Natively Interpretable *t*-SNE

Edouard Couplet[1], Pierre Lambert[1], Michel Verleysen[1],[0000−0003−4366−6155], Dounia Mulders[1], John A. Lee[1], and Cyril de Bodt[1][0000−−0003−2347−1756]

1. UCLouvain - ICTEAM/ELEN, Belgium
{edouard.couplet,pierre.h.lambert,john.lee,cyril.debodt}@uclouvain.be

Abstract. The visual exploration of high-dimensional (HD) data has gained popularity through the use of dimensionality reduction (DR) techniques such as *t*-SNE and UMAP. However, the interpretability of low-dimensional (LD) embeddings produced by these nonlinear methods remains a challenge. Conversely, linear methods such as PCA are natively interpretable but fall behind regarding DR quality. To circumvent this trade-off, post-hoc interpretability methods have been introduced, where simpler models are used *a posteriori* to explain LD positions in terms of HD features. While these approaches can provide explanations for nonlinear DR methods without compromising DR quality, their downside is that they rely on approximations of the original LD embeddings which can lead to misinterpretations. In this paper, we propose a novel solution to the trade-off between DR quality and interpretability: a natively interpretable version of *t*-SNE. The key idea is to express the coordinates of each LD point as individual linear combinations of HD features and use regularization to promote local coherence of the various linear combination weights across the embedding. Experimental results demonstrate the effectiveness of our method in preserving HD structures while providing LD embeddings that are interpretable by design.

Keywords: *t*-SNE · Interpretability · Nonlinear dimensionality reduction · Parametric mappings · Data visualization · Explainability.

1 Introduction

Visual exploration of high-dimensional (HD) data is nowadays popularly conducted thanks to dimensionality reduction (DR) methods such as *t*-SNE [25] and UMAP [26]. Faithfulness of low-dimensional (LD) embeddings with respect to HD coordinates is recognized as being related to the reproduction of HD neighborhoods in the LD space [35]. Several paradigms exist to create LD representations of HD coordinates [22]; for instance, principal component analysis (PCA) [15] and classical metric multidimensional scaling (MDS) [7] compute linear projections of HD data, while methods like stress-based multidimensional scaling [31] and locally linear embedding [30] formalize nonlinear mappings through weighted distance preservation and HD affinity matrices. More recent methods of neighbor embedding (NE) [14], however, often outperform other paradigms

stunningly in tasks of HD data visualization thanks to their shift invariance property [19]. This property alleviates the norm concentration phenomenon [13] that heavily affects DR quality of more traditional approaches [23] [32]. A large number of NE techniques were hence introduced, including t-SNE [25], UMAP [26], perplexity-relieved extensions [20] [3], methods optimizing arbitrary divergences [9], heavy-tailed similarities [38], hybrid NE-MDS schemes [17], hierarchical approaches [29], supervised variants [12] [4], scalable accelerations [24] [6] [33] [39], missing data treatment [5], fast optimizations [36] [28] [10], etc.

Although yielding particularly relevant LD representations of HD samples, NE algorithms are genuinely nonlinear, hindering the interpretability of the embeddings; it is indeed usually difficult or even impossible to relate the computed LD axes to the HD dimensions [37]. This challenge arises as soon as one employs a nonlinear DR (NLDR) method that is not natively interpretable; while linear projections naturally provide weights that enable interpretation of LD coordinates in terms of HD features, nonlinear transformations of HD dimensions inevitably result in less intuitive LD components. On the other hand, methods of NLDR are typically largely better than linear mappings at reproducing HD structures in LD embeddings [23]; some sort of trade-off between (NL)DR quality and interpretability is therefore unavoidable.

Interpretability in DR is highly sought-after as it may lead to meaningful insights and allows for informed analysis of high-dimensional data representations. This is why methods emerged to analyze NLDR results a posteriori [2] [18] [1], commonly referred to as *post-hoc* interpretability or explanation: given a set of LD coordinates determined thanks to NLDR, one relies on several simpler, directly interpretable models such as decision trees or linear regressions, to locally explain the LD positions in terms of HD coordinates. The convenience of such an approach stems from its decoupling of the above trade-off; the responsibility of ensuring faithfulness of the LD embedding with respect to the HD data is left to the method of NLDR alone, while interpretability is sought afterwards. The downside is, however, immediate, as one only interprets an approximation of the NLDR result, instead of the actual LD positions directly. Accuracy of explanations is obviously degraded for LD points with large approximation errors. Increasing the number of local explanations, each with a reduced span in the LD space, bounds the amplitude of local approximation errors, but as contiguous explanations do not necessarily agree, their interpretation as a whole becomes difficult, which conflicts with the initial goal of interpreting the LD embedding. This other kind of trade-off, balancing accuracy and essentiality (i.e., limited number) of local explanations, brings an insightful view on *post-hoc* approaches: given a LD embedding with fixed NLDR quality as input, an additional approximation step is conducted, in which accuracy of local explanations is balanced with their essentiality.

One may then wonder whether users should instead be able to directly trade-off NLDR quality with essentiality of local explanations, hence skipping the additional approximation stage and the concept of accuracy of local explanations with respect to approximation errors. Such an observation calls for NLDR meth-

ods that are natively interpretable; coherence (leading to essentiality) of local explanations across the LD embedding should be explicitly accounted for by the NLDR method itself when tuning the LD coordinates to maximize DR quality. When facing real-world, complex data visualization tasks, maximizing DR quality usually provides LD representations with highly varying local explanations, hence less coherent across the LD embedding. On the other hand, enforcing coherence of local explanations, favoring their essentiality, hurts DR quality. A natively interpretable method of NLDR would then enable users to consciously balance between the desired DR quality and explanation coherence.

This paper therefore introduces a natively interpretable version of a perplexity-free extension of t-SNE, known as multi-scale t-SNE (Ms t-SNE) [6]. Multi-scale t-SNE is a non-parametric NE algorithm that efficiently preserves both local and global HD data structures in LD embeddings. When seeking to develop a natively interpretable version of multi-scale t-SNE, one can first notice that any non-parametric method of NLDR may become parametric using any explicit mapping from the HD to LD space [8]: instead of optimizing the LD embedding with respect to the LD coordinates, one simply needs to optimize it with respect to the mapping weights. In the context of interpretability, in particular, a directly interpretable mapping may be chosen, such as a linear regression for instance; such a parametric version of Ms t-SNE is denoted by Ms t-SNE $[W]$, with W the weights of the linear mapping from the HD to LD space. Being easily interpretable, Ms t-SNE $[W]$ further improves HD neighborhood preservation in the LD embedding compared to DR methods that also compute a linear projection from the HD to LD space, like PCA. However, genuine, non-parametric Ms t-SNE largely outperforms Ms t-SNE $[W]$, especially regarding the preservation of local HD structures, as demonstrated in the experiments of this paper. Such a behavior is natural; Ms t-SNE $[W]$ is in fact absolutist at the left end of the 'coherence' - 'DR quality' spectrum: defining a single W over the entire LD embedding forces perfect coherence of explanations across the whole LD space, at the expense of DR quality, as sketched above.

In order to span the full 'coherence' - 'DR quality' spectrum, we present Ms t-SNE $[W_i]$, as a natively interpretable version of Ms t-SNE. For each HD data point, indexed by i, linear regression weights W_i are computed to determine the associated LD coordinates. Coherence of W_i for i across the LD embedding is tailored by a dedicated coherence term, added to the cost function of Ms t-SNE, to explicitly encode the trade-off between DR quality and explanation coherence. The hyper-parameter binding these two terms in the cost function of Ms t-SNE $[W_i]$, enables users to explore the 'coherence' - 'DR quality' spectrum from end to end. A L1 regularization is further considered for weights W_i, to favor sparse explanations. To the best of our knowledge, Ms t-SNE $[W_i]$ is the first natively interpretable method of NE.

An illustrative example of the proposed Ms t-SNE $[W_i]$ applied on the `adult mouse cortex` dataset [34] is depicted in Fig. 1. One can first observe that the obtained 2D embedding of cellular biology data is consistent in terms of displayed HD structures: non-neurons are separated from neurons; inhibitory

neural cells can be distinguished from excitatory ones, and similar colors are generally grouped together. This suggests that Ms t-SNE $[W_i]$ can achieve high DR quality. It is noteworthy that the considered data set is preprocessed as in [16]; in particular, a PCA is first performed and only the first 50 principal components are provided to Ms t-SNE $[W_i]$. Interpretations, in terms of linear weights W_i, are displayed for 3 randomly selected data points, along with the associated HD features, which are principal components. As the 2D embedding gets initialized with the first two principal components of the data set, PC1 (resp. PC2) is always the most important feature according to $W_\star^{(x)}$ (resp. $W_\star^{(y)}$), as expected. The remaining most important HD features vary from one selected data point to another, highlighting that the delivered explanations are adapted according to the actual position of the selected LD point and are hence not constant across the entire LD embedding.

This paper is structured as follows: Section 2 first reviews t-SNE and multi-scale t-SNE, in their original, non-parametric versions, as well as convenient DR quality criteria. Section 3 then details our proposed Ms t-SNE $[W_i]$ algorithm, while Section 4 reports experimental results assessing its behavior and quantifying its sensitivity with respect to the coherence and L1 regularization terms in its cost function. Section 5 finally draws conclusions and outlines further works.

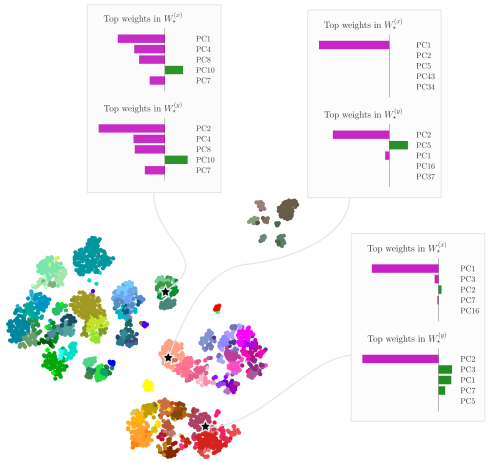

Fig. 1. 2D embedding of the `adult mouse cortex` dataset [34] as produced by our interpretable Ms t-SNE $[W_i]$. Cluster assignments and colors are taken from the original publication [34]. Warm colors correspond to inhibitory neurons, cold colors to excitatory neurons, and brown/greyish colors to non-neural cells. Explanations are provided for randomly selected points, depicted as stars. For each of these points, we only display the top five weights (in absolute value) of the local linear mapping $W_\star$ from HD features to the LD coordinates, and the associated HD features. $W_\star^{(x)}$ denotes the weights that map to the x-coordinate; $W_\star^{(y)}$ the weights that map to the y-coordinate. Negative and positive weights are represented with magenta bars to the left and green bars to the right, respectively. The length of these bars is proportional to the absolute value of the weight, i.e., the importance of the corresponding feature in the explanation.

2 t-SNE, Multi-scale t-SNE, and DR quality criteria

Given N points $\{\boldsymbol{\xi}_i\}_{i=1}^N$ in M dimensions, DR aims at representing them as $\{\mathbf{x}_i\}_{i=1}^N$ in P dimensions, with $P \leq M$. Distance between $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ (resp. $\mathbf{x}_i$ and $\mathbf{x}_j$) is denoted by δ_{ij} (resp. d_{ij}). Instead of preserving distances, the main idea of t-SNE is to preserve neighborhoods of points, modeled through the definition of pairwise similarities [25]: for $i \in \mathcal{I} := \{1, \ldots, N\}$ and $j \in \mathcal{I} \backslash \{i\}$,

$$\sigma_{ij*} = \frac{\exp\left(-\pi_{i*}\delta_{ij}^2/2\right)}{\sum_{k\in\mathcal{I}\backslash\{i\}} \exp\left(-\pi_{i*}\delta_{ik}^2/2\right)}, \quad t_{ij} = \frac{\left(1+d_{ij}^2\right)^{-1}}{\sum_{k\in\mathcal{I},l\in\mathcal{I}\backslash\{k\}} \left(1+d_{kl}^2\right)^{-1}} \tag{1}$$

where precision π_{i*} is determined to secure a user-identified perplexity K_*, related to the granularity of HD structures one wishes to preserve [14].

The LD coordinates are then adjusted through gradient-based minimization of the t-SNE cost function:

$$\{\mathbf{x}_i\}_{i=1}^N : \min_{\mathbf{x}_i, i\in\mathcal{I}} C_* = \sum\nolimits_{i\in\mathcal{I}, j\in\mathcal{I}\backslash\{i\}} \tau_{ij*} \log\left(\tau_{ij*}\,/\,t_{ij}\right) \tag{2}$$

where $\tau_{ij*} = \left(\sigma_{ij*} + \sigma_{ji*}\right)/\left(2N\right)$ are symetrized and normalized HD similarities.

Defining HD similarities that rely on a single perplexity value K_* limits the NE method capability to capture both local and global HD structures in the LD embedding [20]. Multi-scale t-SNE hence computes HD similarities with exponentially increasing perplexities [6]: for $h = 1, \ldots, H = \lfloor \log_2\left(N/2\right) \rceil$, with H the number of scales and $\lfloor\cdot\rceil$ denoting rounding,

$$\sigma_{ijh} = \frac{\exp\left(-\pi_{ih}\delta_{ij}^2/2\right)}{\sum_{k\in\mathcal{I}\backslash\{i\}} \exp\left(-\pi_{ih}\delta_{ik}^2/2\right)} \tag{3}$$

with π_{ih} set as previously thanks to perplexities $K_h = 2^h$ [20], targeting local to global data scales. Averages across scales then enables retrieving both local and global properties of HD data:

$$\sigma_{ij} = H^{-1} \sum\nolimits_{h=1}^{H} \sigma_{ijh}, \;\; \tau_{ij} = \left(\sigma_{ij} + \sigma_{ji}\right)/\left(2N\right) \quad . \tag{4}$$

Multi-scale t-SNE minimizes $C = \sum_{i\in\mathcal{I}} C_i$, with $C_i = -\sum_{j\in\mathcal{I}\backslash\{i\}} \tau_{ij} \log t_{ij}$ [3].

Besides tuning LD embeddings, assessing their quality typically amounts to evaluating their reproduction of HD neighborhoods [11] [27] [35]. Established criteria [23] quantify this preservation, by first computing the sets ν_i^K and n_i^K of K nearest neighbors of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, for $i \in \mathcal{I}$. Their average agreement is

$$Q_{\mathrm{NX}}\left(K\right) = \left(NK\right)^{-1} \sum\nolimits_{i\in\mathcal{I}} \left|\nu_i^K \cap n_i^K\right| \in [0,1] \quad . \tag{5}$$

As $\mathbb{E}\left[Q_{\mathrm{NX}}\left(K\right)\right] = K/\left(N-1\right)$ for random LD coordinates, rescaling $Q_{\mathrm{NX}}\left(K\right)$,

$$R_{\mathrm{NX}}\left(K\right) = \left(\left(N-1\right) Q_{\mathrm{NX}}\left(K\right) - K\right)/\left(N-1-K\right) \quad , \tag{6}$$

eases performance comparison for distinct K. Log-scale for K is employed to depict $R_{\mathrm{NX}}\left(K\right)$, to emphasize smaller neighborhoods [21]. The area AUC under the obtained curve is proportional to DR quality, assessed at all scales, with a particular attention on smaller ones [20].

3 Natively interpretable Multi-scale *t*-SNE

Following our approach to interpretable neighbor embedding, we focus on the key ideas and subsequent design choices that underpin the mathematical formulation of our model. We aim at enabling interpretability of a NLDR method, specifically Ms. t-SNE, without relying on post-hoc explanation techniques. In this work, any method of DR qualifies as *interpretable* if it allows LD embedding coordinates to be expressed in terms of HD features. Additionally, we require this explicit relationship between HD features and LD coordinates to be easily understandable and reasoned about by fellow humans [1].

Notice that linear models, unlike nonlinear ones, serve this purpose very well: they are interpretable by design and the weights of the model are naturally simple explanations indeed for the model output in terms of its inputs. In the context of DR, PCA stands out as the most widely used linear method. For 2-D visualizations, the data is projected on the first two principal components such that the variance is maximally preserved. Coordinates in the LD embedding are thus linear combinations of HD features and the weights of these linear combinations, which can be viewed as *parameters*, are what makes PCA natively interpretable. A first natural idea for achieving our objective would then be to introduce *parameters* in the otherwise non-parametric and non-natively interpretable Ms t-SNE. Bunte et al. showed in [8] that any non-parametric DR method can be made *parametric* by defining a parametric mapping from the HD space to the LD space and optimize for the parameters of this mapping instead of directly optimizing for the LD coordinates. Hence, following our previous idea, we can express LD coordinates as a linear combination of the HD features:

$$\mathbf{x}_i = W^T \boldsymbol{\xi}_i \quad \forall i \in \mathcal{I} \tag{7}$$

and optimize the standard cost function of Ms t-SNE with respect to W. The weight matrix W can be initialized at random, or conveniently, with easy-to-compute PCA weights. While optimizing the cost function of Ms t-SNE in this setting would likely produce better result than PCA in terms of neighborhood preservation, this first parametric method of NE remains a simple linear projection. The global weight matrix W, uniformly applicable to all points, imposes excessive constraints that completely undermine the local nature of Ms t-SNE. To embrace local aspects of Ms t-SNE while maintaining interpretability, we propose to define an individual weight matrix for each sample:

$$\mathbf{x}_i = W_i^T \boldsymbol{\xi}_i \quad \forall i \in \mathcal{I} \ . \tag{8}$$

As before, we optimize the standard cost function of Ms. t-SNE:

$$C = \sum_{i \in \mathcal{I}} C_i \ . \tag{9}$$

[1] This is a stronger requirement than *interpretability* alone, and it is rather referred to as *explainability* in the literature. In this work, however, we do not sharply distinguish between the two concepts. We use a collection of local linear models; our method is thus *interpretable* by design, and the different weights can be understood as *explanations* for the LD embedding.

This simple, yet key idea of individual linear projection is our main contribution. It allows us to construct an interpretable and efficient version of Ms t-SNE. This benefit comes at a cost, though, and a new potential difficulty arises: while a single weight matrix led to an *under-parametrized* optimization problem, individual weight matrices yield an *over-parametrized* problem. This can jeopardize convergence in the numerical optimization problem and cause interpretability issues: with so many degrees of freedom, the model may provide arbitrary, irrelevant, or incomprehensible explanations. We address these issues and avoid ill-posed optimisation problems with regularization.

On the one hand, to ensure sensible interpretation of the visualization obtained through Ms t-SNE, we would like to have *coherent* explanations: two points that are similar in the LD space should have similar explanations. In mathematical terms, if two points $\mathbf{x}_i$ and $\mathbf{x}_j$ are such that t_{ij} is big, then the difference between W_i and W_j should be small. We enforce this by adding a first regularization term to the standard cost function of Ms t-SNE:

$$C = \sum_{i \in \mathcal{I}} \left(C_i + \alpha \cdot \sum_{j \in \mathcal{I} \setminus \{i\}} t_{ij} ||W_i - W_j||_F \right) \ , \tag{10}$$

where $|| \cdot ||_F$ is the Frobenius norm and α is a weighting factor that allows users to tune the strength of the regularization.

On the other hand, we may also desire *simple* explanations. Here, *simple* means that LD coordinates can be explained based only on a few significant HD features. In mathematical terms, this corresponds to sparse explanations, characterized by weight matrices W_i with a low L1 norm. We enforce this by adding a second regularization term to the cost function of Ms t-SNE:

$$C = \sum_{i \in \mathcal{I}} \left(C_i + \alpha \cdot \sum_{j \in \mathcal{I} \setminus \{i\}} t_{ij} ||W_i - W_j||_F + \beta \cdot ||W_i||_1 \right) \ , \tag{11}$$

where β is a weighting factor that controls the sparsity of the explanations.

Equation (11) provides the final formulation of our natively interpretable Ms t-SNE. The user can adjust the hyper-parameters α and β to obtain different embeddings with different characteristics and hopefully gain more insights through multiple views of the data. Once the model is fitted, explanations for each point in the embedding can be accessed in real time, with no further computations needed. One may finally wonder about the essentiality of the local explanations. To reason about this concept, we introduce the coherence coefficient:

$$\kappa_{ij} = ||W_i - W_j||_F \ . \tag{12}$$

Given some point i, if $\kappa_{ij} \approx 0$ for all points j in a consistent region around $\mathbf{x}_i$, we consider that the explanations within this region are locally coherent and we call the region a *coherence region*. Having larger coherence regions with fewer differences between them is synonymous with explanation essentiality.

4 Experiments

We now conduct a series of three experiments to assess the quality of neighborhood preservation and gauge the effective model interpretability under various model configurations. Furthermore, the goal of the first experience in Section 4.1 is to provide empirical validation for the reasoning presented in Section 3 and convince the reader that the design choices are sound. In the second and third experiences, we evaluate the influence of *coherence* hyper-parameter α and *sparsity* hyper-parameter β, in Sections 4.2 and 4.3, respectively.

We conduct all three experiments on the `adult mouse cortex` dataset from [34]. This dataset contains gene expression levels for 23822 cells from adult mouse cortex. Each data point can be classified into one of three main classes (and many other subclasses): excitatory neurons, inhibitory neurons, and non-neural cells. The data was preprocessed as in [16], including a normalization step, a feature selection step, a log transformation, and a PCA retaining only the 50 first principal components (mainly for computational efficiency). Additionally, we randomly subsample the data to obtain a final dataset of 3000 points. We implement our interpretable Ms t-SNE model following equations from Sections 2 and 3, and we take advantage of `pytorch` for automatic gradient computation. For all experiments, we use the `Adam` optimizer with a learning rate of 0.01 and carry out optimization for 1000 iterations. The weight matrices are all initialized with PCA weights. If not stated otherwise, the default values for α and β are 10^6 and 10^2, respectively.

4.1 Balancing DR quality and interpretability

In this first experiment, we compare the different methods described throughout Section 3: PCA, Ms t-SNE $[W]$ with a unique weight matrix, Ms t-SNE $[W_i]$ with one weight matrix per sample, and the original, non-parametric version of Ms t-SNE $[-]$. These are all key components of our rationale and our goal is to verify that our theoretical arguments hold in practice. If it is the case, then we can be more confident in the effectiveness of our proposed approach. Embeddings generated by the various methods are displayed in Fig. 2, and $R_{\mathrm{NX}}(K)$ curves are drawn in Fig. 3.

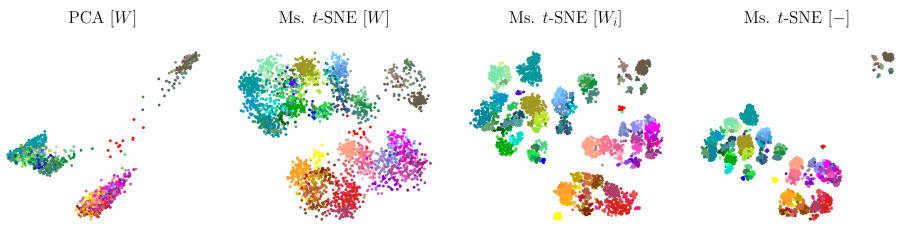

Fig. 2. 2-D embeddings of the `adult mouse cortex` dataset. From left to right : PCA, Ms t-SNE with a unique weight matrix, Ms t-SNE with one weight matrix per sample, standard Ms t-SNE. Cluster assignments and colors are taken from [34].

PCA demonstrates the best global neighborhood preservation. It is clear both from the dominating curve in the rightmost part of the $R_{\mathrm{NX}}(K)$ plot, and from the visualization of the embedding in which excitatory neurons, inhibitory neurons, and non-neural cells are well separated. While the unique linear mapping makes PCA natively interpretable and provides straightforward explanations, it strongly hinders the preservation of local neighborhoods and is responsible for a phenomenon of overlapping points [22]. The idea behind Ms t-SNE $[W]$ was to take inspiration from the linear mapping of PCA, for its straighforward interpretability, but to optimize a standard Ms t-SNE cost function in the hope of achieving better local DR quality. Ms t-SNE [W] mitigates the overlap phenomenon to some extent and it better preserves mid-size neighborhoods, around $K = 10^2$ on the $R_{\mathrm{NX}}(K)$ plot. Nevertheless, as with PCA, its unique linear mapping appears to be the limiting factor for preserving even more local neighborhoods.

The next logical step was thus to put aside this limiting unique linear mapping, and introduce an individual weight matrix per sample. This key aspect allows Ms t-SNE $[W_i]$ to show drastic improvement in local neighborhood preservation, without significant loss in global neighborhood preservation. It retains native interpretability, though the pointwise explanations may not be as straightforward as a unique and global explanation. Finally, as expected, the original version of Ms t-SNE performs best for overall neighborhood preservation, but it is not natively interpretable. All these results back our rationale up and they validate our design choices empirically. Moreover, we observe that interpretable Ms t-SNE $[W_i]$ performs almost on par with the standard, non-natively intepretable Ms t-SNE in terms of $R_{\mathrm{NX}}(K)$. This motivates further experiments to understand in which conditions this happens and, in particular, how the choice of α and β influences performance and explanations.

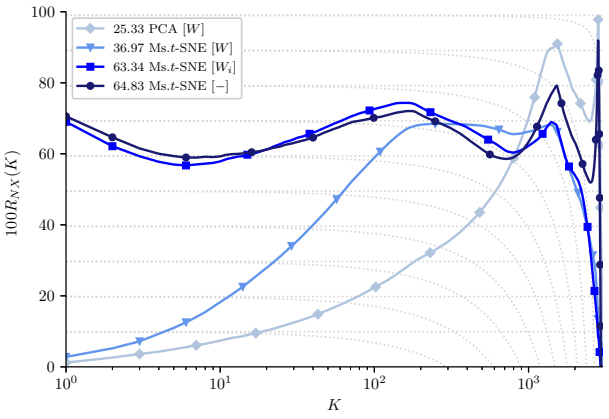

Fig. 3. For each compared method, the $R_{\mathrm{NX}}(K)$ curve quantifies DR quality in terms of average K-ary neighborhoods agreement in HD and LD. The higher $R_{\mathrm{NX}}(K)$, the better. The AUCs stand in the legend in front of each method name.

4.2 Balancing DR quality and explanation essentiality

In this second experiment, we look at the effect of the *coherence* hyper-parameter α on the embedding, both in terms of DR quality and explanations. We fix β to 10^2 and test the following increasing values for α: $\{10^6, 10^7, 10^8, 10^9\}$. All other hyper-parameters are held constant. Quality curves are drawn in Fig. 4, while embeddings and explanations are displayed in Fig. 5. All explanations are provided for a randomly selected data point $\star$. In the left column, we color every other point j in the embedding according to the value of $\kappa_{\star j}$ (normalized within the interval [0,1]). The coherence region of $\mathbf{x}_\star$ is formed by the points colored in brown. In the middle column, we display the top 5 weights along each axis; these are the weights in $W_\star^{(x)}$ (resp. $W_\star^{(y)}$) with maximum absolute value. For each axis, we also select one of the top weights and its associated HD feature, and we represent in the right column how they are distributed in the embedding.

The results illustrate well the trade-off between DR quality and explanation essentiality, and how users could decide where they want to be across the spectrum by tuning α. As we increase α, we notice a significant decrease of the $R_{\text{NX}}(K)$ for low values of K indicating a worse preservation of local neighborhoods. Concurrently, we observe on the 2D visualizations larger coherence regions, leading to essentiality, with an embedding shape evolving towards what Ms t-SNE[W] and PCA would have yielded. This is expected, as a large α enforces each W_i to converge to a similar and unique value W acting as a global explanation for the embedding. For smaller values of α, different HD features are responsible for explaining local structures in different parts of the embedding. As an example, in the case $\alpha = 10^6$, we see that PC8 and PC10 are almost only important in the cluster around point $x_\star$. Notice, however, that PC1 and PC2 remain the most important HD features across all values of α, highlighting the fact that our method can capture a hierarchy of explanations.

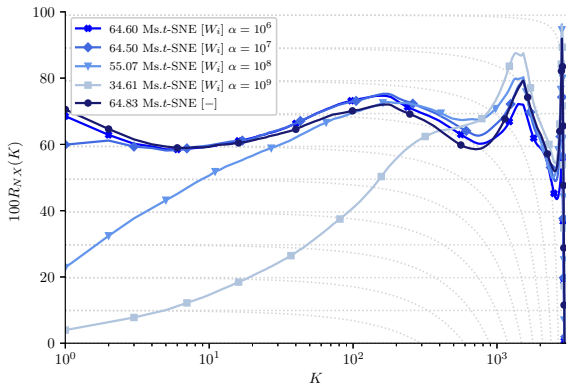

Fig. 4. Quality curves for Ms t-SNE[W_i] with different values of α. Each $R_{\text{NX}}(K)$ curve quantifies DR quality in terms of average K-ary neighborhoods agreement in HD and LD. The AUCs stand in the legend in front of each model configuration.

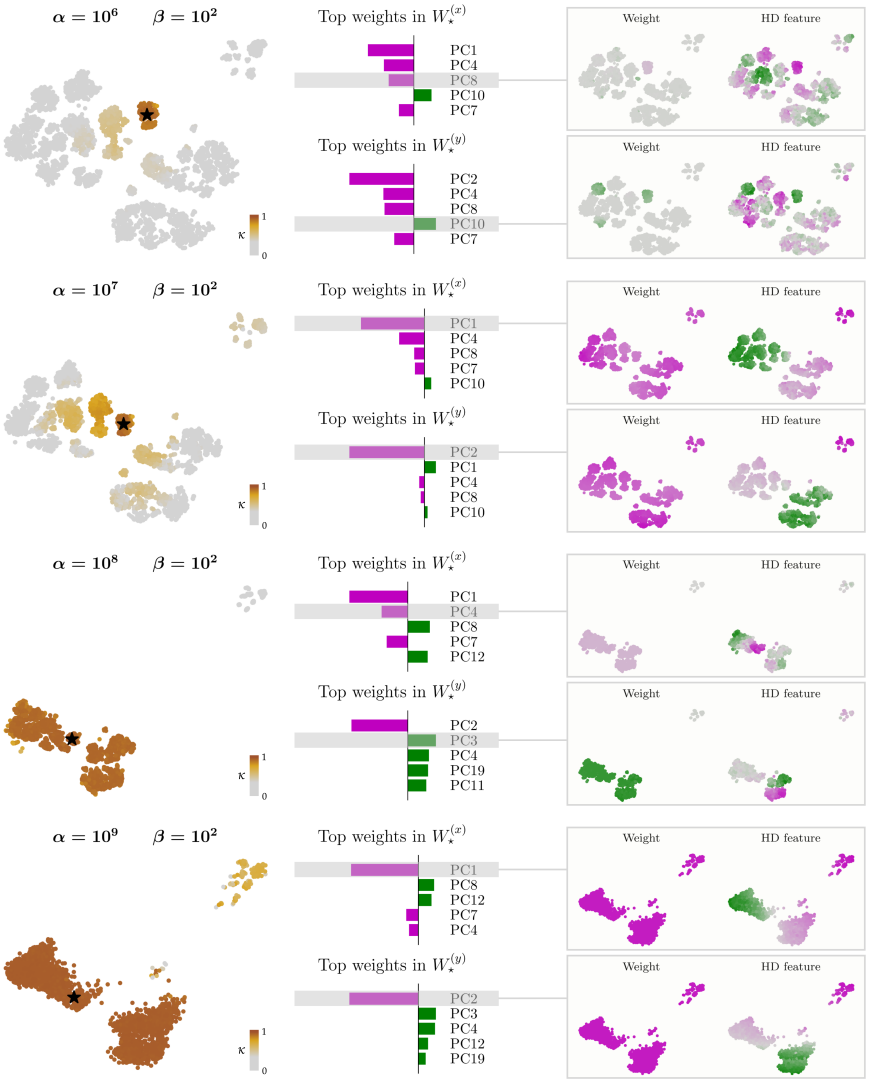

Fig. 5. Embeddings of the `adult mouse cortex` dataset [34] as produced by our interpretable Ms t-SNE $[W_i]$ for various values of the *coherence* hyper-parameter α, and explanations for a selected point $\star$. In the left column, we color the coherence region of $\mathbf{x}_\star$ in brown. In the middle column, we display the top five explanations along each axis, and the corresponding HD features. In the right column, for each axis, we visualize how a chosen feature and its corresponding weight are distributed in the embedding. These small visualizations should help understand in which region of the embedding a given HD feature is the most important to explain the position of LD points. Magenta corresponds to negative values and green to positive values.

4.3 Balancing DR quality and explanation simplicity

In this third experiment, we assess the effect of the *sparsity* hyper-parameter β on the embedding, both in terms of DR quality and explanations. We fix α to 10^6 and test the following increasing values for β: $\{10^1, 10^2, 10^3, 10^4\}$. All other hyper-parameters are held constant. Quality curves are drawn in Fig. 6, while embeddings and explanations are displayed in Fig. 7. The results illustrate an additional trade-off between DR quality and explanation sparsity/simplicity. As we increase β, we notice a significant decrease of the $R_{\text{NX}}(K)$ for low values of K. This is expected as less freedom for adjusting W_i hinders the ability to preserve local neighborhoods. On the other hand, we note that fewer weights are non-zero, simplifying the explanations.

Another behavior can be highlighted: for larger values of β, we also observe a decrease in $R_{\text{NX}}(K)$ for high values of K and a small increase in mid-range values of K, corresponding roughly to the size of the clusters. This translates into a 'leveling' effect, whereby the mean/global explanations get somehow substracted from local explanations, and the focus is shifted towards cluster specificities. This is striking for the case $\beta = 10^4$ where the most important feature for explaining both x and y LD coordinates is PC3, whereas it was only the second most important in all other configuration, behind the globally important PC1 and PC2. An alternative way to think about this effect is that for reasonable values of α, the model has enough flexibility to produce local explanations, although a high β makes it focus on very few features, which limits its capacity to accurately preserve local neighborhoods. The best compromise for DR quality is thus to focus on mid-size neighborhoods (i.e., clusters), and explain these with local information. While these model configurations do not lead to the best DR quality, they may still be worth exploring for insights about individual cluster structures.

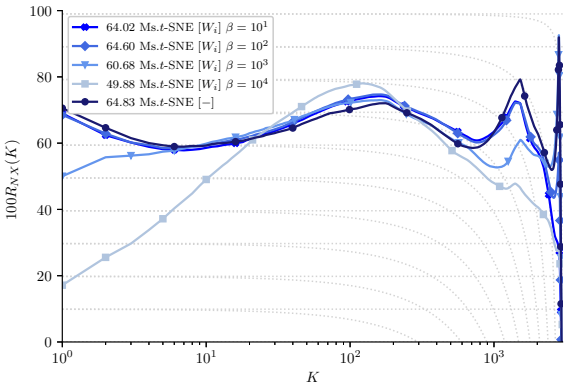

Fig. 6. Quality curves for Ms t-SNE[W_i] with different values of β. Each $R_{\text{NX}}(K)$ curve quantifies DR quality in terms of average K-ary neighborhoods agreement in HD and LD. The AUCs stand in the legend in front of each model configuration.

Fig. 7. Embeddings of the `adult mouse cortex` dataset [34] as produced by our interpretable Ms t-SNE $[W_i]$ for various values of the *sparsity* hyper-parameter β, and explanations for a selected point $\star$. In the left column, we color the coherence region of $\mathbf{x}_\star$ in brown. This region may consist in several non-contiguous smaller parts when groups of data points have similar weights for a set of HD features, but very different values for those same features. In the middle column, we display the top five explanations along each axis, and the corresponding HD features. In the right column, for each axis, we visualize how a chosen feature and its corresponding weight are distributed in the embedding. Magenta corresponds to negative values and green to positive values.

5 Conclusion

In this work, we tackle the challenge of interpretability in methods of nonlinear dimensionality reduction for visualization of high-dimensional data. We introduce a novel, natively interpretable version of multi-scale t-SNE. This perplexity-free neighbor embedding method is non-parametric in its original formulation, but the key to our approach is precisely to incorporate a linear parametric mapping from each HD data point to its LD counterpart. We then optimize the cost function with respect to the parameters of the mappings instead of the LD coordinates directly. Using linear mappings makes our multi-scale t-SNE natively interpretable and using one mapping per point offers greater flexibility in comparison to other natively interpretable approaches that use a single global mapping, such as PCA or our other variant of Multi-scale t-SNE [W] with a unique weight matrix. This flexibility translates in better DR quality, but necessitates to add a regularization term in the cost function to enforce coherence of the local explanations across the embedding, and thus essentiality of explanations. By tuning this regularization term, the user can explore the full spectrum of the trade-off between DR quality and explanation essentiality. L1 regularization can also be applied in order to encourage sparse explanations. Experiments demonstrate the effectiveness of our approach: for an appropriate tuning of the regularization terms, we are able to provide an embedding with coherent and simple explanations while sacrificing very little in DR quality.

In future research, we aim to gain a deeper understanding of the capabilities and limitations of our approach by applying it to a broader range of datasets, as well as comparing it with various other methods using diverse evaluation metrics. A normalized version of the regularization hyper-parameters would greatly facilitate this more comprehensive assessment. Other potential directions of research include investigating alternative regularization schemes and optimization heuristics, like dynamic weighting of the different terms in the cost function, for further improving DR quality and explanations. Ultimately, one may also consider quantifying the trustworthiness of the provided explanations themselves and whether or not they are relevant in the domain of application. Devising criteria measuring relevance would indeed enable users to trust their analyses of LD embeddings. In such an endeavor, value would emerge eventually from applying our method to real-world datasets and seeking feedback from domain experts about the explanations that our natively interpretable multi-scale t-SNE provides.

Acknowledgements EC is supported by a FSR grant (UCLouvain). PL is a FRIA grantee of the Fonds de la Recherche Scientifique - FNRS. JAL is a Research Director with the F.R.S.-FNRS. CdB is supported by Service Public de Wallonie Recherche under grant n°2010235-ARIAC by DIGITALWALLONIA4.AI.

References

1. Bibal, A., Clarinval, A., Dumas, B., Frénay, B.: Ixvc: An interactive pipeline for explaining visual clusters in dimensionality reduction visualizations with decision trees. Array **11**, 100080 (2021)
2. Bibal, A., Vu, V.M., Nanfack, G., Frénay, B.: Explaining t-sne embeddings locally by adapting lime. In: ESANN. pp. 393–398 (2020)
3. de Bodt, C., Mulders, D., Verleysen, M., Lee, J.A.: Perplexity-free t-SNE and twice Student tt-SNE. In: ESANN. pp. 123–128 (2018)
4. de Bodt, C., Mulders, D., Sánchez, D.L., Verleysen, M., Lee, J.A.: Class-aware t-sne: cat-sne. In: ESANN. pp. 409–414 (2019)
5. de Bodt, C., Mulders, D., Verleysen, M., Lee, J.A.: Nonlinear dimensionality reduction with missing data using parametric multiple imputations. IEEE Transactions on Neural Networks and Learning Systems **30**(4), 1166–1179 (2018)
6. de Bodt, C., Mulders, D., Verleysen, M., Lee, J.A.: Fast multiscale neighbor embedding. IEEE Transactions on Neural Networks and Learning Systems (2020)
7. Borg, I., Groenen, P.J.F.: Modern Multidimensional Scaling: Theory and applications. Springer Science & Business Media (2005)
8. Bunte, K., Biehl, M., Hammer, B.: A general framework for dimensionality-reducing data visualization mapping. Neural Computation **24**(3), 771–804 (2012)
9. Bunte, K., Haase, S., Biehl, M., Villmann, T.: Stochastic neighbor embedding (sne) for dimension reduction and visualization using arbitrary divergences. Neurocomputing **90**, 23–45 (2012)
10. Carreira-Perpinán, M.A.: The elastic embedding algorithm for dimensionality reduction. In: ICML. vol. 10, pp. 167–174 (2010)
11. Chen, L., Buja, A.: Local multidimensional scaling for nonlinear dimension reduction, graph drawing, and proximity analysis. J AM STAT ASSOC **104**(485), 209–219 (2009)
12. Colange, B., Peltonen, J., Aupetit, M., Dutykh, D., Lespinats, S.: Steering distortions to preserve classes and neighbors in supervised dimensionality reduction. Advances in neural information processing systems **33**, 13214–13225 (2020)
13. Francois, D., Wertz, V., Verleysen, M.: The concentration of fractional distances. IEEE Transactions on Knowledge and Data Engineering **19**(7), 873–886 (2007)
14. Hinton, G., Roweis, S.: Stochastic neighbor embedding. In: NIPS. vol. 15, pp. 833–840 (2002)
15. Jolliffe, I.T.: Principal component analysis and factor analysis. In: Principal component analysis, pp. 115–128. Springer (1986)
16. Kobak, D., Berens, P.: The art of using t-sne for single-cell transcriptomics. Nature communications **10**(1), 5416 (2019)
17. Lambert, P., de Bodt, C., Verleysen, M., Lee, J.A.: Squadmds: A lean stochastic quartet mds improving global structure preservation in neighbor embedding like t-sne and umap. Neurocomputing **503**, 17–27 (2022)
18. Lambert, P., Marion, R., Albert, J., Jean, E., Corbugy, S., de Bodt, C.: Globally local and fast explanations of t-SNE-like nonlinear embeddings. In: AIMLAI (2022)
19. Lee, J.A., Verleysen, M.: Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants. Procedia Computer Science **4**, 538–547 (2011)
20. Lee, J.A., Peluffo-Ordóñez, D.H., Verleysen, M.: Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. Neurocomputing **169**, 246–261 (2015)

21. Lee, J.A., Renard, E., Bernard, G., Dupont, P., Verleysen, M.: Type 1 and 2 mixtures of kullback–leibler divergences as cost functions in dimensionality reduction based on similarity preservation. Neurocomputing **112**, 92–108 (2013)
22. Lee, J.A., Verleysen, M.: Nonlinear dimensionality reduction. Springer Science & Business Media (2007)
23. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing **72**(7-9), 1431–1443 (2009)
24. Linderman, G.C., Rachh, M., Hoskins, J.G., Steinerberger, S., Kluger, Y.: Fast interpolation-based t-sne for improved visualization of single-cell rna-seq data. Nature methods **16**(3), 243–245 (2019)
25. Van der Maaten, L., Hinton, G.: Visualizing data using t-sne. Journal of machine learning research **9**(11) (2008)
26. McInnes, L., Healy, J., Melville, J.: Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018)
27. Mokbel, B., Lueks, W., Gisbrecht, A., Hammer, B.: Visualizing the quality of dimensionality reduction. Neurocomputing **112**, 109–123 (2013)
28. Peltonen, J., Georgatzis, K.: Efficient optimization for data visualization as an information retrieval task. In: 2012 IEEE International Workshop on Machine Learning for Signal Processing. pp. 1–6. IEEE (2012)
29. Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E., Vilanova, A.: Hierarchical stochastic neighbor embedding. In: Computer Graphics Forum. vol. 35, pp. 21–30. Wiley Online Library (2016)
30. Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. science **290**(5500), 2323–2326 (2000)
31. Sammon, J.W.: A nonlinear mapping for data structure analysis. IEEE Transactions on computers **100**(5), 401–409 (1969)
32. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural computation **10**(5), 1299–1319 (1998)
33. Tang, J., Liu, J., Zhang, M., Mei, Q.: Visualizing large-scale and high-dimensional data. In: Proceedings of the 25th international conference on world wide web. pp. 287–297 (2016)
34. Tasic, B., Yao, Z., Graybuck, L.T., Smith, K.A., Nguyen, T.N., Bertagnolli, D., Goldy, J., Garren, E., Economo, M.N., Viswanathan, S., et al.: Shared and distinct transcriptomic cell types across neocortical areas. Nature **563**(7729), 72–78 (2018)
35. Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. Journal of Machine Learning Research **11**(2) (2010)
36. Vladymyrov, M., Carreira-Perpinan, M.: Entropic affinities: Properties and efficient numerical computation. In: International conference on machine learning. pp. 477–485. PMLR (2013)
37. Wattenberg, M., Viégas, F., Johnson, I.: How to use t-sne effectively. Distill **1**(10), e2 (2016)
38. Yang, Z., King, I., Xu, Z., Oja, E.: Heavy-tailed symmetric stochastic neighbor embedding. Advances in neural information processing systems **22** (2009)
39. Yang, Z., Peltonen, J., Kaski, S.: Scalable optimization of neighbor embedding for visualization. In: International conference on machine learning. pp. 127–135. PMLR (2013)