ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

Rank-based quality assessment of nonlinear dimensionality reduction

John A. Lee[1] and Michel Verleysen[2] *

1- Université catholique de Louvain - Molecular imaging and experimental radiotherapy
Avenue Hippocrate, 54/69, B-1200 Brussels - Belgium

2- Université catholique de Louvain - Machine learning group
Place du Levant, 3, B-1348 Louvain-la-Neuve - Belgium

Abstract. Nonlinear dimensionality reduction aims at providing low-dimensional representions of high-dimensional data sets. Many new methods have been proposed in the recent years, but the question of their assessment and comparison remains open. This paper reviews some of the existing quality measures that are based on distance ranking and K-ary neighborhoods. Many quality criteria actually rely on the analysis of one or several sub-blocks of a co-ranking matrix. The analogy between the co-ranking matrix and a Shepard diagram is highlighted. Finally, a unifying framework is sketched, new measures are proposed and illustrated in a short experiment.

1 Introduction

Dimensionality reduction (DR) gathers techniques that provide a meaningful low-dimensional representation of a high-dimensional data set. Linear DR is well known, with techniques such as principal component analysis. On the other hand, nonlinear dimensionality reduction [1] (NLDR) emerged later and has deeply evolved for the past twenty five years, with neural approaches [2, 3, 4] and spectral techniques [5, 6, 7]. Modern NLDR encompasses the domain of manifold learning and is also closely related to graph embedding [8].

In the most general setting, 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 $\mathbf{X} = [\mathbf{x}_i]_{1 \leq i \leq N}$. In manifold learning, it is assumed that the vectors in $\boldsymbol{\Xi}$ are sampled from a manifold. NLDR aims at providing a low-dimensional representation that is meaningful in some sense. Most often, the goal is to preserve the structure of the data set, which is indicated by proximities, similarities, or neighborhood relationships. These proximities can be obtained by measuring pairwise distances in $\boldsymbol{\Xi}$ with some metric.

As a matter of fact, the scientific community has been focusing on the design of new NLDR methods and the question of quality assessment remains mostly unanswered. As most NLDR methods optimize a given objective function, a simplistic way to assess the quality is to look at the value of the objective function after convergence. Obviously, this allows us to compare several runs with e.g. different parameter values, but makes the comparison of different methods unfair. Furthermore, objective functions often fulfill some requirements, such

*JAL is a Postdoctoral Researcher with the Belgian National Fund for Scientific Research.

as being continuous and differentiable, in order to be easily optimized. In contrast, a quality criterion just needs to be evaluated and so these constraints are no longer relevant. This allows the development of quality criteria that are potentially more complex and truly assess the preservation of the data set structure. First attempts in this direction can be found in the particular case of self-organizing maps [2]; see for instance the topographic product [9] and the topographic function [10]. More recently, new criteria for quality assessment have been proposed, with a broader applicability, such as the trustworthiness and continuity measures [11], the local continuity meta-criterion [12], and the mean relative rank errors [1]. 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.

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

This paper is organized as follows. Section 2 introduces the notations for distances, ranks, and neighborhoods. Section 3 reviews existing rank-based criteria. Section 4 unifies the different approaches and proposes new ones. Section 5 shows some experimental results. Finally, Section 6 draws the conclusions.

2 Distances, ranks, and neighborhoods

Most NLDR techniques involve distances in a more or less direct way. The symbol δ_{ij} denotes the distance from $\boldsymbol{\xi}_i$ to $\boldsymbol{\xi}_j$ in the high-dimensional space. Similarly, d_{ij} is the distance from $\mathbf{x}_i$ to $\mathbf{x}_j$ in the low-dimensional space. Notice that we assume that $\delta_{ij} = \delta_{ji}$ and $d_{ij} = d_{ji}$, although this hypothesis is not always required. For instance, it does not hold true if δ_{ij} and δ_{ji} come from distinct experimental measures. 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 $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } k < j)\}|$. 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 $\{1, \ldots, N-1\}$. 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 can then be defined as $\mathbf{Q} = [q_{kl}]_{1 \leq k,l \leq N-1}$ with $q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|$. The co-ranking matrix is the joint histogram of the ranks and is actually a sum of N permutation matrices of size $N-1$. Matrix $\mathbf{Q}$ can be divided into four blocks that separate the first K rows and columns. If we define $\mathbb{F}_K = \{1, \ldots, K\}$ and $\mathbb{L}_K = \{K+1, \ldots, N-1\}$, the index sets of the four blocks are $\mathbb{UL}_K = \mathbb{F}_K \times \mathbb{F}_K$, $\mathbb{UR}_K = \mathbb{F}_K \times \mathbb{L}_K$, $\mathbb{LL}_K = \mathbb{L}_K \times \mathbb{F}_K$, and

$\mathbb{LR}_K = \mathbb{L}_K \times \mathbb{L}_K$. Similarly, the block covered by $\mathbb{UL}_K$ can be split into its main diagonal $\mathbb{D}_K = \{(i,i) : 1 \leq i \leq K\}$ and lower and upper triangles $\mathbb{LT}_K = \{(i,j) : 1 < i \leq K \text{ and } j < i\}$ and $\mathbb{UT}_K = \{(i,j) : 1 \leq i < K \text{ and } j > i\}$. The co-ranking matrix can also be displayed and interpreted in a similar way as a Shepard diagram [13]. (Shepard's scatterplot shows the distances δ_{ij} with respect to the distances d_{ij}, for all pairs (i,j), with $i \neq j$.)

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. For instance, the trustworthiness and continuity (T&C) measures [11] are defined as:

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

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

where the normalizing factor $G_K = N \min\{K(2N-3K-1), (N-K)(N-K-1)\}$ considers the worst case, i.e. ranks are reversed in the low-dimensional space [14]. Notice that the two criteria distinguish two kinds of errors: faraway points that becomes neighbors and neighbors that are embedded faraway from each other.

The mean relative rank errors [1] (MRREs) rely on the same principle and are defined as

$$E_n(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in n_i^K} \frac{|\rho_{ij} - r_{ij}|}{\rho_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{UL}_K \cup \mathbb{LL}_K} \frac{|k-l|}{l} q_{kl} \ , \quad (3)$$

$$E_\nu(K) = \frac{1}{H_K} \sum_{i=1}^{N} \sum_{j \in \nu_i^K} \frac{|\rho_{ij} - r_{ij}|}{r_{ij}} = \frac{1}{H_K} \sum_{(k,l) \in \mathbb{UL}_K \cup \mathbb{UR}_K} \frac{|k-l|}{k} q_{kl} \ , \quad (4)$$

where the normalizing factor $H_K = N \sum_{k=1}^{K} |N-2k|/k$ considers the worst case. 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 $\mathbf{Q}$.

The local continuity meta-criterion [12] (LCMC) is defined as

$$U_{\text{LC}}(K) = \frac{1}{NK} \sum_{i=1}^{N} \left(|n_i^K \cap \nu_i^K| - \frac{K^2}{N-1} \right) = \frac{K}{1-N} + \frac{1}{NK} \sum_{(k,l) \in \mathbb{UL}_K} q_{kl} \ , \quad (5)$$

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 $\mathbf{Q}$. Elements q_{kl} in the block are not weighted in the sum.

ESANN'2008 proceedings, European Symposium on
Artificial Neural Networks - Advances in Computational Intelligence and Learning
Bruges (Belgium), 23-25 April 2008, d-side publi., ISBN 2-930307-08-0.

4 Unifying framework

The quality measures described in the previous section can be related to the concepts of *precision* and *recall* (P&R) in the domain of information retrieval. The precision is the proportion of relevant items among the retrieved ones, whereas the recall is the proportion of retrieved items among the relevant ones. For rank-based criteria, relevant items are the indices that belong to ν_i^K, whereas n_i^K contains the retrieved indices. The P&R are themselves related to the concepts of false positive and false negative in classification. False positive decrease the precision and false negatives decrease the recall.

If we compare the retrieved neighborhoods to the relevant ones, the blocks of $\mathbf{Q}$ covered by $\mathbb{UL}_K$, $\mathbb{LL}_K$, $\mathbb{UR}_K$, and $\mathbb{LR}_K$ contain the true positives, the false positives, the false negatives, and the true negatives, respectively. Hence, the LCMC quantifies the true positives, the T&C focus on the false positives and false negatives, and the MRREs encompass the positives (true and false) and negatives (true and false). Obviously, as n_i^K and ν_i^K have the same size, the numbers of false positives and false negatives are the same. Each element of ν_i^K that is missed in n_i^K (a false negative) is replaced with an incorrect neighbor (a false positive). Formally, as $\mathbf{Q}$ is a sum of N permutation matrix, we can see that $\sum_{l=1}^{N-1} q_{kl} = N$ and $\sum_{k=1}^{N-1} q_{kl} = N$. Therefore,

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

This shows that we would have $W_{\text{T}}(K) = W_{\text{C}}(K)$ and $E_\nu(K) = E_n(K)$, without an appropriate weighting of the elements q_{kl}. On the other hand, the absence of weighting in the LCMC is 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, the interest does not lie in the average *number* of false positives/negatives in K-ary neighborhoods, but rather in *how bad* they are misranked. This suggests that meaningful criteria should be computed on both sides of the diagonal of $\mathbf{Q}$. This resembles the usual interpretation of a Shepard diagram. The two regions to consider are thus $\mathbb{LT}_K \cup \mathbb{LL}_K$ and $\mathbb{UT}_K \cup \mathbb{UR}_K$. They respectively gather elements of n_i^K for which $\rho_{ij} > r_{ij}$ and elements of ν_i^K for which $r_{ij} > \rho_{ij}$. In the same way as the existing criteria, we can distinguish weighted and unweighted sums of elements q_{kl}. For instance, weighted averages can be written as

$$W_{\text{LT}}^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in\mathbb{LT}_K\cup\mathbb{LL}_K} \frac{(k-l)^v}{k^w}\, q_{kl} \ , \tag{6}$$

$$W_{\text{UT}}^{v,w}(K) = \frac{1}{C_K} \sum_{(k,l)\in\mathbb{UT}_K\cup\mathbb{UR}_K} \frac{(l-k)^v}{l^w}\, q_{kl} \ , \tag{7}$$

where $C_K = N\sum_{k=1}^{K} \max\{0, (N-2k)^w/k^v\}$. The exponents w and v can be adjusted in order to emphasize large rank differences, relatively to the reference

rank. Unweighted average can be written as

$$U_{\text{UT}}(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{UT}_K} q_{kl} \ , \quad U_{\text{LT}}(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{LT}_K} q_{kl} \ , \tag{8}$$

and $U_{\text{D}}(K) = \frac{1}{KN} \sum_{(k,l)\in \mathbb{D}_K} q_{kl}$. Notice that $W_{\text{LT}}^{0,0} - W_{\text{UT}}^{0,0} = U_{\text{LT}}(K) - U_{\text{UT}}(K)$ and $U_{\text{LC}}(K) = U_{\text{D}}(K) + U_{\text{UT}}(K) + U_{\text{LT}}(K) - K/(N-1)$. The quantity $U_{\text{D}}(K)$ indicates the average proportion of vectors that keep the same rank in both ν_i^K and n_i^K. Similarly, $U_{\text{LT}}(K)$ and $U_{\text{UT}}(K)$ are the average proportions of vectors that are respectively "promoted" and "downgraded", but still remain in both ν_i^K and n_i^K. If $U_{\text{LT}}(K) \approx U_{\text{UT}}(K) \gg U_{\text{D}}(K)$, then many permutations occur in the K-ary neighborhoods (due to noise flattening, for instance). On the other hand, if the difference $|U_{\text{LT}}(K) - U_{\text{UT}}(K)|$ is large, then vectors that remain in the K-ary neighborhoods are promoted due to the escape of some others.

5 Experiments

For this brief experimental section, the test manifold is a (hollow) unit sphere. Thousand randomly drawn points of the manifold are available; Gaussian noise is added, with a standard deviation equal to 0.05. The manifold has been embedded with Sammon's nonlinear mapping [15] (NLM) and curvilinear component analysis [3] (CCA). Due to space limitations, we show only the value of $U_{\text{LT}}(K)$, $U_{\text{UT}}(K)$, $U_{\text{D}}(K)$, and their sum in Fig. 1. 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). The results show that these two antagonist behaviors can be markedly distinguished using unweighted averages of elements of the co-ranking matrix.

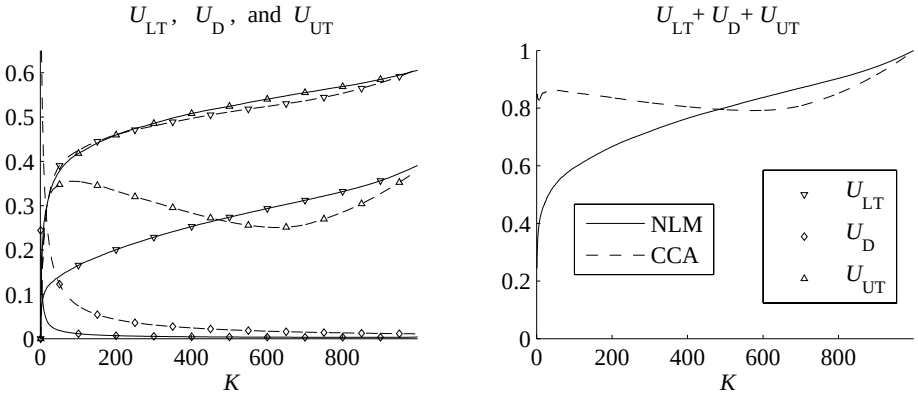

Fig. 1: Values of $U_{\text{LT}}(K)$, $U_{\text{UT}}(K)$, $U_{\text{D}}(K)$, and their sum, for an hollow unit sphere (1000 points with Gaussian noise; $\sigma = 0.05$). The two compared methods are Sammon's NLM and CCA.

6 Conclusions

This paper has reviewed several quality criteria for the assessment of nonlinear dimensionality reduction. All of them rely on distance rankings and the definition of a co-ranking matrix allows us to cast them within a unifying framework. The literature emphasizes the connection of these rank-based criteria with fundamental concepts taken from information retrieval (precision and recall) or classification (false positives and false negatives). Properties of the co-ranking matrix show that these analogies should however be considered carefully.

In constrast, we see that the co-ranking matrix can be interpreted in the same way as a Shepard diagram. This means that quality crieria should focus on the density on both sides of the diagonal of the co-ranking matrix. An experiment with these densities reveals the antagonist behavior of two well known NLDR techniques such as Sammon's NLM and CCA.

References

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

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

[3] 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.

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

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

[6] 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.

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

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

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

[10] 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.

[11] 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.

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

[13] 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.

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

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