JMLR: Workshop and Conference Proceedings 4: 21-35 New challenges for feature selection

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

John A. Lee[*] JOHN.LEE@UCLOUVAIN.BE

Michel Verleysen MICHEL.VERLEYSEN@UCLOUVAIN.BE

Machine Learning Group, Université catholique de Louvain,
Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium.
http://www.ucl.ac.be/mlg/

Editor: Saeys et al.

Abstract

Nonlinear dimensionality reduction aims at providing low-dimensional representions of high-dimensional data sets. Many new methods have been recently proposed, 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 comparison of the ranks in the high- and low-dimensional spaces leads to the definition of the co-ranking matrix. Rank errors and concepts such as neighborhood intrusions and extrusions can be associated with different blocks of the co-ranking matrix. The considered quality criteria are then cast within this unifying framework and the blocks they involve are identified. The same framework allows us to propose simpler criteria, which quantify two aspects of the embedding, namely its overall quality and its tendency to favor either intrusions or extrusions. Eventually, a simple experiment illustrates the soundness of the approach.

1. Introduction

Dimensionality reduction (DR) encompasses techniques that can provide a meaningful low-dimensional representation of high-dimensional data. 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). In contrast, nonlinear dimensionality reduction (Lee and Verleysen, 2007) (NLDR) emerged later, with nonlinear variants of multidimensional scaling (Shepard, 1962; Kruskal, 1964), such as Sammon's nonlinear mapping (Sammon, 1969). For the past twenty five years, this field of research has deeply evolved and after some interest in neural approaches (Kohonen, 1982; Kramer, 1991; Oja, 1991; Mao and Jain, 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; Weinberger and Saul, 2006). Modern NLDR includes the domain of manifold learning and is also closely related to graph embedding (Di Battista et al., 1999) and spectral clustering (Bengio et al., 2003; Saerens et al., 2004; Brand and Huang, 2003).

In the most general context, dimensionality reduction transforms a set of N high-dimensional vectors, denoted $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1 \leq i \leq N}$, into N low-dimensional vectors, denoted

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

$\mathbf{X} = [\mathbf{x}_i]_{1 \leq i \leq N}$. In manifold learning, it is assumed that the vectors in $\boldsymbol{\Xi}$ are sampled from a smooth manifold. Under this hypothesis, the goal of NLDR is then to re-embed the manifold in a space of lower dimensionality, without modifying its topological properties. For this purpose, the embedding theorem (Whitney, 1936) can help deduce the lowest embedding dimensionality, which is related to the manifold intrinsic dimensionality (Fukunaga, 1982).

In practice however, neither the intrinsic dimensionality nor the topological properties can be easily identified, starting from a set of points. Therefore, the goal of NLDR is most often to preserve simpler geometrical properties of the data set, which are indicated for instance by some sort of neighborhood relationships (Kohonen, 1982), 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 (Sammon, 1969; Demartines and Hérault, 1997) in data set $\boldsymbol{\Xi}$ with some metric. Sometimes the coordinates in $\boldsymbol{\Xi}$ are unknown and the collected data is already expressed as pairwise distances. If the data set does not specify all distances, then the problem can elegantly be modeled using a graph. In this case, edges are present for known entries of the pairwise distance matrix and 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 (Roweis and Saul, 2000) or to approximate geodesic distances (Tenenbaum et al., 2000; Lee and Verleysen, 2004) 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 designing new NLDR methods and the question of quality assessment remains mostly unanswered. As many 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 one to compare several runs with e.g. different settings, but it 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 $E_{\text{rec}} = \mathrm{E}\{(\boldsymbol{\xi} - \mathcal{M}^{-1}(\mathcal{M}(\boldsymbol{\xi})))^2\}$. The reconstruction error is a universal quality criterion, but it requires the availability of both $\mathcal{M}$ 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 parametric methods, such as PCA and nonlinear auto-encoders (Kramer, 1991; Oja, 1991). Still another approach mentioned in the literature consists in using an indirect performance index. In the case of labeled data, the classification error is a typical choice; see for instance (Saul and Roweis, 2003) and other references in (Venna, 2007). 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 geometrical properties. 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 an avenue to potentially more complex quality criteria that more faithfully assess the preservation of

the data set properties. First attempts in this direction can be found in the particular case of self-organizing maps (Kohonen, 1982); see for instance 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), the local continuity meta-criterion (Chen and Buja, 2006), and the mean relative rank errors (Lee and Verleysen, 2007). All these criteria analyze what happens in K-ary neighborhoods, for a varying size K. In practice, these neighborhoods result from the ranking of distance measures. This is a fundamental difference, compared to older quality criteria that typically quantify the preservation of pairwise distances, by means of a "stress" function (Kruskal, 1964; Sammon, 1969).

The first aim of this paper is to review some of the recently proposed criteria that rely on ranks and K-ary neighborhoods. Next, the definition of a co-ranking matrix (Lee and Verleysen, 2008) allows us to compare them and to establish a unifying framework. Eventually, this framework also provides us with arguments for proposing new criteria that combine both simplicity and efficiency.

This paper is organized as follows. Section 2 introduces the notations for distances, ranks, and neighborhoods. Section 3 reviews several rank-based criteria found in the literature. Section 4 unifies the different approaches and proposes new criteria. Section 5 gives a few experimental results. Finally, Section 6 draws the conclusions.

2. Distances, ranks, and neighborhoods

Most NLDR techniques involve distances in some way or another. 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* (Lee and Verleysen, 2008) can then be defined as

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

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 (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 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, as illustrated in Fig. 1. 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 $\mathbb{F}_K = \{1, \ldots, K\}$ 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 given by $\mathbb{UL}_K = \mathbb{F}_K \times \mathbb{F}_K$, $\mathbb{UR}_K = \mathbb{F}_K \times \mathbb{L}_K$, $\mathbb{LL}_K = \mathbb{L}_K \times \mathbb{F}_K$, 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.

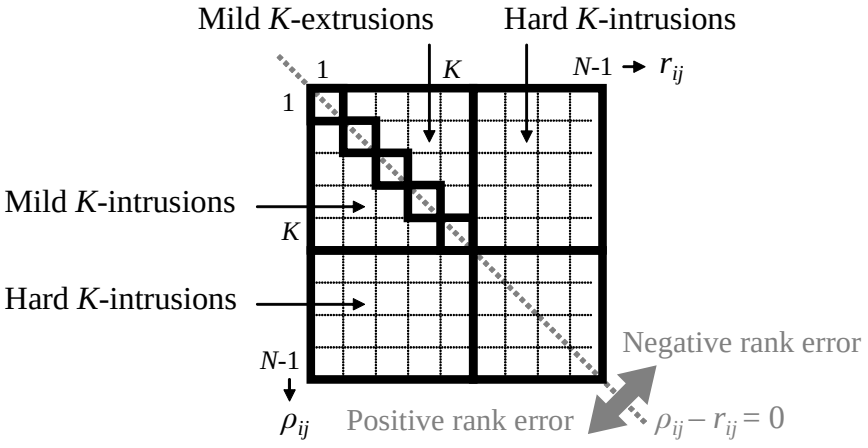

Figure 1: Block division of the co-ranking matrix, showing the different types of intrusions and extrusions, and their relationship with the rank error.

3. Review of existing quality criteria

This section reviews some 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 (Venna and Kaski, 2001, 2006) are defined as:

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

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

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 (Venna, 2007), i.e. ranks are reversed in the low-dimensional space and the co-ranking matrix is anti-diagonal. Both T&C 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 decrease the trustworthiness, whereas neighbors that are embedded faraway from each other decrease the continuity. Eventually, 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 (Lee and Verleysen, 2007) (MRREs) rely on the same principle as the trustworthiness and continuity. They are defined as

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

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

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-extrusions and the mild K-intrusions.

The local continuity meta-criterion (Chen and Buja, 2006) (LCMC) is defined as

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

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 trivial.

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. But, 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 (Venna and Kaski, 2007). 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 positives 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 $M_{\mathrm{T}}(K) = M_{\mathrm{C}}(K)$ and $W_\nu(K) = W_n(K)$. 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 with caution. Hence, T&C do not aim at counting the *fraction* of false positives/negatives in K-ary neighborhoods. Instead, their goal rather consists in estimating *how bad* data vectors are misranked, by means of some weighting. 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. Keeping the principle of weighted fractions, we can account for all

K-intrusions and K-extrusions by defining

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

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

where $C_K = N \sum_{k=1}^{K} \max\{0, N-2k\}^w / k^v$ normalizes with respect to the worst case. 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. On the other hand, based on the observation that the numbers of hard K-intrusions and hard K-extrusions are equal, unweighted fractions seem to be useless at first sight. However, if we follow the same idea as that behind the LCMC, we can focus on what happens inside K-ary neighborhoods and write (Lee and Verleysen, 2008)

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

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

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

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

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

This quantity indicates the "behavior" of an NLDR method, that is, whether it tends to produce an "intrusive" ($B_{\mathrm{NX}}(K) > 0$) or "extrusive" ($B_{\mathrm{NX}}(K) < 0$) embedding. Notice that (8) guarantees that $B_{\mathrm{NX}}(K)$ 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 $B_{\mathrm{NX}}(K) = W_{\mathrm{N}}^{0,0} - W_{\mathrm{X}}^{0,0}$.

In the same spirit as $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$, the distinction between overall quality and behavior can be extended to the other criteria. For this purpose, one can consider the following quantities:

- $Q_{\mathrm{TC}}(K) = M_{\mathrm{T}}(K) + M_{\mathrm{C}}(K)$ and $B_{\mathrm{TC}}(K) = M_{\mathrm{C}}(K) - M_{\mathrm{T}}(K)$ for T&C,
- $Q_{n\nu}(K) = 2 - W_n(K) - W_\nu(K)$ and $B_{\mathrm{n}\nu}(K) = W_n(K) - W_\nu(K)$ for the MRREs, and
- $Q^{v,w}_{\mathrm{wNX}}(K) = 2 - W^{v,w}_{\mathrm{N}}(K) - W^{v,w}_{\mathrm{X}}(K)$ and $B^{v,w}_{\mathrm{wNX}}(K) = W^{v,w}_{\mathrm{N}}(K) - W^{v,w}_{\mathrm{X}}(K)$ for the weighted fractions of intrusions and extrusions.

Curves with respect to the neighborhood size K can be drawn in simple diagrams, as shown in the next section.

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 (Sammon, 1969) (NLM) and curvilinear component analysis (Demartines and Hérault, 1997) (CCA). Notice that we have implemented the version of CCA described in (Hérault et al., 1999), which proves to be more robust against noise. The literature indicates (Lee and Verleysen, 2007; Venna and Kaski, 2006) 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:

- $Q_{n\nu}(K)$ and $B_{n\nu}(K)$ in Fig. 2,
- $Q^{1,1}_{\mathrm{wNX}}(K)$ and $B^{1,1}_{\mathrm{wNX}}(K)$ in Fig. 3,
- $Q_{\mathrm{TC}}(K)$ and $B_{\mathrm{TC}}(K)$ in Fig. 4,
- $Q^{1,0}_{\mathrm{wNX}}(K)$ and $B^{1,0}_{\mathrm{wNX}}(K)$ in Fig. 5, and
- $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ in Fig. 6.

In this way, each pair of curves includes an overall quality criterion and a behavior indicator. All figures contain as many pairs of curves as there are methods or data sets to compare. The left diagram of each figure shows the whole curves, for $1 \leq K \leq N-1$; the second diagram in the upper right corner focuses on the quality criterion for small values of K, whereas the third diagram in the lower right corner does the same for the behavior indicator. In Fig. 6, the dotted ascending line represents the LCMC baseline and highlights the connection between $U_{\mathrm{LC}}(K)$ and $Q_{\mathrm{NX}}(K)$.

As can be seen, all five pairs of curves show that (i) CCA outperforms NLM and (ii) these two methods have antagonist behaviors, as previously mentioned. Looking specifically at quantities that involve a weighting of the co-ranking matrix elements, we can confirm that for small values of K similarities exist between pairs $\{Q_{n\nu}(K), B_{n\nu}(K)\}$ and $\{Q^{1,1}_{\mathrm{wNX}}(K), B^{1,1}_{\mathrm{wNX}}(K)\}$ on the one hand, and between pairs $\{Q_{\mathrm{TC}}(K), B_{\mathrm{TC}}(K)\}$ and

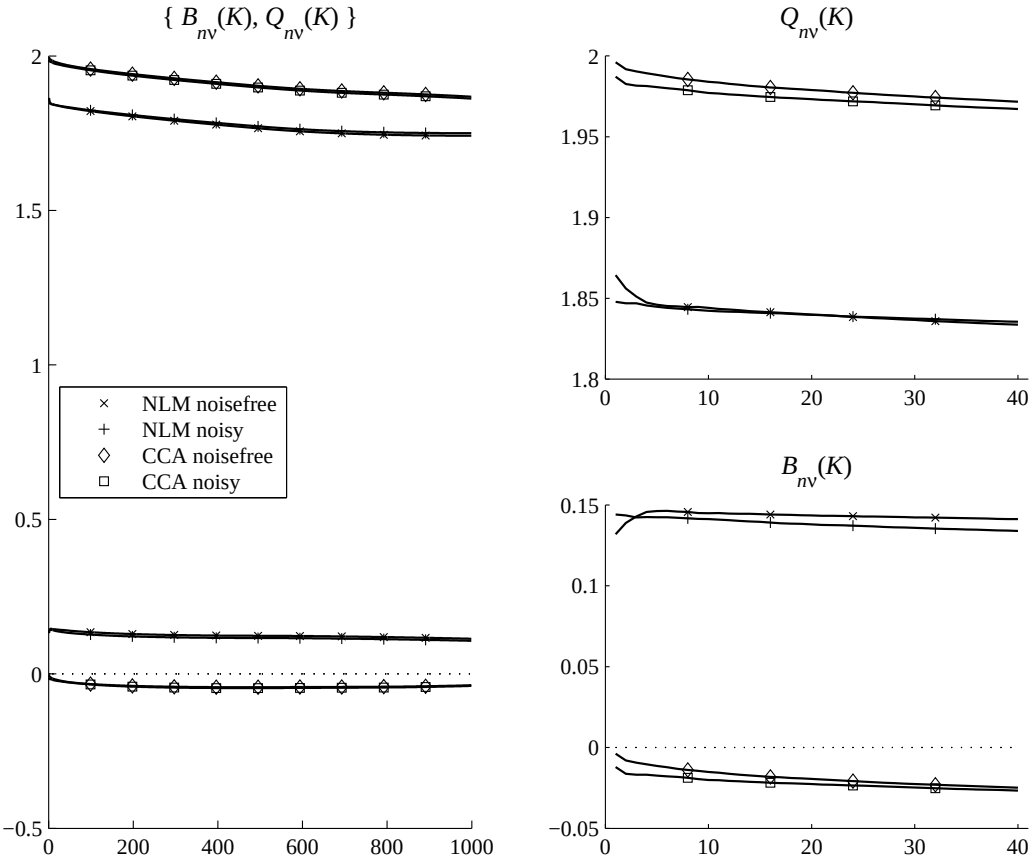

Figure 2: Quality assessment of the hollow sphere embedding: $Q^{1,1}_{n\nu}(K)$ and $B^{1,1}_{n\nu}(K)$ for NLM and CCA, for noisefree as well as noisy data.

$\{Q^{1,0}_{\mathrm{wNX}}(K), B^{1,0}_{\mathrm{wNX}}(K)\}$ on the other hand. For larger values, we can see that the common weighting shared by $\{Q_{n\nu}(K), B_{n\nu}(K)\}$ and $\{Q^{1,1}_{\mathrm{wNX}}(K), B^{1,1}_{\mathrm{wNX}}(K)\}$ gives a higher importance to local errors; as a consequence, the curves essentially remain flat when K grows. In contrast, the similar weightings of $\{Q_{\mathrm{TC}}(K), B_{\mathrm{TC}}(K)\}$ and $\{Q^{1,0}_{\mathrm{wNX}}(K), B^{1,0}_{\mathrm{wNX}}(K)\}$ put all ranks errors on the same footing. This explains why for these criteria the curves of NLM and CCA rejoin or cross each other as K grows. It is noteworthy that the curves associated with T&C drop because the considered blocks start shrinking as soon as $K \geq N/2$, whereas the triangles involved in $\{Q^{1,0}_{\mathrm{wNX}}(K), B^{1,0}_{\mathrm{wNX}}(K)\}$ keep growing. As to noise, its absence or presence has little influence on the four pairs of weighted fractions, although a slight difference can be observed in favor of the noisefree data set.

At this point, an important result is the ability of $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ to distinguish the antagonist behaviors of NLM and CCA without any (arbitrary) weighting of the coranking matrix elements. For instance, $Q_{\mathrm{NX}}(K)$ shows that if CCA succeeds in preserving local neighborhoods better than NLM, this is at the expense of sacrifying the preservation of the global manifold shape. This is illustrated by the crossing of CCA and NLM curves for

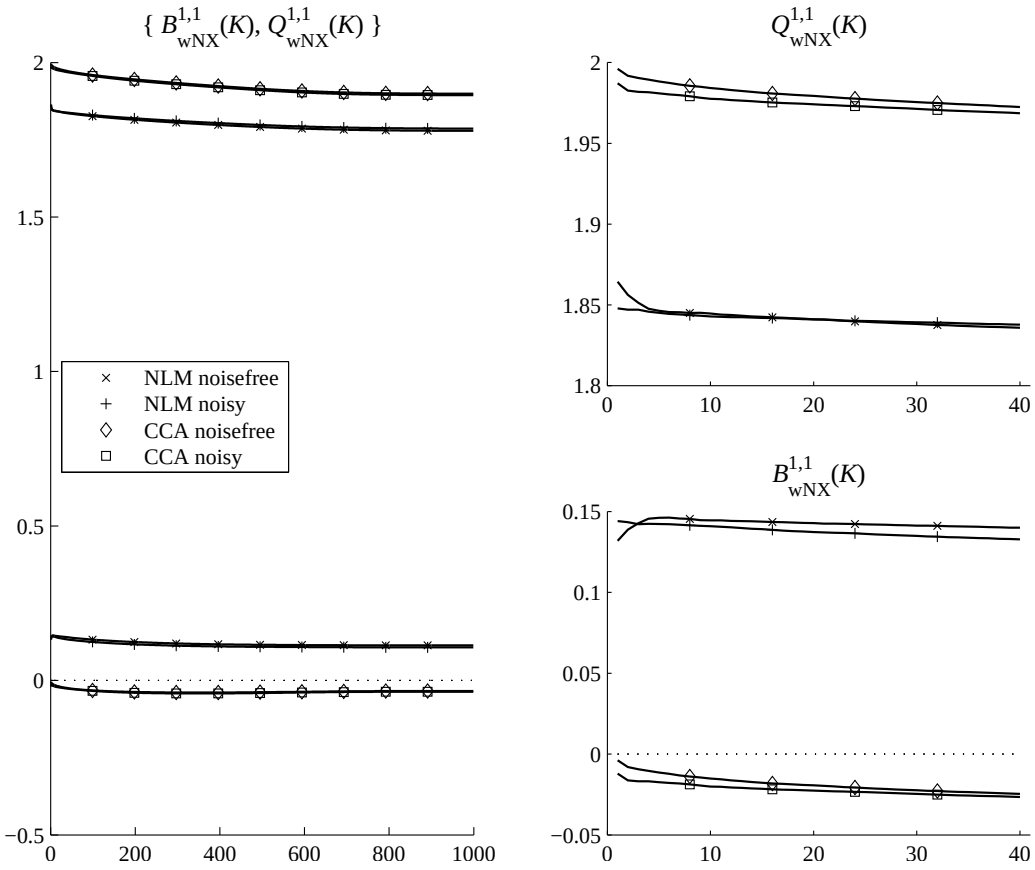

Figure 3: Quality assessment of the hollow sphere embedding: $Q^{1,1}_{\mathrm{wNX}}(K)$ and $B^{1,1}_{\mathrm{wNX}}(K)$ for NLM and CCA, for noisefree as well as noisy data.

$K \approx 500$ in Fig. 6. Unweighted fractions 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 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 $B_{\mathrm{NX}}(K)$ 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 $B_{\mathrm{NX}}(K)$, 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 fractions give too much importance to intrusions or extrusions associated with large rank errors.

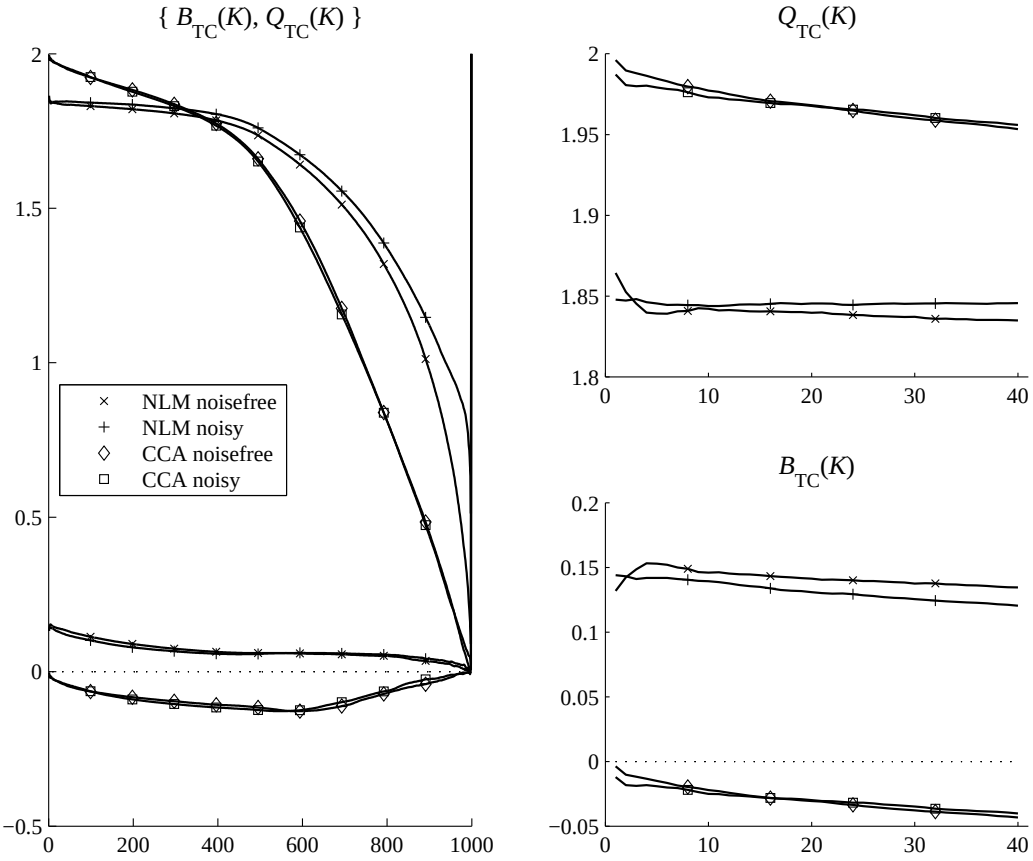

Figure 4: Quality assessment of the hollow sphere embedding: $Q_{\mathrm{TC}}(K)$ and $B_{\mathrm{TC}}(K)$ for NLM and CCA, for noisefree as well as noisy data.

6. Conclusions

This paper has reviewed several quality criteria for the assessment of nonlinear dimensionality reduction. All of them rely on the comparison of distance rankings and K-ary neighborhoods that are computed in both the high- and low-dimensional spaces. The definition of the co-ranking matrix allows us to cast them all within a unifying framework. It is noteworthy that 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). However, properties of the co-ranking matrix conduce to consider these analogies with caution. In constrast, we show that the co-ranking matrix can instead be usefully 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 fractions that are computed over various blocks or triangles of the co-ranking matrix. Experiments show that unweighted fractions of the co-ranking matrix elements are sufficient and that any weighting inevitably turns out to be

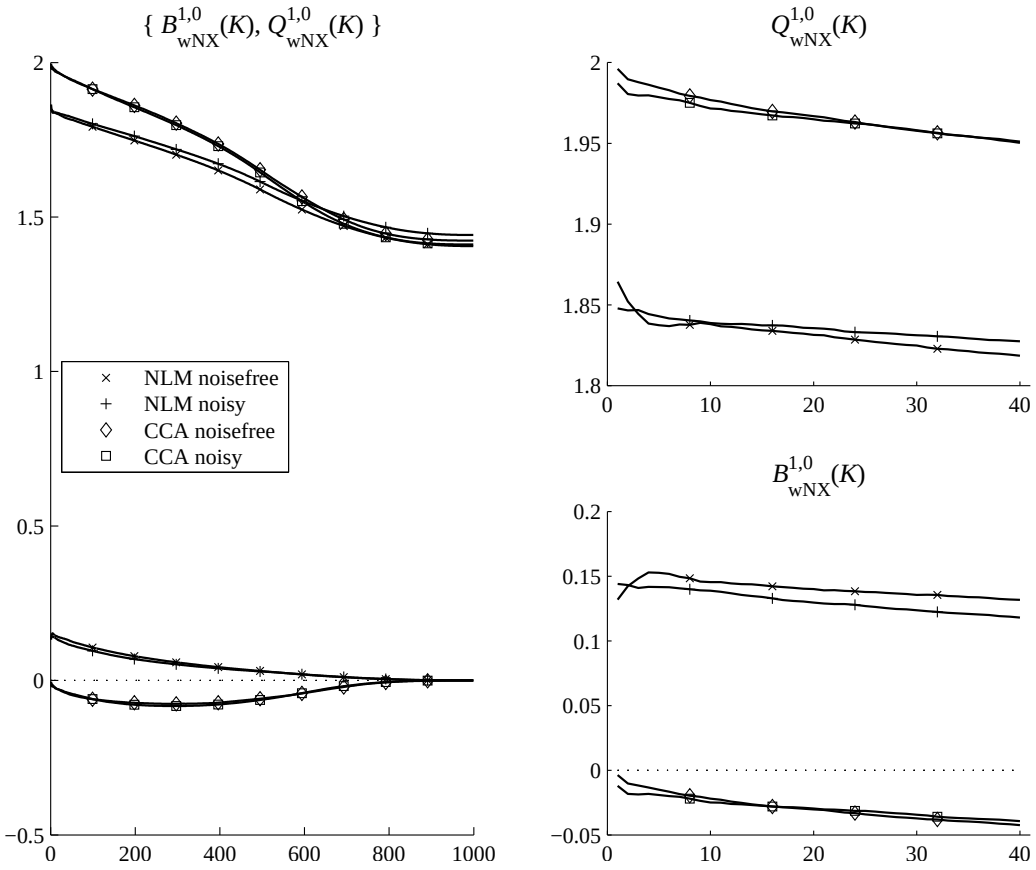

Figure 5: Quality assessment of the hollow sphere embedding: $Q^{1,0}_{\mathrm{wNX}}(K)$ and $B^{1,0}_{\mathrm{wNX}}(K)$ for NLM and CCA, for noisefree as well as noisy data.

arbitrary. Weighted fractions actually tend to emphasize some types of embedding errors and can fail to detect others.

References

H.-U. Bauer and K.R. Pawelzik. Quantifying the neighborhood preservation of self-organizing maps. *IEEE Transactions on Neural Networks*, 3:570–579, 1992.

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

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

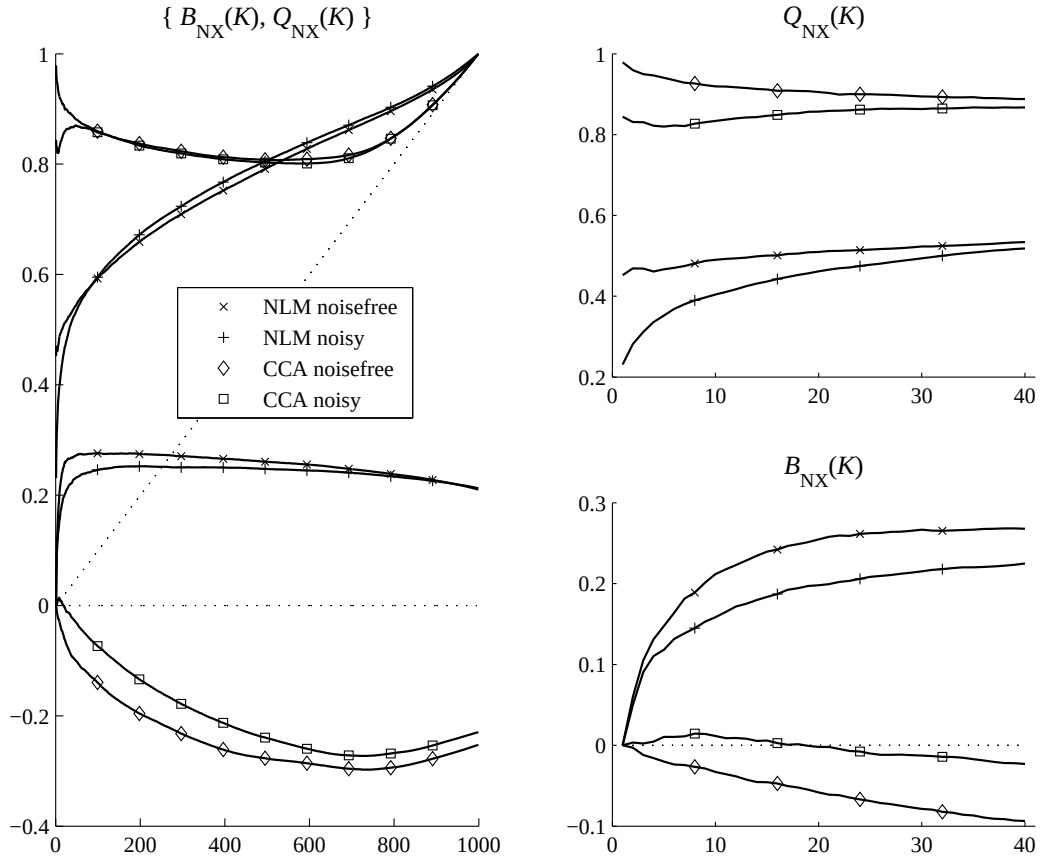

Figure 6: Quality assessment of the hollow sphere embedding: $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ for NLM and CCA, for noisefree as well as noisy data.

M. Brand and K. Huang. A unifying theorem for spectral embedding and clustering. In C.M. Bishop and B.J. Frey, editors, *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS'03)*. Key West, FL, January 2003.

L. Chen and A. Buja. *Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis*. PhD thesis, University of Pennsylviana, July 2006.

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

G. Di Battista, P. Eades, R. Tamassia, and I.G. Tollis. *Graph drawing: Algorithms for the visualization of graphs*. Prentice-Hall, 1999.

K. Fukunaga. Intrinsic dimensionality extraction. In P.R. Krishnaiah and L.N. Kanal, editors, *Classification, Pattern Recognition and Reduction of Dimensionality, Volume 2 of Handbook of Statistics*, pages 347–360. Elsevier, Amsterdam, 1982.

J. Hérault, C. Jaussions-Picaud, and A. Guérin-Dugué. Curvilinear component analysis for high dimensional data representation: I. Theoretical aspects and practical use in the presence of noise. In J. Mira and J.V. Sánchez, editors, *Proceedings of IWANN'99*, volume II, pages 635–644. Springer, Alicante, Spain, June 1999.

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

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

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

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

J.A. Lee and M. Verleysen. Curvilinear distance analysis versus Isomap. *Neurocomputing*, 57:49–76, March 2004.

J.A. Lee and M. Verleysen. Rank-based quality assessment of nonlinear dimensionality reduction. In M. Verleysen, editor, *Proceedings of ESANN 2008, 16th European Symposium on Artificial Neural Networks*, pages 49–54. d-side, Bruges, April 2008.

J.A. Lee and M. Verleysen. *Nonlinear dimensionality reduction*. Springer, 2007.

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

E. Oja. Data compression, feature extraction, and autoassociation in feedforward neural networks. In T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, editors, *Artificial Neural Networks*, volume 1, pages 737–745. Elsevier Science Publishers, B.V., North-Holland, 1991.

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

M. Saerens, F. Fouss, L. Yen, and P. Dupont. 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*, pages 371–383, Pisa, Italy, 2004.

J.W. Sammon. A nonlinear mapping algorithm for data structure analysis. *IEEE Transactions on Computers*, CC-18(5):401–409, 1969.

L.K. Saul and S.T. Roweis. Think globally, fit locally: Unsupervised learning of nonlinear manifolds. *Journal of Machine Learning Research*, 4:119–155, June 2003.

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

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

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

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

J. Venna. *Dimensionality reduction for visual exploration of similarity structures*. PhD thesis, Helsinki University of Technology, Espoo, Finland, June 2007.

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

J. Venna and S. Kaski. Nonlinear dimensionality reduction as information retrieval. *Journal of Machine Learning Research: Workshop and Conference Proceedings*, 2:572–579, 2007.

J. Venna and S. Kaski. Neighborhood preservation in nonlinear projection methods: An experimental study. In G. Dorffner, H. Bischof, and K. Hornik, editors, *Proceedings of ICANN 2001*, pages 485–491. Springer, Berlin, 2001.

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

K.Q. Weinberger and L.K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006. In Special Issue: Computer Vision and Pattern Recognition-CVPR 2004 Guest Editor(s): A. Bobick, R. Chellappa, L. Davis.

H. Whitney. Differentiable manifolds. *Annals of Mathematics*, 37(3):645–680, 1936.

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