Pattern Recognition Letters 31 (2010) 2248–2257

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Scale-independent quality criteria for dimensionality reduction

John A. Lee [a,*,1], Michel Verleysen [b,c]

[a] Molecular Imaging and Experimental Radiotherapy Department, Avenue Hippocrate, 54, B-1200 Bruxelles, Belgium
[b] Machine Learning Group, Université catholique de Louvain, Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium
[c] SAMOS-MATISSE, Université Paris I Panthéon Sorbonne, Rue de Tolbiac, 90, 75634 Paris Cedex 13, France

ARTICLE INFO

Article history:
Available online 22 April 2010

Keywords:
Dimensionality reduction
Embedding
Manifold learning
Quality assessment

ABSTRACT

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, in order to facilitate their visual interpretation. Many techniques exist, ranging from simple linear projections to more complex nonlinear transformations. The large variety of methods emphasizes the need of quality criteria that allow for fair comparisons between them. This paper extends previous work about rank-based quality criteria and proposes to circumvent their scale dependency. Most dimensionality reduction techniques indeed rely on a scale parameter that distinguish between local and global data properties. Such a scale dependency can be similarly found in usual quality criteria: they assess the embedding quality on a certain scale. Experiments with various dimensionality reduction techniques eventually show the strengths and weaknesses of the proposed scale-independent criteria.

1. Introduction

The interpretation of high-dimensional data remains a difficult task, mainly because human vision is not used to deal with spaces whose dimensionality is higher than three. Part of this inability stems from the curse of dimensionality, a convenient expression that encompasses all weird and unexpected properties of high-dimensional spaces. If visualization is difficult in high-dimensional space, perhaps an (almost) equivalent representation in a lower-dimensional space could improve the readability of data. This is precisely the idea that lies underneath the field of dimensionality reduction (DR in short). This domain includes various techniques that are able to construct meaningful data representations in a space of given dimensionality. Linear DR is well known, with techniques such as principal component analysis (Jolliffe, 1986) and classical metric multidimensional scaling (Young and Householder, 1938; Torgerson, 1952). On the other hand, nonlinear dimensionality reduction (Lee and Verleysen, 2007) (NLDR) emerged later, with nonlinear variants of multidimensional scaling (Shepard, 1962; Kruskal, 1964; Takane et al., 1977), such as Sammon's nonlinear mapping (Sammon, 1969). For the past 25 years, research around NLDR has deeply evolved and after some interest in neural approaches (Kohonen, 1982; Kramer, 1991; Oja, 1991; Demartines and Hérault, 1993; Mao et al., 1995), the community has recently focused on spectral techniques (Schölkopf et al., 1998; Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2003; Donoho and Grimes, 2003; Weinberger and Saul, 2006). Modern NLDR is sometimes referred to as manifold learning; it is also tightly connected with graph embedding (Di Battista et al., 1999) and spectral clustering (Bengio et al., 2003; Saerens et al., 2004; Nadler et al., 2006; Brand and Huang, 2003).

In the most general setting, DR transforms a set of N high-dimensional vectors, denoted by $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1 \leqslant i \leqslant N}$, into N low-dimensional vectors, denoted by $\mathbf{X} = [\mathbf{x}_i]_{1 \leqslant i \leqslant N}$. Of course, the low-dimensional representation has to be meaningful in some sense. Usually, the general idea is to embed close neighbors next to each other, while maintaining large distances between faraway data items. In practice, the goal of DR is then to preserve as well as possible simple properties such as soft or hard neighborhoods (Kohonen, 1982), proximities, similarities, or ranks (Shepard, 1962; Kruskal, 1964). A straighter way to construct an embedding is to preserve pairwise distances (Sammon, 1969; Demartines and Hérault, 1993, 1997) measured in $\boldsymbol{\Xi}$, with some appropriate metric. These approaches remain valid if the coordinates in $\boldsymbol{\Xi}$ are unknown, that is, when the data set consists of pairwise distances. If not all distances are specified, then the problem can elegantly be modeled using a graph, in which edges are present for known entries of the pairwise distance matrix. The edge weights can be binary- or real-valued, depending on the data nature. Some NLDR techniques also involve a graph even if all pairwise distance are available. For instance, a graph can be used to focus on small neighborhoods (Roweis and Saul, 2000) or to approximate geodesic distances (Tenenbaum et al., 2000; Lee and Verleysen, 2004) with weighted

* Corresponding author. Fax: +32 10472598.
E-mail addresses: john.lee@uclouvain.be (J.A. Lee), michel.verleysen@uclouvain.be (M. Verleysen).
[1] J.A.L. is a Research Fellow with the Belgian National Fund for Scientific Research (FNRS).

0167-8655/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.patrec.2010.04.013

shortest paths. This illustrates the close relationship between NLDR and graph embedding.

As to manifold learning, one commonly assumes that the vectors in Ξ are sampled from a smooth manifold. Under this hypothesis, one seeks to re-embed the manifold in a space of the lowest possible dimensionality, without modifying its topological properties. As these properties cannot easily be identified starting from a set of Cartesian coordinates, the above-mentioned approaches based on distances, neighborhoods, etc. are followed as well.

As a matter of fact, the scientific community has been mainly focusing on the design of new NLDR methods and the question of quality assessment remains mostly unanswered. As most NLDR methods optimize a given objective function, a simplistic way to assess quality is to look at the value of the objective function after convergence. Obviously, this allows us to compare several runs with e.g. different parameter values, but makes the comparison of different methods unfair. Still, objective functions that assess the preservation of pairwise distances, such as the stress or strain used in various versions of MDS, have been very popular (Venna, 2007).

Another obvious quality criterion is the reconstruction error. If a NLDR technique provides us with a mapping $\mathcal{M}$ such that $\mathbf{x} = \mathcal{M}(\xi)$, then this error can be written as the expectation $E_{\text{rec}} = E\{(\xi - \mathcal{M}^{-1}(\mathcal{M}(\xi)))^2\}$. The reconstruction error is a universal quality criterion, but it requires the availability of $\mathcal{M}$ and $\mathcal{M}^{-1}$ in closed form, whereas most NLDR methods are nonparametric (they merely provide values of $\mathcal{M}$ for the known vectors ξ_i). The minimization of the reconstruction error is the approach that is followed by PCA and nonlinear auto-encoders (Kramer, 1991; Oja, 1991).

```
Procedure Q ← coranking(Δ, D)
  Π ← I                                  Initialize the permutations in the HDS
  P ← I                                  Initialize the permutations in the LDS
  For j from 1 to N                      For each column...
    (δ_j, π_j) ← sort(δ_j)               Sort the distances in the HDS
    (d_j, p_j) ← sort(d_j)               Sort the distances in the LDS
  End
  R ← [0]_{1≤i,j≤N}                      Initialize ranks in the LDS
  For j from 1 to N                      For each column...
    For i from 1 to N                    For rank i in the LDS...
      r_{p_ij} ← i                       Find the associated row
    End
  End
  Q ← [0]_{1≤i,j≤N}                      Initialize the co-ranking matrix
  For j from 1 to N                      For each column...
    For i from 1 to N                    For rank i in the HDS...
      k ← r_{π_ij}                       Find the corresponding rank in the LDS
      q_ik ← q_ik + 1                    Increase the associated element of Q
    End
  End
  Q ← [q_ij]_{2≤i,j≤N}                   Remove first row and column
```

Fig. 1. Procedure to compute co-ranking matrix **Q**, starting from the matrices of pairwise distances in the high- and low-dimensional spaces (HDS and LDS in short). These matrices are defined by $\mathbf{\Delta} = [\delta_{ij}]_{1\leqslant i,j\leqslant N}$ and $\mathbf{D} = [d_{ij}]_{1\leqslant i,j\leqslant N}$. Symbols $\boldsymbol{\delta}_j$ and $\mathbf{d}_j$ denote the jth column of $\mathbf{\Delta}$ and $\mathbf{D}$, respectively. Function $(\mathbf{v}, \mathbf{p}) \leftarrow \text{sort}(\mathbf{u})$ sorts the elements of vector $\mathbf{u}$. Output vector $\mathbf{v}$ is a permutation of $\mathbf{u}$ such that it is sorted in ascending order. Output vector $\mathbf{p}$ results from the application of the same permutation to vector $[1,\ldots,N]^T$. The most expensive step in the procedure is the sorting of each column of $\mathbf{\Delta}$ and $\mathbf{D}$. The time complexity of the whole procedure is thus $\mathcal{O}(N^2 \log N)$.

Still another approach mentioned in the literature consists in using an indirect performance index, such as a classification error (see for instance (Saul et al., 2003; Weinberger et al., 2004) and other references in (Venna, 2007)). Obviously, such an index can be used only with labeled data.

Eventually, a last possibility consists in sticking to the intrinsic goal of DR, by trying to assess the preservation of proximity relationships: are close neighbors embedded near each other and are dissimilar items lying far from each other? As our goal is quality assessment, we can translate this idea into a quantitative criterion without caring about typical constraints that come with the design of an objective function, such as continuity and differentiability. This opens to way to potentially complex quality criteria that more faithfully assess the preservation of the data set structure. First attempts in this direction can be found in the particular case of self-organizing maps (Kohonen, 1982), such as the topographic product (Bauer and Pawelzik, 1992) and the topographic function (Villmann et al., 1997). More recently, new criteria for quality assessment have been proposed, with a broader applicability, such as the trustworthiness and continuity measures (Venna and Kaski, 2001; Venna, 2007), the local continuity metacriterion (Chen, 2006; Chen and Buja, 2009), the mean relative rank errors (Lee and Verleysen, 2007), and the quality/behavior curves (Lee and Verleysen, 2008a; Lee and Verleysen, 2009). All these criteria involve ranks of sorted distances and analyze K-ary neighborhoods before and after dimensionality reduction, for a varying value of K. This is a major improvement over a measurement of distance preservation, as the use of ranks allows distances to grow or to shrink, provided their order does not change. In the case of manifold learning, such distance scalings are often necessary in order to unfold and flatten the manifold.

A unifying framework for quality criteria relying on ranks and K-ary neighborhoods has been proposed in (Lee and Verleysen, 2008a; Lee and Verleysen, 2009), along with a pair of new criteria. As a main advantage, they avoid any scale-dependent weighting that is present in almost all other criteria and that inevitably turns out to be somewhat arbitrary. On the other hand, these criteria keep being functions of K, the neighborhood size, and therefore yield curves that must be scrutinized on several scales. Within this framework, this paper aims at summarizing each curve into a single scalar value, thus enabling simple and direct comparisons of DR methods. An experimental section illustrates the use of the scalar criteria and compares various NLDR techniques applied to several data sets.

This paper is organized as follows. Section 2 introduces the notations for distances, ranks, and neighborhoods. Section 3 reviews existing rank-based criteria. Section 4 describes scalar quality criteria that are scale independent. Section 5 illustrates them in experiments with various DR methods and data sets. Finally, Section 6 draws the conclusions.

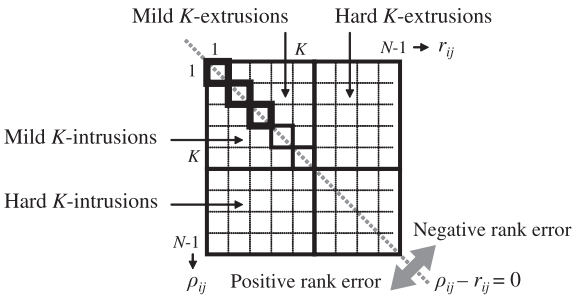

Fig. 2. Block division of the co-ranking matrix, showing the different types of intrusions and extrusions, and their relationship with the rank error.

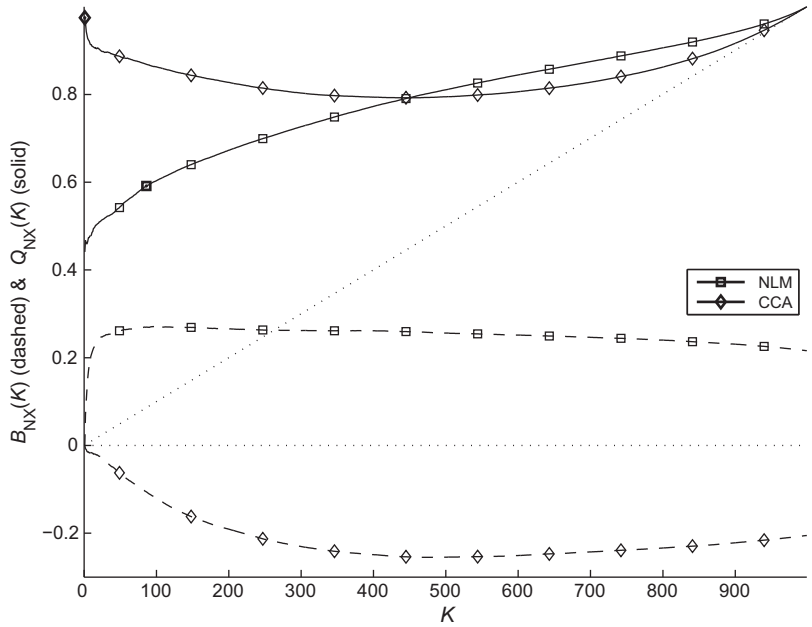

Fig. 3. Criteria $Q_{NX}(K)$ and $B_{NX}(K)$ for two embeddings of a hollow sphere (1000 points). The embeddings are computed with NLM and CCA. The NLM produces an intrusive embedding of average quality, whereas CCA's ability to yield an extrusive embedding leads to a better result. The bold markers on the $Q_{NX}(K)$ curves correspond to the points $[K_{max}, Q_{NX}(K_{max})]^T$ (see Section 4).

Table 1
Scalar quality criteria corresponding to the curves in Fig. 3. The average values of $Q_{NX}(K)$ and $B_{NX}(K)$ are denoted by Q_{avg} and B_{avg}. The 'localness' is given by L, whereas Q_{local} and Q_{global} are the average values of $Q_{NX}(K)$ below and above K_{max}.

	Q_{avg}	B_{avg}	L	Q_{global}	Q_{local}
NLM	0.7895	0.2505	0.9149	0.8134	0.5341
CCA	0.8440	−0.2112	1.0000	0.8440	0.9750

2. Distances, ranks, and neighborhoods

Most NLDR techniques involve distances in some way. Symbol δ_{ij} denotes the distance from ξ_i to ξ_j in the high-dimensional space. Similarly, d_{ij} is the distance from $\mathbf{x}_i$ to $\mathbf{x}_j$ in the low-dimensional space. Notice that we assume that $\delta_{ij} = \delta_{ji}$ and $d_{ij} = d_{ji}$, although this hypothesis is not always required. For instance, it does not hold true if δ_{ij} and δ_{ji} stem from distinct experimental measurements. Starting from distances, we can compute ranks.

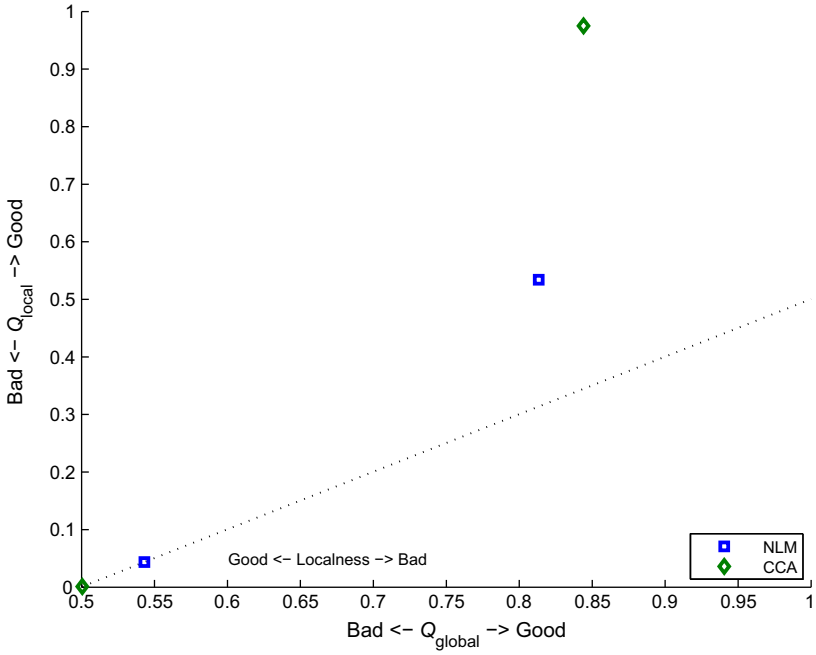

Fig. 4. Performance diagram summarizing the curves in Fig. 3. Each embedding of the hollow sphere is associated with two markers. Their coordinates are $[Q_{global}, Q_{local}]^T$ and $[(2N-1)/(N-1)-L, (N+1)/(N-1)-L]^T/2$, respectively. For each embedding, the coordinates of the second marker corresponds to the values of Q_{global} and Q_{local} in the case of a random embedding with the same value of K_{max}. In this toy example, CCA clearly outperforms NLM.

The rank of ξ_j with respect to ξ_i in the high-dimensional space is written as $\rho_{ij} = |\{k: \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leqslant k < j \leqslant N)\}|$, where $|A|$ denotes the cardinality of set A. Similarly, the rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ in the low-dimensional space is $r_{ij} = |\{k: d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leqslant k < j \leqslant N)\}|$. Hence, reflexive ranks are set to zero ($\rho_{ii} = r_{ii} = 0$) and ranks are unique, i.e. there are no *ex aequo* ranks: $\rho_{ij} \neq \rho_{ik}$ for $k \neq j$, even if $\delta_{ij} = \delta_{ik}$. This means that nonreflexive ranks belong to $\{1,\ldots,N-1\}$. The nonreflexive K-ary neighborhoods of ξ_i and $\mathbf{x}_i$ are denoted by $\nu_i^K = \{j : 1 \leqslant \rho_{ij} \leqslant K\}$ and $n_i^K = \{j : 1 \leqslant r_{ij} \leqslant K\}$, respectively.

The *co-ranking matrix* (Lee and Verleysen, 2008b) can then be defined as

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

In practice, the procedure given in Fig. 1 computes $\mathbf{Q}$ in the most efficient way. The co-ranking matrix is the joint histogram of the ranks and is actually a sum of N permutation matrices of size $N-1$. With an appropriate gray scale, the co-ranking matrix can also be displayed and interpreted in a similar way as a Shepard diagram (Shepard, 1962). Historically, this scatterplot has often been used to assess results of multidimensional scaling and related methods (Demartines and Hérault, 1997); it shows the distances δ_{ij} with respect to the corresponding distances d_{ij}, for all pairs (i,j), with $i \neq j$. The analogy between the co-ranking matrix and Shepard's diagram suggests that meaningful criteria should focus on the upper and lower triangle of the co-ranking matrix $\mathbf{Q}$. Following this line, we define the rank error to be the difference $\rho_{ij} - r_{ij}$. We call an *intrusion* the event of a positive rank error for some pair (i,j). Similarly, an *extrusion* denotes the event of a negative rank error. The amplitude of an intrusion or extrusion is the absolute value of the corresponding rank error.

In order to focus on K-ary neighborhoods, we also define a K-intrusion (resp. K-extrusion) to be the conjunction of an intrusion (resp. extrusion) for some pair (i,j) with the event $r_{ij} < K$ (resp. $\rho_{ij} < K$). We can further distinguish mild and hard K-intrusions. The former correspond to the event $r_{ij} < \rho_{ij} \leqslant K$, whereas the latter is associated with the event $r_{ij} \leqslant K < \rho_{ij}$. Similar definitions for mild and hard K-extrusions can be deduced. Intuitively, mild K-intrusions and mild K-extrusions correspond to vectors that are respectively "promoted" and "downgraded", but still remain in both ν_i^K and n_i^K.

The various types of intrusions and extrusions can be associated with different blocks of the co-ranking matrix, as illustrated in Fig. 2. The idea is to concentrate on K-ary neighborhoods and thus on the four blocks that separate the first K rows and columns. Therefore, if we define $\mathbb{F}_K = \{1,\ldots,K\}$ (index set for the K first ele-

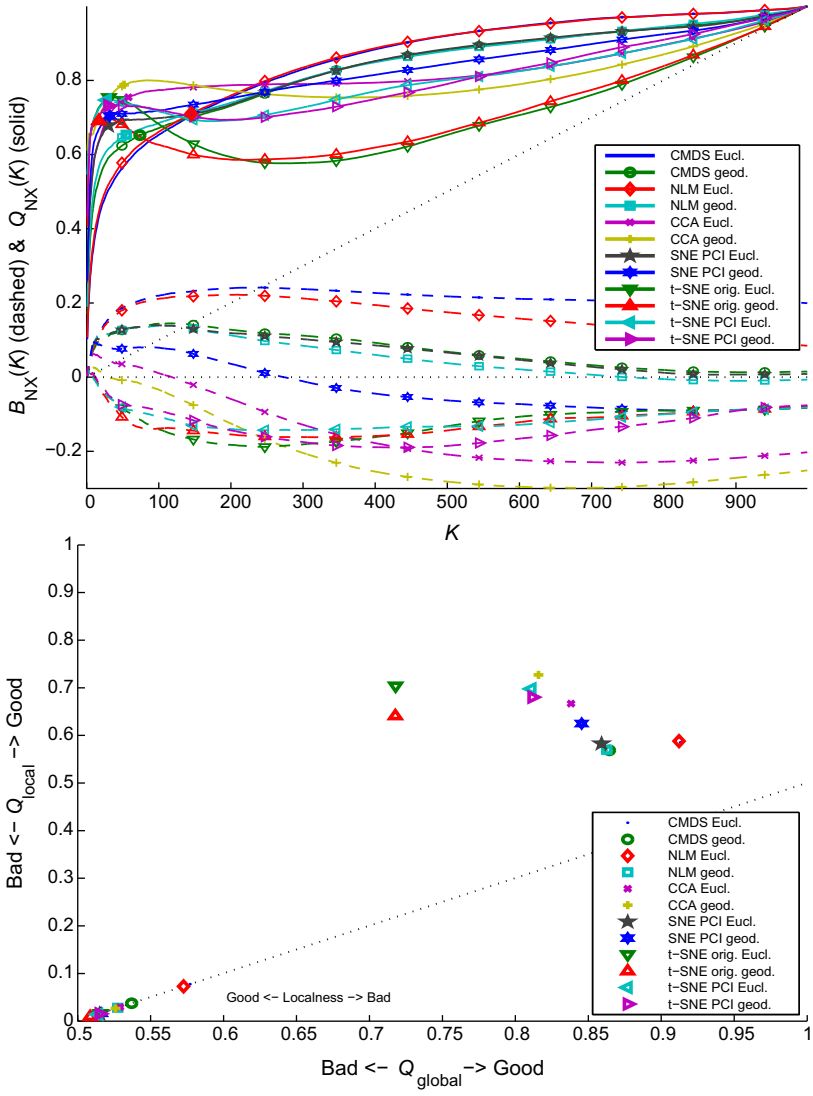

Fig. 5. Quality and behavior curves for embeddings of the noisy Swiss roll in six dimensions.

ments) and $\mathbb{S}_K = \{K+1,\ldots,N-1\}$ (index set for the subsequent ones), the index sets of the upper-left, upper-right, lower-left, and lower-right blocks are given by $\mathbb{UL}_K = \mathbb{F}_K \times \mathbb{F}_K$, $\mathbb{UR}_K = \mathbb{F}_K \times \mathbb{S}_K$, $\mathbb{LL}_K = \mathbb{S}_K \times \mathbb{F}_K$, and $\mathbb{LR}_K = \mathbb{S}_K \times \mathbb{S}_K$. In addition, the block covered by $\mathbb{UL}_K$ can be split into its main diagonal $\mathbb{D}_K = \{(i,i) : 1 \leqslant i \leqslant K\}$ and lower and upper triangles $\mathbb{LT}_K = \{(i,j) : 1 \leqslant j < i \leqslant K\}$ and $\mathbb{UT}_K = \{(i,j) : 1 \leqslant i < j \leqslant K\}$. According to this splitting, K-intrusions and K-extrusions are located in the lower and upper trapezes, respectively (i.e. $\mathbb{LT}_K \cup \mathbb{LL}_K$ and $\mathbb{UT}_K \cup \mathbb{UR}_K$). Hard K-intrusions and K-extrusions are found in $\mathbb{LL}_K$ and $\mathbb{UR}_K$, respectively. In a similar way, mild K-intrusions and K-extrusions are counted in the triangles $\mathbb{LT}_K$ and $\mathbb{UT}_K$, respectively.

3. Weighted and non-weighted rank-based quality criteria

The co-ranking matrix contains all the necessary information about how ranks are preserved in a given low-dimensional representation, but its readability is rather poor. To overcome this issue, most existing rank-based criteria summarize the information by considering the various blocks mentioned in the previous section. The general approach consists in computing weighted sums over some blocks, for a given value of K. Criteria usually come by pair, in order to account for what happens on both sides of $\mathbf{Q}$'s main diagonal. For instance, the trustworthiness and continuity (Venna and Kaski, 2001; Venna, 2007) (T&C) focus on the blocks $\mathbb{LL}_K$ and $\mathbb{UR}_K$, respectively, whereas the mean relative rank errors (Lee and Verleysen, 2007) (MRREs) cover the overlapping blocks $\mathbb{UL}_K \cup \mathbb{LL}_K$ and $\mathbb{UL}_K \cup \mathbb{UR}_K$, respectively (Lee and Verleysen, 2009). The T&C as well as the MRREs rely on a weighting that raises normalization issues (Lee and Verleysen, 2008a). For criteria that involve blocks $\mathbb{LL}_K$ and $\mathbb{UR}_K$, a weighting turns out to be necessary because the co-ranking matrix is such that

$$\sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{UR}_K} q_{kl} = KN \tag{2}$$

and

$$\sum_{(k,l)\in\mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in\mathbb{UR}_K} q_{kl}. \tag{3}$$

Formally, this can also be demonstrated by observing that $\mathbf{Q}$ is a sum of N permutation matrices, whose row-wise as well as column-wise sums are all equal to one (Lee and Verleysen, 2008a). Hence, without an appropriate weighting of the terms in the left and right sums in (3), defining a pair of criteria makes no sense: their values over blocks $\mathbb{LL}$ and $\mathbb{UR}$ are equal. On the other hand, any weighting scheme turns out to involve a somewhat arbitrary choice.

In contrast to the above-mentioned criteria, the LCMC covers a single block of $\mathbf{Q}$, namely $\mathbb{UL}_K$. This eliminates the need for any weighting, at the expense of loosing the other criteria's ability to distinguish between intrusions and extrusions. Such drawback is easily overcome by the pair of criteria proposed in (Lee and Verleysen, 2008a, 2009). They are defined as

$$Q_{\mathrm{NX}}(K) = \frac{1}{KN} \sum_{(k,l)\in\mathbb{UL}_K} q_{kl} \tag{4}$$

and

$$B_{\mathrm{NX}}(K) = \frac{1}{KN}\left(\sum_{(k,l)\in\mathbb{UT}_K} q_{kl} - \sum_{(k,l)\in\mathbb{LT}_K} q_{kl} \right). \tag{5}$$

The first criterion assesses the overall quality of the embedding, it varies between 0 and 1, and measures the preservation of K-ary neighborhoods in a straightforward way. There is a close relationship with the LCMC, which can be written as

$$\mathrm{LCMC}(K) = Q_{\mathrm{NX}}(K) - \frac{K}{N-1}, \tag{6}$$

where the second term is a baseline that accounts for the expected overlap between the initial K-ary neighborhoods and those in a random embedding (Chen, 2006; Lee and Verleysen, 2008a). The second proposed criterion is the difference between the rates of mild K-intrusions and mild K-extrusions. By virtue of equality (3), it also corresponds to the difference between all (hard and mild) K-intrusions and K-extrusions. Hence, the sign of $B_{\mathrm{NX}}(K)$ indicates the 'behavior' of the considered embedding, that is, it indicates whether the embedding is rather intrusive or extrusive.

Fig. 3 shows a simple example of how the proposed quality criteria can be used. The data set consists of 1000 points uniformly sampled from a (hollow) unit sphere. As this manifold is intrinsically two-dimensional, we attempt to embed it in a plane with two different methods, namely Sammon's nonlinear mapping (Sammon, 1969) and curvilinear component analysis (Demartines and Hérault, 1997). The plot shows $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ with respect to K. Baselines are given for both criteria (zero for $B_{\mathrm{NX}}(K)$ and $K/(N-1)$ for $Q_{\mathrm{NX}}(K)$). Looking at the curves for $Q_{\mathrm{NX}}(K)$ shows that CCA better succeeds than NLM in embedding the sphere in a two-dimensional space (CCA's curve is noticibly higher). This better result stems from the ability of CCA to 'tear' the sphere and to embed two adjacent half spheres. In contrast, NML crushes and superimposes the two hemispheres. The opposite signs of $B_{\mathrm{NX}}(K)$ account for this fundamental behavior difference.

Table 2
Scalar quality criteria derived from the curves in Fig. 5, for the noisy Swiss roll in six dimensions. Methods are ranked according to Q_{local} (ranks are between parentheses).

	Q_{avg}	B_{avg}	L	Q_{global}	Q_{local}	
CMDS Eucl.	0.8627	0.2132	0.8468	0.9131	0.5851	(9)
CMDS geod.	0.8428	0.0713	0.9269	0.8645	0.5686	(12)
NLM Eucl.	0.8654	0.1650	0.8559	0.9122	0.5881	(8)
NLM geod.	0.8467	0.0493	0.9459	0.8626	0.5697	(11)
CCA Eucl.	0.8283	−0.1540	0.9429	0.8382	0.6669	(5)
CCA geod.	0.8112	−0.2178	0.9489	0.8158	0.7270	(1)
SNE Eucl.	0.8510	0.0658	0.9710	0.8591	0.5827	(10)
SNE geod.	0.8385	−0.0339	0.9690	0.8454	0.6248	(7)
tSNE Eucl.	0.7174	−0.1236	0.9700	0.7179	0.7040	(2)
tSNE geod.	0.7164	−0.1234	0.9840	0.7177	0.6411	(6)
tSNE Eucl. PCI	0.8079	−0.1151	0.9730	0.8111	0.6978	(3)
tSNE geod. PCI	0.8078	−0.1358	0.9710	0.8117	0.6803	(4)

4. Scalar quality criteria

Interpreting the quality criteria such as those described in Eqs. (4) and (5) and illustrated in Fig. 3 raises two questions:

- How can the user easily figure out which embedding among the compared ones is performing the best?
- Which is the optimal value of K to be looked at?

These two questions turn out to be closely related to the scale issue that underlies the field of dimensionality reduction. As most manifolds cannot be embedded in a low-dimensional space without being somewhat distorted, we have to decide which properties are local and which are global (Saul et al., 2003; Roweis et al., 2002). This distinction allows the DR methods to give a higher priority to the preservation of local properties and to relax the requirements about the global ones. For that purpose, most DR methods have a scale parameter that can be for instance:

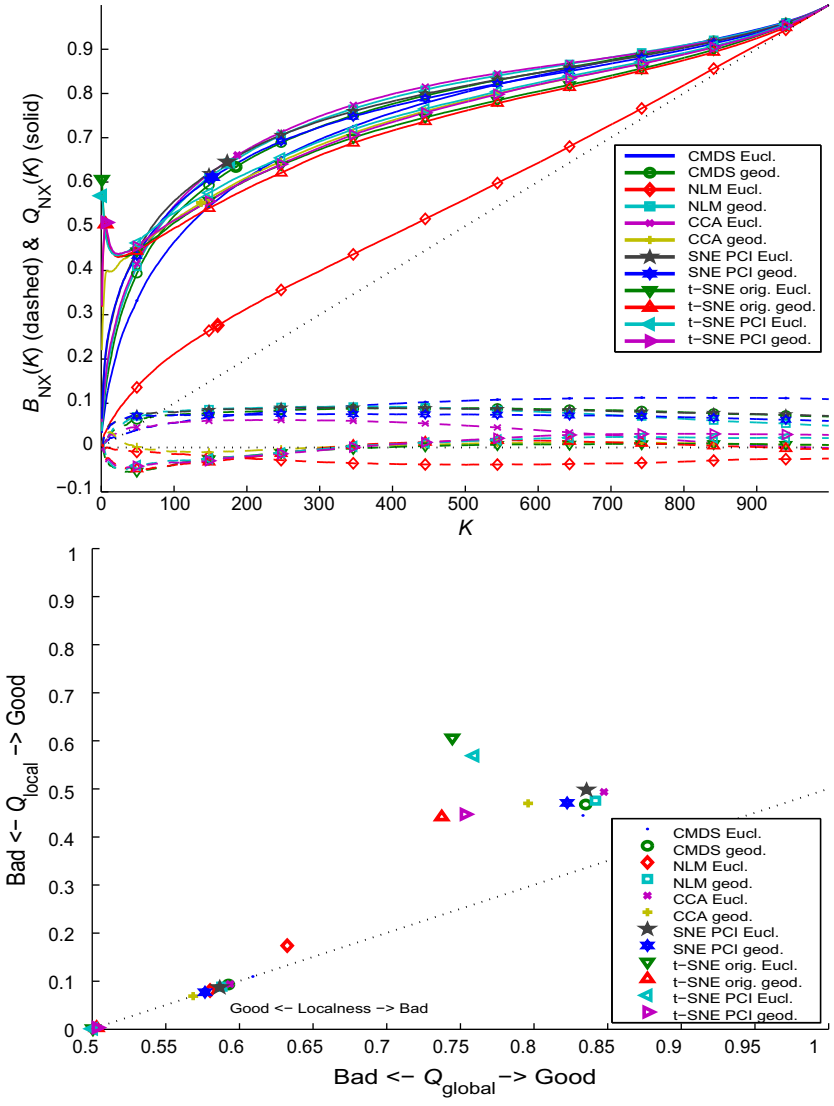

Fig. 6. Quality and behavior curves for embeddings of 1000 images taken from the MNIST database of handwritten digits.

- a number of neighbors (in methods such as Isomap or LLE, which involve K-ary neighborhoods),
- a neighborhood width or radius (such as in CCA and SOMs), or
- a more complex parameterization (such as the perplexity in tSNE).

If local properties are more important than global ones, we deduce that the left part of the curve representing $Q_{NX}(K)$ is likely to be more important than the right part. A good DR method should thus yield a high $Q_{NX}(K)$ for low values of K. Of course, the same method will perform even better if it keeps the curve as high as possible for all values of K. For this reason, quality criteria $Q_{NX}(K)$ and $B_{NX}(K)$ can be summarized in an obvious way by looking at their average values

$$Q_{avg} = \frac{1}{N-1}\sum_{K=1}^{N-1} Q_{NX}(K) \tag{7}$$

and

$$B_{avg} = \frac{1}{N-1}\sum_{K=1}^{N-1} B_{NX}(K). \tag{8}$$

These quantities range from 0 to 1, and from −1 to 1, respectively. They indicate how well a DR method perform, regardless of the scale. For a perfect embedding, we would have $Q_{avg} = 1$ and $B_{avg} = 0$. Undoubtedly these quantities convey an interesting piece of information but they give the same importance to all points of the curves. Hence, they fail to reflect the emphasis that should be put on the preservation of small ranks, which corresponds to the left part of the curves. In the case of $Q_{NX}(K)$, we can split the curve into left and right parts by looking at

$$K_{max} = \arg\max_K \mathrm{LCMC}(K) = \arg\max_K \left(Q_{NX}(K) - \frac{K}{N-1} \right), \tag{9}$$

which gives the neighborhood size for which some method or embedding performs best as compared to a random embedding. Since $Q_{NX}(K)$ trivially attains its maximum for $K = N-1$, baseline $K/(N-1)$ corresponding to neighborhood overlap in a random embedding must be subtracted from $Q_{NX}(K)$. Starting from K_{max}, we can consider a 'localness' indicator defined as

$$L = \frac{N - K_{max}}{N-1}, \tag{10}$$

which assesses how local the best performance is; L varies between $1/(N-1)$ (nonlocal at all) and 1 (fully local). Two other quantities of interest are the average values of $Q_{NX}(K)$ below and above of K_{max}, which are written as

$$Q_{\text{local}} = \frac{1}{K_{\max}} \sum_{K=1}^{K_{\max}} Q_{\text{NX}}(K) \tag{11}$$

and

$$Q_{\text{global}} = \frac{1}{N - K_{\max}} \sum_{K=K_{\max}}^{N-1} Q_{\text{NX}}(K), \tag{12}$$

respectively. Like L, Q_{local} and Q_{global} range from 0 (worst) to 1 (best). They also own the advantage of being scalar without relying on a value of K (arbitrarily) fixed by the user. The value of K is automatically determined by $K_{\max}$. In the case of the hollow sphere manifold, the values corresponding to the curves in Fig. 3 are reported in Table 1.

We suggest that any method or embedding be assessed as follows. First, we advise looking at Q_{local}. The preservation of small neighborhoods emerges as a consensus in the domain of dimensionality reduction (Lee and Verleysen, 2007; Saul et al., 2003; Roweis et al., 2002; Venna and Kaski, 2006) and is thus of prime importance indeed. In case of a tie, the embedding with the highest value of Q_{global} wins. Eventually, L gives a clue about the relative size for which K-ary neighborhoods are best preserved. The last three criteria can be summarized in a simple diagram where markers are plotted for each embedding, with coordinates $[Q_{\text{global}}, Q_{\text{local}}]^T$. Such a diagram is shown in Fig. 4 in the case of the hollow sphere. In order to visualize L within the same diagram, we consider the line that corresponds to random embeddings for a varying value of $K_{\max}$. In this particular case, the ordinate is given by

$$Q_{\text{local}} = \frac{1}{2}\left(\frac{1}{N-1} + \frac{K_{\max}}{N-1}\right) = \frac{N+1}{2(N-1)} - \frac{L}{2}, \tag{13}$$

whereas the corresponding abscissa is

$$Q_{\text{global}} = \frac{1}{2}\left(\frac{K_{\max}}{N-1} + \frac{N-1}{N-1}\right) = \frac{2N-1}{2(N-1)} - \frac{L}{2}. \tag{14}$$

Additional markers for each embedding can then be plotted on this line, according to their respective value of L. The closer to the bottom left corner the marker lies, the higher L is. Furthermore, the horizontal and vertical shifts between the two markers associated with an embedding also convey some information. They indicate how the considered embedding improves Q_{local} and Q_{global} with respect to a random embedding that has the same value of $K_{\max}$.

As to the embeddings of the hollow sphere, CCA outperforms the NLM (CCA's main marker is higher than NLM's one). CCA also achieves a better preservation of large neighborhoods (CCA's main marker is on the right of NLM's one). Finally, the secondary markers located on the baseline indicate that CCA's value of localness L is higher than NLM's one (CCA's secondary marker is closer to the bottom left color). The next section presents comparison with more DR methods on more difficult data sets.

5. Experiments

This section aims at embedding several data sets in a two-dimensional space, for visualization purposes, regardless of the intrinsic data dimensionality. Several methods are used and compared with the proposed quality criteria.

5.1. Methods

The experiments compare the following methods:

- Classical metric multidimensional scaling (Young and Householder, 1938; Torgerson, 1952) (CMDS).
- Sammon's nonlinear mapping (Sammon, 1969) (NLM).
- Curvilinear component analysis (Demartines and Hérault, 1997; Hérault et al., 1999) (CCA).
- Stochastic neighbor embedding (Hinton and Roweis, 2003) (SNE).
- t-Distributed stochastic neighbor embedding (van der Maaten and Hinton, 2008) (tSNE).

Two versions of tSNE are compared. The first one is the implementation provided by the authors of (van der Maaten and Hinton, 2008). The second version relies on a simpler gradient descent (without momentum and 'early exaggeration'). Moreover, it does not randomly initialize the embedding as in the first implementation. Instead, scaled principal components are used.[2] The implementation of SNE relies on the same initialization. The NLM and CCA are initialized with principal components as well.

All methods are used with both Euclidean distances and geodesic ones (Tenenbaum, 1998; Bernstein et al., 2000). The geodesic distance are approximated by computing shortest paths in the Euclidean graph that is associated with 6-ary neighborhoods. Combining CMDS and CCA with geodesic distances amounts to implementing Isomap (Tenenbaum et al., 2000) and CDA (Lee et al., 2000; Lee and Verleysen, 2004), respectively.

Parameters of the various DR methods are set to typical values, with no further optimization, as the point of this paper is to illustrate the use of quality critria, not to claim the superiority of one or another method.

Table 3
Scalar quality criteria derived from the curves in Fig. 6, for the MNIST database of handwritten digits. Methods are ranked according to Q_{local} (ranks are between parentheses).

	Q_{avg}	B_{avg}	L	Q_{global}	Q_{local}	
CMDS Eucl.	0.7486	0.0939	0.7828	0.8332	0.4446	(10)
CMDS geod.	0.7672	0.0801	0.8158	0.8351	0.4675	(8)
NLM Eucl.	0.5593	−0.0294	0.8408	0.6324	0.1740	(12)
NLM geod.	0.7774	0.0767	0.8248	0.8417	0.4756	(5)
CCA Eucl.	0.7815	0.0389	0.8138	0.8475	0.4936	(4)
CCA geod.	0.7515	0.0051	0.8639	0.7960	0.4698	(7)
SNE Eucl.	0.7774	0.0820	0.8278	0.8356	0.4981	(3)
SNE geod.	0.7688	0.0709	0.8478	0.8225	0.4709	(6)
tSNE Eucl.	0.7445	−0.0046	1.0000	0.7445	0.6060	(1)
tSNE geod.	0.7358	−0.0026	0.9950	0.7373	0.4419	(11)
tSNE Eucl. PCI	0.7594	0.0048	1.0000	0.7594	0.5690	(2)
tSNE geod. PCI	0.7513	0.0075	0.9950	0.7529	0.4472	(9)

5.2. Data sets and results

The first data set contains a sample of 1000 points drawn from a Swiss roll (Tenenbaum et al., 2000) with uniform distribution. Its equation is written as

$$\xi = \left[\sqrt{u}\cos(3\pi\sqrt{u}), \sqrt{u}\sin(3\pi\sqrt{u}), \pi v, 0, 0, 0\right]^T, \tag{15}$$

where random parameters u and v have uniform distributions between 0 and 1. The three last coordinates are kept constant and Gaussian noise with standard deviation 0.1 is added to all six dimensions. Fig. 5 and Table 2 summarize the results. Embedding

[2] As tSNE involves nonscaled Student's t distributions in the embedding space, it is scaling invariant, meaning that scaled data always lead to the same embedding. For this reason, the initialization must rely on whitened (that is, nonscaled) components instead of principal components. An additional scaling factor (1e−4) ensures that the embedded data points are not initialized too far away from each other (this prevents the gradient descent to get stuck in poor local minima).

the Swiss roll provided as a first data set clearly entails some difficulties: the variance of the noise that pollutes the six dimensions is quite high. The values of Q_{local} in the last column of Table 2 provide a ranking of the methods. Geodesic distances improve the result of SNE and CCA; all other methods work better with Euclidean distances. The best methods are those that provide extrusive embeddings (B_{avg} is negative). The worst methods are intrusive but tend to better preserve large neighborhoods (the values of Q_{global} are higher). Strong correlations exist between Q_{avg} and Q_{global} and between L and Q_{local}. Initializing tSNE with principal components slightly decreases Q_{local}. However, a significantly larger value of Q_{global} compensates for this loss.

The second data set includes 1000 images from the MNIST digit database (LeCun et al., 1998). Each image is 28 pixels wide and 28 pixels high, leading to 784-dimensional vectors after concatenation. The first 100 images associated with each digit from 0 to 9 are gathered in the data set. The results are shown in Fig. 6 and Table 3. The subset of the MNIST database is embedded best by tSNE, which

Fig. 7. Some faces randomly drawn from the database.

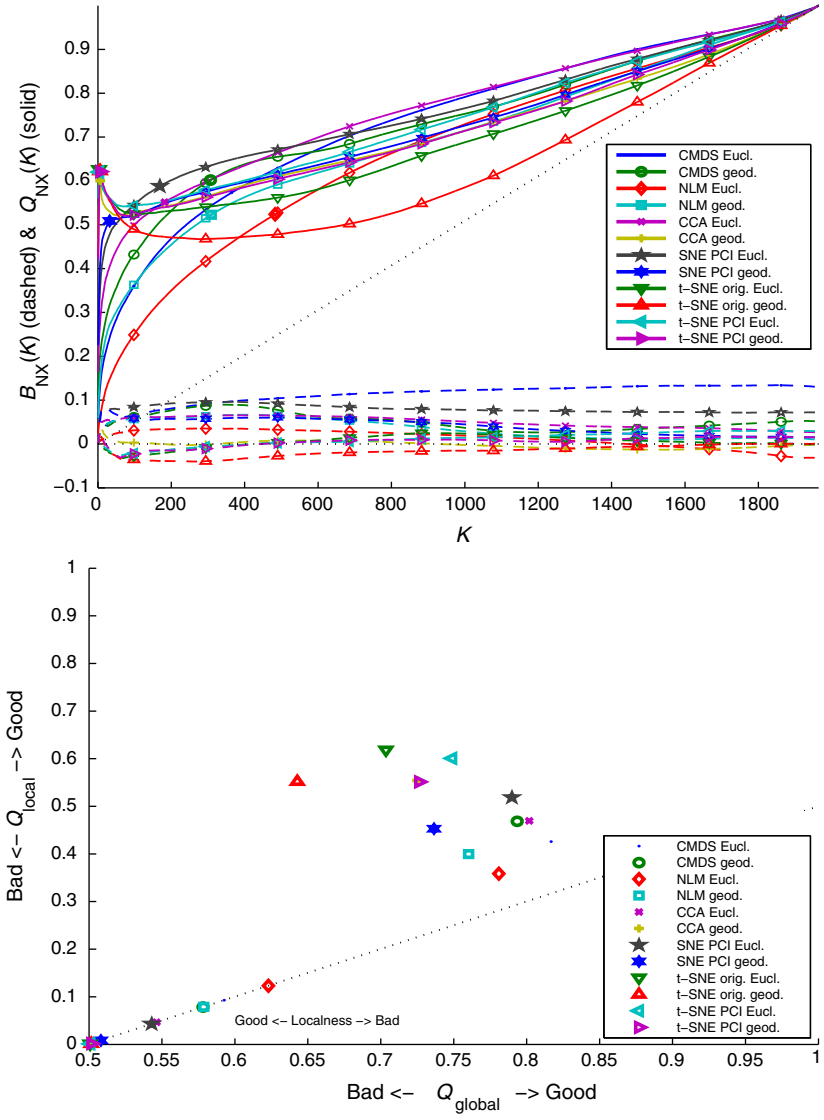

Fig. 8. Quality and behavior curves for embeddings of B. Frey's faces bank.

Table 4
Scalar quality criteria derived from the curves in Fig. 8, for B. Frey's faces. Methods are ranked according to Q_{local} (ranks are between parentheses).

	Q_{avg}	B_{avg}	L	Q_{global}	Q_{local}	
CMDS Eucl.	0.7445	0.1140	0.8152	0.8168	0.4259	(10)
CMDS geod.	0.7428	0.0512	0.8442	0.7936	0.4685	(8)
NLM Eucl.	0.6772	0.0129	0.7546	0.7809	0.3585	(12)
NLM geod.	0.7036	0.0348	0.8432	0.7602	0.3997	(11)
CCA Eucl.	0.7711	0.0498	0.9078	0.8018	0.4695	(7)
CCA geod.	0.7245	−0.0022	0.9975	0.7249	0.5543	(3)
SNE Eucl.	0.7667	0.0802	0.9145	0.7899	0.5187	(6)
SNE geod.	0.7320	0.0411	0.9842	0.7365	0.4529	(9)
tSNE Eucl.	0.7035	0.0055	0.9990	0.7036	0.6184	(1)
tSNE geod.	0.6425	−0.0162	0.9975	0.6428	0.5521	(4)
tSNE Eucl. PCI	0.7488	0.0116	0.9985	0.7491	0.6009	(2)
tSNE geod. PCI	0.7257	0.0051	0.9975	0.7262	0.5514	(5)

is known to perform very well with clustered and very high-dimensional data (van der Maaten and Hinton, 2008). The version of tSNE initialized with principal components takes the second place and slightly improves Q_{global}. Though usually successful in manifold learning, geodesic distances prove to be useless with this 10-cluster data set. Sammon's NLM performs badly, especially with Euclidean distances: any two-dimensional embedding requires inter-cluster distances to be distorted in a way that is incompatible with the weighting scheme of its objective function.

The third data set contains 1965 pictures of B.J. Frey's face (Roweis and Saul, 2000). Each face is 20 pixels wide and 28 pixels high, leading to 560-dimensional vectors after concatenation. Some face poses are illustrated in Fig. 7. Fig. 8 and Table 4 summarize the results. As the MNIST data set, Frey's face bank contains vectorized images. The dimensionality is very high as well, although the data cloud owns a different structure. Since the face pictures are drawn from a movie featuring the same person, there are smooth transitions between the various face expressions. In other words, clusters of the dataset (if any) are likely to be distributed on a smooth manifold. As a consequence, one expects geodesic distances to be useful for distance-preserving methods. Values of Q_{local} for CMDS, NLM, and CCA confirm this hypothesis. In contrast, geodesic distances do not improve the results of similarity-preserving methods such as SNE and tSNE. With a principal component initialization, tSNE yields a higher value of Q_{global} than with a random initialization, at the expense of a small decrease of Q_{local}.

6. Conclusions

The question of quality assessment for dimensionality reduction methods has remained unanswered for a long time. Recently, several publications have proposed quality criteria that are based on ranks and neighborhoods. These are for instance the trustworthiness and continuity, the mean relative rank errors, the local continuity metacriterion, and the quality and behavior criteria. Relying on ranks rather than distances makes these criteria more pertinent, as ranks are almost invariant to the dilations or contractions that are often required to embed complex data sets in low-dimensional spaces. Yet, these criteria all leaves the user with a free parameter: the observation scale, that is, the size of the K-ary neighborhoods to be considered.

This paper suggests that information provided by some of these scale-dependent criteria be summarized into a single scalar value. For this purpose, we first compute the local continuity metacriterion and the closely related quality criterion $Q_{\text{NX}}(K)$ for all admissible values of K. Next, for a given embedding, we determine the value of K where the local continuity metacriterion attains its maximum value. This splits the range of K into two intervals. Averaging $Q_{\text{NX}}(K)$ over both intervals yields Q_{local} and Q_{global}, which assess the preservation of small and large neighborhoods, respectively. We suggest Q_{local} as a unique and scalar quality criterion, in agreement with the widely admitted consensus that dimensionality reduction should focus on the preservation of local data properties.

A quantity such as Q_{local} obviously inherits the main advantages and shortcomings of the rank-based criteria it is based upon, namely $Q_{\text{NX}}(K)$ and the local continuity metacriterion. In spite of their qualities, ranks that come out of a distance sorting process still depend in a straightforward way on some underlying metric. Rank-based criteria leave this responsibility to the user. On the positive side, Q_{local} elegantly circumvents the question of the observation scale. The user is provided with a single figure that allows him/her to compare embeddings or DR methods in a straightforward way.

References

Bauer, H.-U., Pawelzik, K., 1992. Quantifying the neighborhood preservation of self-organizing maps. IEEE Trans. Neural Networks 3, 570–579.

Belkin, M., Niyogi, P., 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15 (6), 1373–1396.

Bengio, Y., Vincent, P., Paiement, J.-F., Delalleau, O., Ouimet, M., Le Roux, N., 2003. Spectral clustering and kernel PCA are learning eigenfunctions. Tech. rep. 1239, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, Montréal.

Bernstein, M., de Silva, V., Langford, J., Tenenbaum, J., 2000. Graph approximations to geodesics on embedded manifolds. Tech. rep., Stanford University, Palo Alto, CA.

Brand, M., Huang, K., 2003. A unifying theorem for spectral embedding and clustering. In: Bishop, C., Frey, B. (Eds.), Proc. Internat. Workshop on Artificial Intelligence and Statistics (AISTATS'03).

Chen, L., 2006. Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis. Ph.D. Thesis, University of Pennsylviana.

Chen, L., Buja, A., 2009. Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. J. Amer. Statist. Assoc. 101 (485), 209–219.

Demartines, P., Hérault, J., 1993. Vector Quantization and Projection Neural Network. Lecture Notes in Computer Science, vol. 686. Springer-Verlag, New York. pp. 328–333.

Demartines, P., Hérault, J., 1997. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. IEEE Trans. Neural Networks 8 (1), 148–154.

Di Battista, G., Eades, P., Tamassia, R., Tollis, I., 1999. Graph Drawing: Algorithms for the Visualization of Graphs. Prentice-Hall.

Donoho, D., Grimes, C., 2003. Hessian eigenmaps: locally linear embedding techniques for high-dimensional data. In: Proc. National Academy of Arts and Sciences, vol. 100. pp. 5591–5596.

Hérault, J., Jaussions-Picaud, C., Guérin-Dugué, A., 1999. Curvilinear component analysis for high dimensional data representation: I. Theoretical aspects and practical use in the presence of noise. In: Mira, J., Sánchez, J. (Eds.), Proc. IWANN'99, vol. II. Springer, Alicante, Spain, pp. 635–644.

Hinton, G., Roweis, S., 2003. Stochastic neighbor embedding. In: Becker, S., Thrun, S., Obermayer, K. (Eds.), Advances in Neural Information Processing Systems (NIPS 2002), vol. 15. MIT Press, pp. 833–840.

Jolliffe, I., 1986. Principal Component Analysis. Springer-Verlag, New York, NY.

Kohonen, T., 1982. Self-organization of topologically correct feature maps. Biological Cybernet. 43, 59–69.

Kramer, M., 1991. Nonlinear principal component analysis using autoassociative neural networks. AIChE J. 37 (2), 233–243.

Kruskal, J., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–28.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., 1998. Gradient-based learning applied to document recognition. Proc. IEEE 86 (11), 2278–2324.

Lee, J., Verleysen, M., 2004. Curvilinear distance analysis versus isomap. Neurocomputing 57, 49–76.

Lee, J., Verleysen, M., 2007. Nonlinear Dimensionality Reduction. Springer.

Lee, J., Verleysen, M., 2008a. Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods. In: Saeys, Y., Liu, H., Inza, I., Wehenkel, L., Van de Peer, Y. (Eds.), JMLR Workshop and Conf. Proc. (New Challenges for Feature Selection in Data Mining and Knowledge Discovery), vol. 4. pp. 21–35.

Lee, J., Verleysen, M., 2008b. Rank-based quality assessment of nonlinear dimensionality reduction. In: Verleysen, M. (Ed.), Proc. ESANN 2008, 16th European Symposium on Artificial Neural Networks. d-side, Bruges, pp. 49–54.

Lee, J., Verleysen, M., 2009. Quality assessment of dimensionality reduction: rank-based criteria. Neurocomputing 72 (7–9), 1431–1443.

Lee, J., Lendasse, A., Donckers, N., Verleysen, M., 2000. A robust nonlinear projection method. In: Verleysen, M. (Ed.), Proc. ESANN 2000, 8th European Symposium on Artificial Neural Networks. D-Facto public., Bruges, Belgium, pp. 13–20.

Mao, J., Jain, A., 1995. Artificial neural networks for feature extraction and multivariate data projection. IEEE Trans. Neural Networks 6 (2), 296–317.

Nadler, B., Lafon, S., Coifman, R., Kevrekidis, I., 2006. Diffusion maps, spectral clustering and eigenfunction of Fokker–Planck operators. In: Weiss, Y., Schölkopf, B., Platt, J. (Eds.), Advances in Neural Information Processing Systems (NIPS 2005), vol. 18. MIT Press, Cambridge, MA.

Oja, E., 1991. Data compression, feature extraction, and autoassociation in feedforward neural networks. In: Kohonen, T., Mäkisara, K., Simula, O., Kangas, J. (Eds.), Artificial Neural Networks, vol. 1. Elsevier Science Publishers, B.V., North-Holland, pp. 737–745.

Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290 (5500), 2323–2326.

Roweis, S., Saul, L., Hinton, G., 2002. Global coordination of local linear models. In: Dietterich, T., Becker, S., Ghahramani, Z. (Eds.), Advances in Neural Information Processing Systems (NIPS 2001), vol. 14. MIT Press, Cambridge, MA.

Saerens, M., Fouss, F., Yen, L., Dupont, P., 2004. The principal components analysis of a graph, and its relationships to spectral clustering. In: Proc. 15th European Conf. on Machine Learning (ECML 2004), pp. 371–383.

Sammon, J., 1969. A nonlinear mapping algorithm for data structure analysis. IEEE Trans. Comput. CC-18 (5), 401–409.

Saul, L., Roweis, S., 2003. Think globally, fit locally: unsupervised learning of nonlinear manifolds. J. Machine Learn. Res. 4, 119–155.

Schölkopf, B., Smola, A., Müller, K.-R., 1998. Nonlinear component analysis as a kernel eigenvalue problem. Neural Comput. 10, 1299–1319.

Shepard, R., 1962. The analysis of proximities: multidimensional scaling with an unknown distance function (parts 1 and 2). Psychometrika 27. pp. 125–140, 219–249.

Takane, Y., Young, F., de Leeuw, J., 1977. Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. Psychometrika 42, 7–67.

Tenenbaum, J., 1998. Mapping a manifold of perceptual observations. In: Jordan, M., Kearns, M., Solla, S. (Eds.), Advances in Neural Information Processing Systems (NIPS 1997), vol. 10. MIT Press, Cambridge, MA, pp. 682–688.

Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290 (5500), 2319–2323.

Torgerson, W., 1952. Multidimensional scaling, I: Theory and method. Psychometrika 17, 401–419.

van der Maaten, L., Hinton, G., 2008. Visualizing data using t-SNE. J. Machine Learn. Res. 9, 2579–2605.

Venna, J., 2007. Dimensionality reduction for visual exploration of similarity structures. Ph.D. Thesis, Helsinki University of Technology, Espoo, Finland.

Venna, J., Kaski, S., 2001. Neighborhood preservation in nonlinear projection methods: an experimental study. In: Dorffner, G., Bischof, H., Hornik, K. (Eds.), Proc. ICANN 2001, Springer, Berlin, pp. 485–491.

Venna, J., Kaski, S., 2006. Local multidimensional scaling. Neural Networks 19, 889–899.

Villmann, T., Der, R., Herrmann, M., Martinetz, T., 1997. Topology preservation in self-organizing feature maps: exact definition and measurement. IEEE Trans. Neural Networks 8 (2), 256–266.

Weinberger, K., Saul, L., 2006. Unsupervised learning of image manifolds by semidefinite programming. Internat. J. Comput. Vision 70 (1), 77–90.

Weinberger, K., Sha, F., Saul, L., 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In: Proc. 21st Internat. Conf. on Machine Learning (ICML-04). Banff, Canada, pp. 839–846.

Young, G., Householder, A., 1938. Discussion of a set of points in terms of their mutual distances. Psychometrika 3, 19–22.