Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 4 (2011) 538–547

Procedia Computer Science

International Conference on Computational Science, ICCS 2011

Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants

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

[a]*Department of Molecular Imaging, Radiotherapy, and Oncology, Université catholique de Louvain, Brussels (Belgium)*
[b]*Machine Learning Group, ICTEAM institute Université catholique de Louvain, Louvain-la-Neuve (Belgium)*

Abstract

Dimensionality reduction aims at representing high-dimensional data in low-dimensional spaces, mainly for visualization and exploratory purposes. As an alternative to projections on linear subspaces, nonlinear dimensionality reduction, also known as manifold learning, can provide data representations that preserve structural properties such as pairwise distances or local neighborhoods. Very recently, similarity preservation emerged as a new paradigm for dimensionality reduction, with methods such as stochastic neighbor embedding and its variants. Experimentally, these methods significantly outperform the more classical methods based on distance or transformed distance preservation. This paper explains both theoretically and experimentally the reasons for these performances. In particular, it details (i) why the phenonomenon of distance concentration is an impediment towards efficient dimensionality reduction and (ii) how SNE and its variants circumvent this difficulty by using similarities that are invariant to shifts with respect to squared distances. The paper also proposes a generalized definition of shift-invariant similarities that extend the applicability of SNE to noisy data.

Keywords: Dimensionality reduction, data visualization, norm concentration, similarity preservation, stochastic neighbor embedding

1. Introduction

The interpretation of high-dimensional data remains a difficult task, mainly because human vision cannot deal with spaces having more than three dimensions. Part of this difficulty stems from the curse of dimensionality, a convenient expression that encompasses all weird and unexpected properties of high-dimensional spaces. Dimensionality reduction (DR) aims at constructing a low-dimensional representation of data, in order to improve readability and interpretability. Of course, this low-dimensional representation must be meaningful and faithful to the genuine data. In practice, the representation must preserve important structural properties of the data set, such as relative proximities, similarities or dissimilarities. The general idea is that dissimilar data items should be represented far from each

Email addresses: John.Lee@uclouvain.be (John A. Lee), Michel.Verleysen@uclouvain.be (Michel Verleysen)
[1]Corresponding author

1877–0509
doi:10.1016/j.procs.2011.04.056

other, whereas similar ones should appear close to each other. Dimensionality reduction applies to other purposes than just data visualization. For instance, DR can be used in data compression and denoising. Dimensionality reduction can also preprocess data, with the hope that a simplified representation can accelerate any subsequent processing or improve its outcome.

Linear DR is well known, with techniques such as principal component analysis [1] and classical metric multidimensional scaling [2, 3]. The former tries to preserve the covariances in the low-dimensional space, whereas the latter attempts to reproduce the Gram matrix of pairwise dot products. Nonlinear dimensionality reduction [4] (NLDR) emerged later, with nonlinear variants of multidimensional scaling [5, 6, 7], such as Sammon's nonlinear mapping [8] (NLM), and curvilinear component analysis [9, 10] (CCA). Most of these methods are based on the preservation of pairwise distances. The eighties and early nineties saw the advent of methods inspired by artificial neural networks and soft-computing. Auto-encoders with multilayer perceptrons [11] and Kohonen's self-organizing maps [12] are the most prominent examples in this trend. After the seminal paper describing kernel PCA [13], spectral embedding has met a growing interest. Isomap [14], locally linear embedding [15], maximum variance unfolding [16], and other manifold learners based e.g. on diffusion in graphs [17] are only a few examples of spectral methods [18, 19]. Spectral methods provide the guarantee of finding the global optimum of their cost function. In contrast, methods based on other optimization techniques generally do not offer this advantage. However, they usually compensate for this drawback by their capability of handling a broader range of cost functions, which are potentially more relevant for NLDR. Recent and successful methods following this approach are stochastic neighbor embedding (SNE) [20] and its variants, t-distributed SNE (t-SNE) [21] and NeRV (standing for neighborhood retrieval and visualization) [22].

All these methods attempt to match so-called similarities, which are basically decreasing functions of the pairwise distances. Such functions were already used in the cost functions of methods like Sammon's NLM [8] and CCA [10]. In terms of results, t-SNE and NeRV significantly outperform most of the older methods, as shown in [21, 22, 23, 24] for instance. However, the reasons as to why these methods behave so well still remain obscure.

This paper aims at investigating them pragmatically. It is indeed of utmost importance to explain why features that seem at first sight to be details of design, change fundamentally the behavior of NLDR algorithms, in order to exploit these changes in future developments in the field. For this purpose, we show that SNE and its variants rely on a specific formulation of the similarities, which allows these methods to fight the curse of dimensionality. More specifically, the similarities involve a softmax ratio that has an important property: the normalization of the exponential function makes it invariant to shifts applied to its argument. Such a shift invariance facilitates the NLDR process by circumventing the fact that the phenomenon of norm concentration [25] manifests itself differently in the high-dimensional data space and the low-dimensional visualization space. This difference also explains why NLDR with a naive distance preservation principle is an ill-posed problem.

The remainder of this paper is organized as follows. Section 2 introduces similarity preservation as a way to achieve NLDR and also briefly describes SNE and its variants. Section 3 deals with the phenomenon of norm and distance concentration. Section 4 investigates the property of shift-invariance and proposes a generalized and consistent definition of similarities. Section 5 gathers the experimental results. Finally, Section 6 draws the conclusions and sketches some perspectives for the near future.

2. Stochastic neighbor embedding and its variants

In order for a low-dimensional visualization to be faithful to the data it has to represent, some structural properties must be preserved or reproduced. These properties can be pairwise distances, more general dissimilarities, local neighborhoods, or similarities. The basic principle that drives SNE, t-SNE, and NeRV is similarity preservation. For two points in some space, a similarity is defined as a decreasing function of their distance.

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1\le i\le N}$ denote a set of N points in some M-dimensional space. Similarly, let $\mathbf{X} = [\mathbf{x}_i]_{1\le i\le N}$ be its representation in a P-dimensional space, with $P \le M$. The distances between the ith and jth points are given by $\delta_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2$ and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ in the high- and low-dimensional spaces respectively. The corresponding similarities in SNE are defined for $i \neq j$ by

$$\sigma_{ij} = \frac{\exp(-\delta_{ij}^2/(2\lambda_i^2))}{\sum_{k,k\neq i}\exp(-\delta_{ik}^2/(2\lambda_i^2))} \quad \text{and} \quad s_{ij} = \frac{\exp(-d_{ij}^2/2)}{\sum_{k,k\neq i}\exp(-d_{ik}^2/2)}\ , \tag{1}$$

where λ_i is a bandwidth parameter. If $i = j$, then $\sigma_{ij} = s_{ij} = 0$ by convention. Similarity preservation could be implemented in various ways, such as e.g. a sum of squared similarity differences. However, SNE takes advantages of the fact that the similarities are normalized (they sum to 1). Instead of using cost functions based on squared differences between σ_{ij} and s_{ij}, SNE defines for each point i a Kullback-Leibler divergence $E_i(\mathbf{X}; \boldsymbol{\Xi}, \lambda_i) = \sum_j \sigma_{ij} \log(\sigma_{ij}/s_{ij})$. The resulting cost function can be written as

$$E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}) = \sum_i E_i(\mathbf{X}; \boldsymbol{\Xi}, \lambda_i) = \sum_{i,j} \sigma_{ij} \log(\sigma_{ij}/s_{ij}) \ , \tag{2}$$

and can be minimized with respect to $\mathbf{X}$ by means of a gradient descent. This requires the parameters λ_i in $\boldsymbol{\Lambda}$ to be fixed. For this purpose, let us notice that each $E_i(\mathbf{X}; \boldsymbol{\Xi}, \lambda_i) = \sum_j \sigma_{ij} \log(\sigma_{ij}/s_{ij})$ consists of a constant part that is the entropy of $\boldsymbol{\sigma}_i = [\sigma_{ij}]_{1 \le j \le N}$ and a variable part that is the cross-entropy between $\boldsymbol{\sigma}_i$ and $\mathbf{s}_i = [s_{ij}]_{1 \le j \le N}$. In SNE, the scaling parameters λ_i are adjusted in order to equalize all entropies, namely, $H = \sum_j \sigma_{ij} \log(\sigma_{ij})$ for all i. The user specifies a perplexity value $\exp(H)$ that is trivially converted into the targeted entropy value. The equalization actually ensures that each data point is given the same weight in the cost function. In the computation of its gradient, the combination of logarithms in the divergences and the exponential functions in the similarities yields a very simple update formula: $\mathbf{x}_i \leftarrow \mathbf{x}_i + \alpha \sum_j (\sigma_{ij} - s_{ij} + \sigma_{ji} - s_{ji})(\mathbf{x}_i - \mathbf{x}_j)$, where α is the step size or learning rate.

In NeRV, the cost function also involves the dual KL divergence, which leads to

$$E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}) = \sum_{i,j} (1-\beta)\sigma_{ij} \log(\sigma_{ij}/s_{ij}) + \beta s_{ij} \log(s_{ij}/\sigma_{ij}) \ , \tag{3}$$

where β is a balancing factor. The bandwidth parameters in the similarities are adjusted in the same way as in SNE. The resulting update for the gradient descent is as simple as the one of regular SNE.

In t-SNE, the modifications concern mainly the similarities in the low-dimensional space, which are defined as

$$s_{ij} = \frac{(1 + d_{ij}^2)^{-1}}{\sum_{k,l,k \neq l} (1 + d_{kl}^2)^{-1}} \ . \tag{4}$$

The name of the method stems from the replacement of the Gaussian shape in (1) with an expression that is closely related to the probability density function of a Student t distribution with a single degree of freedom. Another noticible change is the normalization, which runs over both indices instead of one only. The update for the gradient descent becomes $\mathbf{x}_i \leftarrow \mathbf{x}_i + \alpha \sum_j (\sigma_{ij} - s_{ij} + \sigma_{ji} - s_{ji})(\mathbf{x}_i - \mathbf{x}_j)/(1 + d_{ij}^2)$ The discrepancy between the Gaussian similarities in the high-dimensional space and the heavy-tailed ones in the low-dimensional space amounts to applying an exponential transformation to δ_{ij} to obtain d_{ij} [26]. In practice, this transformation stretches the distances and allows t-SNE to yield less cluttered data representations than regular SNE; separations between clusters are also reinforced. In the paper describing t-SNE, this transformation accounts for the superior results of t-SNE, as compared to those of regular SNE. Stretched distances are assumed to circumvent a so-called 'crowding problem' [21], which intuitively refers to fact NLDR requires data distributed in vast (hyper-)volumes to be 'packed' and displayed on a limited surface in the low-dimensional space. This assumption is however questioned in [22], where NeRV is shown to perform as well as t-SNE without using heavy-tailed similarities, and thus without exponential transformation of the distances. These contradictory results motivate a thorough investigation in order to clearly elucidate why similarity preservation can be so successful in SNE, t-SNE and NeRV, and to exploit the successful features of these methods in further developments of the field.

3. The phenomenon of norm and distance concentration

High-dimensional spaces have unexpected and counter-intuitive properties. The difficulty to cope with them has been coined the curse of dimensionality [27, 28]. Among many other aspects, the dimensionality of the space affects the statistical distribution of norms and distances [25]. Let us consider a hypothetical case where vector $\boldsymbol{\xi} \in \mathbb{R}^M$ has a Gaussian distribution, namely, $\boldsymbol{\xi} \sim G(\mathbf{0}, \nu\mathbf{I})$. In this case, we have $\|\boldsymbol{\xi}\|_2^2/\nu^2 = \boldsymbol{\xi}^T\boldsymbol{\xi}/\nu^2 \sim \chi_M^2$ and $\|\boldsymbol{\xi}\|_2/\nu \sim \chi_M$. Figure 1 shows the probability density function of several χ_M^2 distributions for various values of M. For an increasing dimensionality M, the mode situated at $\max(0, M-2)$ drifts to the right without sufficient thickening, since the standard

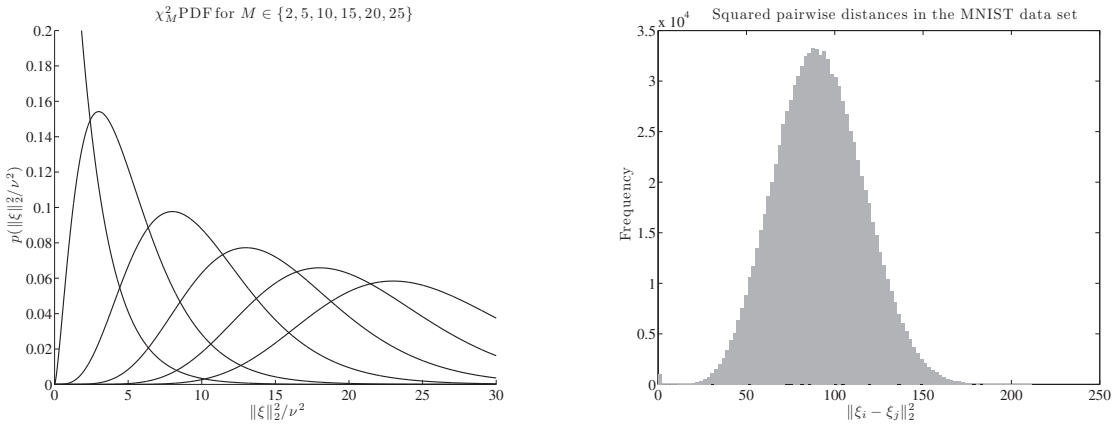

Figure 1: (left) The probability density function of the χ^2_M distribution for $M \in \{2, 5, 10, 15, 20, 25\}$ (curves from left to right). (right) Histogram of squared pairwise Euclidean distances for 1000 points in the MNIST data set of handwritten digit images. Notice the small secondary mode at zero, caused by reflexive distances.

deviation divided by the mean is equal to $\sqrt{2M}/M$ and tends to 0. This shows that the squared Euclidean norm concentrates [25]. For two vectors $\boldsymbol{\xi}_1$ and $\boldsymbol{\xi}_2$ drawn independently from $G(\boldsymbol{\mu}, \nu\mathbf{I})$, we know that $(\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2) \sim G(\mathbf{0},\ \sqrt{2}\nu\mathbf{I})$ and thus $\|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_2^2/(2\nu^2) \sim \chi^2_M$. The case of pairwise distances computed within a finite set of vectors is very similar up to a subtle difference. Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1\le i\le N}$ be a finite set of vectors, with $\boldsymbol{\xi}_i \in \mathbb{R}^M$ and $\boldsymbol{\xi}_i \sim G(\boldsymbol{\mu}, \nu\mathbf{I})$. The statement $\delta^2_{ij}/(2\nu) \sim \chi^2_M$ holds only for nonreflexive distances, that is, for $j \neq i$. Reflexive distances have a different and trivial distribution, namely the delta distribution, which can be written as $\delta^2_{ii} \sim G(\mathbf{0}, 0\mathbf{I})$. The probability density function (PDF) of squared pairwise Euclidean distances δ^2_{ij} can thus be written as

$$p\,(u; M, N) = \frac{1}{N}\delta(0) + \frac{N-1}{N}\frac{u^{M/2-1}\exp(-u/2)}{2^{M/2-1}\Gamma(M/2)} \quad , \tag{5}$$

where $\delta(u)$ denotes Dirac's delta function.

The concentration phenomenon applies to many distributions of points [25]. The right side of Fig. 1 shows the histogram of pairwise distances in a real data set [29]. Concentration phenomenon is clearly visible, as well as the small spike near zero that accounts for the reflexive distances δ_{ii}.

The fact that distance concentration varies with the dimensionality explains one of the most prominent difficulties of DR, when it relies on distance preservation. A necessary (but still insufficient) requirement is that the shapes of the distance distributions in both spaces approximately match each other. Figure 1 shows that this is not the case: the shapes of a χ^2_2 and e.g. a χ^2_{20} are irremediably different. This also means that DR with a cost function that involves differences of (squared) distances [8, 10] makes little sense. In other words, distance preservation should be replaced with another paradigm that takes into account the dimensionalities of both spaces. An effective idea is to transform the distances with a function that cancels the concentration, regardless of the dimensionality. The simplest transformation is linear, with a shift and a rescaling. Visually, this would allow the modes of all distributions in Fig. 1 to shift and to maximize their overlap. This idea is further supported by Fig. 2 that shows the loci $\{(u, v)\} \subset \mathbb{R}^2$ with u and v such that the cumulative distribution function (CDF) of a χ^2_M equals that of a χ^2_2. Distance shifting and scaling can diminish the discrepancy between the distance distributions in the high- and low-dimensional spaces. However, the shift cannot be applied to the reflexive distances, which would otherwise become negative. If the reflexive distances are not transformed, we can rely on the fact that in a finite data set the probability to actually observe nonreflexive distances within the first percentiles of the χ_M distribution is very low. The maximal shift amplitude then corresponds to the minimal nonreflexive distance, i.e. $\min_{j,j\neq i}\delta^2_{ij}$, and hopefully suffices to bring the curve elbow near zero in Fig. 2. In practice, however, excluding the diagonal of the matrix of pairwise distances might not suffice for two reasons. First, the probability to observe a distance within the first percentiles is very low but not exactly null. Second, real

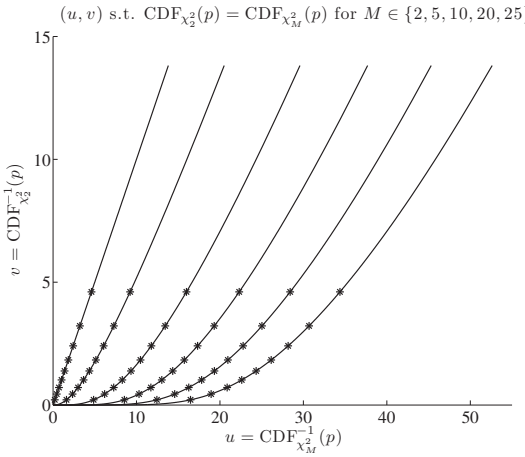

Figure 2: Loci for which a χ^2_2 CDF equals a χ^2_M CDF, showing that DR driven by naive distance preservation proves to be a bad idea. Distance preservation would be possible up to a scaling if the curves were straight lines passing through the origin. The percentiles $\{10, \ldots, 90\}$ indicate that the first flat segment of each curve is sparsely populated and likely to be empty in a finite sample. If this segment is neglected, a linear approximation of each curve suggests that the most basic transformation of the squared distances consists of a shift to the left and a scaling.

data could contain duplicates and/or near-duplicates. Typically, one can imagine that a document collection contains several almost identical copies (the original and scanned versions, for instance). These two arguments suggest that small, unexpected distances should be discarded like reflexive ones, in order to allow for a maximal shift.

So far, the provided theoretical framework overlooks the density differences at each data point. If the density of points around $\boldsymbol{\xi}_i$ is lower than that around $\boldsymbol{\xi}_j$, then $\min_{k,k\neq i} \delta^2_{ik}$ is likely to be much larger than $\min_{k,k\neq j} \delta^2_{jk}$. Hence, distances δ_{ik} from the ith vector and δ_{jk} from the jth one, with $i \neq j$ and $1 \leq k \leq N$, should be shifted differently. In other words, a smart way to determine the best shift and scaling for each data point must be found.

4. Shift-invariant similarities

In similarities defined as softmax ratios, shift invariance results from the equality

$$\sigma_{ij} = \frac{\exp(-\delta^2_{ij}/(2\lambda_i^2))}{\sum_{k,k\neq i} \exp(-\delta^2_{ik}/(2\lambda_i^2))} = \sigma_{ij} \frac{\exp(S_i^2)}{\exp(S_i^2)} = \frac{\exp(S_i^2 - \delta^2_{ij}/(2\lambda_i^2))}{\sum_{k,k\neq i} \exp(S_i^2 - \delta^2_{ik}/(2\lambda_i^2))} \ , \tag{6}$$

where S_i is a shift to the left and λ_i a scaling factor. Working with squared distances δ^2_{ij} or with shifted quantities $\delta^2_{ij} - 2S_i^2\lambda_i^2$ has thus no impact on the final similarity values. The similarities in the high- and low-dimensional spaces can therefore match without knowing the value of S_i. As a matter of fact, this matching is possible for two reasons already mentioned in the previous section. First, reflexive distances must remain equal to zero and are thus not considered in the similarity definition (the case $k = i$ is excluded in the normalization factor). Second, the main mode of a distance distribution has a left tail that is very thin. This is visible in Fig. 1 (right), where several of the first bins between the peak of reflexive distances and the main mode are empty. These empty bins are of the utmost importance since positivity of the shifted distances limits the shift amplitude to $S_i \leq \min_{k,k\neq i} 2^{-1/2}\delta_{ik}/\lambda_i$. The last inequality also confirms that the null reflexive distance must be excluded in the normalization of the similarity, as done in (1) and (4) for SNE and t-SNE respectively. Allowing the case $k = i$ would lead trivially to $\min_k \delta_{ik} = 0$ and thus $S_i \leq 0$. Any shift to the left would therefore be impossible.

In practice, one might conjecture that the constraint $k \neq i$ in the normalization factor could be insufficient in some specific cases. For instance, with a very large data set, the probability to observe small distances in the left tail is increased. Such a small distance could limit the shift amplitude and therefore would almost annihilate the invariance to shift. Another case where unexpected small distances occur is the presence of duplicated data items, resulting in

$\delta_{ij} = 0$ for $i \neq j$. If duplicates are easy to eliminate, near duplicates are more of an issue. They can be associated with particular noise patterns that thicken the left tail of the distance distribution and annihilates shift invariance.

A typical example would be to consider text and image documents. Databases are likely to include duplicates of the same document. In the case of photocopies, scanned documents, watermarked copies, or even character recognition typos, the copies slightly differ from the genuine document. In this example, the distribution of pairwise distances consists of three components: reflexive distances, spurious small distances caused by noise, and eventually large distances related to the video content. Only the third component is of interest and a good visualization requires the shift to be large enough in order to overlook the spurious distances.

From a completely different viewpoint, one might point out a lack of consistency in the similarity definition as it is proposed in SNE and its variants. The constraint $k \neq i$ in the normalization is equivalent to imposing $\sigma_{ii} = 0$ for $\delta_{ii} = 0$, which is not really compatible with the requirement that similarities should be positive decreasing functions of the distances. On the other hand, a small but nonzero distance leads to a similarity that almost attains its maximal value according to the definition in (1). This jump from zero to a large similarity value is also difficult to reconcile with our claim that small distances are likely to be spurious and should therefore be ignored just as reflexive distances. In order to address this consistency issue, we propose to modify and rewrite the definitions of the similarities in the high- and low-dimensional spaces as

$$\sigma_{ij} = \frac{\exp(-\max(\delta_{ij}, \tau_i)^2/(2\lambda_i^2))}{\sum_k \exp(-\max(\delta_{ik}, \tau_i)^2/(2\lambda_i^2))} \quad \text{and} \quad s_{ij} = \frac{\exp(-\max(d_{ij}, t_i)^2/2)}{\sum_k \exp(-\max(d_{ik}, t_i)^2/2)} \ , \tag{7}$$

where τ_i and t_i are thresholds. The exclusion of reflexive distances previously embodied by the constraint $k \neq i$ in the normalization can be replaced here with $\tau_i = \min_{k,k\neq i} \delta_{ik}$. This guarantees shift invariance in the high-dimensional space. As shift-invariance in the low-dimensional space is not necessary, t_i can be null. In the new definition, similarities are indeed decreasing functions of the distances. Moreover, the new formulation is more flexible than the previous one, as threshold τ_i can take on any value. With $\tau_i > 0$, all zero distances are discarded (reflexive ones and those related to duplicates). Instead of setting τ_i to the shortest non-reflexive distance, one can also equate it with the distance to the Tth nearest neighbor of $\boldsymbol{\xi}_i$, for any $T \geq 1$. Spurious distances can be ignored in this way with a well chosen value of T. At the same time, this TNN-like strategy also ensures that the value of τ_i is driven by data and adjusted in an adaptive way.

Eventually, we can also compare graphically the shapes of the various similarity definitions and relate them with the distribution of distances. In Fig. 3, the non-squared non-reflexive pairwise Euclidean distances follow a χ_M distribution, depicted by the dotted curve. The second curve (solid black line) corresponds to the complementary CDF (CCDF) of this distribution, which is arguably the optimal similarity in terms of contrast for the considered distance distribution [30, 25]. The third curve (dashed gray line) is drawn according the similarity definition used in SNE and its variants; the discontuinity at zero is represented. The fourth and last curve (dashed black line) is the truncated similarity that we propose: it is constant on the left side of threshold τ_i and Gaussian on the right side. As a matter of fact, the truncated similarity has mostly the same shape as the CCDF of the actual distribution. In contrast, the left part of the similarity as used in SNE totally differs but it is unlikely to be exploited as the probability to observe a distance in this region is very low in practice. Eventually, one might wonder why we should not prefer the CCDF to the truncated similarity as it is theoretically optimal. The reason is that the CCDF depends on several parameters, such as the number of degrees of freedom and the scaling, whose value is not known and difficult to estimate. Compared to the CCDF, the truncated similarity might look as a poor approximation. However, it is computationally much simpler and its parameterization is not as intricate as in the CCDF. As indicated above, there are straightforward and adaptive ways to adjust parameters τ_i and λ_i.

5. Experiments

This section aims to verify experimentally that shift invariant similarities are the key ingredient that allows SNE and its variants to outperform other NLDR approaches. For this purpose, we show in various cases that the embedding quality depends on an effective treatment of both reflexive and spurious distances. Any limitation of the shift caused by these distances degrades the embedding quality. In practice, we compare the similarities as defined in (1) with those that we propose in (7). The former definition discards only reflexive distances. The latter can show what happens

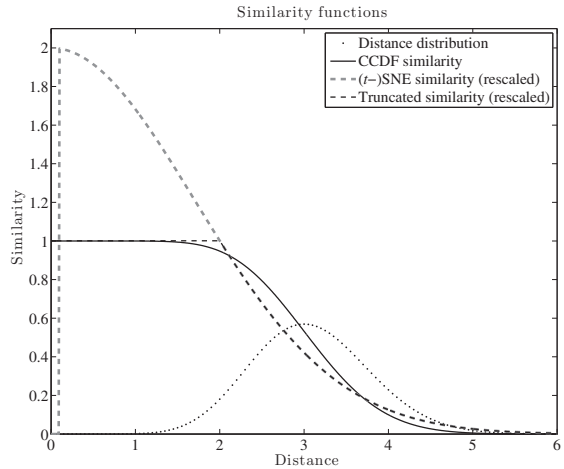

Figure 3: The shape of several similarity functions. The dotted curve is the PDF of non-reflexive Euclidean distances drawn from a 10-dimensional Gaussian distribution (thus a χ_{10} PDF). The solid black curve corresponds to the CCDF of the distance distribution, i.e. the most discriminant similarity for that particular distribution. The thick gray dashed curve shows the shift-invariant pseudo-similarity used in (*t*-)SNE and NeRV. The dashed black curve illustrates the truncated shift-invariant softmax similarity that we propose. The last two similarity functions are rescaled to match the CCDF similarity; the actual maximum value depends on the normalization denominator.

if all distances are considered ($\tau_i = 0$), if only reflexive distances are ignored (τ_i equal to the distance to the nearest neighbor), and if other spurious distances are eliminated (τ_i equal to the distance to the Tth nearest neighbor).

Three data sets are considered. The first one is an academic example: 1000 points are drawn from a 30-dimensional Gaussian distribution and we wish to obtain the best 2D visualization of the sample. Unlike the full Gaussian distribution, a finite sample can be expected to have some random structure that DR methods could reveal. The second data set contains real data; it is a randomly drawn subset of the MNIST data base of handwritten digits [29]. This data set comprises 1000 28-by-28 gray-level images whose pixels are rearranged in 784-dimensional vectors. A few examples are shown on the left of Fig. 4. A 2D visualization is sought here too. The third data set is closely derived from the second one. Half of the vectorized images are duplicated and speckle noise is added to all pairs of copies. If p denotes a pixel intensity in the noisefree image, then we know that $0 \leq p \leq 1$ and its noisy value can be written as $\min(1, p + \text{ind}(u < 0.01)v/2)$, where ind is an indicator function and u and v are drawn from a uniform distribution between 0 and 1. The third data set is intended to investigate the effect of spurious distances. All visualizations are computed with the same cost function as in *t*-SNE (perplexity set to 50, 300 iterations). To provide comparison points, the results of Torgerson's classical metric MDS [3] (CMMDS) and Demartines' curvilinear component analysis [10] (CCA) are also shown. The former is linear and equivalent to PCA, while the latter is an advanced nonlinear method.

In order to assess the quality of dimensionality reduction, we use one of the performance indices decribed in [31]. It measures the average preservation of K-ary neighborhoods around all data points. The formulation of this performance index requires ranks of sorted pairwise distances to be defined. The rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ in the high-dimensional space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}|$, where $|A|$ denotes the cardinality of set A. Similarly, the rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ in the low-dimensional space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. 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 the sets defined by $\nu_i^K = \{j : 1 \leq \rho_{ij} \leq K\}$ and $n_i^K = \{j : 1 \leq r_{ij} \leq K\}$, respectively. Eventually, the performance index can be written as $Q_{\text{NX}}(K) = \sum_{i=1}^{N} |\nu_i^K \cap n_i^K|/(KN)$. The index measures the average normalized agreement between corresponding K-ary neighborhoods in the high- and low-dimensional spaces. It varies between 0 and 1; for a random embedding, $Q_{\text{NX}}(K) \approx K/(N-1)$ [31].

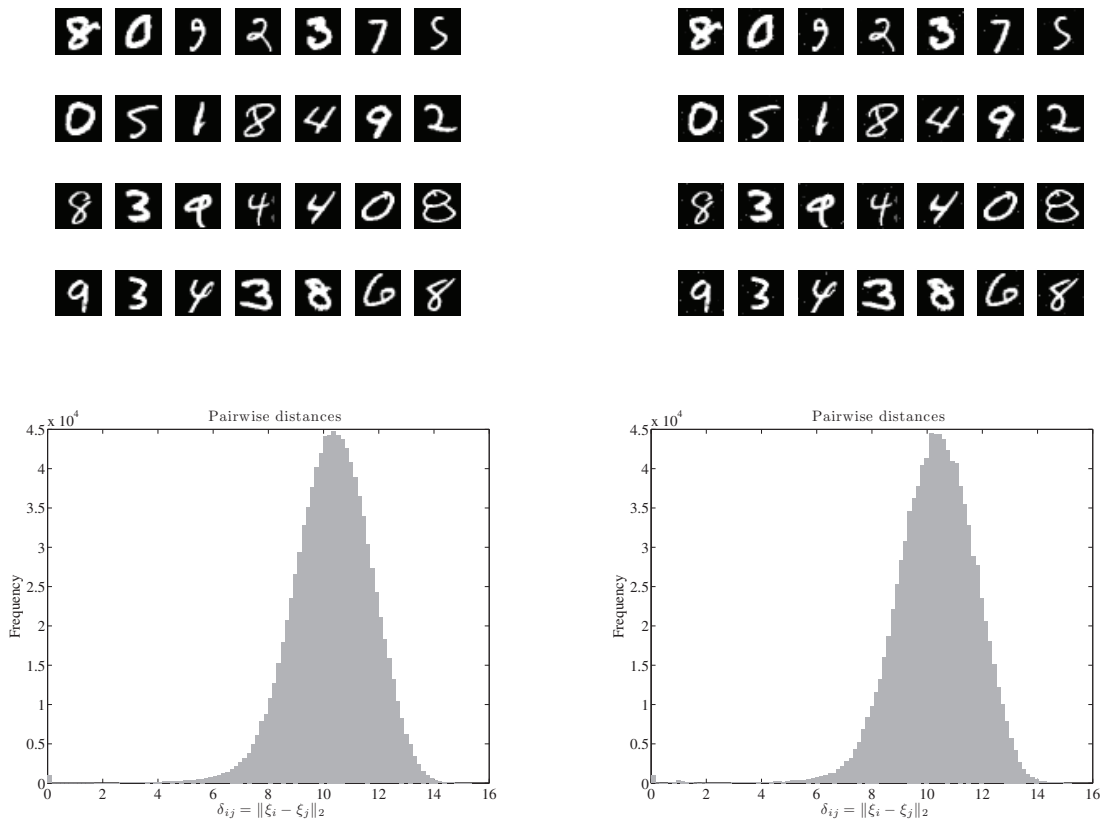

Figure 4: Typical images randomly drawn from the MNIST data base of handwritten digits. Noisefree images are on the left, whereas images with speckle noise are on the right. Histograms of all pairwise Euclidean distances are shown in the bottom row. Notice the appearance of a tiny tertiary mode in the histogram on the right, close to the peak of reflexive distances and caused by near-duplicates in the data set.

The quality assessment curves of the 2D visualization of the 30-dimensional sample is shown in Fig. 5. Quite obviously, linear DR yields the poorest results: the curve for MDS is the lowest one. Distance preservation with CCA performs hardly better. As expected, the behavior of t-SNE significantly depends on the way similarities are implemented. A similarity normalization that excludes the terms associated with reflexive distances, such as originally proposed in SNE and its variants performs well in this case. However, we see that our alternative definition with threshold τ_i can improve quality. In this example, τ_i is adaptively made equal to the distance to the Tth nearest neighbor. With $T = 0$, $\tau_i = 0$ and then no shift is possible without generating negative distances; this leads to a relatively bad performance. With $T = 1$, reflexive distances are discarded in SNE's pseudo-similarities, but performance are slightly improved, especially for small neighborhood sizes ($K \leq 10$). With $T = 2$, the similarity value is equal for the first and second neighbors and rank distinction between them is lost. Therefore, the quality index is much lower for $K = 1$ than for $K = 2$ (the quality index is insensitive to random permutations within the K-ary neighborhoods). As there are no spurious distances in this example, $T = 2$ leads to too large a value for τ_i. Nevertheless, the visualization quality remains good.

Figure 6 shows the quality assessment curves for the MNIST data. For the noisefree data (on the left), the conclusions remain mostly the same as in the previous academic example, namely, shift invariance is the key to good results and our similarity definition performs equally well or sometimes even better than SNE's pseudo-similarity. For the images with pairs of noisy duplicated images, the analysis is a bit more complicated, as t-SNE must cope with spurious distances. All quality assessment curves start very high: as each image is duplicated, its closest neighbor is

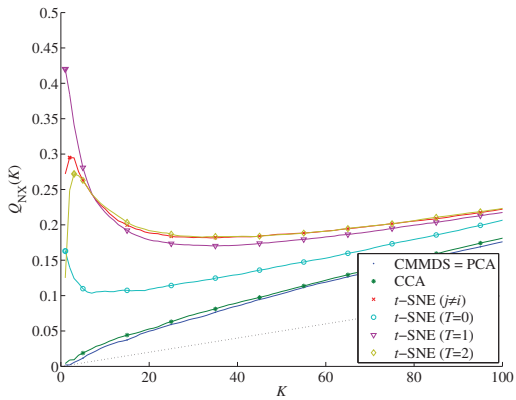

Figure 5: Quality assessment curves for the Gaussian data set. Each curve indicates the average normalized agreement between corresponding K-ary neighborhoods in the high- and low-dimensional spaces. The higher the curve, the better the performance (the thin dotted ascending line indicates the performance level of a random embedding).

pretty obvious, provided noise does not induce too large a distance between them. For K larger than 1, quality falls. All similarities that discard merely reflexive distances lead to rather poor performance. Only our definition with $T = 2$ is able to maintain a good quality as the neighborhood size grows.

6. Conclusions

Nonlinear dimensionality reduction driven by the principle of similarity preservation has recently yielded impressive experimental results in the literature. However, the reasons of this breakthrough have remained mostly unidentified. This paper has shown both theoretically and experimentally that a proper definition of the similarities brings useful properties, such as the invariance to shifts. This property allows the similarities to fight the curse of dimensionality and in particular to circumvent the phenomenon of distance concentration. Therefore, the comparison and matching of shift-invariant similarities computed in spaces of different dimensionalities make sense. On the contrary, naive distance preservation without shift is likely to fail.

This paper has also proposed a modified and more consistent similarity definition that generalizes previous ones. In particular, it introduces a parameter that controls the maximal shift amplitude and widens the applicability of similarity-based NLDR. For instance, it can cope with specific noise models and maintain a good level of performance, in contrast to the regular similarity definition.

The identification of shift-invariance as a key property in the success of similarity-based NLDR will help us to design better visualization methods, with more effective cost functions and improved robustness.

References

[1] I. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, NY, 1986.
[2] G. Young, A. Householder, Discussion of a set of points in terms of their mutual distances, Psychometrika 3 (1938) 19–22.
[3] W. Torgerson, Multidimensional scaling, I: Theory and method, Psychometrika 17 (1952) 401–419.
[4] J. Lee, M. Verleysen, Nonlinear dimensionality reduction, Springer, 2007.
[5] R. Shepard, The analysis of proximities: Multidimensional scaling with an unknown distance function (parts 1 and 2), Psychometrika 27 (1962) 125–140, 219–249.
[6] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–28.
[7] 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.
[8] J. Sammon, A nonlinear mapping algorithm for data structure analysis, IEEE Transactions on Computers CC-18 (5) (1969) 401–409.
[9] P. Demartines, J. Hérault, Vector quantization and projection neural network, Vol. 686 of Lecture Notes in Computer Science, Springer-Verlag, New York, 1993, pp. 328–333.

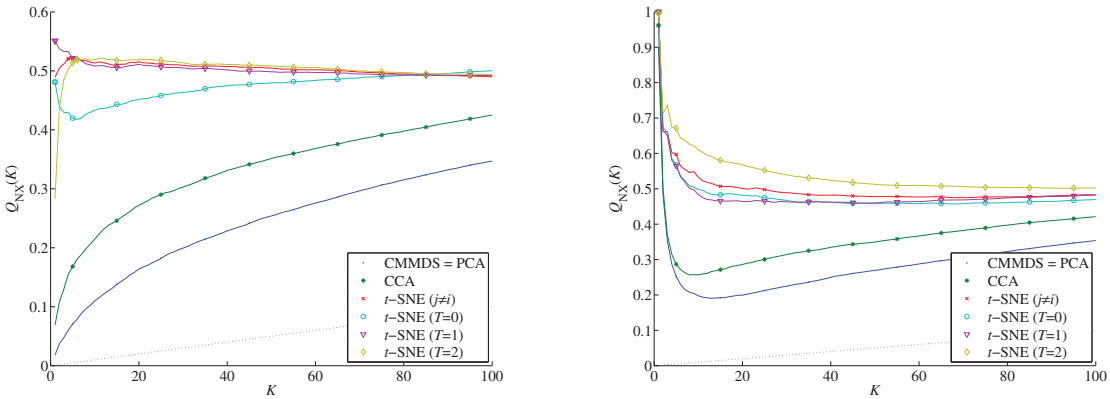

Figure 6: Quality assessment curves for MNIST data. Left: noisefree images. Right: images with speckle noise.

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

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

[12] T. Kohonen, Self-organization of topologically correct feature maps, Biological Cybernetics 43 (1982) 59–69.

[13] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Computation 10 (1998) 1299–1319, also available as technical report 44 at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany, December 1996.

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

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

[16] K. Weinberger, L. Saul, Unsupervised learning of image manifolds by semidefinite programming, International Journal of Computer Vision 70 (1) (2006) 77–90.

[17] 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.

[18] 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), 2003.

[19] L. Xiao, J. Sun, S. Boyd, A duality view of spectral methods for dimensionality reduction, in: Proceedings of the 23rd International Conference on Machine Learning, Pittsburg, PA, 2006, pp. 1041–1048.

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

[21] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, Journal of Machine Learning Research 9 (2008) 2579–2605.

[22] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, Journal of Machine Learning Research 11 (2010) 451–490.

[23] D. Erhan, P.-A. Manzagol, Y. Bengio, S. Bengio, V. P., The difficulty of training deep architectures and the effect of unsupervised pre-training, in: Journal of Machine Learning Research, Workshop and Conference Proceedings, Vol. 5, 2009, pp. 153–160.

[24] Z. Yang, I. King, Z. Xu, E. Oja, Heavy-tailed symmetric stochastic neighbor embedding, in: Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta (Eds.), Advances in Neural Information Processing Systems 22, 2009, pp. 2169–2177.

[25] D. François, V. Wertz, M. Verleysen, The concentration of fractional distances, IEEE Transactions on Knowledge and Data Engineering 19 (7) (2007) 873–886.

[26] J. Lee, M. Verleysen, On the role and impact of the metaparameters in t-distributed stochastic neighbor embedding, in: Y. Lechevallier, G. Saporta (Eds.), Proc. 19th COMPSTAT, Paris (France), 2010, pp. 337–348.

[27] R. Bellman, Adaptative Control Processes: A Guided Tour, Princeton University Press, Princeton, NJ, 1961.

[28] D. Donoho, High-Dimensional Data Analysis: The Curse and Blessings of Dimensionality, aide-mémoire for a lecture for the American Math. Society "Math. Challenges of the 21st Century" (2000).

[29] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[30] J. Lee, M. Verleysen, Simbed: similarity-based embedding, in: Proc. ICANN 2009, Limassol, Cyprus, 2009.

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