Quality assessment of nonlinear dimensionality reduction based on K-ary neighborhoods

John A. Lee* and Michel Verleysen

Machine Learning Group, Université catholique de Louvain
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium
`http://www.ucl.ac.be/mlg/`
`john.lee@uclouvain.be` & `michel.verleysen@uclouvain.be`

Abstract. Nonlinear dimensionality reduction aims at providing low-dimensional represenions of high-dimensional data sets. Many new methods have been proposed in the recent years, but the question of their assessment and comparison remains open. This paper reviews some of the existing quality measures that are based on distance ranking and K-ary neighborhoods. In this context, the definition of the co-ranking matrix allows the comparison between the ranks in the initial data and the low-dimensional embedding. Rank errors and concepts such as intrusions and extrusions can be associated with different sub-blocks of the co-ranking matrix. Several quality criteria can be cast within this unifying framework and they are shown to involve one or several of these blocks. Following this line, simple criteria are proposed, which quantify two aspects of the embedding quality. A simple experiment illustrates the soundness of the approach.

1 Introduction

Dimensionality reduction (DR) gathers techniques that provide a meaningful low-dimensional representation of a high-dimensional data set. Linear DR is well known, with techniques such as principal component analysis [1] and classical metric multidimensional scaling [2, 3]. On the other hand, nonlinear dimensionality reduction [4] (NLDR) emerged later, with nonlinear variants of multidimensional scaling [5, 6], such as Sammon's nonlinear mapping [7]. For the past twenty five years, this field of research has deeply evolved and after some interest in neural approaches [8–11], the community has recently focused on spectral techniques [12–16]. Modern NLDR encompasses the domain of manifold learning and is also closely related to graph embedding [17] and spectral clustering [18–20].

In the most general setting, dimensionality reduction transforms a set of N high-dimensional vectors, denoted $\mathbf{\Xi} = [\boldsymbol{\xi}_i]_{1\leq i\leq N}$, into N low-dimensional vectors, denoted $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$. In manifold learning, it is assumed that the vectors

* J.A.L. is a Postdoctoral Researcher with the Belgian National Fund for Scientific Research (FNRS).

in Ξ are sampled from a smooth manifold. Under this hypothesis, the goal of NLDR is then to re-embed the manifold in a space of the lowest possible dimensionality, without modifying its topological properties. For this purpose, the embedding theorem [21] can help deduce the lowest embedding dimensionality, which is related to the manifold intrinsic dimensionality [22].

In practice however, neither the intrinsic dimensionality nor the topological properties can easily be identified, starting from a set of points. Therefore, the goal of NLDR is most often to preserve the structure of the data set, which is indicated for instance by some sort of neighborhood relationships [8], such as proximities or similarities. In other words, NLDR provides some low-dimensional representation that is meaningful in some sense, with respect to those specific relationships. As a well known example, proximities can be obtained by measuring pairwise distances [7, 23] in the data set Ξ, with some metric. Sometimes the coordinates in Ξ are unknown and the collected data consist of pairwise distances. If the data set does not specify all distances, 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 distances are available. For instance, a graph can be used to focus on small neighborhoods [14] or to approximate geodesic distances [13, 24] with weighted shortest paths. This illustrates that NLDR and graph embedding share many similarities.

As a matter of fact, the scientific community has been 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 the 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. Another obvious criterion is the reconstruction error. If a NLDR technique provides us with a mapping $\mathcal{M}$ such that $\mathbf{x} = \mathcal{M}(\boldsymbol{\xi})$, then the reconstruction error can be evaluated as the expectation $E_{\mathrm{rec}} = \mathrm{E}\{(\boldsymbol{\xi} - \mathcal{M}^{-1}(\mathcal{M}(\boldsymbol{\xi})))^2\}$. 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 $\boldsymbol{\xi}_i$). The minimization of the reconstruction error is the approach that is followed by PCA and nonlinear auto-encoders [9, 10]. Still another approach mentioned in the literature consists in using an indirect performance index, such as a classification error (see for instance [25] and other references in [26]). Obviously, this works only for labeled data. Eventually, a last possibility consists in sticking to the intrinsic goal of NLDR and we can try to assess the preservation of the data set structure. Quality assessment then relies on the same principles as those that guide the design of an objective function. However, as the objective function needs to be optimized, it must fulfill some requirements, such as being continuous and differentiable. In contrast, these constraints can be relaxed in the definition of a quality criterion, as it just needs to be evaluated. This opens the way to potentially more 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 [8]; see for instance the topographic product [27] and the topographic function [28]. More recently, new criteria for quality assessment have been proposed, with a broader applicability, such as the trustworthiness and continuity measures [29], the local continuity meta-criterion [30], and the mean relative rank errors [4]. All these criteria analyze what happens in K-ary neighborhoods, for a varying value of K. In practice, these neighborhoods result from the ranking of distance measures. This is a fundamental difference, compared to older quality criteria that classically quantify the preservation of pairwise distances, with a stress function [6, 7].

The first aim of this paper is to review some of these recent rank-based criteria. Next, the definition of a co-ranking matrix [31] allows us to compare them from a theoretical point of view, so that a unifying framework can emerge. Eventually, this framework also provides us with arguments to propose new measures.

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 unifies the different approaches and proposes new ones. Section 5 shows some experimental results. Finally, Section 6 draws the conclusions.

2 Distances, ranks, and neighborhoods

Most NLDR techniques involve distances in a more or less direct way. The symbol δ_{ij} denotes the distance from $\boldsymbol{\xi}_i$ to $\boldsymbol{\xi}_j$ 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.

The rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ in the high-dimensional space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. Similarly, the rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ in the low-dimensional space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. 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 $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ are denoted by $\nu_i^K = \{j : 1 \leq \rho_{ij} \leq K\}$ and $n_i^K = \{j : 1 \leq r_{ij} \leq K\}$, respectively.

The *co-ranking matrix* [31] can then be defined as

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

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 [5]. Historically, this scatterplot has often been used to assess results of multidimensional scaling and related methods [23]; 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 with a Shepard 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). In other words, for values of K such that $r_{ij} \leq K < \rho_{ij}$, the jth vector is an intruder in the K-ary neighborhood n_i^K, with respect to the genuine neighborhood ν_i^K. 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 $r_{ij} < \rho_{ij} \leq K$, whereas the latter is associated with the event $r_{ij} \leq K < \rho_{ij}$. 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. For this purpose, we divide the co-ranking matrix into four blocks that separate the first K rows and columns. If we define $\mathbb{F}_K = \{1, \ldots, K\}$ and $\mathbb{L}_K = \{K+1, \ldots, N-1\}$, the index sets of the upper-left, upper-right, lower-left, and lower-right blocks are $\mathbb{UL}_K = \mathbb{F}_K \times \mathbb{F}_K$, $\mathbb{UR}_K = \mathbb{F}_K \times \mathbb{L}_K$, $\mathbb{LL}_K = \mathbb{L}_K \times \mathbb{F}_K$, and $\mathbb{LR}_K = \mathbb{L}_K \times \mathbb{L}_K$. Similarly, the block covered by $\mathbb{UL}_K$ can be split into its main diagonal $\mathbb{D}_K = \{(i, i) : 1 \leq i \leq K\}$ and lower and upper triangles $\mathbb{LT}_K = \{(i, j) : 1 < i \leq K \text{ and } j < i\}$ and $\mathbb{UT}_K = \{(i, j) : 1 \leq i < K \text{ and } j > i\}$. According to this division, 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 the blocks $\mathbb{LL}_K$ 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 Review of quality criteria

This section reviews some of the recently published criteria that rely on ranks and K-ary neighborhoods. Beside the definition found in the literature, we give an equivalent expression in terms of the co-ranking matrix.

The trustworthiness and continuity (T&C) measures [29, 32] are defined as:

$$W_{\mathrm{T}}(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in n_i^K \setminus \nu_i^K} (\rho_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{LL}_K} (k - K) q_{kl} \ , \quad (2)$$

$$W_{\mathrm{C}}(K) = 1 - \frac{2}{G_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K \setminus n_i^K} (r_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{UR}_K} (l - K) q_{kl} \ , \quad (3)$$

where the normalizing factor

$$G_K = \begin{cases} NK(2N-3K-1) & \text{if } K < N/2 \\ N(N-K)(N-K-1) & \text{if } K \geq N/2 \end{cases} \tag{4}$$

considers the worst case [26], i.e. ranks are reversed in the low-dimensional space and the co-ranking matrix is anti-diagonal. Both the trustworthiness and continuity can theoretically vary between 0 and 1, although the worst case is seldom encountered in practice. Notice that the embedding quality is described by two criteria, which distinguish two types of errors. Faraway vectors that become neighbors increase the trustworthiness error measure $E_{\mathrm{T}}(K)$, whereas neighbors that are embedded faraway from each other increase the continuity error measure $E_{\mathrm{C}}(K)$. As can be seen, the reformulation in terms of the co-ranking matrix shows that the trustworthiness is related to the hard K-intrusions, whereas the continuity involves the hard K-extrusions, with some weighting.

The mean relative rank errors [4] (MRREs) rely on the same principle as the trustworthiness and continuity. They are defined as

$$E_n(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in n_i^K} \frac{|\rho_{ij} - r_{ij}|}{\rho_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{UL}_K \cup \mathbb{LL}_K} \frac{|k-l|}{l} q_{kl} \enspace , \tag{5}$$

$$E_\nu(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K} \frac{|\rho_{ij} - r_{ij}|}{r_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{UL}_K \cup \mathbb{UR}_K} \frac{|k-l|}{k} q_{kl} \enspace , \tag{6}$$

where the normalizing factor $H_K = N \sum_{k=1}^{K} |N - 2k|/k$ considers the worst case, like that of T&C. The differences between the MRREs and the T&C hold in the weighting of the elements q_{kl} and the blocks of $\mathbf{Q}$ that are covered. The MRREs involve the first K rows and colums of $\mathbf{Q}$. Hence, the first error involves all K-intrusions (hard and mild), along with the mild K-extrusions. The second error takes into account all K-extrusion and the mild K-intrusions.

The local continuity meta-criterion [30] (LCMC) is defined as

$$U_{\mathrm{LC}}(K) = \frac{1}{NK} \sum_{i=1}^{N} \left(|n_i^K \cap \nu_i^K| - \frac{K^2}{N-1} \right) = \frac{K}{1-N} + \frac{1}{NK} \sum_{(k,l) \in \mathbb{UL}_K} q_{kl} \enspace , \tag{7}$$

where the subtracted term is a "baseline" that corresponds to the expected overlap between two subsets of K elements out of $N-1$. In contrast to the MRREs and T&C, the LCMC yields a single quantity that is computed over the block $\mathbb{UL}_K$ of $\mathbf{Q}$. Notice also that the elements q_{kl} in the block $\mathbb{UL}_K$ are not weighted in the sum and that the normalization is simpler.

From an intuitive point of view, T&C and MRREs try to detect what goes wrong in a given embedding, whereas the LCMC accounts for things that work well. The prominent strength of T&C and MRREs is their ability to distinguish two sorts of undesired events. On the other hand, in contrast to the LCMC, they cannot directly express the overall performance of an NLDR method by means of a single scalar.

4 Unifying framework

The error and quality measures described in the previous section can be related to the concepts of *precision* and *recall* (P&R) in the domain of information retrieval. The precision is the proportion of relevant items among the retrieved ones, whereas the recall is the proportion of retrieved items among the relevant ones. For rank-based criteria, relevant items are the indices that belong to ν_i^K, whereas n_i^K contains the retrieved indices. The P&R are themselves related to the concepts of false positive and false negative in classification. False positive decrease the precision and false negatives decrease the recall. If we compare the retrieved neighborhoods to the relevant ones, the blocks of $\mathbf{Q}$ covered by $\mathbb{UL}_K$, $\mathbb{LL}_K$, $\mathbb{UR}_K$, and $\mathbb{LR}_K$ contain the true positives, the false positives, the false negatives, and the true negatives, respectively. Hence, the LCMC quantifies the true positives, the T&C focus on the false positives and false negatives, and the MRREs encompass the positives (true and false) and negatives (true and false). Obviously, as n_i^K and ν_i^K have the same size, the numbers of false positives and false negatives are the same. Each element of ν_i^K that is missed in n_i^K (a false negative) is replaced with an incorrect neighbor (a false positive). Formally, as $\mathbf{Q}$ is a sum of N permutation matrices, we can see that $\sum_{l=1}^{N-1} q_{kl} = N$ and $\sum_{k=1}^{N-1} q_{kl} = N$. As we compute ranks starting from N reference points, we have always N kth neighbors. Therefore, we have

$$\sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in\mathbb{UL}_K\cup\mathbb{UR}_K} q_{kl} = KN \text{ and } \sum_{(k,l)\in\mathbb{LL}_K} q_{kl} = \sum_{(k,l)\in\mathbb{UR}_K} q_{kl} \ . \tag{8}$$

This shows that the numbers of hard K-intrusions and hard K-extrusions are equal. As a corollary, without an appropriate weighting of the elements q_{kl}, we would end up with the equalities $W_{\mathrm{T}}(K) = W_{\mathrm{C}}(K)$ and $E_\nu(K) = E_n(K)$. On the other hand, the absence of weighting in the LCMC is obviously not critical.

At this point, we see that the analogy between T&C on one side, and false positives and negatives on the other side, must be interpreted carefully. Hence, T&C do not aim at counting the average *number* of false positives/negatives in K-ary neighborhoods. Instead, the goal consists in estimating *how bad* data vectors are misranked. This suggests that meaningful criteria should be computed on both sides of the diagonal of the co-ranking matrix $\mathbf{Q}$, in order to optimally reveal the dominance of either intrusions or extrusions. For instance, weighted averages that account for all K-intrusions and K-extrusions can be written as

$$W_{\mathrm{N}}^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in\mathbb{LT}_K\cup\mathbb{LL}_K} \frac{(k-l)^v}{k^w} q_{kl} \ , \tag{9}$$

$$W_{\mathrm{X}}^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in\mathbb{UT}_K\cup\mathbb{UR}_K} \frac{(l-k)^v}{l^w} q_{kl} \ , \tag{10}$$

where $C_K = N\sum_{k=1}^{K} \max\{0, (N-2k)^w/k^v\}$. The exponents v and w can be adjusted in order to emphasize large rank differences, relatively to the reference rank. Choosing $v = 1$ and $w = 1$ gives the same weighting as in MRREs, whereas the combination $v = 1$ and $w = 0$ leads to a similar weighting as that of T&C. Looking at the blocks they are covering, the two proposed criteria occupy an intermediate position between T&C and MRREs: they involve more elements than the former, but fewer than the latter.

As a matter of fact, quantities such as $W_{\mathrm{N}}^{v,w}(K)$ and $W_{\mathrm{X}}^{v,w}(K)$ rely on a more or less arbitrary weighting. Based on the observation that the numbers of hard K-intrusions and hard K-extrusions are equal, unweighted averages seem to be useless at first sight. However, if we follow the same idea as that behind the LCMC, we can instead focus on what happens inside K-ary neighborhoods and write [31]

$$U_{\mathrm{N}}(K) = \frac{1}{KN} \sum_{(k,l)\in\mathbb{UT}_K} q_{kl} \ , \qquad U_{\mathrm{X}}(K) = \frac{1}{KN} \sum_{(k,l)\in\mathbb{LT}_K} q_{kl} \ , \tag{11}$$

and

$$U_{\mathrm{P}}(K) = \frac{1}{KN} \sum_{(k,l)\in\mathbb{D}_K} q_{kl} \ . \tag{12}$$

The first two quantities correspond to the fractions of mild K-intrusions and mild K-extrusions, respectively. The quantity $U_{\mathrm{P}}(K)$ indicates the fraction of vectors that keep the same rank in both ν_i^K and n_i^K. The sum of these three fractions is closely related to the LCMC (up to the baseline term); it can be written as

$$Q(K) = U_{\mathrm{P}}(K) + U_{\mathrm{N}}(K) + U_{\mathrm{X}}(K) = U_{\mathrm{LC}}(K) + \frac{K}{N-1} \tag{13}$$

and quantifies the overall quality of an embedding. On the other hand, the difference of the two fractions $U_{\mathrm{N}}(K)$ and $U_{\mathrm{X}}(K)$ can be denoted by

$$B(K) = U_{\mathrm{N}}(K) - U_{\mathrm{X}}(K) \ . \tag{14}$$

This quantity indicates the "behavior" of an NLDR method, that is, whether it tends to produce an "intrusive" ($B(K) > 0$) or "extrusive" ($B(K) < 0$) embedding. Notice that (8) guarantees that $B(K)$ is equal to the difference between the fractions of all K-intrusions and all K-extrusions (both mild and hard ones). This can be formally written as $B(K) = W_{\mathrm{N}}^{0,0} - W_{\mathrm{X}}^{0,0}$.

5 Experiment: the hollow sphere

In order to illustrate the different quality criteria, thousand points are randomly drawn from a simple manifold, namely a hollow sphere whose radius is equal to one. A first data set includes the noisefree points, whereas the second is formed by adding Gaussian noise with standard deviation equal to 0.05 to the same points. Next, the manifold has been embedded in a two-dimensional space with Sammon's nonlinear mapping [7] (NLM) and curvilinear component analysis [23]

(CCA). Notice that we have implemented the version of CCA described in [33], which proves to be more robust against noise. The literature indicates [4, 32] that NLM is known to "crush" the manifold (faraway points can become neighbors), whereas CCA can "tear" the manifold (some close neighbors can be embedded faraway from each other). In other words, this means that NLM tends to produce "intrusive" embeddings whereas CCA rather works in an "extrusive" way.

In order to present results that can be easily compared, the following quantities are displayed: $\{Q(K), B(K)\}$ in Fig. 1, $\{2 - W_n(K) - W_\nu(K), W_n(K) - W_\nu(K)\}$ in Fig. 2, and $\{W_\mathrm{T}(K) + W_\mathrm{C}(K), W_\mathrm{C}(K) - W_\mathrm{T}(K)\}$ in Fig. 3. Each

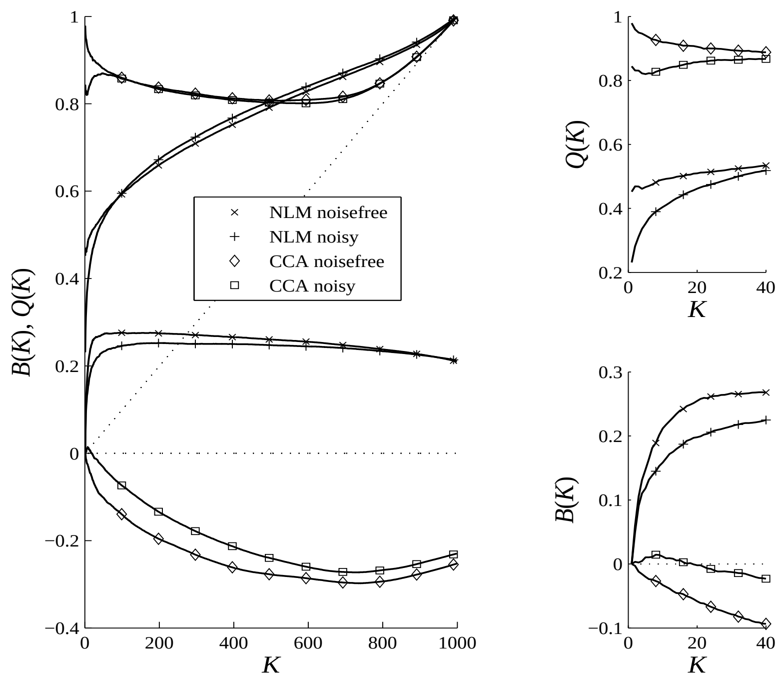

Fig. 1. Quality assessment of the hollow sphere embedding: $Q(K)$ and $B(K)$ for NLM and CCA, for noisefree as well as noisy data.

figure thus includes as many pairs of curves as there are methods to compare. Each pair of curves refers to an overall quality criterion and a behavior indicator, in the same spirit as $Q(K)$ and $B(K)$. In each figure, the left diagram shows the whole curves, for $1 \leq K \leq N - 1$; the upper right diagram focuses on the quality criterion for small values of K, whereas the last one does the same for the behavior indicator. In Fig. 1, the dotted ascending line reprensents the LCMC baseline and highlights the connection with $Q(K)$.

As can be seen, all three pairs of curves show that (i) CCA outperforms NLM and (ii) these two methods have antagonist behaviors, as previously mentioned.

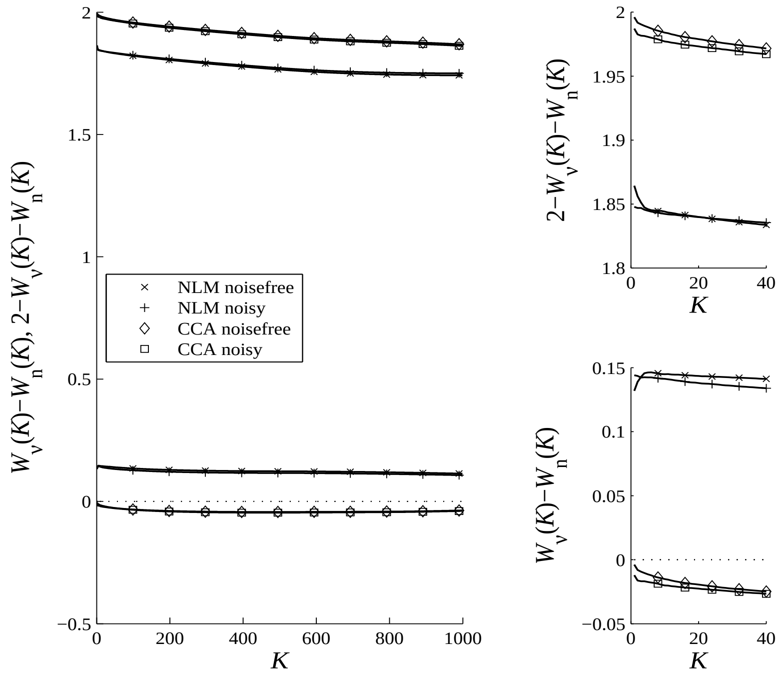

Fig. 2. Quality assessment of the hollow sphere embedding: $2 - W_n(K) - W_\nu(K)$ and $W_n(K) - W_\nu(K)$ for NLM and CCA, for noisefree as well as noisy data.

Looking specifically at quantities that involve a weighting of the co-ranking matrix elements, we can confirm that for small values of K both the MRREs and T&C provide similar results. For larger values, we can see that the common weighting used in the MRREs gives a higher importance to local errors; as a consequence, the curves essentially remain flat when K grows. On the other hand, the curves of T&C drop to zero, because of the particular weighting and since the area of the involved blocks decreases as K grows. This explains why for those criteria the curves of NLM and CCA can rejoin or cross each other as K grows. As to noise, its absence or presence has little influence on the four pairs of weighted averages, although a slight difference can be observed in favor of the noisefree data set.

At this point, an important result is the ability of $Q(K)$ and $B(K)$ to distinguish the antagonist behaviors of NLM and CCA without any (arbitrary) weighting of the co-ranking matrix elements. For instance, $Q(K)$ shows that if CCA succeeds in preserving local neighborhoods better than NLM, this is at the expense of sacrificing the preservation of the global manifold shape. This is illustrated by the crossing of CCA and NLM curves for $K \approx 500$ in Fig. 1. Unweighted averages also clearly identify the effect of noise. For NLM as well as CCA and for small values of K, a marked gap separates the curves associated

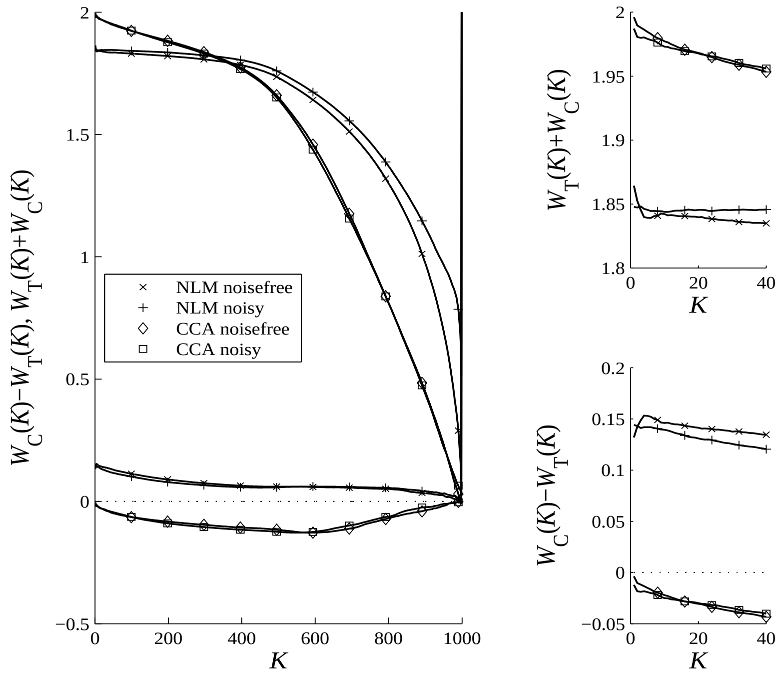

Fig. 3. Quality assessment of the hollow sphere embedding: $W_{\mathrm{T}}(K) + W_{\mathrm{C}}(K)$ and $W_{\mathrm{C}}(K) - W_{\mathrm{T}}(K)$ for NLM and CCA, for noisefree as well as noisy data.

with the noisy and noisefree data sets. This gap then vanishes as K grows. This is expected and corresponds to noise flattening on small scales. In particular, the evolution of $B(K)$ for the noisy data set embedded with CCA conveys interesting information. This method is known to be "extrusive" and it indeed tears the sphere. Locally however, noise must be flattened, what corresponds to an intrusive behavior. Such a behavior reversion is nicely rendered by $B(K)$, not by the other criteria. The explanation resides in the fact that noise flattening generates many small-amplitude intrusions, whereas tearing a manifold generally causes a few large-amplitude extrusions. Hence, depending on the weighting of the rank errors, the contributions of either intrusions or extrusions can dominate. Obviously, weighted averages give too much importance to intrusions or extrusions associated with large rank errors.

6 Conclusions

This paper has reviewed several quality criteria for the assessment of nonlinear dimensionality reduction. All of them rely on distance rankings in both the high- and low-dimensional spaces. The definition of the co-ranking matrix allows us to cast them within a unifying framework. The literature emphasizes the connection of these rank-based criteria with fundamental concepts taken from information retrieval (precision and recall) or classification (false positives and false negatives). Properties of the co-ranking matrix show however that these analogies should however be considered carefully. In constrast, the co-ranking matrix can be interpreted in a similar way as a Shepard diagram. Therefore quality criteria should focus on the rank errors that are distributed on both sides of the co-ranking matrix diagonal, namely intrusions and extrusions. According to this observation we have proposed weighted and unweighted averages that are computed on various blocks or triangles of the co-ranking matrix. Experiments show the soundness of the approach based on the co-ranking matrix. In particular, they show that unweighted averages of the co-ranking matrix elements are sufficient and that any weighting inevitably turns out to be arbitrary. More importantly, weighted averages tend to emphasize some types of embedding errors and can fail to detect others.

References

1. Jolliffe, I.: Principal Component Analysis. Springer-Verlag, New York, NY (1986)
2. Young, G., Householder, A.: Discussion of a set of points in terms of their mutual distances. Psychometrika **3** (1938) 19–22
3. Torgerson, W.: Multidimensional scaling, I: Theory and method. Psychometrika **17** (1952) 401–419
4. Lee, J., Verleysen, M.: Nonlinear dimensionality reduction. Springer (2007)
5. Shepard, R.: The analysis of proximities: Multidimensional scaling with an unknown distance function (parts 1 and 2). Psychometrika **27** (1962) 125–140, 219–249
6. Kruskal, J.: Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika **29** (1964) 1–28
7. Sammon, J.: A nonlinear mapping algorithm for data structure analysis. IEEE Transactions on Computers **CC-18**(5) (1969) 401–409
8. Kohonen, T.: Self-organization of topologically correct feature maps. Biological Cybernetics **43** (1982) 59–69
9. Kramer, M.: Nonlinear principal component analysis using autoassociative neural networks. AIChE Journal **37**(2) (1991) 233–243
10. Oja, E.: Data compression, feature extraction, and autoassociation in feedforward neural networks. In Kohonen, T., Mäkisara, K., Simula, O., Kangas, J., eds.: Artificial Neural Networks. Volume 1. Elsevier Science Publishers, B.V., North-Holland (1991) 737–745
11. Mao, J., Jain, A.: Artificial neural networks for feature extraction and multivariate data projection. IEEE Transactions on Neural Networks **6**(2) (1995) 296–317
12. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural Computation **10** (1998) 1299–1319
13. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for nonlinear dimensionality reduction. Science **290**(5500) (December 2000) 2319–2323
14. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**(5500) (2000) 2323–2326
15. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. Neural Computation **15**(6) (June 2003) 1373–1396

16. Weinberger, K., Saul, L.: Unsupervised learning of image manifolds by semidefinite programming. International Journal of Computer Vision **70**(1) (2006) 77–90 In Special Issue: Computer Vision and Pattern Recognition-CVPR 2004 Guest Editor(s): A. Bobick, R. Chellappa, L. Davis.
17. Di Battista, G., Eades, P., Tamassia, R., Tollis, I.: Graph drawing: Algorithms for the visualization of graphs. Prentice-Hall (1999)
18. Bengio, Y., Vincent, P., Paiement, J.F., Delalleau, O., Ouimet, M., Le Roux, N.: Spectral clustering and kernel PCA are learning eigenfunctions. Technical Report 1239, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, Montréal (July 2003)
19. Saerens, M., Fouss, F., Yen, L., Dupont, P.: The principal components analysis of a graph, and its relationships to spectral clustering. In: Proceedings of the 15th European Conference on Machine Learning (ECML 2004). Volume 3201 of Lecture notes in Artificial Intelligence., Pisa, Italy (2004) 371–383
20. Brand, M., Huang, K.: A unifying theorem for spectral embedding and clustering. In Bishop, C., Frey, B., eds.: Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS'03), Key West, FL (January 2003)
21. Whitney, H.: Differentiable manifolds. Annals of Mathematics **37**(3) (1936) 645–680
22. Fukunaga, K.: Intrinsic dimensionality extraction. In Krishnaiah, P., Kanal, L., eds.: Classification, Pattern Recognition and Reduction of Dimensionality, Volume 2 of Handbook of Statistics. Elsevier, Amsterdam (1982) 347–360
23. Demartines, P., Hérault, J.: Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks **8**(1) (January 1997) 148–154
24. Lee, J., Verleysen, M.: Curvilinear distance analysis versus Isomap. Neurocomputing **57** (March 2004) 49–76
25. Saul, L., Roweis, S.: Think globally, fit locally: Unsupervised learning of nonlinear manifolds. Journal of Machine Learning Research **4** (June 2003) 119–155
26. Venna, J.: Dimensionality reduction for visual exploration of similarity structures. PhD thesis, Helsinki University of Technology, Espoo, Finland (June 2007)
27. Bauer, H.U., Pawelzik, K.: Quantifying the neighborhood preservation of self-organizing maps. IEEE Transactions on Neural Networks **3** (1992) 570–579
28. Villmann, T., Der, R., Herrmann, M., Martinetz, T.: Topology preservation in self-organizing feature maps: Exact definition and measurement. IEEE Transactions on Neural Networks **8**(2) (1997) 256–266
29. Venna, J., Kaski, S.: Neighborhood preservation in nonlinear projection methods: An experimental study. In Dorffner, G., Bischof, H., Hornik, K., eds.: Proceedings of ICANN 2001. Springer, Berlin (2001) 485–491
30. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis. PhD thesis, University of Pennsylviana (July 2006)
31. Lee, J., Verleysen, M.: Rank-based quality assessment of nonlinear dimensionality reduction. In Verleysen, M., ed.: Proceedings of ESANN 2008, 16th European Symposium on Artificial Neural Networks. d-side, Bruges (April 2008) 49–54
32. Venna, J., Kaski, S.: Local multidimensional scaling. Neural Networks **19** (2006) 889–899
33. Hérault, J., Jaussions-Picaud, C., Guérin-Dugué, A.: 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.: Proceedings of IWANN'99. Volume II. Springer, Alicante, Spain (June 1999) 635–644