On the Role and Impact of the Metaparameters in t-distributed Stochastic Neighbor Embedding

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

[1] Imagerie Moléculaire et Radiothérapie Expérimentale
Avenue Hippocrate 54, B-1200 Brussels, Belgium *john.lee@uclouvain.be*
[2] Machine Learning Group - DICE, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium *michel.verleysen@uclouvain.be*

Abstract. Similarity-based embedding is a paradigm that recently gained interest in the field of nonlinear dimensionality reduction. It provides an elegant framework that naturally emphasizes the preservation of the local structure of the data set. An emblematic method in this trend is t-distributed stochastic neighbor embedding (t-SNE), which is acknowledged to be an efficient method in the recent literature. This paper aims at analyzing the reasons of this success, together with the impact of the two metaparameters embedded in the method. Moreover, the paper shows that t-SNE can be interpreted as a distance-preserving method with a specific distance transformation, making the link with existing methods. Experiments on artificial data support the theoretical discussion.

Keywords: similarity-based embedding, dimensionality reduction, nonlinear projection, manifold learning, t-SNE

1 Introduction

Dimensionality reduction is the task of finding faithful, low-dimensional representations of high-dimensional data. Although the case of clustered data can be considered too, dimensionality reduction usually relies on the assumption that the data are sampled from a smooth manifold. Methods such as principal component analysis (PCA) or classical metric multidimensional scaling (MDS) (Young and Householder (1938)) can be successfully applied when the manifold is a linear subspace. However, when the manifold is curved or folded (Tenenbaum et al. (2000)), one should use adapted nonlinear dimensionality reduction (Lee and Verleysen (2007)) (NLDR). Nonmetric MDS (Shepard (1962), Kruskal (1964)) and Sammon's nonlinear mapping (SNLM) (Sammon (1969)) are early methods generalizing MDS, based on the principle of distance preservation. Spectral embedding (Saul et al. (2006)) has emerged since the seminal paper describing kernel PCA (Scholkopf et al. (1998)). Isomap (Tenenbaum et al. (2000)), locally linear embedding (Roweis and Saul (2000)), Laplacian eigenmaps (Belkin and Niyogi (2002)), and maximum variance unfolding (MVU) (Weinberger and Saul (2006)) are among

Y. Lechevallier, G. Saporta (eds.), *Proceedings of COMPSTAT'2010*,
DOI 10.1007/978-3-7908-2604-3_31,

the most representative methods in this category. 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 the capability of handling a broader range of cost functions. Successful nonspectral methods are among others curvilinear component analysis (CCA) (Demartines and Herault (1997)), stochastic neighbor embedding (SNE) (Hinton and Roweis (2003)), and its variant t-SNE (Van der Maaten and Hinton (2008)). CCA has long been considered to be a distance-preserving method that can be related to SNLM. In contrast, SNE and t-SNE seek to match similarities, which are basically decreasing functions of the pairwise distances. This reformulation provides a more natural way to formalise the importance of preserving the local structure of data. t-SNE is nowadays considered as an efficient method for visualizing high-dimensional data (see for example Erhan et al. (2009), Parviainen and Vehtari (2009)).

This paper aims at analysing the behavior of t-SNE and the key influence of its two metaparameters, namely the so-called perplexity and the number of degrees of freedom. It addition, it shows that t-SNE can be cast within the framework of distance preservation, by means of a distance tranformation; we identify this transformation, and compare it to other methods.

The remainder of this paper is organized as follows. Section 2 briefly reviews SNE and t-SNE. Section 3 weaves the connection between distance preservation and similarity matching. Sections 4 and 5 provide and discuss the experimental results. Finally, Section 6 draws the conclusions.

2 Stochastic Neighbor Embedding

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1 \leq i \leq N}$ denote a data set of N vectors picked in an M dimensional space. Symbol δ_{ij} denotes the pairwise distance between data vectors $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$. The similarity between $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ is defined in SNE as:

$$p_{j|i}(\lambda_i) \doteq \begin{cases} 0 & \text{if } i = j \\ g(\delta_{ij}/\lambda_i) \Big/ \sum_{k \neq i} g(\delta_{ik}/\lambda_i) & \text{otherwise} \end{cases},$$

where $g(u) = \exp(-u^2/2)$. In Hinton and Roweis (2003), $p_{j|i}$ is referred to as a conditional probability and represents the empirical probability of $\boldsymbol{\xi}_j$ to be a neighbor of $\boldsymbol{\xi}_i$. The softmax denominator indeed guarantees that $\sum_{j=1}^{N} p_{i|j} = 1$. Probabilities $p_{j|i}(\lambda_i)$ and $p_{i|j}(\lambda_j)$ are not equal since they involve kernels with individual widths. A user-defined metaparameter, the perplexity PPXT, induces all widths λ_i through the equation $2^{H(p_{j|i})} = \text{PPXT}$, where $H(p_{j|i}) \doteq \sum_{j=1}^{N} p_{j|i} \log_2 p_{j|i}$ is the entropy of $p_{j|i}$. Intuitively, the perplexity allows the Gaussian kernels to adapt their width to the local density of data points.

Within this framework, a symmetric similarity function between $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ can be defined by $p_{ij}(\boldsymbol{\lambda}) \doteq \frac{1}{2N}\left(p_{j|i}(\lambda_i) + p_{i|j}(\lambda_j)\right)$, where $\boldsymbol{\lambda} = [\lambda_i]_{1 \leq i \leq N}$

and p_{ij} is referred to as a joint probability. As for the conditional probability, we have $p_{ii}(\boldsymbol{\lambda}) = 0$ and $\sum_{i,j=1}^{N} p_{ij}(\boldsymbol{\lambda}) = 1$.

In the low-dimensional space, let $\mathbf{X} = [\mathbf{x}_i]_{1 \le i \le N}$ denote the embedding to be found by SNE. If d_{ij} is the Euclidean distance $\|\mathbf{x}_i - \mathbf{x}_j\|_2$, then pairwise similarities in the low-dimensional space can be written as

$$q_{ij}(n) \doteq \begin{cases} 0 & \text{if } i = j \\ t(d_{ij}, n) \Big/ \sum_{k \neq l} t(d_{kl}, n) & \text{otherwise} \end{cases} , \tag{1}$$

where $t(u, n) = (1 + u^2/n)^{-(n+1)/2}$. Function $t(u, n)$ is proportional to the probability density of a Student's t-distributed variable with n degrees of freedom; n controls the tail thickness of the similarity kernel. It is noteworthy that $\lim_{n \to \infty} t(u, n) = g(u)$. Hence, SNE (Hinton and Roweis (2003) corresponds to the case $n \to \infty$ whereas t-SNE (Van der Maaten and Hinton (2008), Van der Maaten (2009)) turns out to be the case $n = 1$. Like $p_{ij}(\boldsymbol{\lambda})$, $q_{ij}(n)$ is referred to as a joint probability, although it is differently defined. Two important differences are the kernel shape and the absence of scaling parameter. Another difference is that $q_{ij}(n)$ can be interpreted as a (non-conditional) probability, thanks to the single softmax denominator.

The t-SNE method compares $p_{ij}(\boldsymbol{\lambda})$ and $q_{ij}(n)$ by means of a (discrete) Kullback-Leibler divergence:

$$E(\boldsymbol{\Xi}, \mathbf{X}, \boldsymbol{\lambda}, n) = D_{\mathrm{KL}}(p \| q) \doteq \sum_{i,j=1}^{N} p_{ij}(\boldsymbol{\lambda}) \log \frac{p_{ij}(\boldsymbol{\lambda})}{q_{ij}(n)} \ .$$

The minimization of E can be achieved by gradient descent. In t-SNE, the gradient of $E(\boldsymbol{\Xi}, \mathbf{X}, \boldsymbol{\lambda}, n)$ with respect to $\mathbf{x}_i$ can be written as (Van der Maaten and Hinton (2008), Van der Maaten (2009))

$$\frac{\partial E}{\partial \mathbf{x}_i} = \frac{2n+2}{n} \sum_{j=1}^{N} \frac{p_{ij}(\boldsymbol{\lambda}) - q_{ij}(n)}{1 + d_{ij}^2/n} (\mathbf{x}_i - \mathbf{x}_j) \ . \tag{2}$$

It is easy to verify that

$$\lim_{n \to \infty} \frac{\partial E}{\partial \mathbf{x}_i} = 2 \sum_{j=1}^{N} (p_{ij}(\boldsymbol{\lambda}) - q_{ij}(n))(\mathbf{x}_i - \mathbf{x}_j) \ ,$$

which corresponds to the gradient of SNE. In the context of a gradient descent, three factors can be identified in each term of (2). Factor $(\mathbf{x}_i - \mathbf{x}_j)$ is a vector that allows $\mathbf{x}_i$ to move towards $\mathbf{x}_j$. Factor $(p_{ij}(\boldsymbol{\lambda}) - q_{ij}(n))$ varies between -1 and $+1$; it is proportional to the similarity error and adjusts the length and direction (inwards/outwards) of the movement. Finally, factor $(1 + d_{ij}^2/n)^{-1}$ varies between 0 and 1 and damps the movement, especially if $\mathbf{x}_i$ lies far away from $\mathbf{x}_j$. A similar factor can be found in the gradient of CCA

(Demartines and Herault (1997)). It provides the capability of 'tearing' the manifold to be embedded. Additional details about t-SNE and the gradient descent of E can be found in van der Maaten and Hinton (2008).

The discrepancy between the kernels in the high- and low-dimensional spaces is intended to address the curse of dimensionality. Let us take the exemple of a curved P-dimensional manifold embedded in an M-dimensional space (with $M > P$). It is easy to see that the Euclidean distance between any two manifold points shrinks as the curvature increases, especially if they lie far away from each other. Hence, a strict isometry will not succeed in embedding the manifold in a low-dimensional space: a poor unfording with several regions superimposed would result (Hinton and Roweis (2003)). Based on this observation, the similarity kernels in t-SNE have heavier tails in the embedding space, to force large distances to grow in order to attain the same similarity value as in the data space.

3 Connection between similarity and distance preservation

Let us assume that t-SNE finds an embedding that cancels the gradient of its objective function. A sufficient condition to obtain such a solution is that one of the terms in (2) vanishes. A trivial (useless) solution is when $d_{ij} = 0$ for all i and j. Another trivial solution is $d_{ij} \to \infty$ for all i and j, because the damping factor tends to zero. The useful solution consists in satisfying $p_{ij}(\boldsymbol{\lambda}) = q_{ij}(n)$. In this last case, let us approximate the above definition of $p_{ij}(\boldsymbol{\lambda})$ with $p_{ij} \approx p_{j|i}/N$. Using this approximation in conjuntion with the definition of $q_{ij}(n)$ in (1) allows us to write

$$d_{ij} \approx f(\delta_{ij}) \doteq \sqrt{nR_i^{\frac{2}{n+1}} \exp\left(\frac{\delta_{ij}^2}{(n+1)\lambda_i^2}\right) - n} \ , \tag{3}$$

where $R_i = N \sum_{k \neq i} g(\delta_{ik}/\lambda_i) / \sum_{k \neq l} t(d_{kl}, n)$ is the ratio of the softmax denominators. If we get rid of the difficulty raised by the softmax denominators, namely if we assume that $R_i \approx 1$, then we can see SNE and t-SNE as NLDR methods that preserve transformed distances. The transformation has an exponentially increasing shape; its key properties are

- $f(0) = 0$ and f is monotonically increasing on $\mathbb{R}_+$,
- $\lim_{n\to\infty} f(\delta_{ij}) = \delta_{ij}/\lambda_i$,
- if $\delta_{ij} \ll \lambda_i$, then $d_{ij} = f(\delta_{ij}) \approx \delta_{ij}/(\lambda_i\sqrt{n+1})$.

Intuitively, t-SNE tries to preserve stretched distances; the stretch is exponential. In the case of SNE, the transformation degenerates and distances are merely locally scaled by λ_i. The second property shows that a similar scaling occurs when $\delta_{ij} \ll \lambda_i$. What is important to note is that λ_i and n act more or

less in the same way in transformation (3), namely they modulate the exponential growth. However, the differences are that (i) n is global whereas the λ_i can fluctuate locally, (ii) changing the perplexity (which approximately amounts to multiplying all λ_i with approximately the same factor) impacts the scale of the embedding, while changing n does not have this effect.

The relationship between n and the intrinsic dimensionality of the manifold that is put forward in Van der Maaten (2009) is questioned by the above analysis. First, because n and the perplexity cannot be studied separately and, second, because n depends neither on the embedding dimensionality nor on the data dimensionality. In addition, the optimal distance transformation depends on the manifold shape: in the above example, changing the curvature of the manifold should have an impact on the required stretch, hence on n. This motivates the experiments described in the next section.

4 Experiments

The experiments rely on the widely used Swiss roll (Tenenbaum et al. (2000)) benchmark manifold. A dataset of 750 noisefree vectors is sampled from $\boldsymbol{\xi} = [\sqrt{u}\cos(3\pi\sqrt{u}),\ \sqrt{u}\sin(3\pi\sqrt{u}),\ \pi v]^T$, where u and v have uniform distributions in $[0, 1]$. Two reasons justify using the Swiss roll. First, it is a widespread benchmark that has however not been used in (Van der Maaten and Hinton (2008), Van der Maaten (2009)). Second, it is a Euclidean manifold, which implies that a linear projection (such as in metric MDS) suffices to obtain a perfect embedding of the Swiss roll, provided geodesic distances are used (Tenenbaum et al. (2000), Lee and Verleysen (2007)).

The experiments compare t-SNE, CCA, SNLM, and classical metric MDS, whose result serves as baseline. Each method is used with both Euclidean and geodesic distances. The latter are approximated with graph distances, that is, with shortest paths in a Euclidean graph that stems from 6-ary neighborhoods around each data point. MDS is equivalent to PCA with Euclidean distances, and to Isomap (Tenenbaum et al. (2000)) with geodesic distances. CCA with geodesic distances is known as Curvilinear Distances Analysis (Lee and Verleysen (2004)). The implementation of t-SNE is provided by the authors of Van der Maaten and Hinton (2008); the only extension concerns the possibility to vary the number of degrees of freedom.

Performance assessment is achieved by means of the criteria proposed in (Lee and Verleysen (2009). These criteria look at K-ary neighborhoods around each vector in the data space as well as in the embedding space. Criterion $Q_{\mathrm{NX}}(K)$ reflects the overall quality of the embedding; its value corresponds to the average percentage of identical neighbors in both spaces. Criterion $B_{\mathrm{NX}}(K)$ measures to what extend a NLDR method can be 'intrusive' or 'extrusive'. A positive $B_{\mathrm{NX}}(K)$ corresponds to an intrusive embedding, wherein many distant points are embedded close to each other; a negative value corresponds to many extrusions, i.e. to close neighbors em-

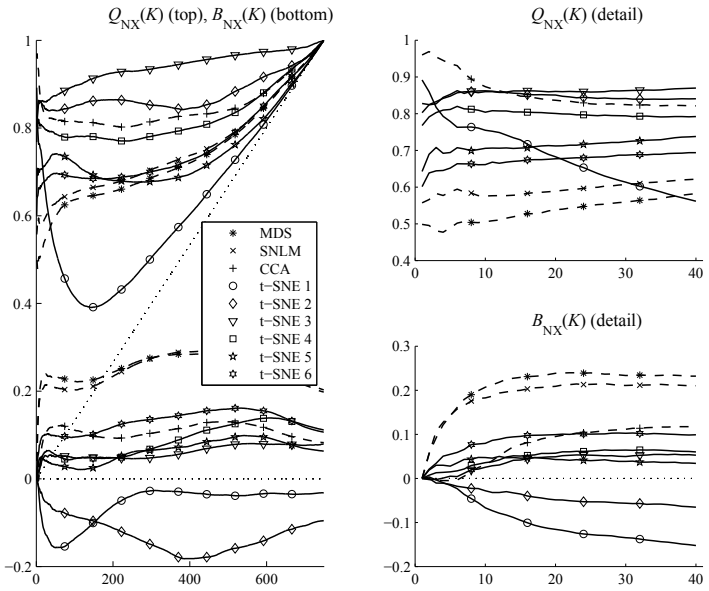

Fig. 1. Quality assessment of the embeddings using Euclidean distances. The numbered curves for t-SNE refer to perplexity values equal to 4, 25, 64, 121, 196, and 289 respectively. See text for details.

bedded far away from each other. Both $Q_{\mathrm{NX}}(K)$ and $B_{\mathrm{NX}}(K)$ are shown in specific diagrams that consist of three panels. The first one spans the interval $1 \leq K \leq N-1$, whereas the small ones on the right focus on small values of K, for each criterion separately. The quality and behavior of the various methods are depicted in Figs. 1 and 2. Figure 3 shows the evolution of $Q_{\mathrm{NX}}(K)$ with respect to the perplexity, for $K = \{5, 15, 50, 150\}$.

5 Discussion

Unfolding the Swiss roll with Euclidean distances is a difficult task, as shown by the low values of $Q_{\mathrm{NX}}(K)$ produced by MDS, SNLM, and CCA in Fig. 1. The result of t-SNE largely depends on the value of the perplexity; it ranges from poorer than MDS to excellent. Looking at the curves for $B_{\mathrm{NX}}(K)$ shows that t-SNE tends to be extrusive or intrusive, depending on the perplexity, whereas all other methods are rather intrusive. An illustration of the embeddings provided by t-SNE with different perplexity values is given in Fig. 4.

Replacing Euclidean distances with geodesic ones obviously facilitates the task. All methods achieve good results in Fig. 2, the best being CCA. The variability in t-SNE's results remains, but one sees that increasing the perplexity leads to better performances (see also Fig. 3). As expected, geodesic

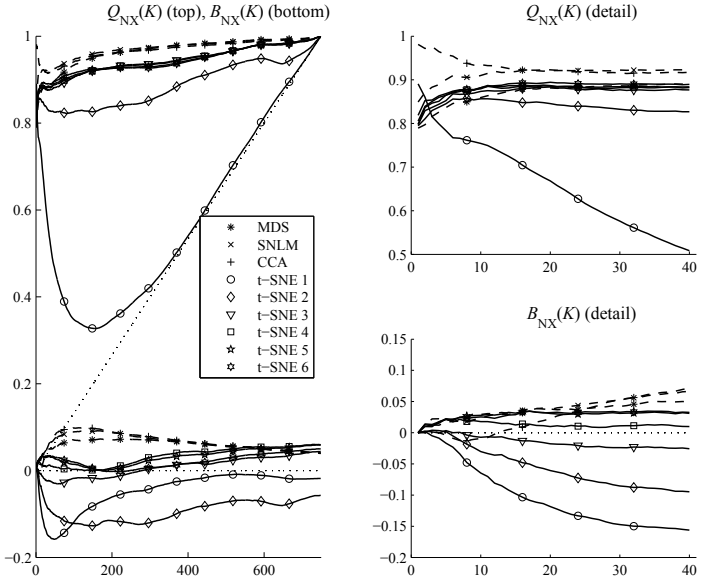

Fig. 2. Quality assessment of the embeddings using geodesic distances. The numbered curves for t-SNE refer to perplexity values equal to 4, 25, 64, 121, 196, and 289 respectively. See text for details.

distances do not need to be stretched, what can be achieved with a high (infinite) perplexity. For Euclidean distances, we observe a peak in $Q_{\mathrm{NX}}(K)$: there exist an optimal value of the perplexity, such that the transformed distances approximate as closely as possible the geodesic distances.

At this point, we can state that the distance transformation that is implicitly achieved by t-SNE is not always optimal and that its parameters must be carefully tuned. More specifically, the perplexity controls the way distances are stretched. If stretching distances is fundamentally a pertinent idea when one wishes to unfold a manifold, t-SNE cannot always approximate the optimal transformation, which can be much more complex than in the Swiss roll. A positive point for t-SNE is that its gradient includes a damping factor that diminishes the importance of large distances, whose transformed value could be inappropriate. Nevertheless, this does not address the issue raised by non-Euclidean manifolds, such as an half (hollow) sphere. Near the pole, small distances should shrink or remain unchanged, whereas a stronger and stronger stretch is required when moving away from the pole. The situation gets obviously much more favorable with clustered data, as stretching large distances improves the separation between the clusters; examples can be found in Van der Maaten and Hinton (2008) and in Van der Maaten (2009). There is a risk however that too small a value of the perplexity could lead to an embedding with spurious clusters.

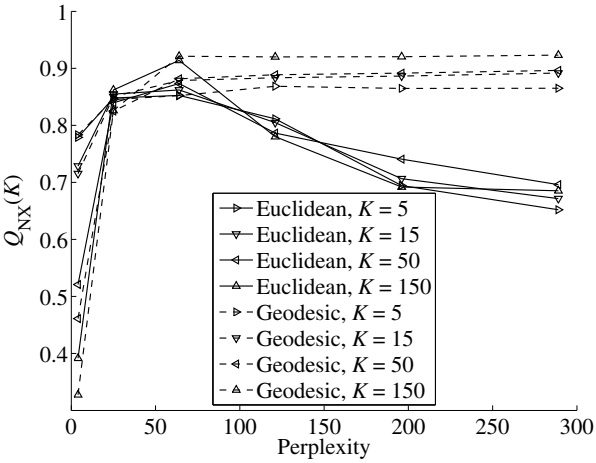

Fig. 3. Some values of $Q_{\mathrm{NX}}(K)$ taken from Figs. 1 and 2, shown with respect to the perplexity.

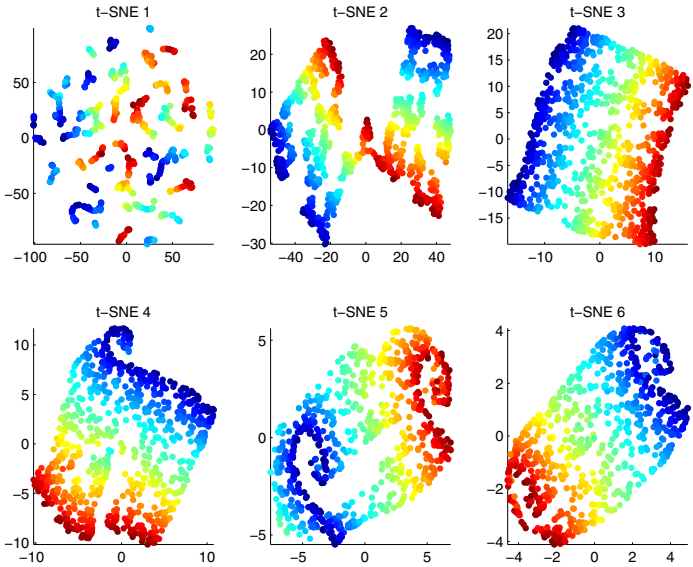

Fig. 4. Embeddings provided by t-SNE with Euclidean distances, for perplexity values equal to 4, 25, 64, 121, 196, and 289 (from top left to bottom right).

Finally, the paradigm of distance preservation can be used to compare t-SNE to other NLDR methods. The comparison is straightforward for many spectral methods that explicitly use distances, like Isomap and MVU. For other methods, such as those involving the bottom eigenvectors of some Gram

matrix, duality (Xiao et al. (2006)) can be invoked in order to first build a virtual matrix of pairwise distances, on which classical MDS (isometric embedding) is applied. For example, Laplacian eigenmaps and related methods can be shown to involve commute time distances or random walks in a graph (Saerens et al. (2004)). Focusing on the transformation we see for instance that Isomap, MVU, and t-SNE all stretch distances. In this respect however, MVU proves to be more powerful than Isomap, which in turn is superior to t-SNE. The transformations in Isomap and MVU are indeed data-driven: any distance value depends on the shape of the underlying manifold. Moreover, the semidefinite programming step in MVU adjusts long distances in order to minimize the embedding dimensionality. In contrast, the transformation achieved by t-SNE marginally depends on the data density (when the individualized widths in $\boldsymbol{\lambda}$ are computed from the perplexity), not on the manifold shape. Apart from this and a minor impact of the softmax normalizations, t-SNE achieves an 'a priori' distance transformation, which is not data driven.

6 Conclusions

Many methods of nonlinear dimensionality reduction rely on distance preservation. Recent works reveal however a growing interest in similarity matching; an emblematic method that follows this trend is undoubtedly t-SNE. This contribution aims at analyzing t-SNE's properties. Casting t-SNE within the framework of distance preservation allows a better understanding of its behaviour. Specifically, it has been shown that t-SNE can be considered to preserve transformed distance and the transformation has been identified to be an exponential stretch. The slope of the transformation turns out to be controlled by the main two metaparameters of t-SNE, namely the perplexity and the number of degrees of freedom in the Student similarity functions. Such an exponential stretch increases the separation between clusters and this explains why t-SNE performs so well with clustered data. On the other hand, the transformation shape can be suboptimal for manifold data. Experiments are performed on the Swiss roll, a manifold for which an optimal distance transformation is known and consists in replacing Euclidean distances with geodesic distances. In this case, t-SNE requires a careful parameter adjustment and cannot outperform basic methods that preserve geodesic distances.

References

BELKIN, M. and NIYOGI, P. (2002): Laplacian eigenmaps and spectral techniques for embedding and clustering. In: T.G. Dietterich, S. Becker, Z. Ghahramani (Eds.): *NIPS 2001 proc., 14*. MIT Press, 585-591.

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

ERHANY D., MANZAGOL P.-A., BENGIO Y., BENGIO S. and VINCENT P. (2009): The Difficulty of Training Deep Architectures and the Effect of Unsupervised Pre-Training *Journal of Machine Learning Research Proc., 5, 153-160.*

HINTON, G. and ROWEIS, S.T. (2003): Stochastic Neighbor Embedding. In: S. Becker, S. Thrun and K. Obermayer (Eds.): *Advances in NeuralInformation Processing Systems (NIPS 2002), 15.* MIT Press, 833-840.

KRUSKAL, J.B. (1964): Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika, 29, 1-28.*

LEE, J.A. and VERLEYSEN, M. (2004): Curvilinear Distance Analysis versus Isomap. *Neurocomputing, 57, 49-76.*

LEE, J.A. and VERLEYSEN, M. (2007): *Nonlinear dimensionality reduction.* Springer, New York.

LEE, J.A. and VERLEYSEN, M. (2009): Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing, 72 (7-9), 1431-1443.*

PARVIAINEN E. and VEHTARI A. (2009): Features and metric from a classifier improve visualizations with dimension reduction In: C. Alippi, M. Polycarpou, C. Panayiotou, G. Ellinas (Eds.): *ICANN 2009 proc.* Springer, LNCS 5769, 225-234.

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

SAERENS, M., FOUSS, F., YEN, L. and DUPONT, P. (2004): The principal components analysis of a graph, and its relationships to spectral clustering. In: J.-F. Boulicaut, F. Esposito, F. Giannotti, D. Pedreschi (Eds.): *ECML 2004 proc.*. Springer, LNCS 3201, 371-383.

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

SAUL, L.K., WEINBERGER, K.Q., HAM, J.H., SHA, F. and LEE, D.D. (2006): Spectral methods for dimensionality reduction. In: O. Chapelle, B. Schoelkopf, B. and A. Zien, A. (Eds.): *Semisupervised Learning.* MIT Press, 293-308.

SCHOLKOPF, B., SMOLA, A. and MULLER, K.-R. (1998): Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation, 10 ,1299–1319.*

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

TENENBAUM, J.B., DE SILVA, V. and LANGFORD, J.C. (2000): A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science, 290 (5500), 2319-2323.*

VAN DER MAATEN, L. and HINTON, G. (2008): Visualizing Data using t-SNE. *Journal of Machine Learning Research, 9, 2579-2605.*

VAN DER MAATEN, L.J.P. (2009): Learning a Parametric Embedding by Preserving Local Structure. In: D. van Dyk and M. Welling (Eds.): *Proc. 12th Int. Conf. on Artificial Intel. and Statistics, Clearwater Beach, FL.* JMLR Proceedings 5, 384-391.

XIAO, L. , SUN, J. and BOYD, S. (2006): A Duality View of Spectral Methods for Dimensionality Reduction. In: W. Cohen and A. Moore (Eds.): *ICML proc., Pittsburg (PA).* Omni Press, 1041-1048.

WEINBERGER K.Q. and SAUL, L.K. (2006): Unsupervised Learning of Image Manifolds by Semidefinite Programming. *International Journal of Computer Vision, 70 (1), 77-90.*