ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Can MDS rival with t-SNE by using the symmetric Kullback-Leibler divergence across neighborhoods as a pseudo-distance?

John A. Lee[12], Pierre Lambert[1], Edouard Couplet[1], Pierre Merveille[1]
Ludovic Journaux[3], Dounia Mulders[1], Cyril de Bodt[1], and Michel Verleysen[1] *

1- UCLouvain - IREC/MIRO
Avenue Hippocrate 55, 1200 Brussels, Belgium

2- UCLouvain - ICTEAM/ELEN
Place du Levant 3, 1348 Louvain-la-Neuve, Belgium

3- Institut Agro Dijon - Laboratoire d'Informatique de Bourgogne
Boulevard Docteur Petitjean 26, 21079 Dijon, France

Abstract. Local methods of dimensionality reduction like neighborhood embedding (NE) and t-SNE in particular outperform older global approaches such as stress-based multi-dimensional scaling (MDS). Stochastic neighborhoods are less sensitive than distances to statistical variations between spaces with strongly different dimensionalities, making a match across them very difficult. Here, we take inspiration from those stochastic neighborhoods in order to devise a pseudo-distance that is less prone to concentration than the Euclidean distance. For two points in the high-dimensional data space, it is defined as the **s**ymmetrized **K**ullback-**L**eibler divergence **a**cross the (stochastic) **n**eighborhoods of the two points (SKLAN). Plugging the SKLAN in a method of stress-based MDS, we compare quantitatively t-SNE, MDS with all Euclidean distances, and MDS with SKLAN & Euclidean distances on several data sets. The results show that SKLAN allows MDS to perform competitively with t-SNE.

1 Dimensionality reduction and motivation

Dimensionality reduction (DR) aims at representing high-dimensional (HD) data with low-dimensional (LD) embeddings, in which salient features of data are preserved or even highlighted. Such features can be for instance variance in principal component analysis (PCA) [1], distances in multidimensional scaling (MDS) [2], or similarities in neighbor embedding (NE) [3]. Most of the current, state-of-the-art methods of DR stem from the family of neighbor embedding, with stochastic neighbor embedding (SNE) [4] as their common but forgotten ancestor. The celebrity in the family remains undoubtedly t-SNE [5], probably because of its capability to amplify cluster gaps in the embeddings, on top of overall good performance at DR. Neighbor embedding in general and t-SNE in particular have dusted previous paradigms of DR, and notably so for MDS. Stress-based MDS consists in finding embedding coordinates such that LD distances match those in HD [2]. In contrast, NE determines the embedding by

*J.A.Lee is a Research Director with the Belgian F.R.S.-FNRS. P.Lambert is funded by the F.R.S.-FNRS FRIA scholarship #1.E013.23.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

matching normalized similarities, also known as (entropic) affinities [6], which can be interpreted as probabilities of points to be neighbors of one another. The key difference appears to be that such similarities are less sensitive to broad dimensionality gaps between the data and embedding spaces, while distances are known to concentrate more or less in HD or LD, respectively.

This paper investigates whether MDS can be revived to some extent by replacing the Euclidean distance in the HD space with a (pseudo-)distance that is more robust to norm concentration [7]. For two points i and j in the data set, the proposed distance involves entropic affinities just like in NE, for a given perplexity, to reflect probabilities of being neighbors, also known as 'stochastic neighborhoods' [4]. Then, the pseudo-distance is computed as the symmetrized Kullback-Leibler divergence between the distributions of those probabilities around i and j, over all points k, excluding i and j themselves. In spite of not fulfilling the triangle inequality, this distance measured in HD space can be matched to Euclidean distances in a LD embedding with stress-based MDS. Results on a few typical benchmark data sets show that MDS with the proposed metric can compete with NE and t-SNE, except for cluster gap magnification.

The rest of this paper is organized as follows. Section 2 is a short reminder of NE and t-SNE. Section 3 details the proposed pseudo-distance and integrates it in stress-based MDS . Experimental results are reported and discussed in Section 4. Section 5 concludes and sketches perspectives.

2 Neighbor embedding and t-SNE

Most methods of NE involve stochastic neighborhoods, one around each data point, which are discrete probabilities of point j to be a neighbor of point i. In SNE, these soft neighborhoods are softmax ratios, i.e., Gaussian functions that are normalized into discrete neighborhood probabilities in both the data and embedding spaces (HD & LD), whose mismatch is measured with Kullback-Leibler divergences. In t-SNE, the HD Gaussian affinities are symmetrized, while the LD affinities are Student t hyperbolic functions that are normalized jointly. If $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$ and $\mathbf{Y} = [\mathbf{y}_i]_{1\leq i\leq N}$ denote the HD data and their LD embedding, then the pairwise Euclidean distances be shortened as $\mathrm{d}\mathbf{x}_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$ and $\mathrm{d}\mathbf{y}_{ij} = \|\mathbf{y}_i - \mathbf{y}_j\|_2$ in HD and LD, respectively. Next, the HD entropic and symmetric affinities are

$$p_{j|i} = \frac{\exp(-\mathrm{d}\mathbf{x}_{ij}^2/2\sigma_i^2)}{\sum_{1\leq k\leq N, k\neq i}\exp(-\mathrm{d}\mathbf{x}_{ik}^2/2\sigma_i^2)}\,, \quad p_{i|i} = 0\,, \quad \text{and } p_{ij} = \frac{p_{j|i} + p_{i|j}}{2N}\ , \tag{1}$$

where bandwidth σ_i is such that entropy $H_i = \log K_\star = -\sum_{1\leq j\leq N, j\neq i} p_{j|i}\log p_{j|i}$ is the same around all $\mathbf{x}_i$ and set by perplexity $K_\star$. In the LD embedding space, the symmetrized entropic affinities are

$$q_{ij} = \frac{(1+\mathrm{d}\mathbf{y}_{ij}^2)^{-1}}{\sum_{1\leq k,l\neq N, k\neq l}(1+\mathrm{d}\mathbf{y}_{kl}^2)^{-1}}\,, \quad q_{ii} = 0\ . \tag{2}$$

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

Then, t-SNE tries to match p_{ij} with q_{ij} by minimizing the joint KL divergence $\mathrm{KL}(P||Q) = \sum_{1\leq i,j\leq N, i\neq j} p_{ij} \log(p_{ij}/q_{ij})$ with respect to $\mathbf{Y}$. Minimization is carried out with gradient descent and (Nesterov) momentum.

3 Symmetric KL divergences across neighborhoods

A pitfall of stress-based MDS is that most stress functions proceed with direct comparison of HD and LD distances between two points i and j, as in $S(\mathbf{X};\mathbf{Y}) = \sum_{1\leq i<j\leq N} w_{ij}(\mathrm{d}\mathbf{x}_{ij} - \mathrm{d}\mathbf{y}_{ij})^2$, where w_{ij} are weights, typically giving the shorter distances more priority. However, the statistical distributions of distances differ significantly in shape and properties, depending on the space dimensionality. As part of the curse of dimensionality, most norms concentrate as the dimensionality grows: the ratio expectation over standard deviation increases [7]. Therefore, $\mathrm{d}\mathbf{x}_{ij}$ and $\mathrm{d}\mathbf{y}_{ij}$ can have inherently different and thus irreconcilable distributions if the dimensionality gap is too broad. Part of the success of NE and t-SNE stems from replacing distances with affinities or similarities, not affected as strongly as distances by space dimensionality. Inspired by NE, we propose a pseudo-distance that address this issue of excessive sensitivity to dimensionality. Intuitively, for two point $\mathbf{x}_i$ and $\mathbf{x}_j$, it measures how the stochastic neighborhoods around those differ. The symmetric Kullback-Leibler divergence across neighborhoods (SKLAN) is defined as

$$\eth\mathbf{x}_{ij} \triangleq \frac{1}{2}\mathrm{KL}(p_{\cdot|i}||p_{\cdot|j}) + \frac{1}{2}\mathrm{KL}(p_{\cdot|j}||p_{\cdot|i}) \tag{3}$$

$$\approx \frac{1}{2}\sum_{1\leq k\leq N, k\neq j} p_{k|i}\log(p_{k|i}/p_{k|j}) + \frac{1}{2}\sum_{1\leq k\leq N, k\neq i} p_{k|j}\log(p_{k|j}/p_{k|i}) \tag{4}$$

$$\approx \frac{1}{2}\sum_{1\leq k\leq N, i\neq k\neq j} (p_{k|i} - p_{k|j})(\log p_{k|i} - \log p_{k|j}) \geq 0 \tag{5}$$

with, by construction, $\eth\mathbf{x}_{ij} = \eth\mathbf{x}_{ji}$ and $\eth\mathbf{x}_{ii} = 0$. In contrast to positivity and symmetry, triangle inequality is not verified and therefore $\eth\mathbf{x}_{ij}$ is merely a pseudo-distance. The successive approximations (4) and (5) result from the usual though arbitrary convention of setting $p_{i|i} = 0 = p_{j|j}$, with an infinite logarithm, contradicting the alternative intuition that $\log p_{i|i} \propto \mathrm{d}\mathbf{x}_{ii} = 0$. Following this intuition would lead to null terms for $k = i$ and $k = j$ when summing over k. Therefore, those two problematic terms (for $k \neq i$, $k \neq j$) get excluded and preliminary experiments, not reported here, confirmed better results without them.

Next, the SKLAN can easily be plugged in any stress-based MDS, replacing the Euclidean distance in the HD data space. It can fit in Sammon mapping [8]: $S(\mathbf{X};\mathbf{Y}) = \sum_{1\leq i<j\leq N}(\eth\mathbf{x}_{ij} - \mathrm{d}\mathbf{y}_{ij})^2/\eth\mathbf{x}_{ij}$, where the weight $1/\eth\mathbf{x}_{ij}$ is expected to work better than the original $1/\mathrm{d}\mathbf{x}_{ij}$, which concentrates and therefore struggles to discriminate local and global structure. The SKLAN can also fit in curvilinear component analysis (CCA) [9]: $S(\mathbf{X};\mathbf{Y}) = \sum_{1\leq i<j\leq N}(\eth\mathbf{x}_{ij} - \mathrm{d}\mathbf{y}_{ij})^2 H(\lambda - \mathrm{d}\mathbf{y}_{ij})$, where λ is a neighborhood radius and H is Heaviside's step function. The

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

implementation that is used here rely on pointwise λ_i, such that these individualized radii encompass K neighbors around each point, from $K = N$ to $K \approx 2$, with a decrement schedule along iterations. CCA is expected to outperfom Sammon mapping, as it can give up on some attractive force, just like t-SNE.

4 Experiments, results, and discussion

In order to assess the proposed pseudo-distance in MDS, several data sets are embedded in 2D. Sammon mapping and CCA with Euclidean distances and SKLAN are compared to t-SNE. The perplexity $K_\star$ of stochastic neighborhoods is kept the same in both SKLAN and t-SNE; other metaparameters are left to default values. In addition to embeddings, the curves $R_{\mathrm{NX}}(K) = ((\frac{N-1}{KN}\sum_i |\nu_i^K \cap n_i^K|) - K)/(N-1+K)$ [10] are reported, where $1 \leq K \leq N$ is a neighborhood size and ν_i^K and n_i^K are the K-ary (binary, non-stochastic) neighborhoods of $\mathbf{x}_i$ and $\mathbf{y}_i$, respectively. These curves allow inspecting both the local and global structures. The minimum value is 0 (not better than a random embedding on average) and the maximum is 1 (perfect rendering of all K-ary neighborhoods from HD to 2D). The area under the curves (AUCs) compounds local and global.

The experiments embed in 2D eight popular datasets from the UCI repository. Figure 1 shows the results for COIL-20 ($N = 1440$, $K_\star = 32$), COIL-100 ($N = 7200$, $K_\star = 64$), a tenth of MNIST ($N = 6000$, $K_\star = 32$), and Brendan Frey's faces ($N = 1965$, $K_\star = 32$). Each data set is embedded with Euclidean Sammon mapping and CCA, t-SNE, as well as SKLAN Sammon and SKLAN CCA. Figure 2 shows the result for Phoneme ($N = 4501$, $K_\star = 64$), Google ($N = 5456$, $K_\star = 64$), Abalone ($N = 4177$, $K_\star = 64$) and a third of MouseRNA ($N = 7941$, $K_\star = 32$). Visual inspection and quantitative checks reveal that MDS with SKLAN becomes competitive again with state-of-the-art methods of NE like t-SNE. Small neighborhoods are best preserved with SKLAN CCA, which also achieves best AUC for 5 data sets out of 8. At this stage of the proof of concept, MDS with SKLAN has cubic complexity for initial distance computation (N^2 distances over $N-2$ neighbors), followed by quadratic MDS iterations, which are simpler than those of non-accelerated t-SNE.

5 Conclusions and perspectives

The results confirm indirectly that stress-based MDS, as it is usually applied, namely, by trying to match Euclidean distances between the data and embedding spaces, gets affected by the difference in distance concentration between these two spaces, reflecting their dimensionality gap. However, in the same spirit as t-SNE, this issue is overcome by treating the two spaces differently. Here, the distance in the embedding space is kept Euclidean for mathematical convenience while the distance in the data space is the SKLAN, the symmetric KL divergence across neighborhoods. Those stochastic neighborhoods, originally devised for SNE [4], show little sensitivity to distance concentration and measuring their symmetrized divergence brings a pseudo-distance that can be plugged in MDS.

ESANN 2025 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 23-25 April 2025, i6doc.com publ., ISBN 9782875870933. Available from http://www.i6doc.com/en/.

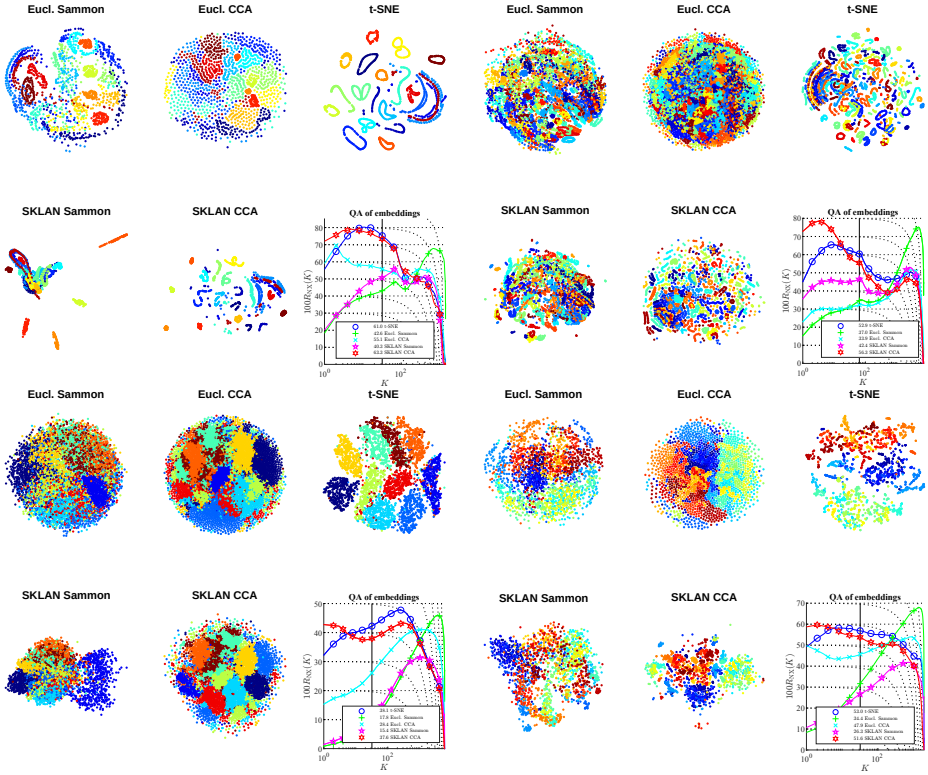

Fig. 1: Embeddings & DR quality curves: (top) COIL-20 & COIL-100, (bottom) MNIST & Frey faces.

Two variants of stress-based MDS are compared here, Sammon mapping and the more advanced CCA. The SKLAN quantitatively improves local neighborhood preservation for both, marginally for Sammon, strongly for CCA, the latter often rivaling or even outperforming t-SNE with a simpler algorithm. Future work will investigate how accelerated versions of the SKLAN and CCA can compete with Barnes-Hut t-SNE [3] and UMAP [11].

References

[1] I.T. Jolliffe. *Principal Component Analysis*. Springer Verlag, 1986.

[2] I. Borg and P.J.F. Groenen. *Modern Multidimensional Scaling Theory and Applications*. Springer, 2005.

[3] Z. Yang, J. Peltonen, and S. Kaski. Scalable optimization of neighbor embedding for visualization. In S. Dasgupta and D. McAllester, editors, *Proc. 30th ICML*, volume 28 of *PMLR*, pages 127–135, Atlanta, Georgia, USA, 17–19 Jun 2013.

[4] G.E. Hinton and S. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems*, volume 15. MIT Press, 2002.

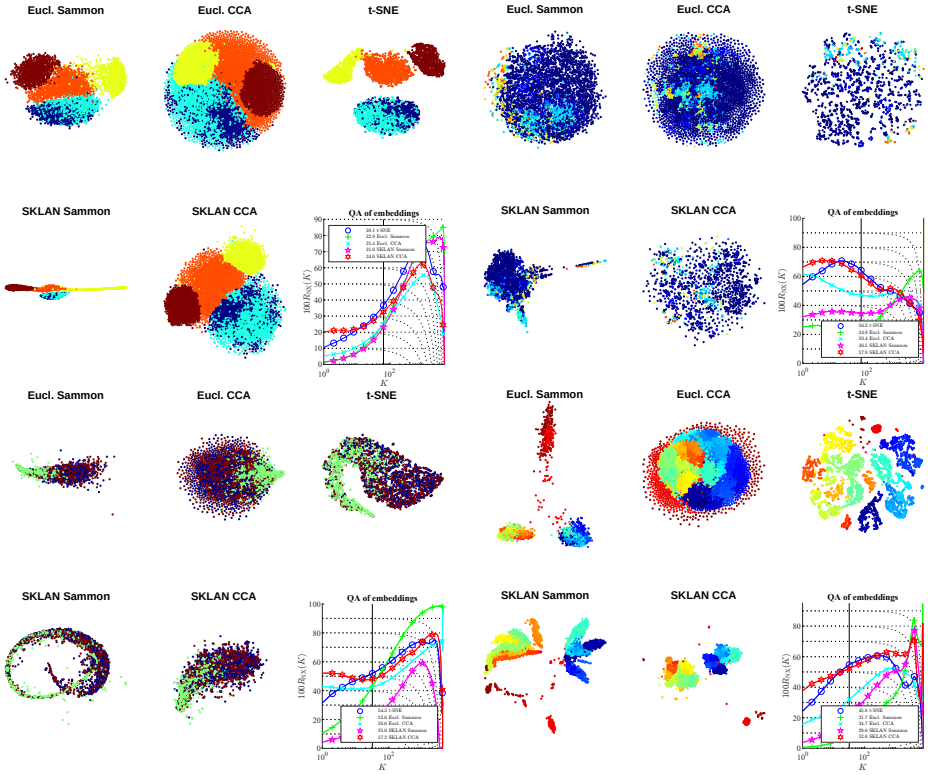

Fig. 2: Embeddings & DR quality curves: (top) Phoneme & Google, (bottom) Abalone & Mouse RNA.

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

[6] M. Vladymyrov and M. Carreira-Perpinan. Entropic affinities: Properties and efficient numerical computation. In S. Dasgupta and D. McAllester, editors, *Proc. 30th ICML*, volume 28 of *PMLR*, pages 477–485, Atlanta, Georgia, USA, 17–19 Jun 2013.

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

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

[9] 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.

[10] 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.

[11] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.