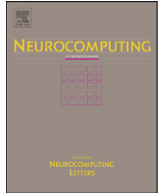




Contents lists available at ScienceDirect

## Neurocomputing

journal homepage: [www.elsevier.com/locate/neucom](http://www.elsevier.com/locate/neucom)

## Quality assessment of dimensionality reduction: Rank-based criteria

John A. Lee<sup>\*1,2</sup>, Michel Verleysen<sup>3</sup>

Machine Learning Group, Université catholique de Louvain, Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium

## ARTICLE INFO

Available online 10 January 2009

## Keywords:

Dimensionality reduction  
 Embedding  
 Quality assessment  
 Co-ranking matrix  
 Trustworthiness and continuity  
 Intrusion and extrusion fractions

## ABSTRACT

Dimensionality reduction aims at providing low-dimensional representations of high-dimensional data sets. Many new nonlinear methods have been proposed for the last years, yet the question of their assessment and comparison remains open. This paper first reviews some of the existing quality measures that are based on distance ranking and  $K$ -ary neighborhoods. Next, the definition of the co-ranking matrix provides a tool for comparing the ranks in the initial data set and some low-dimensional embedding. Rank errors and concepts such as neighborhood intrusions and extrusions can then be associated with different blocks of the co-ranking matrix. Several quality criteria can be cast within this unifying framework; they are shown to involve one or several of these characteristic blocks. Following this line, simple criteria are proposed, which quantify two aspects of the embedding quality, namely its overall quality and its tendency to favor intrusions or extrusions. They are applied to several recent dimensionality reduction methods in two experiments, with both artificial and real data.

© 2009 Elsevier B.V. All rights reserved.

## 1. Introduction

Research about dimensionality reduction (DR) deals with 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 [13] and classical metric multidimensional scaling [44,35]. On the other hand, nonlinear dimensionality reduction (NLDR) [22] emerged later, with nonlinear variants of multidimensional scaling [32,17,33], such as Sammon's nonlinear mapping (NLM) [29]. For the past 25 years, this field of research has deeply evolved and after some interest in neural approaches [14,16,26,7,24], the community has recently focused on spectral techniques [31,34,27,3,10,41]. Modern NLDR encompasses the domain of manifold learning and is also closely related to graph embedding [9] and spectral clustering [4,28,25,5].

In the most general setting, DR transforms a set of  $N$  high-dimensional vectors, denoted  $\Xi = [\xi_i]_{1 \leq i \leq N}$ , 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 in  $\Xi$  are sampled from a smooth manifold. Under this hypothesis, the goal of NLDR is then to re-embed the manifold in a space of the lowest possible

dimensionality, without modifying its topological properties. For this purpose, the embedding theorem [43] can help deduce the lowest embedding dimensionality, which is related to the manifold intrinsic dimensionality [11].

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 [14], 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 [29,7,8] in the data set  $\Xi$ , with some metric. Sometimes the coordinates in  $\Xi$  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 distance are available. For instance, a graph can be used to focus on small neighborhoods [27] or to approximate geodesic distances [34,20] 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

\* Corresponding author. Tel.: +32 2 7644778; fax: +32 10 47 25 98.

E-mail addresses: [john.lee@uclouvain.be](mailto:john.lee@uclouvain.be) (J.A. Lee), [michel.verleysen@uclouvain.be](mailto:michel.verleysen@uclouvain.be) (M. Verleysen).

<sup>1</sup> J.A.L. is a Postdoctoral Researcher with the Belgian National Fund for Scientific Research (FNRS).

<sup>2</sup> Department of Molecular Imaging and Experimental Radiotherapy.

<sup>3</sup> Microelectronics Laboratory.

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 an NLDL technique provides us with a mapping  $\mathcal{M}$  such that  $\mathbf{x} = \mathcal{M}(\xi)$ , then the reconstruction can be written as the expectation

$$E_{\text{rec}} = E\{(\xi - \mathcal{M}^{-1}(\mathcal{M}(\xi)))^2\}. \tag{1}$$

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 NLDL methods are nonparametric (they merely provide values of  $\mathcal{M}$  for the known vectors  $\xi_i$ ). The minimization of the reconstruction error is the approach that is followed by PCA and nonlinear auto-encoders [16,26]. Still another approach mentioned in the literature consists in using an indirect performance index, such as a classification error (see for instance [30,42] and other references in [36]). Obviously, this works only for labeled data. Eventually, a last possibility consists in sticking to the intrinsic goal of NLDL 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 is usually intended to be optimized by typical techniques such as gradient descent, it must fulfill some requirements as to continuity and differentiability. 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 (SOMs) [14]; see for instance the topographic product [1] and the topographic function [40]. More recently, new criteria for quality assessment have been proposed, with a broader applicability, such as the trustworthiness and continuity (T&C) measures [37], the local continuity meta-criterion (LCMC) [6], and the mean relative rank errors (MRREs) [22]. 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 [17,29].

The first aim of this paper is to review some of these recent rank-based criteria. The definition of a co-ranking matrix [23] 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 NLDL techniques involve distances in a more or less direct way. The symbol  $\delta_{ij}$  denotes the distance from  $\xi_i$  to  $\xi_j$  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  $\delta_{ij}$  and  $\delta_{ji}$  stem from distinct experimental measurements. No assumption is made as to the metrics that are associated with the high- and low-dimensional spaces, which can be different. Starting from distances, we can compute ranks.

The rank of  $\xi_j$  with respect to  $\xi_i$  in the high-dimensional space is written as  $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } k < j)\}|$ , where  $|\cdot|$  denotes the set cardinality. Similarly, the rank of  $\mathbf{x}_j$  with respect to  $\mathbf{x}_i$  in the low-dimensional space is  $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } k < j)\}|$ . 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  $\{1, \dots, N - 1\}$ . The nonreflexive  $K$ -ary neighborhoods of  $\xi_i$  and  $\mathbf{x}_i$  are denoted by  $v_i^K = \{j : 1 \leq \rho_{ij} \leq K\}$  and  $n_i^K = \{j : 1 \leq r_{ij} \leq K\}$ , respectively.

The co-ranking matrix [23] can then be defined as

$$\mathbf{Q} = [q_{kl}]_{1 \leq k, l \leq N-1} \quad \text{with} \quad q_{kl} = |\{(i, j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|. \tag{2}$$

Computing  $\mathbf{Q}$  requires  $2N$  sorting operations and therefore the time complexity is  $\mathcal{O}(N^2 \log N)$  with a typical sorting algorithm.

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 [32]. Historically, this scatterplot has often been used to assess results of multidimensional scaling and related methods [8]; it shows the distances  $\delta_{ij}$  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  $j$ th vector is an intruder in the  $K$ -ary neighborhood  $n_i^K$ , with respect to the genuine neighborhood  $v_i^K$ . Similarly, an *extrusion* denotes the event of a negative rank error. The amplitude of an intrusion or extrusion refers to 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 easily be deduced. Intuitively, mild  $K$ -intrusions and mild  $K$ -extrusions correspond to vectors that are, respectively, “promoted” and “downgraded”, but still remain in both  $v_i^K$  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  $\mathbb{F}_K = \{1, \dots, K\}$  and  $\mathbb{L}_K = \{K + 1, \dots, N - 1\}$ , the index sets of the upper-left, upper-right, lower-left, and lower-right blocks are

$$\begin{aligned} \mathbb{U}\mathbb{L}_K &= \mathbb{F}_K \times \mathbb{F}_K, & \mathbb{U}\mathbb{R}_K &= \mathbb{F}_K \times \mathbb{L}_K, \\ \mathbb{L}\mathbb{L}_K &= \mathbb{L}_K \times \mathbb{F}_K, & \mathbb{L}\mathbb{R}_K &= \mathbb{L}_K \times \mathbb{L}_K. \end{aligned} \tag{3}$$

Similarly, the block covered by  $\mathbb{U}\mathbb{L}_K$  can be split into its main diagonal and lower and upper triangles:

$$\mathbb{D}_K = \{(i, i) : 1 \leq i \leq K\}, \tag{4}$$

$$\mathbb{L}\mathbb{T}_K = \{(i, j) : 1 < i \leq K \text{ and } j < i\}, \tag{5}$$

$$\mathbb{U}\mathbb{T}_K = \{(i, j) : 1 \leq i < K \text{ and } j > i\}. \tag{6}$$

According to this division,  $K$ -intrusions and  $K$ -extrusions are located in the lower and upper trapezes, respectively (i.e.  $\mathbb{L}\mathbb{T}_K \cup \mathbb{L}\mathbb{L}_K$  and  $\mathbb{U}\mathbb{T}_K \cup \mathbb{U}\mathbb{R}_K$ ). Hard  $K$ -intrusions and  $K$ -extrusions are found in the blocks  $\mathbb{L}\mathbb{L}_K$  and  $\mathbb{U}\mathbb{R}_K$ , respectively. In a similar way, mild  $K$ -intrusions and  $K$ -extrusions are counted in the triangles  $\mathbb{L}\mathbb{T}_K$  and  $\mathbb{U}\mathbb{T}_K$ , respectively.

### 3. Rank-based criteria

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

The T&C measures [37,38] are defined as

$$M_T(K) = 1 - \frac{2}{G_K} \sum_{i=1}^N \sum_{j \in n_i^K \setminus v_i^K} (\rho_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{L}_K} (k - K)q_{kl}, \quad (7)$$

$$M_C(K) = 1 - \frac{2}{G_K} \sum_{i=1}^N \sum_{j \in v_i^K \setminus n_i^K} (r_{ij} - K) = 1 - \frac{2}{G_K} \sum_{(k,l) \in \mathbb{U}_K} (l - K)q_{kl}, \quad (8)$$

where the normalizing factor

$$G_K = \begin{cases} NK(2N - 3K - 1) & \text{if } K < N/2, \\ N(N - K)(N - K - 1) & \text{if } K \geq N/2 \end{cases} \quad (9)$$

considers the worst case [36], i.e. ranks are reversed in the low-dimensional space and the co-ranking matrix is anti-diagonal. Both the 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. 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 MRREs [22] rely on the same principle as the T&C. They are defined as

$$W_n(K) = \frac{1}{H_K} \sum_{i=1}^N \sum_{j \in n_i^K} \frac{|\rho_{ij} - r_{ij}|}{\rho_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{U}_K \cup \mathbb{L}_K} \frac{|k - l|}{l} q_{kl}, \quad (10)$$

$$W_v(K) = \frac{1}{H_K} \sum_{i=1}^N \sum_{j \in v_i^K} \frac{|\rho_{ij} - r_{ij}|}{r_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{U}_K \cup \mathbb{L}_K} \frac{|k - l|}{k} q_{kl}, \quad (11)$$

where the normalizing factor

$$H_K = N \sum_{k=1}^K \frac{|N - 2k + 1|}{k} \quad (12)$$

considers the worst case, like that of T&C. The differences between the MRREs and the T&C are found 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 columns 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 LCMC [6] is defined as

$$U_{LC}(K) = \frac{1}{NK} \sum_{i=1}^N \left( |n_i^K \cap v_i^K| - \frac{K^2}{N-1} \right) = \frac{K}{1-N} + \frac{1}{NK} \sum_{(k,l) \in \mathbb{L}_K} q_{kl}, \quad (13)$$

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  $\mathbb{U}_K$  of  $\mathbf{Q}$ . Notice also that the elements  $q_{kl}$  in the block  $\mathbb{U}_K$  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 NLDL method by means of a single scalar. All these observations are visually summarized in Fig. 1, which illustrates the blocks of the co-ranking matrix that are covered by the various quality criteria.

### 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)

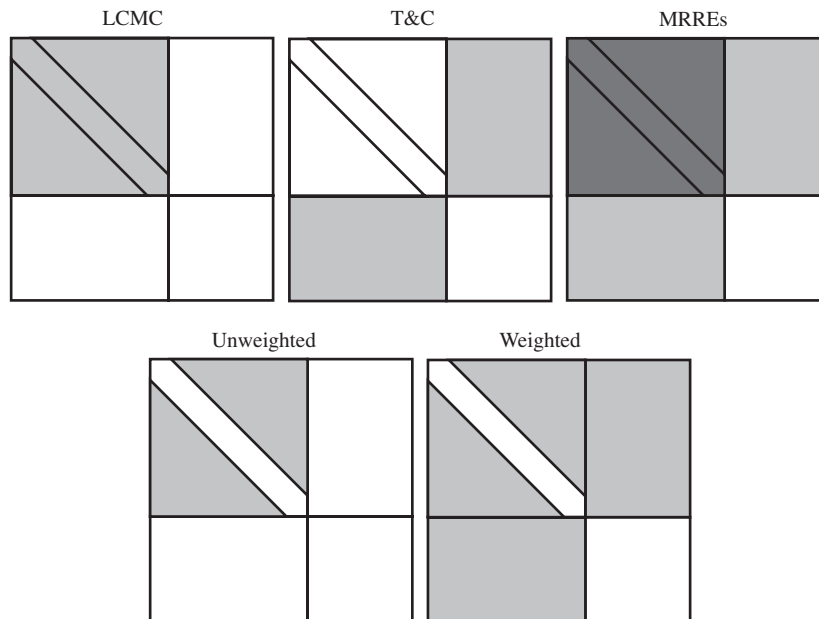


Fig. 1. For all (pairs of) quality criteria, a schematic illustration of the co-ranking matrix is shown: the blocks that are taken into account are shaded. For the MRREs, the block  $\mathbb{U}_K$  is covered twice. The second row corresponds to the pairs  $\{U_N(K), U_X(K)\}$  and  $\{W_N^{n,w}(K), W_X^{n,w}(K)\}$ , which are defined in Section 4.

in the domain of information retrieval [39]. 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  $v_i^K$ , 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  $\mathbb{U}L_K$ ,  $\mathbb{L}L_K$ ,  $\mathbb{U}R_K$ , and  $\mathbb{L}R_K$  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  $v_i^K$  have the same size, the numbers of false positives and false negatives are the same. Each element of  $v_i^K$  that is missed in  $n_i^K$  (a false negative) is replaced with an incorrect neighbor (a false positive). Formally, this can also be demonstrated by observing that  $\mathbf{Q}$  is a sum of  $N$  permutation matrices, and thus

$$\forall k, \sum_{l=1}^{N-1} q_{kl} = N \quad \text{and} \quad \forall l, \sum_{k=1}^{N-1} q_{kl} = N. \tag{14}$$

As we compute ranks starting from  $N$  reference points, we have always  $N$   $k$ th neighbors. Therefore, we have

$$\sum_{(k,l) \in \mathbb{U}L_K \cup \mathbb{L}L_K} q_{kl} = \sum_{(k,l) \in \mathbb{U}L_K \cup \mathbb{U}R_K} q_{kl} = KN \tag{15}$$

and

$$\sum_{(k,l) \in \mathbb{L}L_K} q_{kl} = \sum_{(k,l) \in \mathbb{U}R_K} q_{kl}. \tag{16}$$

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  $M_T(K) = M_C(K)$  and  $W_v(K) = W_n(K)$ . 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 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 take into account all  $K$ -intrusions and  $K$ -extrusions can be written as

$$W_N^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l) \in \mathbb{L}T_K \cup \mathbb{U}L_K} \frac{(k-l)^v}{k^w} q_{kl}, \tag{17}$$

$$W_X^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l) \in \mathbb{U}T_K \cup \mathbb{U}R_K} \frac{(l-k)^v}{l^w} q_{kl}, \tag{18}$$

where

$$C_K^{v,w} = \begin{cases} NK & \text{if } v = w = 0 \\ N \sum_{k=1}^K \frac{\max\{0, N-2k\}^v}{k^w} & \text{if } v \geq 1 \end{cases}. \tag{19}$$

The integer exponents  $v$  and  $w$  are such that  $v \geq w \geq 0$  and can be adjusted in order to emphasize large rank differences, relatively to the reference rank. The normalization is based on the worst case. For  $v \geq 1$ , it corresponds to a co-ranking matrix that is anti-

diagonal. For  $v = w = 0$ , if  $\mathbf{e}_i$  denotes the  $i$ th basis vector, then the normalization considers the two circulant matrices<sup>4</sup> generated by vectors  $N\mathbf{e}_2$  for  $W_N^{v,w}(K)$  and  $N\mathbf{e}_{N-1}$  for  $W_X^{v,w}(K)$ .

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 (see Fig. 1), 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.

In order to obtain a single and global quality criterion, we can define

$$Q_{wNX}^{v,w} = 1 - \frac{W_N^{v,w}(K) + W_X^{v,w}(K)}{2}, \tag{20}$$

which increases if both numbers of intrusions and extrusions diminish. Information about the overall quality can be completed by the quantity

$$B_{wNX}^{v,w} = W_N^{v,w}(K) - W_X^{v,w}(K), \tag{21}$$

which reveals the “behavior” of a DR method. For  $v \geq 1$ , positive or negative values, respectively, indicate that the produced embedding is “intrusive” or “extrusive”. Being intrusive or extrusive is thus related with the occurrence of large rank errors that are highly penalized in either  $W_N^{v,w}(K)$  or  $W_X^{v,w}(K)$ . Nevertheless, the weighting scheme involved in  $W_N^{v,w}(K)$  and  $W_X^{v,w}(K)$  can be questioned. As a matter of fact, choosing values for  $v$  and  $w$  proves to be somewhat arbitrary. This can be shown by considering the combination  $v = w = 0$ . In this case indeed, the weighting factors degenerate, all rank errors are on the same footing, and only the numbers of intrusions and extrusions are relevant. Due to the properties of the co-ranking matrix, we can see that these numbers are actually related. For this purpose, let us consider some neighborhood size  $K \ll N$  and a single permutation matrix  $\mathbf{P}$  among those constituting the co-ranking matrix. Starting from  $\mathbf{P}$  equal to the identity, imagine that we observe a hard  $K$ -intrusion, that is, we have  $r_{ij} \leq K < \rho_{ij}$  for some  $i$  and  $j$ . The resulting permutation matrix compensates for the hard  $K$ -intrusion with one hard  $K$ -extrusion and  $(K - r_{ij})$  mild  $K$ -extrusions, as illustrated in Fig. 2. The numbers of hard  $K$ -intrusions and  $K$ -extrusions are equal, as expected. Intuitively, the intrusion at some rank  $k$  shifts back all subsequent neighbors ( $r_{ij} > k$ ); their rank is increased by one unit and the last neighbor is thus evicted from the  $K$ -ary neighborhood. As a matter of fact, the occurrence of a single hard  $K$ -intrusion translates into several mild  $K$ -extrusions, whose number is proportional to the severity of the  $K$ -intrusion. Hence, on top of the observation that the numbers of hard  $K$ -intrusions and hard  $K$ -extrusions are equal comes the paradoxical conclusion that the number of mild  $K$ -intrusions measures the severity of hard  $K$ -extrusions, and vice versa. If we write the fractions of mild  $K$ -intrusions and mild  $K$ -extrusions as

$$U_N(K) = \frac{1}{KN} \sum_{(k,l) \in \mathbb{L}T_K} q_{kl} \quad \text{and} \quad U_X(K) = \frac{1}{KN} \sum_{(k,l) \in \mathbb{U}T_K} q_{kl} \tag{22}$$

then we can define a behavior indicator as

$$B_{NX}(K) = U_X(K) - U_N(K) = W_X^{0,0} - W_N^{0,0} = -B_{wNX}^{0,0}(K). \tag{23}$$

Its value is positive for intrusive embeddings and negative for extrusive ones. The second and third equalities results from (16) and the definitions of  $W_X^{v,w}$  and  $W_N^{v,w}$ . In particular, the second

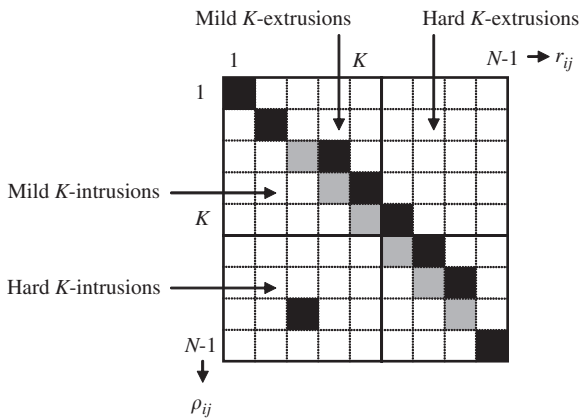
<sup>4</sup> Notice that according to this description  $C_K^{0,0}$  should actually be equal to  $N \min\{N-2, K\}$ . However, we prefer to use the approximation  $NK$ , which allows us to simplify the comparisons with other criteria to be defined in the sequel.

equality guarantees that  $B_{NX}(K)$  is equal to the difference between the fractions of all  $K$ -extrusions and all  $K$ -intrusions (both mild and hard ones). Keeping the focus on quantities computed inside  $K$ -ary neighborhoods and following the same idea as that behind the LCMC, we can write the fraction of vectors that keep the same rank in both neighborhoods  $v_i^K$  and  $n_i^K$  as

$$U_P(K) = \frac{1}{KN} \sum_{(k,l) \in \mathbb{D}_K} q_{kl} \quad (24)$$

and an overall quality criterion as the sum

$$Q_{NX}(K) = U_P(K) + U_N(K) + U_X(K) = U_{LC}(K) + \frac{K}{N-1}, \quad (25)$$



**Fig. 2.** Intuitive illustration of the relationship between the severity of hard  $K$ -intrusions and the number of mild  $K$ -extrusions. Transposition of the matrix easily shows that a similar relationship also exists for hard  $K$ -extrusions and mild  $K$ -intrusions. Shaded cells indicate the elements that were nonzero before the occurrence of the hard  $K$ -intrusion at  $\rho_{ij} = 8$  and  $r_{ij} = 3$ .

where the second equality shows the relationship with the LCMC. Because only hard  $K$ -extrusions and hard  $K$ -intrusions make it decrease, this last quality criterion is more permissive than  $Q_{wNX}$ . Inequality  $Q_{NX}(K) \geq Q_{wNX}^{0,0}(K)$  formally translates this statement.

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

$$Q_{TC}(K) = \frac{M_T(K) + M_C(K)}{2}, \quad (26)$$

$$B_{TC}(K) = M_C(K) - M_T(K) \quad (27)$$

for T&C, and

$$Q_{nv}(K) = 1 - \frac{W_n(K) + W_v(K)}{2}, \quad (28)$$

$$B_{nv}(K) = W_n(K) - W_v(K) \quad (29)$$

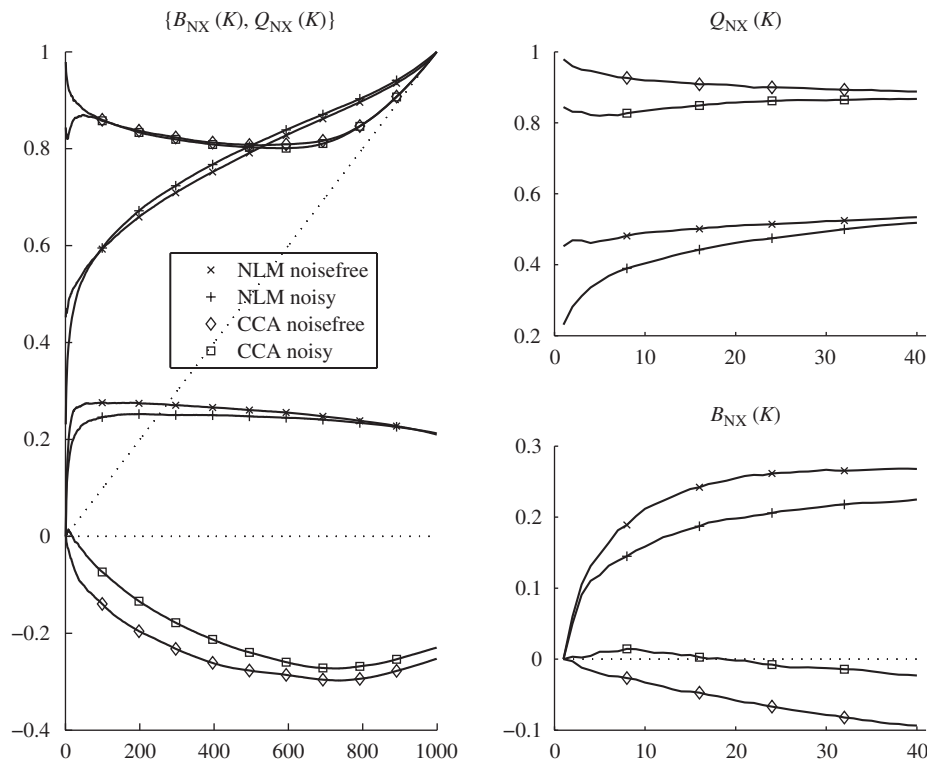
for the MRREs. All quality and behavior criteria can be drawn in simple diagrams as curves with respect to the neighborhood size  $K$ .

## 5. Experiments

The next subsections describe two experiments that illustrate and compare the various above-mentioned quality criteria. The first one involves artificial data, all criteria, and two NLD methods. The second experiment investigates a larger set of methods that are applied to an image database.

### 5.1. Artificial data: the hollow sphere

For this experiment, 1000 points are randomly drawn from a hollow sphere whose radius is equal to one. Two data sets are formed. The first one comprises the noise-free points, whereas



**Fig. 3.** Quality assessment for the embeddings of the hollow sphere:  $Q_{NX}(K)$  and  $B_{NX}(K)$  for NLM and CCA, for noise-free as well as noisy data.

Gaussian noise with standard deviation equal to 0.05 is added in the second set. Next, the manifold has been embedded in a two-dimensional space with Sammon's NLM [29] and curvilinear

component analysis (CCA) [8]. Notice that we have implemented the version of CCA described in [12], which proves to be more robust against noise. The literature indicates [22,38] that NLM is

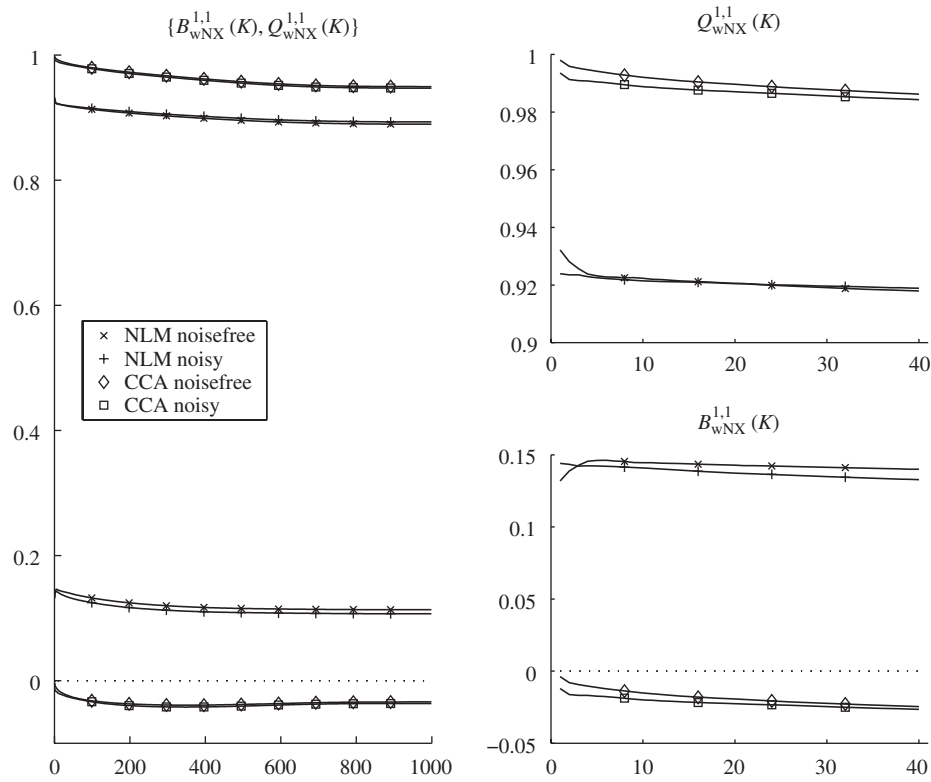


Fig. 4. Quality assessment for the embeddings of the hollow sphere:  $Q_{wNX}^{1,1}(K)$  and  $B_{wNX}^{1,1}(K)$  for NLM and CCA, for noisefree as well as noisy data.

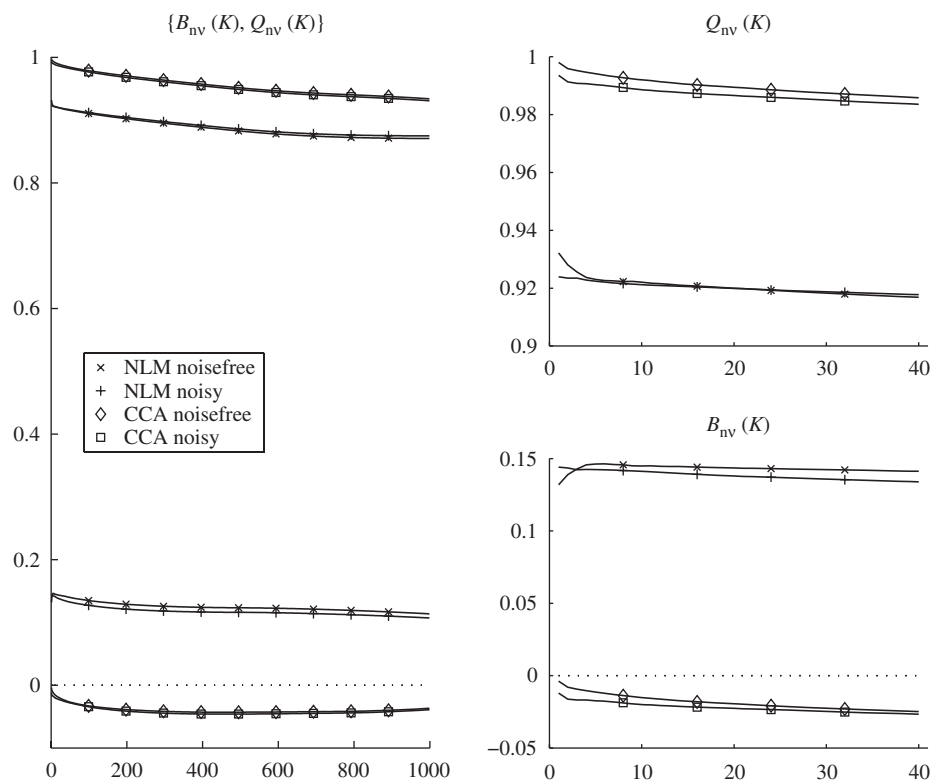


Fig. 5. Quality assessment for the embeddings of the hollow sphere:  $Q_{nv}(K)$  and  $B_{nv}(K)$  for NLM and CCA, for noisefree as well as noisy data.

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.

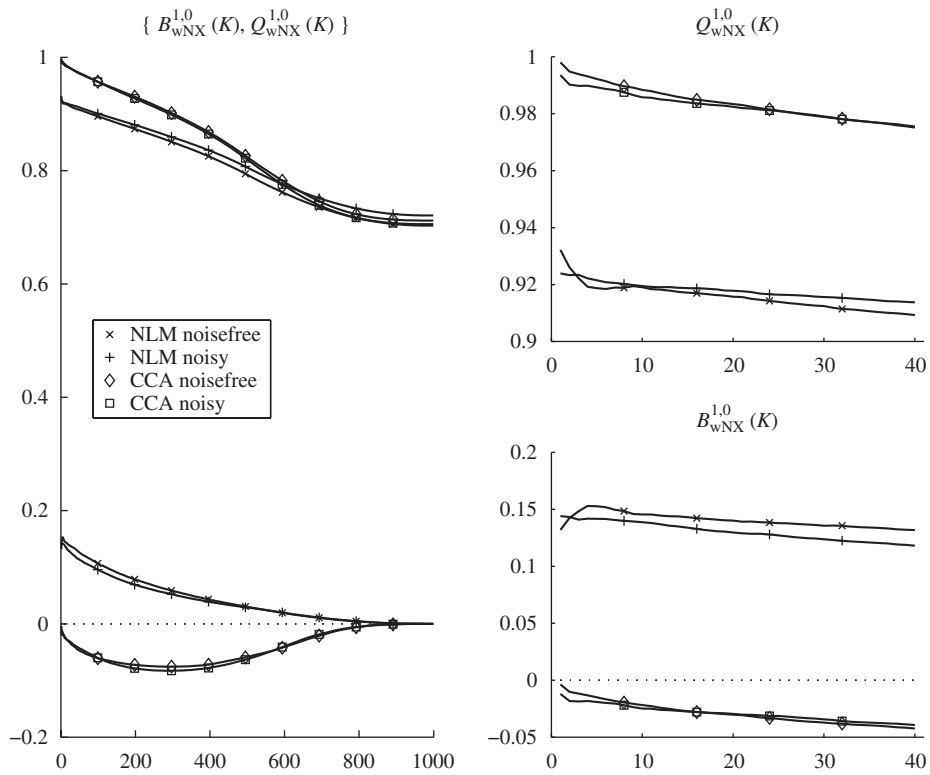


Fig. 6. Quality assessment for the embeddings of the hollow sphere:  $Q_{wNX}^{1.0}(K)$  and  $B_{wNX}^{1.0}(K)$  for NLM and CCA, for noise-free as well as noisy data.

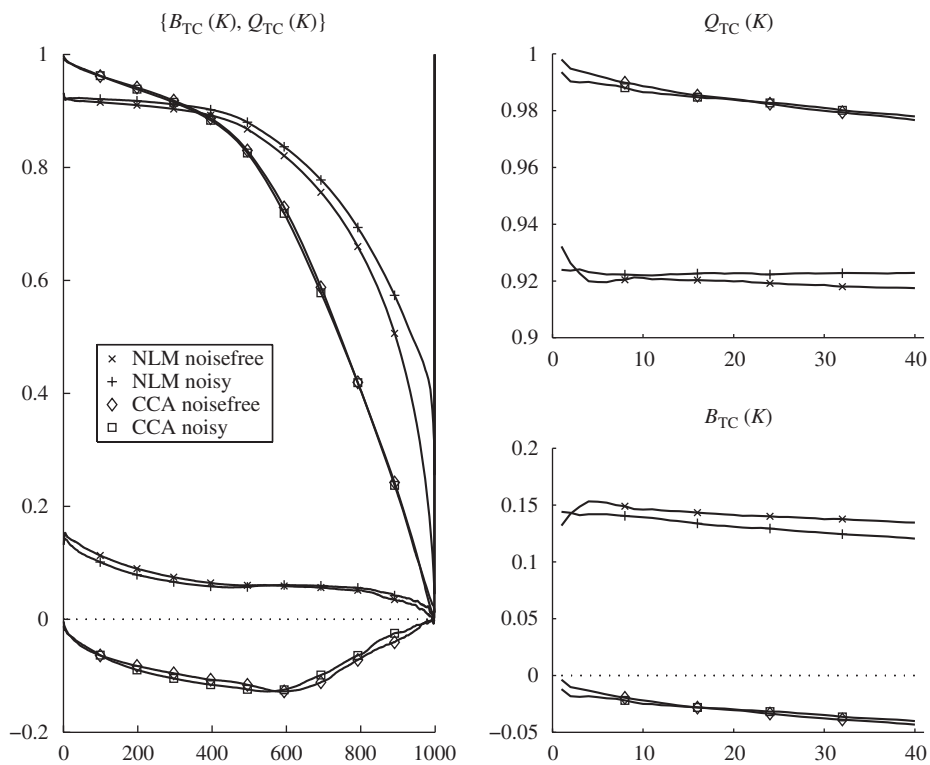


Fig. 7. Quality assessment for the embeddings of the hollow sphere:  $Q_{TC}(K)$  and  $B_{TC}(K)$  for NLM and CCA, for noise-free as well as noisy data.

In order to present results that can be easily compared, the following quantities are displayed:

- Fig. 3:  $Q_{NX}(K)$  and  $B_{NX}(K)$ .
- Fig. 4:  $Q_{wNX}^{1,1}(K)$  and  $B_{wNX}^{1,1}(K)$ .
- Fig. 5:  $Q_{TW}(K)$  and  $B_{TW}(K)$ .
- Fig. 6:  $Q_{wNX}^{1,0}(K)$  and  $B_{wNX}^{1,0}(K)$ .
- Fig. 7:  $Q_{TC}(K)$  and  $B_{TC}(K)$ .

Each figure thus includes as many pairs of curves as there are methods to be compared. Each pair of curves refers to an overall quality criterion and a behavior indicator. 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. 3, the dotted ascending line represents the LCMC baseline and highlights the connection with  $Q_{NX}(K)$ .

Considering that  $K$  is a sort of scale parameter, one can interpret the figures as follows. As DR intends to preserve local vicinities, one method outperforms the other if it has the highest value of the quality criterion for a wide range among the smallest values of  $K$ . Similarly, the smallest absolute value of the behavior indicator is desirable for the lowest values of  $K$ . As to its sign, experiments show that methods that can produce extrusive

embeddings usually attain higher values of the quality criterion. Up to which value of  $K$  should we look at the curves is a difficult question, whose answer depends on properties of the data set, such as its size, (intrinsic) dimensionality, and noise variance.

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 the MRREs and  $\{W_N^{1,1}(K), W_X^{1,1}(K)\}$  on the one hand, and between T&C and  $\{W_N^{1,0}(K), W_X^{1,0}(K)\}$  on the other hand. For larger values, we can see that the common weighting shared by the MRREs and  $\{W_N^{1,1}(K), W_X^{1,1}(K)\}$  gives a higher importance to local errors; as a consequence, the curves essentially remain flat when  $K$  grows. On the other hand, the similar weightings of T&C and  $\{W_N^{1,0}(K), W_X^{1,0}(K)\}$  put all ranks errors on the same footing. This explains why for those criteria the curves of NLM and CCA 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 noise-free data set.

Unreported experiments on the same data sets allow us to briefly describe the influence of parameters  $\nu$  and  $w$  in  $Q_{wNX}^{\nu,w}(K)$  and  $B_{wNX}^{\nu,w}(K)$  as follows. The special case where  $\nu = w = 0$  has been investigated in the previous section. Increasing  $\nu$  or  $w$  flattens the



Fig. 8. Some faces randomly drawn from the database.

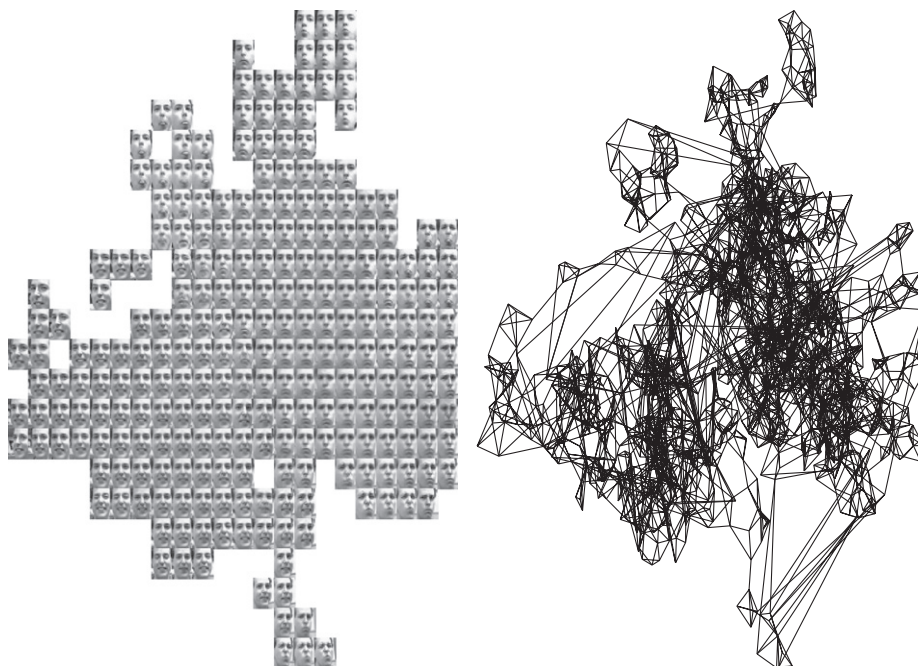


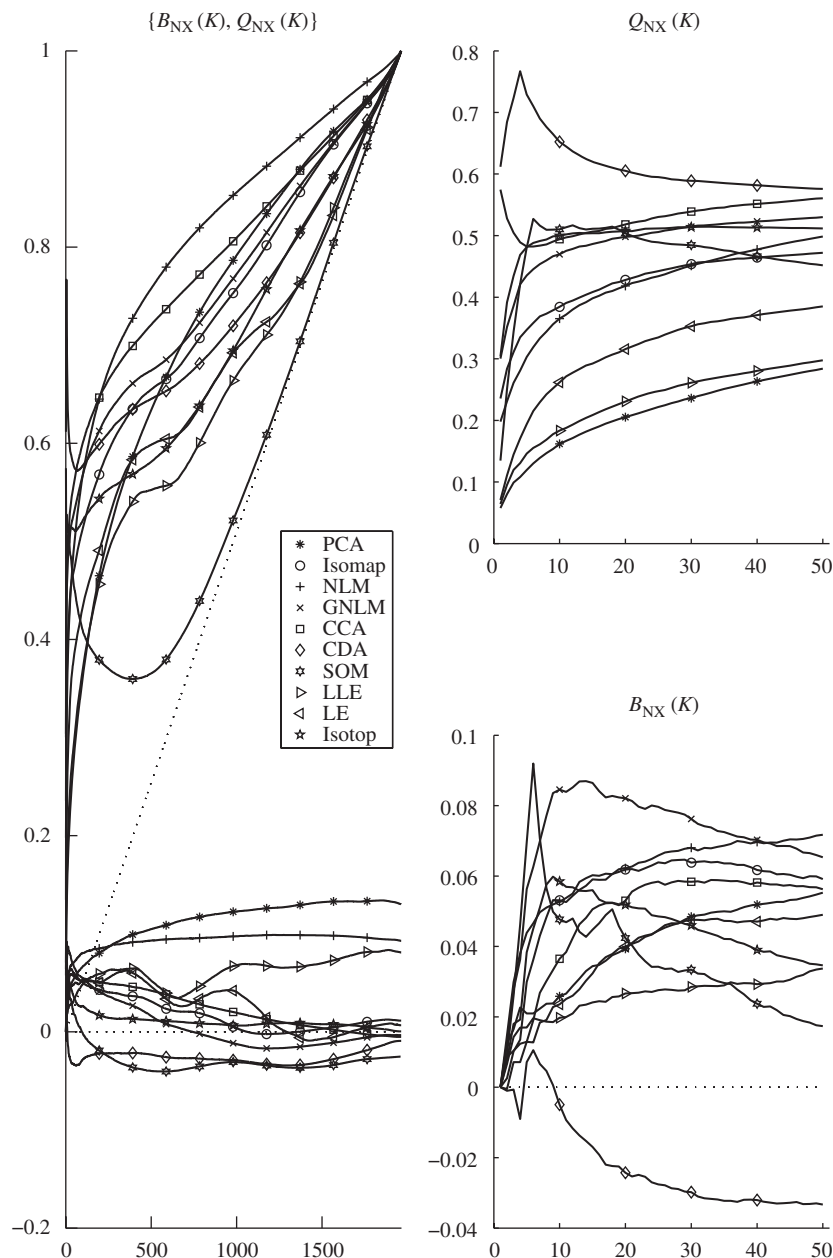
Fig. 9. The face database embedded in a two-dimensional space by CDA. The left panel shows the same embedding with the appropriate faces attached to each point. The right panel is a geometric graph that illustrates the  $K$ -ary neighborhoods ( $K = 4$ ).



curves of  $Q_{wNX}^{\nu,w}(K)$  and  $B_{wNX}^{\nu,w}(K)$ : they span a narrower and narrower interval over the range of  $K$ . This conclusion can easily be drawn by looking at (17) and (18): if  $\nu$  and/or  $w$  are increased, the largest weighting factors tend to concentrate in the lower left and upper right corner of the co-ranking matrix. This means that if  $K$  grows, only smaller and smaller contributions are added to the weighted sums  $Q_{wN}^{\nu,w}(K)$  and  $Q_{wX}^{\nu,w}(K)$ .

At this point, an important result is the ability of  $Q_{NX}(K)$  and  $B_{NX}(K)$  to distinguish the antagonist behaviors of NLM and CCA without any (arbitrary) weighting of the co-ranking matrix elements. For instance,  $Q_{NX}(K)$  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. 3. 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 with the noisy and noise-free 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  $B_{NX}(K)$  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  $B_{NX}(K)$ , 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 averages give too much



**Fig. 10.** Quality assessment for the face database with  $\{Q_{NX}(K), B_{NX}(K)\}$ . The embedding dimensionality is equal to two. The considered NLDL methods are PCA, Isomap, Sammon’s NLM (with Euclidean and geodesic distances), CCA, CDA, Kohonen’s SOM, locally linear embedding, Laplacian eigenmaps, and Isotop. The main plot on the left gives an overview of the whole curves for  $1 \leq K \leq N$ , whereas the smaller plots on the right focus on small values of  $K$ .

importance to intrusions or extrusions associated with large rank errors.

5.2. Real data: face images

For this experiment, the face database has been provided by B.J. Frey. It has already been exploited in [27], for instance. There are 1965 pictures of the same face, with different poses and various expressions. Some of them are illustrated in Fig. 8. Each face is 20 pixels wide and 28 pixels high. Each image is converted into a single 560-dimensional vector that gathers the gray values of all pixels.

The considered NLDR methods are PCA [13], Isomap [34], Sammon's NLM [29] (with both Euclidean and geodesic distances

[19,21], NLM and GNLM), CCA [8], cuvilinear distance analysis (CDA) [19], Kohonen's SOM [15], locally linear embedding (LLE) [27], Laplacian eigenmaps (LE) [2,3], and Isotop [18]. The number of neighbors  $K$  was equal to 4 for GNLM, CDA, and Isotop. For LLE and LE,  $K$  is equal to 12 and 6, respectively. The numbers of iterations are 60 for the SOM, 100 for NLM, GNLM, CCA, and CDA, and 200 for Isotop. The SOM consists of a rectangular grid of 1961 prototypes ( $37 \times 53$ ); the grid rows are shifted so that each prototypes has six equidistant neighbors.

The objective is to embed the database in a two-dimensional space, for visualization purposes. Hence, the embedding dimensionality is chosen regardless of the actual intrinsic dimensionality, which is probably higher, considering that the degrees of freedom of facial expressions are presumably numerous. As an examples, Fig. 9 illustrates the result of CDA. Two representations

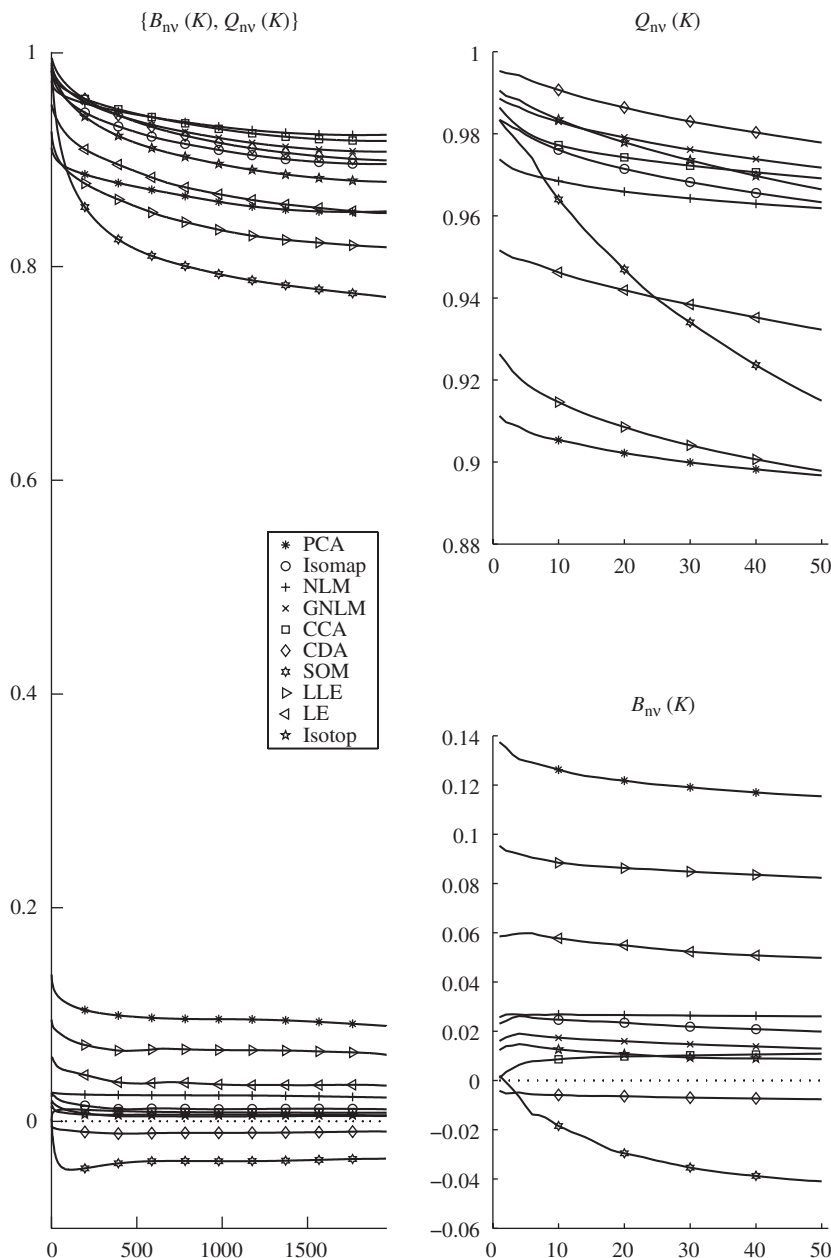
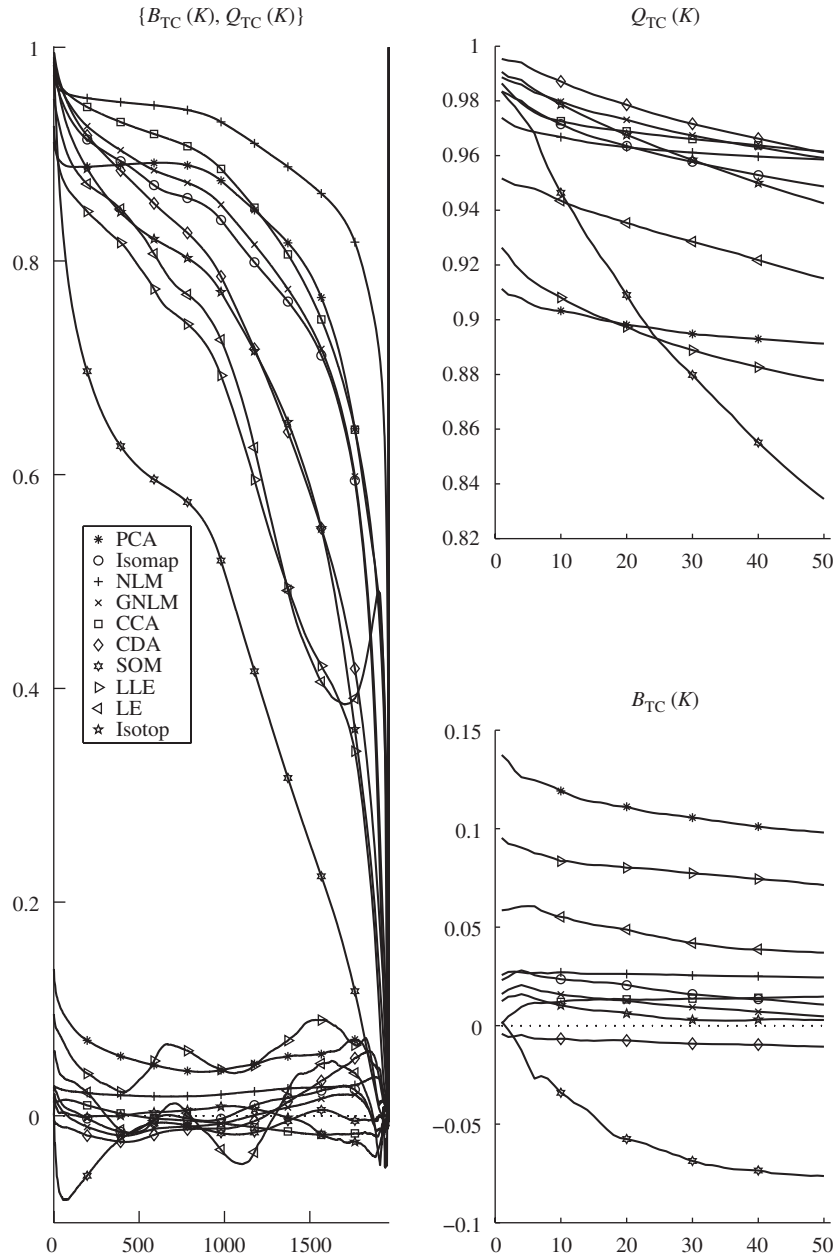


Fig. 11. Quality assessment for the face database with  $\{Q_{nv}(K), B_{nv}(K)\}$ . The embedding dimensionality is equal to two. The considered NLDR methods are PCA, Isomap, Sammon's NLM (with Euclidean and geodesic distances), CCA, CDA, Kohonen's SOM, locally linear embedding, Laplacian eigenmaps, and Isotop. The main plot on the left gives an overview of the whole curves for  $1 \leq K \leq N$ , whereas the smaller plots on the right focus on small values of  $K$ .



**Fig. 12.** Quality assessment for the face database with  $\{Q_{TC}(K), B_{TC}(K)\}$ . The embedding dimensionality is equal to two. The considered NLD methods are PCA, Isomap, Sammon's NLM (with Euclidean and geodesic distances), CCA, CDA, Kohonen's SOM, locally linear embedding, Laplacian eigenmaps, and Isotop. The main plot on the left gives an overview of the whole curves for  $1 \leq K \leq N$ , whereas the smaller plots on the right focus on small values of  $K$ .

are given. The first one is intuitive and includes thumbnails that show the faces corresponding to the embedded data points. The second representation is a geometric graph whose vertex coordinates are determined by the embedding, whereas the edges stem from (symmetrized)  $K$ -ary neighborhoods in the 560-dimensional face space, with  $K = 4$ .

As to the quality assessment, only  $\{Q_{NX}(K), B_{NX}(K)\}$ ,  $\{Q_{nv}(K), B_{nv}(K)\}$ , and  $\{Q_{TC}(K), B_{TC}(K)\}$  are shown in Figs. 10–12. All quality measures agree as to the best and worst methods in this experiment: CDA clearly outperforms all methods, whereas PCA (the only linear method on the stage) delivers the weakest performance. Iterative methods such as NLM, GNLM, CCA, and CDA lead to rather good results; the use of geodesic distances increases the quality for GNLM as well as CDA. Spectral methods (Isomap, LLE, and LE) are not among the best performers. Isomap

works better than LE, which in turn systematically outperforms LLE. All spectral methods actually turn out to be equivalent to classical metric MDS [44,35] with a non-Euclidean and/or data-driven metric. In the case of LLE and LE, this metric is implicitly determined by the construction of a sparse matrix of pairwise similarities. This induces a non-Euclidean metric that can be related to the diffusivity [28,25] across the data set. Such a metric is well suited for clustering [4,5] and this explains why those two methods fail in a visualization task: large clusters are individualized, but smaller neighborhoods are not well preserved. On the other hand, Isomap relies on a metric (i.e. shortest paths in a graph) whose geometrical interpretation in terms of manifold geodesics is straightforward. This simpler and more pragmatic approach proves to be more successful in the considered visualization task.

Compared to the majority of other methods, the SOM produces an atypical picture: small neighborhoods are pretty well recovered, but the performance indexes steeply decrease as  $K$  grows. Isotop shows a similar but somewhat reduced performance drop as the SOM. The explanation holds in the fact that, unlike most other methods that explicitly rely on distance preservation, the SOM as well as Isotop loosely try to preserve small neighborhoods [22]. Hence ranks can change a lot beyond the scope of these small neighborhoods. In the case of the SOM, the stronger performance decrease is also due to the likely discrepancy between the rectangular SOM grid and the irregular shape of the data swarm.

Looking at the behavior of the various methods shows that PCA is the most intrusive. On the other hand, CDA and the SOM to a lesser extent are the most extrusive methods. In the case of the SOM, this behavior directly results from the method principle: once the grid is fitted to the data swarm, some folds can abut and appear away from each other in the two-dimensional representation. Notice, however, that this explanation holds for large folds only and the SOM remains intrusive for the lowest values of  $K$ .

At this point, it seems reasonable to conclude that an extrusive method such as CDA can provide better performance than techniques that tend to be intrusive. Being extrusive allows CDA to better “unroll” the underlying manifold, at the moderate expense of losing a few neighborhood relationships.

Concerning the three different pairs of quality measures, we can see that the unweighted averages  $Q_{NX}(K)$  and  $B_{NX}(K)$  provide useful information over the whole range  $1 \leq K \leq N$ . On the other hand, the weighting involved in the MRRE reduce the variability of the curves, which are nearly flat. The T&C occupies an intermediate place. Notice also that for small values of  $K$ ,  $Q_{NX}(K)$  is more discriminatory than the other corresponding quality criteria.

## 6. Conclusions

This paper has reviewed several quality criteria for the assessment of DR. 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 often reports 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, be considered carefully. Instead, the interpretation of the co-ranking matrix could be based on its similarities with 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.

The experiments have involved embeddings of both synthetic and real data, with some salient characteristics. They have been produced by various DR methods that are known in the literature to show some specific strengths or properties. The efficiency of the various quality criteria has then been assessed by examining their ability to put forward antagonist characteristics of the embeddings. In particular, they show that computing unweighted averages of the co-ranking matrix elements over specific blocks provides detailed information about the embedding quality for different neighborhood sizes. In fact, any weighting in the quality measures inevitably turns out to be arbitrary: weighted averages tend to purposely emphasize some types of embedding errors but therefore they can also fail to detect others. Eventually, the co-ranking matrix proves to be a useful tool to design

simple and discriminatory quality criteria with a straightforward interpretation.

## References

- [1] H.-U. Bauer, K. Pawelzik, Quantifying the neighborhood preservation of self-organizing maps, *IEEE Transactions on Neural Networks* 3 (1992) 570–579.
- [2] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: T. Dietterich, S. Becker, Z. Ghahramani (Eds.), *Advances in Neural Information Processing Systems (NIPS 2001)*, vol. 14, MIT Press, Cambridge, MA, 2002.
- [3] M. Belkin, P. Niyogi, Laplacian eigenmaps for dimensionality reduction and data representation, *Neural Computation* 15 (6) (2003) 1373–1396.
- [4] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, 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.
- [5] M. Brand, K. Huang, A unifying theorem for spectral embedding and clustering, in: C. Bishop, B. Frey (Eds.), *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS'03)*, Key West, FL, 2003, also presented at NIPS 2002 Workshop on Spectral Methods, Available as Technical Report TR2002-042.
- [6] L. Chen, Local multidimensional scaling for nonlinear dimensionality reduction, graph layout, and proximity analysis, Ph.D. Thesis, University of Pennsylvania, July 2006.
- [7] P. Demartines, J. Héroult, Vector quantization and projection neural network, in: *Lecture Notes in Computer Science*, vol. 686, Springer, New York, 1993, pp. 328–333.
- [8] P. Demartines, J. Héroult, Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets, *IEEE Transactions on Neural Networks* 8 (1) (1997) 148–154.
- [9] G. Di Battista, P. Eades, R. Tamassia, I. Tollis, *Graph Drawing: Algorithms for the Visualization of Graphs*, Prentice-Hall, Englewood-Cliffs, NJ, 1999.
- [10] D. Donoho, C. Grimes, Hessian eigenmaps: locally linear embedding techniques for high-dimensional data, *Proceedings of the National Academy of Arts and Sciences*, vol. 100, 2003.
- [11] K. Fukunaga, Intrinsic dimensionality extraction, in: P. Krishnaiah, L. Kanal (Eds.), *Classification, Pattern Recognition and Reduction of Dimensionality*, Handbook of Statistics, vol. 2, Elsevier, Amsterdam, 1982, pp. 347–360.
- [12] J. Héroult, C. Jaussions-Picaud, 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, J. Sánchez (Eds.), *Proceedings of IWANN'99*, vol. II, Springer, Alicante, Spain, 1999, pp. 635–644.
- [13] I. Jolliffe, *Principal Component Analysis*, Springer, New York, NY, 1986.
- [14] T. Kohonen, Self-organization of topologically correct feature maps, *Biological Cybernetics* 43 (1982) 59–69.
- [15] T. Kohonen, *Self-Organizing Maps*, second ed., Springer, Heidelberg, 1995.
- [16] M. Kramer, Nonlinear principal component analysis using autoassociative neural networks, *AIChE Journal* 37 (2) (1991) 233–243.
- [17] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, *Psychometrika* 29 (1964) 1–28.
- [18] J. Lee, C. Archambeau, M. Verleysen, Locally linear embedding versus isotop, in: M. Verleysen (Ed.), *Proceedings of 11th European Symposium on Artificial Neural Networks, ESANN 2003*, d-side, Bruges, Belgium, 2003, pp. 527–534.
- [19] J. Lee, A. Lendasse, M. Verleysen, Curvilinear distances analysis versus isomap, in: M. Verleysen (Ed.), *Proceedings of 10th European Symposium on Artificial Neural Networks, ESANN 2002*, d-side, Bruges, Belgium, 2002, pp. 185–192.
- [20] J. Lee, M. Verleysen, Curvilinear distance analysis versus isomap, *Neurocomputing* 57 (2004) 49–76.
- [21] J. Lee, M. Verleysen, Nonlinear dimensionality reduction of data manifolds with essential loops, *Neurocomputing* 67 (2005) 29–53.
- [22] J. Lee, M. Verleysen, *Nonlinear Dimensionality Reduction*, Springer, Berlin, 2007.
- [23] J. Lee, M. Verleysen, Rank-based quality assessment of nonlinear dimensionality reduction, in: M. Verleysen (Ed.), *Proceedings of 16th European Symposium on Artificial Neural Networks, ESANN 2008*, d-side, Bruges, 2008, pp. 49–54.
- [24] J. Mao, A. Jain, Artificial neural networks for feature extraction and multivariate data projection, *IEEE Transactions on Neural Networks* 6 (2) (1995) 296–317.
- [25] B. Nadler, S. Lafon, R. Coifman, I. Kevrekidis, Diffusion maps, spectral clustering and eigenfunction of Fokker–Planck operators, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), *Advances in Neural Information Processing Systems (NIPS 2005)*, vol. 18, MIT Press, Cambridge, MA, 2006.
- [26] E. Oja, Data compression, feature extraction, and autoassociation in feedforward neural networks, in: T. Kohonen, K. Mäkinen, O. Simula, J. Kangas (Eds.), *Artificial Neural Networks*, vol. 1, Elsevier Science Publishers, B.V., North-Holland, 1991, pp. 737–745.
- [27] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, *Science* 290 (5500) (2000) 2323–2326.
- [28] M. Saerens, F. Fous, L. Yen, 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)*, Lecture Notes in Artificial Intelligence, vol. 3201, Pisa, Italy, 2004.

- [29] J. Sammon, A nonlinear mapping algorithm for data structure analysis, *IEEE Transactions on Computers* CC-18 (5) (1969) 401–409.
- [30] L. Saul, S. Roweis, Think globally, fit locally: unsupervised learning of nonlinear manifolds, *Journal of Machine Learning Research* 4 (2003) 119–155.
- [31] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* 10 (1998) 1299–1319.
- [32] R. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function (parts 1 and 2), *Psychometrika* 27 (1962) 125–140, 219–249.
- [33] Y. Takane, F. Young, J. de Leeuw, Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features, *Psychometrika* 42 (1977) 7–67.
- [34] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, *Science* 290 (5500) (2000) 2319–2323.
- [35] W. Torgerson, Multidimensional scaling, I: theory and method, *Psychometrika* 17 (1952) 401–419.
- [36] J. Venna, Dimensionality reduction for visual exploration of similarity structures, Ph.D. Thesis, Helsinki University of Technology, Espoo, Finland, June 2007.
- [37] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: an experimental study, in: G. Dorffner, H. Bischof, K. Hornik (Eds.), *Proceedings of ICANN 2001*, Springer, Berlin, 2001, pp. 485–491.
- [38] J. Venna, S. Kaski, Local multidimensional scaling, *Neural Networks* 19 (2006) 889–899.
- [39] J. Venna, S. Kaski, Nonlinear dimensionality reduction as information retrieval, in: M. Meila, X. Shen (Eds.), *Proceedings of the 11th International Conference on Artificial Intelligence and Statistics (AISTATS 2007)*, San Juan, Puerto Rico, 2007, pp. 568–575.
- [40] T. Villmann, R. Der, M. Herrmann, T. Martinetz, Topology preservation in self-organizing feature maps: exact definition and measurement, *IEEE Transactions on Neural Networks* 8 (2) (1997) 256–266.
- [41] K. Weinberger, L. Saul, Unsupervised learning of image manifolds by semidefinite programming, *International Journal of Computer Vision* 70 (1) (2006) 77–90, in: A. Bobick, R. Chellappa, L. Davis (Guest Eds.), *Special Issue: Computer Vision and Pattern Recognition-CVPR 2004*.
- [42] K. Weinberger, F. Sha, L. Saul, Learning a kernel matrix for nonlinear dimensionality reduction, in: *Proceedings of the 21st International Conference on Machine Learning (ICML-04)*, Banff, Canada, 2004.
- [43] H. Whitney, Differentiable manifolds, *Annals of Mathematics* 37 (3) (1936) 645–680.
- [44] G. Young, A. Householder, Discussion of a set of points in terms of their mutual distances, *Psychometrika* 3 (1938) 19–22.



**John Aldo Lee** was born in 1976 in Brussels, Belgium. He received the M.Sc. degree in Applied Sciences (Computer Engineering) in 1999 and the Ph.D. degree in Applied Sciences (Machine Learning) in 2003, both from the Université catholique de Louvain (UCL, Belgium). His main interests are nonlinear dimensionality reduction, intrinsic dimensionality estimation, independent component analysis, clustering, and vector quantization.

He is a former member of the UCL Machine Learning Group and is now a Postdoctoral Researcher of the Belgian Fonds National de la Recherche Scientifique (F.N.R.S.). His current work aims at developing specific image enhancement techniques for positron emission tomography in the Molecular Imaging and Experimental Radiotherapy Department of the Saint-Luc University Hospital (Belgium).



**Michel Verleysen** was born in 1965 in Belgium. He received the M.S. and the Ph.D. degrees in Electrical Engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an Invited Professor at the Swiss Ecole Polytechnique Fédérale de Lausanne (E.P.F.L.), Switzerland in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris 1—Panthéon-Sorbonne in 2002–2004.

He is now a Research Director of the Belgian Fonds National de la Recherche Scientifique (F.N.R.S.) and Lecturer at the Université catholique de Louvain. He is Editor-in-Chief of the *Neural Processing Letters* journal, Chairman of the Annual European Symposium on Artificial Neural Networks (ESANN) Conference, Associate Editor of the *IEEE Transactions on Neural Networks* journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He is the author or the co-author of about 200 scientific papers in international journals and books or communications to conferences with reviewing committee.

He is the co-author of the scientific popularization book on artificial neural networks in the series “Que Sais-Je?,” in French. His research interests include machine learning, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high-dimensional data analysis.