ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

Type 1 and 2 mixtures of divergences for stochastic neighbor embedding

John A. Lee [*]

Université catholique de Louvain,
Pole of Molecular Imaging, Radiotherapy, and Oncology
Avenue Hippocrate 55, B-1200 Bruxelles, Belgium

Abstract. Stochastic neighbor embedding (SNE) is a method of dimensionality reduction (DR) that involves softmax similarities measured between all pairs of data points. In order to build a low-dimensional embedding, SNE tries to reproduce the similarities observed in the high-dimensional data space. The capability of softmax similarities to fight the phenomenon of norm concentration has been studied in previous work. This paper investigates a complementary aspect, namely, the cost function that quantifies the mismatch between the high- and low-dimensional similarities. We show experimentally that switching from a simple Kullback-Leibler divergences to mixtures of dual divergences increases the quality of DR. This modification brings SNE to the performance level of its Student t-distributed variant, without the need to resort to non-identical similarity definitions in the high- and low-dimensional spaces. These results allow us to conclude that future improvements in similarity-based DR will likely emerge from better definitions of the cost function.

1 Introduction

Dimensionality reduction (DR) aims at producing faithful and meaningful representations of high-dimensional data into a lower-dimensional space. The general intuition that drives DR is that close or similar data items should be represented near each other, whereas dissimilar ones should be represented far from each other. Through the history of DR, authors have formalized this idea of neighborhood preservation in various ways, using several models for the mapping or embedding of data from the high-dimensional (HD) space to the low-dimensional (LD) one. For instance, principal component analysis (PCA) [1] and classical metric multidimensional (MDS) [2] scaling rely on linear projections that maximize variance preservation and dot product preservation, respectively. Nonlinear variants of metric MDS [3] are based on (weighted) distance preservation. These distances can Euclidean or approximation of geodesic lengths [4]. The use of similarities in DR is quite recent and emerged with methods based on spectral optimization. Methods like Laplacian eigenmaps [5] and locally linear embedding [6] involve sparse matrices of similarities, also called affinity matrices. In spite of a sound theoretical framework, these methods somehow failed to outperform older techniques [7, 8, 9]. A possible explanation is that these methods can

[*] J.A. Lee is a Research Associate with the Belgian National Fund of Scientific Research (FNRS).

be reformulated into classical metric MDS achieved in a feature space. In this case, the definition of the similarities merely determines the implicit nonlinear mapping from the HD data space to the LD feature space [10, 11].

Genuine similarity preservation appeared later with stochastic neighbor embedding [12] (SNE). Interest in this new paradigm grew after the publication of variants such as Student t-distributed SNE [9] (t-SNE) and NeRV [13], standing for neighborhood retrieval and visualization. These variants led to breakthroughs in terms of DR quality, with outstanding results. Nevertheless, the reasons of this performance leap remain partly unknown. The role played by SNE's specific similarities has been investigated in [14], which revealed their capability to fight the phenomenon of norm concentration in HD spaces.

This paper focuses on a complementary aspect of SNE, namely, the definition of its cost function. In the original version and in SNE, the cost function is a sum of Kullback-Leibler (KL) divergences that measure, for each point, the mismatch between the HD and LD similarities with respect to its neighbors. NeRV replaces the asymmetric KL divergence in each term of the sum with a weighted mixture of two 'dual' KL divergences, which turns out to be a type 1 mixture of KL divergences [15]. Here, we use the type 2 mixture KL divergences [15] and we show experimentally that it outperforms both the type 1 mixture and the usual non-blended divergence.

The rest of this paper is organized as follows. Section 2 defines the similarities used in SNE and its variants. Section 3 deals with the considered divergences and cost functions for SNE. Section 4 presents and discusses the experimental results. Finally, Section 5 draws the conclusions.

2 Shift-invariant softmax similarities

Let $\boldsymbol{\Xi} = [\boldsymbol{\xi}_i]_{1\leq i\leq N}$ denote a set of N points in some M-dimensional space. Similarly, let $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$ be its representation in a P-dimensional space, with $P \leq M$. The Euclidean 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 HD and LD 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. An important feature of similarities defined as softmax exponential ratios such as above is the scale invariance of the ratios, which translates into shift-invariance with respect to the squared distances δ_{ij}^2 and d_{ij}^2 [14]. Because null distances are excluded for the sum in the denominators, the shift applicable to δ_{ij}^2 can range from $-\min_{j,j\neq i}\delta_{ij}^2$ to ∞. The lower end of this interval ensures that the shifted distances remain positive. A negative shift close to this lower bound is particularly interesting to fight the phenomenon of norm concentration. One manifestation of this phenomenon is the following: for a finite sample of points $\boldsymbol{\Xi}$, $\min_{j,j\neq i}\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|$ grows faster with M than $\max_j\|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|$. The changing

ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2012, i6doc.com publ., ISBN 978-2-87419-049-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100967420.

shape of distance distributions, depending on the dimensionality, partly explains to failure of DR methods based on distance preservation. The distances in LD spaces are always 'too short' to match those observed in HD spaces. Invariance to shifts in similarities circumvents this problem.

3 Divergences to measure similarity mismatch

Due to normalization, softmax similarities add up to one, i.e. $\sum_j \sigma_{ij} = \sum_j s_{ij} = 1$. Therefore, $\boldsymbol{\sigma}_i = [\sigma_{ij}]_{1 \leq j \leq N}$ and $\mathbf{s}_i = [s_{ij}]_{1 \leq j \leq N}$ can be seen as discrete probability distributions and divergences can be used to assess their mismatch. In SNE, the Kullback-Leibler divergence is used. It is defined as $D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = \sum_j \sigma_{ij} \log(\sigma_{ij}/s_{ij})$. The cost function of SNE [12] can then be written as $E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}) = \sum_i D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$. It can be minimized with respect to $\mathbf{X}$ by means of a gradient descent. This requires the bandwidths in $\boldsymbol{\Lambda} = [\lambda_i]_{1 \leq i \leq N}$ to be fixed. For this purpose, let us notice that each $D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is the varying cross-entropy of $\boldsymbol{\sigma}_i$ and $\mathbf{s}_i$ minus the constant entropy of $\boldsymbol{\sigma}_i$. In SNE, the bandwidths λ_i are adjusted in order to equalize all entropies, namely, $\sum_j \sigma_{ij} \log(\sigma_{ij}) = H$ for all i. The user indirectly specifies the targeted entropy value H by giving the perplexity, which is proportional to $\exp(H)$. 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.

In NeRV [13], the cost function mixes dual KL divergences:

$$D^{\beta}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = (1-\beta) D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) + \beta D_{\mathrm{KL}}(\mathbf{s}_i \| \boldsymbol{\sigma}_i) \ . \tag{2}$$

Parameter β balances both terms. A similar mixture of reciprocated functions was previously studied in the context of distance preservation [16]. The cost function is then $E(\mathbf{X}; \boldsymbol{\Xi}, \boldsymbol{\Lambda}, \beta) = \sum_i D^{\beta}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$. The bandwidth parameters in the similarities are adjusted in the same way as in SNE (the constant term w.r.t. $\mathbf{s}_i$ in the divergence remains the same, multiplied by $1-\beta$). The gradient is however more complicated than in SNE. For $\beta = 1/2$, $D^{1/2}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is symmetric: $D^{1/2}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = D^{1/2}_{\mathrm{KLs1}}(\mathbf{s}_i \| \boldsymbol{\sigma}_i)$. According to [15], $D^{1/2}_{\mathrm{KLs1}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is the type 1 symmetric generalization of the KL divergence.

Another way to combine KL divergence is given by

$$D^{\beta}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i) = (1-\beta) D_{\mathrm{KL}}(\boldsymbol{\sigma}_i \| \mathbf{z}_i) + \beta D_{\mathrm{KL}}(\mathbf{s}_i \| \mathbf{z}_i) \ , \tag{3}$$

where $\mathbf{z}_i = (1-\beta)\boldsymbol{\sigma}_i + \beta \mathbf{s}_i$. For $\beta = 1/2$, $D^{1/2}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is known as the type 2 symmetric KL divergence, or symmetric Jensen-Shannon divergence [15]. To our best knowledge, $D^{\beta}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ has never been used as a substitute for the KL divergence in SNE. The constant term w.r.t. $\mathbf{s}_i$ in $D^{\beta}_{\mathrm{KLs2}}(\boldsymbol{\sigma}_i \| \mathbf{s}_i)$ is again the entropy of $\boldsymbol{\sigma}_i$ times $1-\beta$. The gradient expression is more complicated and the optimization of the cost function is rather slow without a good approximation of at least the diagonal elements of the Hessian. Due to space limitations, the technical details cannot be described here. Let us just mention that the computational complexity remains the same as that of (t-)SNE (i.e. $\mathcal{O}(N^2)$).

4 Experiments and results

The experiments involve two data sets ($N = 1000$). The first one is a spherical shell that can be re-embedded from 3 to 2 dimensions, such as a planisphere. The second set is a random subsample of the MNIST database of handwritten digits; all gray-level images are vectorized ($M = 576$) and we seek a 2D representation.

The quality criterion used to assess the various embeddings evaluates the preservation K-ary neighborhoods [8]. The rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ in the HD 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 LD space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. The 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 $Q_{\mathrm{NX}}(K) = \sum_{i=1}^N |\nu_i^K \cap n_i^K|/(KN)$.

The competing nonlinear DR methods are classical metric MDS, Sammon's MDS [3], curvilinear component analysis [17] (CCA), SNE, NeRV (i.e. SNE with $D_{\mathrm{KLs1}}^{1/2}$), SNE with $D_{\mathrm{KLs2}}^{1/2}$, and eventually t-SNE. The similarity bandwidths are adjusted to attain a perplexity equal to 40. Compared to SNE, t-SNE use non-identical definitions of the similarities in the HD and LD spaces; the latter are given by $s_{ij} = (1 + d_{ij}^2)^{-1}/(\sum_{k,l,k\neq l}(1 + d_{kl}^2)^{-1})$. The resemblance with the p.d.f. of a Student t distribution explains the method name. The different normalization has no noticible effect in experiments. The discrepancy between the Gaussian similarities in the HD space and the heavy-tailed ones in the LD space amounts to applying an exponential transformation to δ_{ij} to obtain d_{ij} [18]. This transformation stretches the long distances and t-SNE yields therefore more clustered embeddings than regular SNE. In [9], this transformation accounts for the superior results of t-SNE, as compared to those of regular SNE.

The quality curves are plotted in Figures 1 and 2. As a toy example, the spherical shell shows that similarity preservation does not outperform distance preservation; CCA remains the best by far, with the highest curve on the left of the diagram. The main advantage of shift-invariant similarities is useless here, since norm concentration is negligible in 3D. The situation changes totally with the MNIST subset: similarity preservation rules the game. In particular, the mixtures of divergences outperform the non-blended KL divergence and the type 2 mixture is better than type 1 (NeRV). In fact, the type 2 mixture performs just as well as t-SNE, if not better, without resorting to heavy-tailed similarities in the LD space. Other non-reported experiments confirm these trends. One can conjecture that blended divergences in the cost function somehow relax similarity preservation by allowing cuts and tears. Composite similarity vector $\mathbf{z}_i$ used in (3) (type 2 mixture) seems to better serve this purpose than mere reciprocation of the divergence like in (2) (type 1 mixture).

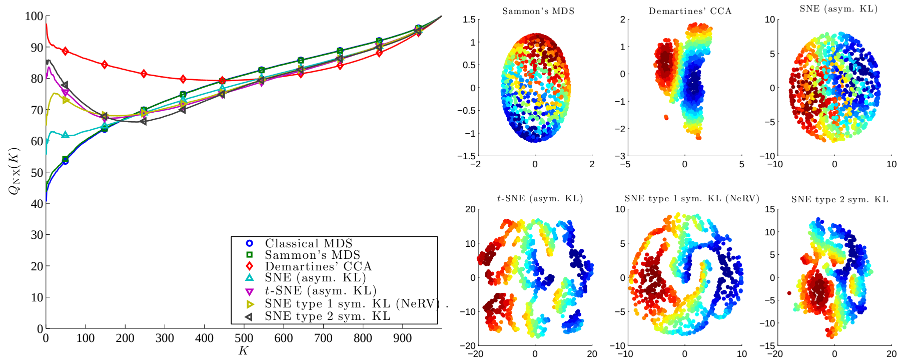

Fig. 1: Left: the quality curves for the spherical shell. Each curve indicates the average normalized agreement between corresponding K-ary neighborhoods in the HD and LD spaces. The higher the left part of the curve, the better the performance. Right: the corresponding embeddings (except for classical MDS).

5 Conclusion

Nonlinear DR methods based on similarity preservation occupy a more and more enviable place in the state of the art. Although the specific similarity definitions used in these methods is certainly one key of their success, other aspects such as the cost function that measures the similarity mismatch cannot be overlooked. Switching from a simple asymmetric KL divergence to symmetric extensions significantly improves the DR results, which then become comparable to those of other SNE variants, such as t-SNE. This shows that other approaches than the use of heavy-tailed similarities work well too. In the near future, we will extend our study to symmetric mixtures of β-divergences, which encompasses the KL divergence and the sum of squared differences as particular cases.

References

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

[2] I. Borg and P. Groenen. *Modern Multidimensional Scaling: Theory and Applications.* Springer-Verlag, New York, 1997.

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

[4] 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.

[5] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In T.G. Dietterich, S. Becker, and Z. Ghahramani, editors, *Advances in Neural Information Processing Systems (NIPS 2001)*, volume 14. MIT Press, 2002.

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

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

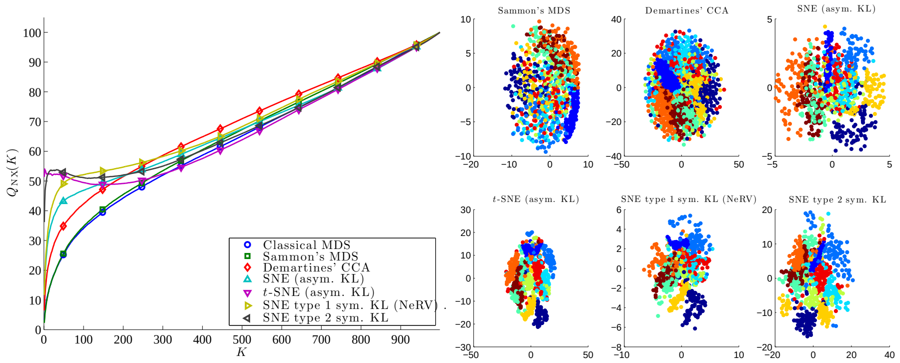

Fig. 2: The quality assessment curves and embeddings for the MNIST subset.

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

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

[10] M. Saerens, F. Fouss, L. Yen, and P. Dupont. The principal components analysis of a graph, and its relationships to spectral clustering. In *Proceddings of the 15th European Conference on Machine Learning (ECML 2004)*, pages 371–383, 2004.

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

[12] G. Hinton and S.T. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems (NIPS 2002)*, volume 15, pages 833–840. MIT Press, 2003.

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

[14] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. In *Proc. International Conference on Computational Science (ICCS 2011)*, Singapore, 2011.

[15] A. Cichocki and S.-i. Amari. Families of alpha- beta- and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12:1532–1568, 2010.

[16] J. Venna and S. Kaski. Local multidimensional scaling with controlled tradeoff between trustworthiness and continuity. In *Proceedings of the 5th Workshop on Self-Organizing Maps (WSOM'05)*, pages 695–702. Paris, September 2005.

[17] 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.

[18] J.A. Lee and M. Verleysen. On the role and impact of the metaparameters in t-distributed stochastic neighbor embedding. In Y. Lechevallier and G. Saporta, editors, *Proc. 19th COMPSTAT*, pages 337–348. Paris (France), August 2010.