckers, M. Verleysen, A robust non-linear projection method., in: ESANN, 2000, pp. 13–20.
[59] J.A. Lee, A. Lendasse, M. Verleysen, Curvilinear Distance Analysis versus Isomap., in: ESANN, 2002, pp. 185–192.
[60] J.A. Lee, E. Renard, G. Bernard, P. Dupont, M. Verleysen, Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation, Neurocomputing 112 (2013) 92–108.
[61] J.A. Lee, M. Verleysen, Nonlinear dimensionality reduction of data manifolds with essential loops, Neurocomputing 67 (2005) 29–53.
[62] J.A. Lee, M. Verleysen, Nonlinear dimensionality reduction, Springer, New York, London, 2007.
[63] J.A. Lee, M. Verleysen, Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods, J. Mach. Learning Res. – Proc. Track 4 (2008) 21–35.
[64] J.A. Lee, M. Verleysen, Rank-based quality assessment of nonlinear dimensionality reduction, in: ESANN, 2008, pp. 49–54.
[65] J.A. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomput 72 (2009) 1431–1443.
[66] J.A. Lee, M. Verleysen, Scale-independent quality criteria for dimensionality reduction, Pattern Recogn. Lett. 31 (2010) 2248–2257.
[67] J.M. Lee, Introduction to Topological Manifolds (Graduate Texts in Mathematics), Springer, 2000.
[68] S. Lespinats, M. Aupetit, CheckViz: Sanity Check and Topological Clues for Linear and Non-Linear Mappings, Comput. Graph. Forum 30 (2011) 113–125.
[69] M.A. Little, P.E. McSharry, E.J. Hunter, J.L. Spielman, L.O. Ramig, Suitability of dysphonia measurements for telemonitoring of parkinson's disease, IEEE Trans. Biomed. Eng. 56 (2009) 1015–1022.
[70] M.A. Little, P.E. Mcsharry, S.J. Roberts, D.A.E. Costello, I.M. Moroz, Exploiting nonlinear recurrence and fractal scaling properties for voice disorder detection, BioMed. Eng. OnLine 6 (2007) 23+.
[71] L. van der Maaten, The Matlab Toolbox for Dimensionality Reduction, 2012.
[72] L. van der Maaten, G. Hinton, Visualizing high-dimensional data using t-SNE, J. Mach. Learning Res. 9 (2008) 2579–2605.
[73] L.J.P. van der Maaten, U. Maastricht, An Introduction to Dimensionality Reduction Using Matlab, 2007.
[74] L.J.P. Van der Maaten, E.O. Postma, H.J. van den Herik, Dimensionality Reduction: A Comparative, Review, 2007.
[75] O.L. Mangasarian, W.N. Street, W.H. Wolberg, Breast cancer diagnosis and prognosis via linear programming, Oper. Res. 43 (1995) 570–577.
[76] K.V. Mardia, J.T. Kent, J.M. Bibby, Multivariate Analysis, Academic Press, Orlando, 1979.
[77] D. Meng, Y. Leung, Z. Xu, A new quality assessment criterion for nonlinear dimensionality reduction, Neurocomputing 74 (2011) 941–948.
[78] B. Mokbel, W. Lueks, A. Gisbrecht, B. Hammer, Visualizing the quality of dimensionality reduction, Neurocomputing 112 (2013) 109–123.
[79] A. Musa, A comparison of 1-regularizion, PCA, KPCA and ICA for dimensionality reduction in logistic regression, Int. J. Mach. Learning Cybern. (2013) 1–13.
[80] B. Nadler, S. Lafon, R.R. Coifman, I.G. Kevrekidis, Diffusion maps, spectral clustering and reaction coordinates of dynamical systems, Appl. Comput. Harmonic Anal. 21 (2006) 113–127.
[81] R. Ohbuchi, J. Kobayashi, A. Yamamoto, T. Shimizu, Comparison of dimension reduction methods for database-adaptive 3D model retrieval, in: Adaptive Multimedial Retrieval: Retrieval, User, and Semantics, Lecture Notes in Computer Science, vol. 4918, Springer, Berlin, Heidelberg, 2008, pp. 196–210.
[82] G. Pölzlbauer, Survey and comparison of quality measures for self-organizing maps, in: Proceedings of the Fifth Workshop on Data Analysis (WDA'04), Elfa Academic Press, Sliezsky dom, Vysoké Tatry, Slovakia, 2004, pp. 67–82.
[83] B. Qian, I. Davidson, Semi-Supervised Dimension Reduction for Multi-Label Classification, in: AAAI, AAAI Press, 2010.
[84] W.M. Rand, Objective criteria for the evaluation of clustering methods, J. Am. Stat. Assoc. 66 (1971) 846–850.
[85] S.T. Roweis, L.K. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (2000) 2323–2326.
[86] J.W. Sammon, A nonlinear mapping for data structure analysis, IEEE Trans. Comput. C-18 (1969).
[87] L.K. Saul, S.T. Roweis, Think globally, low dimensional manifolds, J. Mach. Learning Res. 4 (2003) 119–155.
[88] B. Schölkopf, A. Smola, K.R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319.
[89] B. Schölkopf, A.J. Smola, K.R. Müller, Kernel Principal Component Analysis, MIT Press, Cambridge, MA, USA, 1999. pp. 327–352.
[90] G. Seber, Multivariate Observations, Wiley Series in Probability and Statistics, Wiley, 2004.
[91] G.A.F. Seber, Multivariate Observations, Wiley, New York, 1984.
[92] S.R. Setlur, K.D. Mertz, Y. Hoshida, F. Demichelis, M. Lupien, S. Perner, A. Sboner, Y. Pawitan, O. Andrén, L.A. Johnson, J. Tang, H.O. Adami, S. Calza, A.M. Chinnaiyan, D. Rhodes, S. Tomlins, K. Fall, L.A. Mucci, P.W. Kantoff, M.J. Stampfer, S.O. Andersson, E. Varenhorst, J.E. Johansson, M. Brown, T.R. Golub, M.A. Rubin, Estrogen-dependent signaling in a molecularly distinct subclass of aggressive prostate cancer, J. Nat. Cancer Inst. 100 (2008) 815–825.
[93] R. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function. I, Psychometrika 27 (1962) 125–140.
[94] R. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function, II, Psychometrika 27 (1962) 219–246.
[95] M.A. Shipp, K.N. Ross, P. Tamayo, A.P. Weng, J.L. Kutok, R.C. Aguiar, M. Gaasenbeek, M. Angelo, M. Reich, G.S. Pinkus, T.S. Ray, M.A. Koval, K.W. Last, A. Norton, T.A. Lister, J. Mesirov, D.S. Neuberg, E.S. Lander, J.C. Aster, T.R. Golub, Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning, Nat. Med. 8 (2002) 68–74.
[96] R. Sibson, Studies in the robustness of multidimensional-scaling: procrustes statistics, J.R. Stat. Soc. Ser. B Methodol 40 (1978) 234–238.
[97] R. Sibson, Perturbational analysis of classical scaling, J.R. Stat. Soc. Ser. B Methodol 41 (1979) 217–229.
[98] S. Siegel, N. Castellan, Nonparametric Statistics for the Behavioral Sciences, McGraw–Hill Inc., 1988.
[99] W.N. Street, O.L. Mangasarian, W.H. Wolberg, An Inductive Learning Approach to Prognostic Prediction, in: ICML, 1995, pp. 522–530.
[100] Y. Takane, F.W. Young, J. De Leeuw, Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features, Psychometrika 42 (1977) 7–67.
[101] B. Tang, M. Shepherd, E. Milios, M.I. Heywood, Comparing and combining dimension reduction techniques for efficient text clustering, in: Proceedings of the Workshop on Feature Selection for Data Mining, SIAM Data Mining, 2005.
[102] J. Tenenbaum, Matlab Package for Isomap (MIT), 2000.
[103] J. Tenenbaum, V. Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (2000) 2319–2323.
[104] W. Torgerson, Multidimensional scaling: I. Theory and method, Psychometrika 17 (1952) 401–419.
[105] E.C.C. Tsang, D. Chen, D. Yeung, X.Z. Wang, J. Lee, Attributes reduction using fuzzy rough sets, IEEE Trans. Fuzzy Syst. 16 (2008) 1130–1141.

[106] J. Venna, Dimensionality Reduction for Visual Exploration of Similarity Structures, Dissertations in computer and information science, Helsinki University of Technology, 2007.
[107] J. Venna, S. Kaski, Local multidimensional scaling, Neural Networks 19 (2006) 889–899.
[108] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, J. Mach. Learning Res. 11 (2010) 451–490.
[109] J. Vesanto, J. Himberg, E. Alhoniemi, J. Parhankangas, SOM Toolbox 2.0, 2005.
[110] T. Villmann, R. Der, M. Herrmann, T. Martinetz, Topology preservation in self-organizing feature maps: exact definition and measurement, IEEE Trans. Neural Networks 8 (1997) 256–266.
[111] T. Villmann, R. Der, T. Martinetz, A new quantitative measure of topology preservation in Kohonen's feature maps, in: 1994 IEEE International Conference on Neural Networks, 1994, IEEE World Congress on Computational Intelligence, vol. 2, pp. 645–648.
[112] V. Vinay, I.J. Cox, K.R. Wood, N. Milic-Frayling, A comparison of dimensionality reduction techniques for text retrieval, in: ICMLA, IEEE Computer Society, 2005.
[113] J. Wang, Geometric Structure of High-dimensional Data and Dimensionality Reduction, Higher Education Press, 2012.
[114] K. Weinberger, F. Sha, L.K. Saul, Learning a kernel matrix for nonlinear dimensionality reduction, in: Proceedings of the Twenty First International Conference on Machine Learning (ICML-04), Banff, Canada, pp. 839–846.
[115] K.Q. Weinberger, Matlab package for MVU, 2012.
[116] K.Q. Weinberger, L.K. Saul, An introduction to nonlinear dimensionality reduction by maximum variance unfolding, in: Proceedings of the 21st National Conference on Artificial Intelligence, vol. 2, AAAI 2006, pp. 1683–1686.
[117] K.Q. Weinberger, L.K. Saul, Unsupervised learning of image manifolds by semidefinite programming, Int. J. Comput. Vision 70 (2006) 77–90.
[118] W.H. Wolberg, W.N. Street, D. Heisey, O.L. Mangasarian, Computerized breast cancer diagnosis and prognosis from fine-needle aspirates, Arch. Surg. 130 (1995) 511–516.
[119] P. Zhang, Y. Ren, B. Zhang, A new embedding quality assessment method for manifold learning, Neurocomputing 97 (2012) 251–266.