ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

Multi-Scale Stochastic Neighbor Embedding with Twice Adaptive Bandwidths

John A. Lee[1], Pierre Lambert[1], Edouard Couplet[1], Pierre Merveille[1]
Dounia Mulders[1], Cyril de Bodt[2], and Michel Verleysen[1] *

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

2- UNamur - Dept. of Mathematics & naXys institute
Rue Grafé 2, 5000 Namur, Belgium

Abstract. Neighbor embedding has been a quantum leap in nonlinear dimensionality reduction, revolutionizing the way data can be visualized. Neighbor embedding typically adapts to the local density in the high-dimensional data space with adaptive bandwidths in entropic affinities, while it resolves scale indeterminacies by having unit bandwidths in the low-dimensional embedding space. In this paper, multi-scale stochastic neighbor embedding (Ms.SNE) is improved by allowing it to adapt low-dimensional bandwidths in a data-driven way instead of having fixed ones. In practice, Ms.SNE goes through a multi-scale optimization process; coordinates and bandwidths are optimized separately, in an alternate fashion, to avoid interferences: (i) bandwidths are optimized from previous coordinates and (ii) coordinates are optimized given the new bandwidths. Experimentally, twice adaptive bandwidths improve Ms.SNE's capability to preserve neighborhoods on all scales, i.e., local *and* global data structure; this claim is supported with quantitative results on several benchmarks.

1 Neighbor embedding for data visualization

Dimensionality reduction (DR) [1] yields nonlinear embeddings [2] that allow for visualization and exploratory analysis of data in many domains, such as computational biology [3], to cite just one example. Modern DR involves mostly methods of neighbor embedding (NE) [4], like Student t-distributed stochastic NE (t-SNE) [5] or uniform manifold approximation and projection (UMAP) [6]. These methods are very robust to the curse of dimensionalty [7] and produce *local* embeddings; sparsity of small-size neighborhoods is also the key to accelerate these methods [8, 6, 9, 10]. However, sparsity might also cause the loss of the global structure of data [11, 6, 12, 10, 2, 13, 14]. This depends on how the final embedding does reminisce [12] about its initialization with PCA [15] or Laplacian eigenmaps [16], either due to early stopping [5] or explicit regularization [13]. Another workaround consists in having neighborhoods on two [9, 10] or more scales [11], even though acceleration can become more difficult.

A less investigated feature of NE is a form of uniformization of data density in the low-dimensional (LD) embedding. It results from the use of entropic affinities

*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 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

[17] in the high-dimensional (HD) space, with a bandwidth that adapts to the local data density, whereas the low-dimensional counterpart of these affinities have typically a unit bandwidth to (i) resolve a scale indeterminacy and (ii) avoid interference between updates of the manifold coordinates and the bandwidths that would divide them. A noticible exception is neighborhood retrieval and visualization (NeRV) [18], in which HD bandwidths are copied and used in the LD affinities. Another case where LD bandwidths differ from one is Mutli-scale SNE (Ms.SNE) [11]: due to its multi-scale nature, this method has unit bandwidth for the smallest-size neighborhoods, while larger neighborhoods have bandwidths that grow with the square root of their size K (in 2D).

We propose a modified version of Ms.SNE in which entropic affinities [17] are used in the HD data space as well as in the LD embedding, i.e., twice-adaptive bandwidths, to increase the data-driven flexibility and expressiveness of the method. In practice, the method optimizes first HD bandwiths and next both LD coordinates and LD bandwidths, in an alternate scheme. Entropy on scale h in LD is the same as in HD: $H_h^{\mathrm{LD}} = H_h^{\mathrm{HD}} = \log K_{h\star}$, where $K_{h\star} = 2^h \leq N$ is known as the perplexity and N is the data size. Experimentally, we show both visually and quantitatively the difference between regular NE and scale-adaptive NE with data-driven affinity bandwidths in both HD and LD spaces.

2 Single-scale stochastic neighbor embedding

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 embedding in a P-dimensional space, with $P \ll M$. Often $P = 2$ for visualization purposes. The squared Euclidean distances between the ith and jth points are given by $\delta_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2^2$ and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2^2$ in the HD and LD spaces, respectively.

Starting from pairwise distances, similarities $\boldsymbol{\sigma}_i = [\sigma_{ij}]_{1\leq j\leq N}$ and $\mathbf{s}_i = [s_{ij}]_{1\leq j\leq N}$ can be defined for $i \neq j$ in the M- and P-dimensional spaces as

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

If $i = j$, then $\sigma_{ij} = s_{ij} = 0$ by convention. Symbol $\pi_i = \lambda_i^{-2}$ and $p_i = l_i^{-2}$ denote the precisions of the normalised Gaussian functions centered on $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, that is, the inverse of their squared bandwidths λ_i and l_i.

The positive, normalized softmin ratios in σ_{ij} and s_{ij} can be interpreted as a probabilistic or stochastic membership to soft Gaussian neighborhoods. Like hard neighborhoods, which can be characterised by either their radius or size K, there is a relationship between the bandwidth and a generalised concept of size, in the form of the perplexity. The perplexity K_i of the soft neighborhood around $\boldsymbol{\xi}_i$ depends on the entropy $H(\boldsymbol{\sigma}_i)$ of σ_{ij}: $K_i = \exp(H(\boldsymbol{\sigma}_i))$ and $H(\boldsymbol{\sigma}_i) = -\sum_{j=1}^{N} \sigma_{ij} \log \sigma_{ij}$. The wider the neighborhood is, the smaller precision π_i, the larger size K_i, and the higher entropy $H(\boldsymbol{\sigma}_i)$. In many variants of NE, the user sets a unique perplexity value $K_\star$ that is then used to adjust precision π_i of

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

each HD neighborhood individually. In practice, these methods solve $\log(K_\star) = H(\boldsymbol{\sigma}_i)$ for $1 \leq i \leq N$, in order to have soft neighborhoods with size $K_\star$ around each datum $\boldsymbol{\xi}_i$. Hence, σ_{ij} are (iso-)entropic affinities [17].

Determining bandwidths π_i is achieved by dichotomic search or gradient-based Newton root finding. In the LD embedding space, the coordinates in $\mathbf{x}_i$ and distances d_{ij} are yet to be determined, which makes identifying precisions p_i in the same way as π_i in the HD space not as straightforward. The workaround in (t-)SNE [4, 5] is to use unit precisions ($p_i = 1$ for all i). In contrast, NeRV [18] relies rather on $p_i = \pi_i$. This design comes with no guarantee of optimality and has a side effect on density rendering in the embedding. SNE, t-SNE, PaCMAP [10], and other similar methods that scale distances in HD but not in LD tend to make density uniform within clusters. Knowing $\boldsymbol{\Xi}$ and now also both π_i and p_i, matching HD and LD affinities by minimizing Kulback-Leibler (KL) divergences allows determining the embedding $\mathbf{X}$: $\arg\min_{\mathbf{X}} \sum_{i=1}^{N} \mathrm{KL}(\boldsymbol{\sigma}_i || \mathbf{s}_i)$.

3 Multi-scale SNE with twice adaptive bandwidths

Multi-scale entropic affinities [11] introduce an additional index h to denote scale in definition (1). This leads to single-scale affinities that are now written as

$$\sigma_{hij} = \frac{\exp(-\pi_{hi}\delta_{ij}/2)}{\sum_{k,k\neq i}\exp(-\pi_{hi}\delta_{ik}/2)} \quad \text{and} \quad s_{hij} = \frac{\exp(-p_{hi}d_{ij}/2)}{\sum_{k,k\neq i}\exp(-p_{hi}d_{ik}/2)} \ , \quad (2)$$

where symbols π_{hi} and p_{hi} denote the HD and LD precisions of the ith soft neighborhoods on scale h. If $i = j$, then $\sigma_{hij} = s_{hij} = 0$, just like previously. Compounding those over h leads to multi-scale affinities

$$\sigma_{ij} = \tfrac{1}{L}\sum_{h=L_{\min}}^{L_{\max}} \sigma_{hij} \quad \text{and} \quad s_{ij} = \tfrac{1}{L}\sum_{h=L_{\min}}^{L_{\max}} s_{hij} \ , \quad (3)$$

with $1 \leq L_{\min} \leq h \leq L_{\max} \leq \lfloor \log_2(N) \rfloor$ and $L = L_{\max} - L_{\min} + 1$ is the considered number of scales. Since $\sum_{j=1}^{N} \sigma_{hij} = 1$ for all h and i, $\sum_{j=1}^{N} \sigma_{ij} = 1$ holds for all i, thereby perpetuating the probabilistic interpretation of SNE with soft 'stochastic' neighborhoods. Multi-scale affinities inherit all necessary properties to use KL divergences just as in SNE.

Deploying multi-scale affinities entails determining several precisions π_{hi} and p_{hi} for the neighborhoods around each $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$. In the HD case of π_{hi}, perplexity values that cover the whole range from 1 to nearly N can be selected, without any user intervention, like $K_{h\star} = 2^h$, with h within the same bounds as above. In the LD case of p_{hi}, an educated guess could be $p_{hi} = 2^{-2(h-1)/P}$ [11]. In 2D, this corresponds to the (strong) assumption that $\mathbf{X}$ is uniform and the number K_i of neighbors around $\boldsymbol{\Xi}_i$ grows like a disc area with radius $l_i = p_{hi}^{-1/2}$.

Multi-scale problems can be solved in cascaded (or multi-staged) ways to mitigate the odds of getting stuck in local minima [11]. Precisions $\boldsymbol{\pi}_h = [\pi_{hi}]_{1\leq i\leq N}$ can be determined efficiently using such a cascade:

$$\boldsymbol{\pi}_{h=L_{\max}} \leftarrow \arg\min_{\boldsymbol{\pi}_h | 1/\delta_{ij}} (H(\boldsymbol{\sigma}_{hi}) - K_{L_{\max}\star})^2 \quad (4)$$

$$\text{For } h = L_{\max} - 1 \text{ down to } 1 : \ \boldsymbol{\pi}_h \leftarrow \arg\min_{\boldsymbol{\pi}_h | \boldsymbol{\pi}_{h+1}} (H(\boldsymbol{\sigma}_{hi}) - K_{h\star})^2 \ , \quad (5)$$

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

where $\bar{\delta}_{ij}$ is the average of δ_{ij} and $\min_{u|u_0} f(u)$ minimizes $f(u)$ w.r.t. u starting from initilization $u \leftarrow u_0$. Next, the embedding $\mathbf{X}$ can be determined with

$$\mathbf{X}_{\ell=L_{\max}} \leftarrow \arg\min_{\mathbf{X}_h|\mathrm{PCA}} \sum_{i=1}^{N} \mathrm{KL}(\boldsymbol{\sigma}_i||\mathbf{s}_i) \quad (6)$$

$$\text{For } \ell = L_{\max} - 1 \text{ down to } 1: \; L_{\min} \leftarrow \ell \text{ (to compute (3))} \quad (7)$$

$$\mathbf{X}_{\ell} \leftarrow \arg\min_{\mathbf{X}_\ell|\mathbf{X}_{\ell+1}} \sum_{i=1}^{N} \mathrm{KL}(\boldsymbol{\sigma}_i||\mathbf{s}_i) \; . \quad (8)$$

Instead of having fixed $p_{hi} = 2^{-2(h-1)/P}$ [11], we run (4)–(5) in between (7) and (8), for p_{hi} with $\mathbf{X}$, no longer for π_{hi} with $\boldsymbol{\Xi}$. In this way, we alternate updates of $\mathbf{X}$ and p_{hi} while progressively introducing smaller and smaller scales.

4 Experiments, results, and discussion

In order to assess the proposed multi-scale method, six data sets are embedded in 2D. These are COIL-20 [19] ($N = 1440$), a Swiss roll [20] (N) $= 3072$), (penta-)hierarchical blobs [14] ($N = 4096$), B. Frey's faces [21] ($N = 1965$), MNIST [22] (1/10; $N = 6000$), Rat Brain RNA [23] (1/4; $N = 5956$). As to DR methods, PCA [15] and t-SNE [5] with $K_\star = 32$ serve as benchmarks. In addition to embeddings, the curves $R_{\mathrm{NX}}(K) = ((\frac{N-1}{KN} \sum_i |\nu_i^K \cap n_i^K|) - K)/(N - 1 + K)$ [11] are reported, where $1 \leq K \leq N$ is a neighborhood size and ν_i^K and n_i^K are the K-ary (non-stochastic) neighborhoods of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, 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). For t-SNE, extremes among five repetitions (PCA ϵ-scaling) are shown with a paler sleeve in addition to the average [14]; the other methods are deterministic. The area under the curves (AUCs) compounds local and global into a single scalar score.

The embeddings by PCA, t-SNE, Ms.SNE with fixed and adaptive bandwidths ('fBW' & 'aBW') look visually different in Fig. 1. PCA best preserves data global structure (large K in Fig. 2). As a linear projection, PCA is supposed to also preserve data density to some extent. All NE methods are good at preserving local and/or mid-size structure (low K in Fig. 2). With strong repulsive forces, t-SNE magnifies cluster gaps, unlike Ms.SNE, often leading to losses in the global structure. Both t-SNE and Ms.SNE fBW tend to make densities uniform across clusters. In contrast, Ms.SNE aBW reproduces density differences similar to those of PCA, except for the Swiss roll, sampled uniformly. See, e.g., circular objects in COIL-20 having smaller clusters or zeros and sixes in MNIST being more widely spread. Quantitatively, this comes at no cost compared to Ms.SNE fBW: a minor loss locally is compensated by better global structure preservation, with Ms.SNE even outperforming PCA in that respect.

5 Conclusion

Multi-scale SNE [11] can be fitted with entropic affinities having adaptive bandwidths in HD *and* LD. This additional adjustability allows Ms.SNE 'aBW' to

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

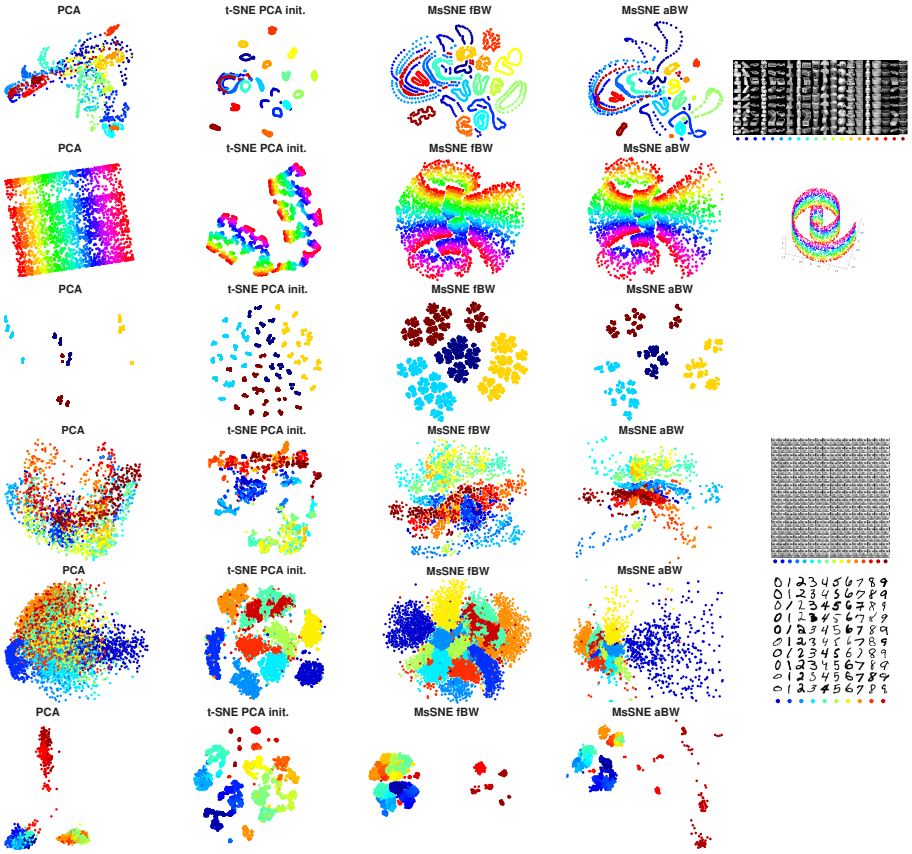

Fig. 1: Embeddings of the six data sets. From top to bottom: COIL-20, Swiss roll, hierarchical blobs, Frey faces, MNIST, rat brain RNA. For Ms.SNE, 'fBW' and 'aBW' denote fixed and adaptive bandwidths, respectively.

more faithfully render data density variations, unlike other methods of neighbor embedding; this may be of interest in some applicative domains. Performance in terms of neighborhood preservation stays on a comparable level, if not increasing.

References

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

[2] C. de Bodt et al. Low-dimensional embeddings of high-dimensional data. https://arxiv.org/abs/2508.15929, 2025.

[3] D. Kobak and Ph. Berens. The art of using t-SNE for single-cell transcriptomics. *Nature Communications*, 10(1):5416, 2019.

[4] G.E. Hinton and S. Roweis. Stochastic neighbor embedding. In S. Becker, S. Thrun, and K. Obermayer, editors, *NIPS*, volume 15. MIT Press, 2002.

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

ESANN 2026 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 22-24 April 2026, i6doc.com publ., ISBN 9782875870964. Available from http://www.i6doc.com/en/.

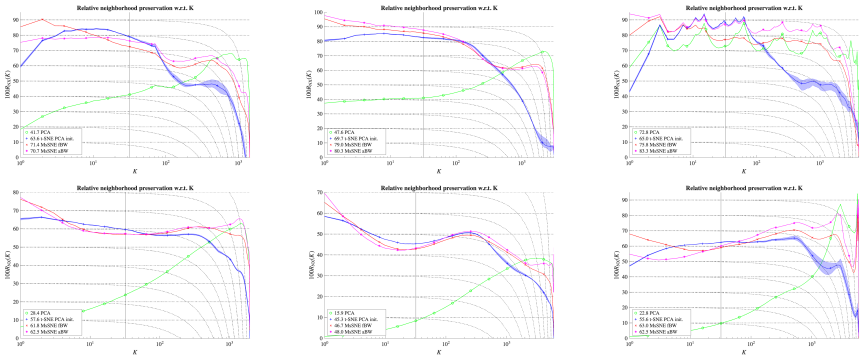

Fig. 2: Neighborhood preservation $R_{\mathrm{NX}}(K)$ for the 6 data sets. In reading order: COIL-20, Swiss roll, hierarchical blobs, Frey faces, MNIST, rat brain RNA.

[6] L. McInnes, J. Healy, and J. Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. http://arxiv.org/abs/1802.03426, 2018.

[7] J. A. Lee and M. Verleysen. Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. In *Proc. ICCS 2011*, Singapore, 2011.

[8] L. van der Maaten. Accelerating t-SNE using tree-based algorithms. *JMLR*, 15(1):3221–3245, jan 2014.

[9] G.C. Linderman et al. Fast interpolation-based t-SNE for improved visualization of single-cell rna-seq data. *Nature Methods*, 16(3):243–245, 2019.

[10] Y. Wang, H. Huang, C. Rudin, and Y. Shaposhnik. Understanding how dimension reduction tools work: An empirical approach to deciphering t-SNE, UMAP, TriMap, and PaCMAP for data visualization. *JMLR*, 22(201):1–73, 2021.

[11] 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.

[12] D. Kobak and G.C. Linderman. Initialization is critical for preserving global data structure in both t-SNE and UMAP. *Nature Biotechnology*, 39(2):156–157, 2021.

[13] N. Kury, D. Kobak, and S. Damrich. DREAMS: Preserving both local and global structure in dimensionality reduction. 2025.

[14] J.A. Lee et al. Improving on early exaggeration in t-SNE: Early hierarchization better preserves global structure. *Neurocomputing*, 660:131882, 2026.

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

[16] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *NIPS*, volume 14. MIT Press, 2002.

[17] M. Vladymyrov and M. Carreira-Perpinan. Entropic affinities: Properties and efficient numerical computation. In *Proc. 30th ICML*, volume 28, pages 477–485, 2013.

[18] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski. Information retrieval perspective to nonlinear dimensionality reduction for data visualization. *JMLR*, 11(13):451–490, 2010.

[19] S.A. Nene, S.K. Nayar, and H. Murase. Columbia Object Image Library (COIL-20). `http://www.cs.columbia.edu/CAVE/software/softlib/coil-20.html`, 1996.

[20] 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.

[21] B.J. Frey, P. Dayan, and R.M. Neal. Brendan Frey's Faces Dataset. `https://www.cs.toronto.edu/$\sim$roweis/data/frey\_rawface.mat`, 1998.

[22] Y. LeCun, C. Cortes, and Ch.J.C. Burges. The MNIST database of handwritten digits. `http://yann.lecun.com/exdb/mnist/`, 1998.

[23] B. Tasic et al. Shared and distinct transcriptomic cell types across neocortical areas. *Nature*, 563(7729):72–78, 2018.