ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Perplexity-free t-SNE and twice Student tt-SNE

Cyril de Bodt[1], Dounia Mulders[1], Michel Verleysen[1] and John A. Lee[2] *

1- Université catholique de Louvain - ICTEAM/ELEN
Place du Levant 3 L5.03.02, 1348 Louvain-la-Neuve - Belgium

2- Université catholique de Louvain - IREC/MIRO
Avenue Hippocrate 55 B1.54.07, 1200 Brussels - Belgium

Abstract. In dimensionality reduction and data visualisation, t-SNE has become a popular method. In this paper, we propose two variants to the Gaussian similarities used to characterise the neighbourhoods around each high-dimensional datum in t-SNE. A first alternative is to use t distributions like already used in the low-dimensional embedding space; a variable degree of freedom accounts for the intrinsic dimensionality of data. The second variant relies on compounds of Gaussian neighbourhoods with growing widths, thereby suppressing the need for the user to adjust a single size or perplexity. In both cases, heavy-tailed distributions thus characterise the neighbourhood relationships in the data space. Experiments show that both variants are competitive with t-SNE, at no extra cost.

1 Introduction

Nonlinear dimensionality reduction (NLDR) aims at representing faithfully high-dimensional (HD) data in low-dimensional (LD) spaces, mainly for exploratory visualisation or to foil the curse of dimensionality [1]. Several paradigms have been studied over time, like the reproduction of distances [2] or neighbourhoods [3, 4] from the data space to the embedding space. As a typical method of multidimensional scaling based on distance preservation, Sammon's nonlinear mapping [5] has been popular for nearly 50 years, with about 3600 citations on Google Scholar at the time of writing. As a more recent neighbourhood-preserving NLDR method, t-distributed stochastic neighbour embedding (t-SNE) [6] has earned more than 3400 citations in merely a single decade. Over these 10 years, t-SNE has raised much interest, owing to its impressive results. Variants with alternative divergences as cost functions [7, 8, 9] and lower-complexity tree-based approximations [10, 11] have been developed. Properties of t-SNE have been investigated to explain why it outperforms blatantly many older NLDR methods [12]. However, some questions remain.

For instance, it remains difficult to justify the use of apparently unrelated neighbourhood distributions in the HD and LD spaces, namely, Gauss versus Student. For very high-dimensional data, the intuition of a crowding problem [6], related to the curse of dimensionality, motivates tails that are heavier in LD than in HD. Mid-range to distant neighbours are then repelled further away, which compensates for the relative lack of volume in LD spaces, compared to HD. This approach tends, however, to over-emphasise clusters or to be too extreme

*DM and CdB are FNRS Research Fellows. JAL is a FNRS Senior Research Associate.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

for data with low intrinsic dimensionality. Adapting the degrees of freedom of the Student function has been a possible workaround [13]. We investigate here the use of twice Student SNE, coined tt-SNE, with t functions in both spaces.

Another question relates to the relevance of the perplexity, the main metaparameter in t-SNE. Following previous work [14], we show that compound Gaussian neighbourhoods, with exponentially growing perplexities, can be used in t-SNE to obtain multi-scale neighbourhoods with intermediate tails, between Gauss and Student. Such compounded Gaussian neighbourhoods span systematically a broad range of perplexities, relieving the user of its adjustment.

Experiments show that the proposed variants are as competitive and efficient as t-SNE and they are applicable in a broader variety of cases.

The rest of this paper is organised as follows. Section 2 is a reminder of SNE and t-SNE. Section 3 describes our contributions. Section 4 reports the experimental results. Section 5 draws the conclusions and sketches perspectives.

2 SNE and t-SNE

Let $\mathbf{\Xi} = [\boldsymbol{\xi}_i]_{i=1}^N$ denote a set of N points in a HD metric space with M features. Let $\mathbf{X} = [\mathbf{x}_i]_{i=1}^N$ represent it in a P-dimensional LD metric space, $P \leq M$. The HD and LD distances between the ith and jth points are noted δ_{ij} and d_{ij}. SNE defines HD and LD similarities, for $i \in \mathcal{I} = \{1, \ldots, N\}$ and $j \in \mathcal{I} \backslash \{i\}$ [4]:

$$\sigma_{ij} = \frac{\exp\left(-\pi_i \delta_{ij}^2/2\right)}{\sum_{k \in \mathcal{I}\backslash\{i\}} \exp\left(-\pi_i \delta_{ik}^2/2\right)}, \; s_{ij} = \frac{\exp\left(-d_{ij}^2/2\right)}{\sum_{k \in \mathcal{I}\backslash\{i\}} \exp\left(-d_{ik}^2/2\right)}, \; \sigma_{ii} = s_{ii} = 0.$$

Precisions π_i are set to reach a user-fixed perplexity $K_\star$ for the distribution $[\sigma_{ij}; j \in \mathcal{I}\backslash\{i\}]$: π_i such that $\log K_\star = -\sum_{j \in \mathcal{I}\backslash\{i\}} \sigma_{ij} \log \sigma_{ij}$. The perplexity is interpreted as the size of the soft Gaussian neighbourhood. SNE then finds the LD positions by minimising the sum of divergences between the HD and LD similarity distributions: $C_{SNE} = \sum_{i \in \mathcal{I}, j \in \mathcal{I}\backslash\{i\}} \sigma_{ij} \log\left(\sigma_{ij} / s_{ij}\right)$.

Besides symmetrising the similarities, t-SNE uses a Student t function with one degree of freedom in the LD space, to cope with 'crowding' problems [6]:

$$\sigma_{ij,s} = \frac{\sigma_{ij} + \sigma_{ji}}{2N}, \; s_{ij,t} = \frac{1}{(1 + d_{ij}^2) \sum_{k \in \mathcal{I}, l \in \mathcal{I}\backslash\{k\}} (1 + d_{kl}^2)^{-1}}, \; s_{ii,t} = 0.$$

The t-SNE cost function uses KL divergences as in SNE. It is minimised by gradient descent, each evaluation requiring $\mathcal{O}\left(N^2\right)$ operations.

The discrepancy between Gauss and Student distributions is arbitrarily fixed in t-SNE, without guarantee that the induced distance transformation is optimal. To address this issue, a variant of t-SNE has been described [13], where $s_{ij,t}$ has an additional parameter α, adjusting the degrees of freedom:

$$s_{ij,t} = \frac{1}{(1 + d_{ij}^2/\alpha)^{(\alpha+1)/2} \sum_{k \in \mathcal{I}, l \in \mathcal{I}\backslash\{k\}} (1 + d_{kl}^2/\alpha)^{-(\alpha+1)/2}} \;, \tag{1}$$

where α can be learned. A relation between α and the intrinsic dimensionality of data is hypothesised. If $\alpha \to \infty$, the Student function tends to a Gaussian.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

3 Heavy tails in the high-dimensional space

A Student with infinite degrees of freedom boils down to a Gaussian. Conversely, a Student turns out to be a compound distribution, involving zero-mean Gaussians with gamma-distributed precisions. Marginalising over the precision leads to a Student. The shape and scale parameters of the gamma determine the degrees of freedom and scaling of the resulting Student. Hence, t-SNE appears as a variant of SNE where both HD and LD neighbourhoods follow Student functions with, respectively, ∞ and 1 degrees of freedom. Let us consider that (i) t-SNE works usually better for very high-dimensional than simple manifolds, and (ii) tweaking the degrees of freedom α in the LD neighbourhoods $s_{ij,t}$ has been tried [13], linking α and the intrinsic data dimensionality M'. From there, we suggest using Student functions in both spaces, with degrees of freedom equal to $M'-1$ and $P-1$, instead of ∞ and 1 in t-SNE. Formally, twice t-SNE, coined tt-SNE, relies thus on $\sigma_{ij,ts} = (\sigma_{ij,t} + \sigma_{ji,t})/(2N)$,

$$\sigma_{ij,t} = \frac{(1+\pi_i d_{ij}^2/M')^{-M'/2}}{\sum_{k\in\mathcal{I}\setminus\{i\}}(1+\pi_i d_{kl}^2/M')^{-M'/2}},\ s_{ij,t} = \frac{(1+d_{ij}^2/P)^{-P/2}}{\sum_{k\in\mathcal{I},l\in\mathcal{I}\setminus\{k\}}(1+d_{kl}^2/P)^{-P/2}}.$$

Unlike in [13], the degrees of freedom are fixed beforehand in both $\sigma_{ij,t}$ and $s_{ij,t}$; an ad hoc estimator of the intrinsic data dimensionality provides M' [14]. The chosen degrees of freedom ensure consistency with genuine t-SNE in the LD space if $P = 2$, whereas the discrepancy between $\sigma_{ij,t}$ and $s_{ij,t}$ adapts with respect to the dimensionality gap between M' and P. Precisions π_i can be determined to reach the user-specified perplexity $K_\star$. Learning the degrees of freedom as in [13] would increase running times, since the precisions should be retuned at each iteration.

If both Gaussian and heavy-tailed distributions like Student's can be used for the HD neighbourhoods, then other intermediate distributions could be considered as well. Relying again on the definition of the Student as a compound of Gaussians with gamma-distributed precisions, we suggest a similar but discretised scheme. We use the multi-scale Gaussian similarities from previous work [14], defined as $\sigma_{ij,s} = (\sigma_{ij} + \sigma_{ji})/(2N)$, with

$$\sigma_{ij} = \frac{1}{H}\sum_{h=1}^{H}\sigma_{hij},\ \text{and}\ \sigma_{hij} = \frac{\exp(-\pi_{hi}\delta_{ij}^2/2)}{\sum_{k\in\mathcal{I}\setminus\{i\}}\exp(-\pi_{hi}\delta_{ik}^2/2)}\ , \tag{2}$$

where $H = \lfloor \log_2(N/2) \rceil$. Precisions π_{hi} are adjusted by imposing exponentially growing perplexities $K_{\star h} = 2^h$ with $1 \leq h \leq H$ (or linearly growing entropies). Precisions are thus sampled in a data-driven way. The resulting compound distribution has a heavier tail than a single Gaussian, although it remains lighter than a Student since the widest Gaussian component σ_{Hij} in σ_{ij} keeps an exponentially decreasing tail. As a key advantage, multi-scale Gaussian neighbourhoods do no longer require the user to set a perplexity. Integrating them in t-SNE leads to multi-scale t-SNE or Ms.t-SNE in short.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

4 Experiments and discussion

The compared NLDR methods are (i) SNE [4] (ii) t-SNE in its genuine 2008 implementation [6], (iii) a modified version where precisions are determined with Newton's method [15] instead of binary searches, (iv) tt-SNE, (v) Ms.t-SNE, and (vi) Ms.SNE [14]. Perplexities in the first four methods are fixed to 32.

The data sets are a spherical manifold ($N = 1000, M = 3, M' = 2.03$) [14], COIL-20 ($N = 1440, M = 16384, M' = 6.5$) [16], a subset of MNIST ($N = 1000, M = 784, M' = 8.07$) [17], and B. Frey's faces ($N = 1965, M = 560, M' = 6.43$). Intrinsic dimensionality was computed as in [14]. Target dimensionality P is two for all data sets. As to DR quality, some studies developed indicators of the HD neighbourhoods preservation in the LD space [18, 14], becoming generally adopted in several publications [7, 14]. Let sets ν_i^K and n_i^K index the K nearest neighbours of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ in the HD and LD space, respectively, with $Q_{\text{NX}}(K) = \sum_{i \in \mathcal{I}} \left| \nu_i^K \cap n_i^K \right| / (KN) \in [0, 1]$ measuring their average normalised agreement. As its expectation amounts to $K/(N-1)$ for random LD points, $R_{\text{NX}}(K) = ((N-1)\, Q_{\text{NX}}(K) - K)/(N-1-K)$ allows comparing different neighbourhood sizes [14]. The curve is displayed on a log-scale for K as close neighbours prevail. Evaluated in $\mathcal{O}\left(N^2 \log N\right)$ time [18], its area $\text{AUC} = \left(\sum_{K=1}^{N-2} R_{\text{NX}}(K)/K\right) / \left(\sum_{K=1}^{N-2} 1/K\right) \in [-1, 1]$ grows with DR quality, quantified at all scales with an emphasis on small ones.

The results are shown in Fig. 1. In terms of AUC, SNE performs poorly compared to all t-SNE variants [12]. Precision computation with either binary search or Newton's method [15] has little impact on the results. Ms.SNE [14] yields the best AUCs. Using heavy tails in the HD space, the proposed methods tt-SNE and Ms.t-SNE produce embeddings where clusters are less over-emphasised, since the HD-LD tail gap is actually tighter for these than for t-SNE. A Student with $\alpha = 2$ or 3 is indeed always closer to another Student (with $\alpha < \infty$) than a Gaussian (a Student with $\alpha = \infty$). Cluster over-emphasis, present in t-SNE and to a lesser extent in tt-SNE and Ms.t-SNE, is slightly detrimental to the AUC, with weaker preservation of large neighbourhoods. Precision computation with t HD neighbourhoods in tt-SNE fails to converge for some points, without impact on the embedding quality, tt-SNE being only second to Ms.SNE and Ms.t-SNE. Ms.t-SNE shows very good 1NN fidelity, without significant loss in larger neighbourhoods, with the advantage for the user of having no perplexity to adjust, like in Ms.SNE. As to running times, all t-SNE variants have the same complexity of $\mathcal{O}\left(N^2\right)$, whereas Ms.SNE runs in $\mathcal{O}\left(N^2 \log N\right)$, explaining the overall best results of the latter. Computation of multi-scale neighbourhoods takes $\mathcal{O}\left(N^2 \log N\right)$ in Ms.t-SNE but it is carried out only once at initialisation and is thus negligible with respect to the iterated gradient evaluation in $\mathcal{O}\left(N^2\right)$.

5 Conclusions and perspectives

Two variants of t-SNE are proposed. The first replaces Gaussian neighbourhoods in the data space with Student t functions, to be more consistent with

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

the t neighbourhoods in the embedding space. The dimensionality gap between both spaces is accounted for by different degrees of freedom, determined in the HD space by the intrinsic data dimensionality. The second variant reuses the compound, multi-scale HD neighbourhoods described in [14] and matches them with LD t neighbourhoods; the user has no longer any perplexity or neighbourhood size to adjust. Experiments show that these variants are competitive with t-SNE and broadens its applicability. Perspectives include (i) the possibility to use multivariate t distributions to better accomodate for taking into account the intrinsic dimensionality, and (ii) adapting precision sampling in Ms.t-SNE to better match the t function in the LD space.

References

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

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

[3] T. Kohonen. *Self-Organizing Maps*. Springer, Heidelberg, 2nd edition, 1995.

[4] G. Hinton and S.T. Roweis. Stochastic neighbor embedding. In *Proc. (NIPS 2002)*, pages 833–840. MIT Press, 2003.

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

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

[7] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR*, 11:451–490, 2010.

[8] K. Bunte, S. Haase, M. Biehl, and T. Villmann. Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences. *Neurocomputing*, 90:23–45, 2012.

[9] J.A. Lee, E. Renard, G. Bernard, P. Dupont, and M. Verleysen. Type 1 and 2 mixtures of Kullback-Leibler divergences as cost functions in dimensionality reduction based on similarity preservation. *Neurocomputing*, 112:92–108, 2013.

[10] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In *Proc. ICML 2013*, pages 127–135. JMLR W&CP, 2013.

[11] L. van der Maaten. Barnes-Hut-SNE. In *Proc. International Conference on Learning Representations*, 2013.

[12] J.A. Lee and M. Verleysen. Two key properties of dimensionality reduction methods. In *Proc. IEEE SSCI – CIDM*, pages 163–170, 2014.

[13] L.J.P. van der Maaten. Learning a parametric embedding by preserving local structure. In *Proc. 12th AISTATS*, pages 384–391, Clearwater Beach, FL, 2009.

[14] J.A. Lee, D.H. Peluffo-Ordóñez, and M. Verleysen. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*, 169:246–261, 2015.

[15] M. Vladymyrov and M.Á. Carreira-Perpiñán. Entropic affinities: Properties and efficient numerical computation. In *Proc. 30th ICML*, JMLR: W&CP. Atlanta, GA, 2013.

[16] S. A. Nene, S. K. Nayar, and H. Murase. Columbia object image library (COIL-20). Technical Report CUCS-005-96, Columbia University, 1996.

[17] Yann LeCun, Corinna Cortes, and Christopher J.C. Burges. The MNIST database of handwritten digits, 1998.

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

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

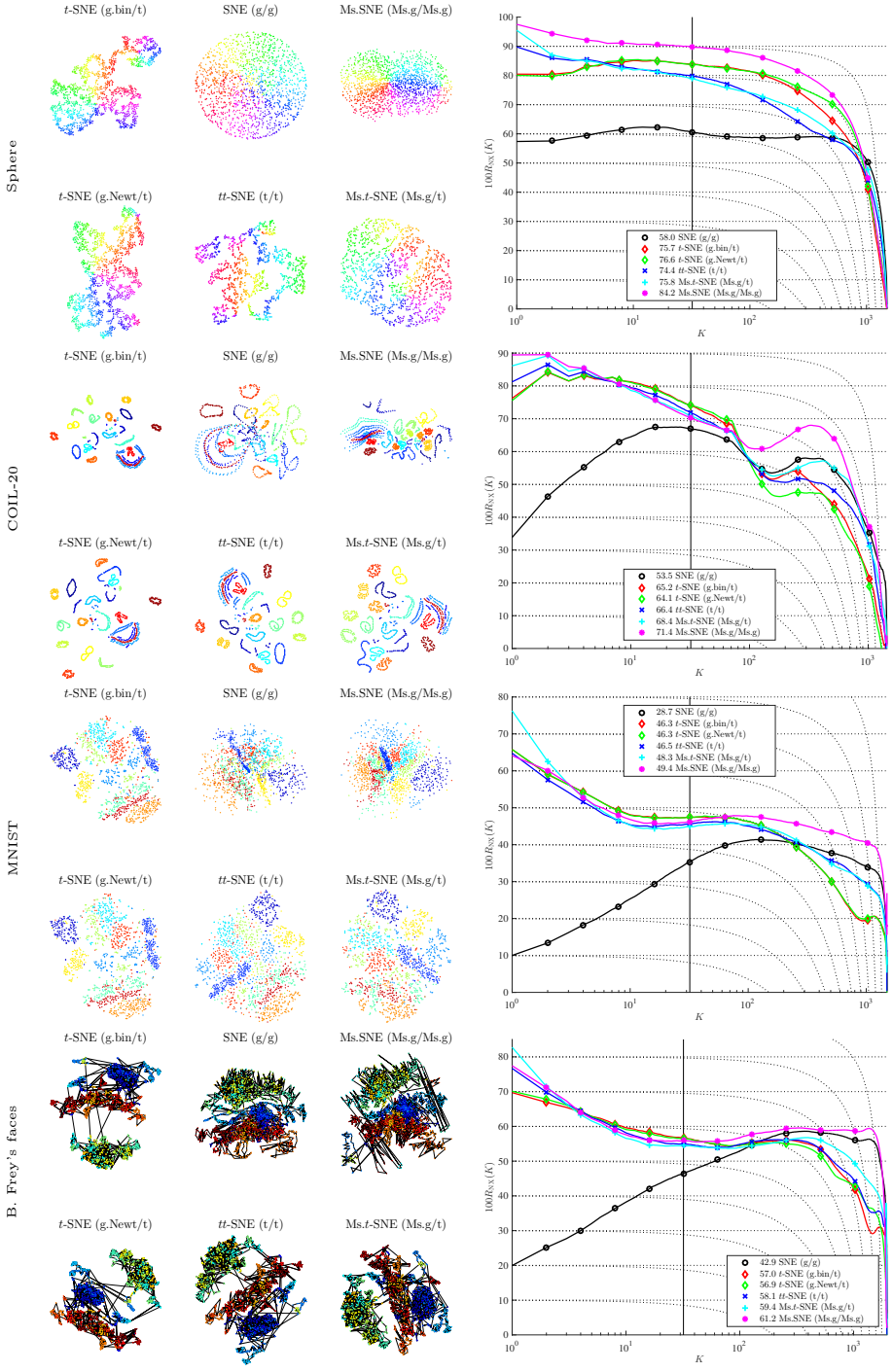

Fig. 1: Embeddings and $R_{\mathrm{NX}}(K)$ quality curves (with their AUC) for Sphere, COIL-20, MNIST, and Frey's faces. 'g' Gaussian, 't' Student, 'Ms.' multi-scale.