Neurocomputing

Semi-Supervised t-SNE with Multi-Scale Neighborhood Preservation

--Manuscript Draft--

Manuscript Number:	
Article Type:	Regular article
Section/Category:	Data analytics
Keywords:	dimensionality reduction; stochastics neighbor embedding; semi-supervised learning; neighborhood preservation; data visualization
Corresponding Author:	Walter Serna, M. Sc Universidad Tecnológica de Pereira Pereira, Risaralda COLOMBIA
First Author:	Walter Serna, M. Sc
Order of Authors:	Walter Serna, M. Sc
	Cyril de Bodt, Ph. D.
	Andres Marino Alvarez, Ph. D.
	John Aldo Lee, Ph. D.
	Michel Verleysen, Ph. D
	Alvaro Angel Orozco, Ph. D.
Abstract:	Unsupervised dimensionality reduction (DR) aims to preserve input data structure in a low-dimensional (LD) space based on neighborhood information. In contrast, supervised DR intends to improve the learning performance, i.e., classification and regression, in an LD representation. Unfortunately, obtaining the complete label outputs of a data set for real-world applications is hard. Here, we introduce a novel DR framework coupling both available class labels and input feature similarities to extend the well-known t-distributed Stochastic Neighbor Embedding (SNE) for semi-supervised scenarios. Our proposal, termed Semi-Supervised t-SNE (SS.t-SNE), properly fixes the widths of Gaussian neighborhoods to reveal the salient local and global data structures in an LD space. Indeed, our approach is presented as a generalization of unsupervised and supervised versions of t-SNE. The results show that our SS.t-SNE outperforms other semi-supervised DR methods in data visualization and classification tasks carried out in LD embeddings.

Highlights

Semi-Supervised t-SNE with Multi-Scale Neighborhood Preservation

Walter Serna Serna, Cyril de Bodt, Andres M. Alvarez, John A. Lee, Michel Verleysen, Alvaro A. Orozco

- A semi-supervised framework based on similarity preservation for dimensionality reduction is presented.
- Class-label and data structure insights are coupled for neighborhood preservation based on the well-known t-SNE approach.
- Our SS. t-SNE is a well-founded generalization of the t-SNE method from multi-scale neighborhood preservation and class-label coupling within a divergence-based loss.
- Visualization, rank, and classification performance criteria are tested on synthetic and real-world datasets devoted to dimensionality reduction and data discrimination.

Semi-Supervised t-SNE with Multi-Scale Neighborhood Preservation

Walter Serna Serna[a], Cyril de Bodt[b], Andres M. Alvarez[c], John A. Lee[b], Michel Verleysen[b], Alvaro A. Orozco[a]

[a] *Automatic Research Group, Universidad Tecnológica de Pereira, Pereira-Risaralda, Colombia, 660001*
[b] *Université Catholique de Louvain - ICTEAM, Louvain-la-Neuve, Belgium, 1348*
[c] *Signal Processing and Recognition Group, Universidad Nacional de Colombia, Manizales-Caldas, Colombia, 170001*

Abstract

Unsupervised dimensionality reduction (DR) aims to preserve input data structure in a low-dimensional (LD) space based on neighborhood information. In contrast, supervised DR intends to improve the learning performance, i.e., classification and regression, in an LD representation. Unfortunately, obtaining the complete label outputs of a data set for real-world applications is hard. Here, we introduce a novel DR framework coupling both available class labels and input feature similarities to extend the well-known t-distributed Stochastic Neighbor Embedding (SNE) for semi-supervised scenarios. Our proposal, termed Semi-Supervised t-SNE (SS.t-SNE), properly fixes the widths of Gaussian neighborhoods to reveal the salient local and global data structures in an LD space. Indeed, our approach is presented as a generalization of unsupervised and supervised versions of t-SNE. The results show that our SS.t-SNE outperforms other semi-supervised DR methods in data visualization and classification tasks carried out in LD embeddings.

Key words: dimensionality reduction, stochastics neighbor embedding,

semi-supervised learning, neighborhood preservation, data visualization

1. Introduction

Machine learning is a transversal science field in charge of providing methodologies to obtain relevant information from data in classification, clustering, and regression tasks [10, 46, 52]. Nevertheless, the input data dimensionality, irrelevant variables, the norm concentration phenomenon, and the high sparsity of samples are important challenges when training machine learning algorithms [2, 10, 37]. Therefore, dimensionality reduction (DR) methods have been introduced to solve the abovementioned issues. The driving principle of DR is to map the high-dimensional (HD) data to a low-dimensional (LD) space, retaining local and global properties from the HD representation. The fundamental DR approach resides in the well-known principal component analysis (PCA) algorithm, which computes a feature subspace maximizing the preservation of the data variance through a linear mapping holding an orthonormal constraint [30]. As a dissimilarity-based DR alternative of the PCA algorithm, the multidimensional scaling (MDS) technique seeks for an LD space preserving the Euclidean distance-based relationships among instances [14]. Yet, linear mapping encodes the global data structure but cannot reflect local patterns in nonlinear data [34].

Regarding nonlinear solutions, spectral DR methods involve sparse similarity matrices, which are understood as a sub-manifold preserved in the LD space [15]. Examples of spectral algorithms are kernel PCA (KPCA) [60], isometric feature mapping (ISOMAP) [63], locally linear embedding (LLE) [58], Laplacian eigenmaps (LE) [7], and diffusion maps [47]. Even though

they have a well-supported theoretical framework, these methods fail to outperform older iterative techniques in visualization tasks [37, 39, 42]. Lee and Verleysen (2009) present a review of DR methods and quality criteria, finding that ISOMAP, LLE, and LE turn out to be equivalent to classical MDS in neighborhood preservation, as well as iterative methods as Sammon's mapping (NLM) [59], curvilinear component analysis (CCA) [19], and curvilinear distance analysis (CDA) [36]. Alike, [25, 42] argue that most of these techniques cannot retain both local and global data structure in a single mapping.

Therefore, the well-known stochastic neighbor embedding (SNE) [26] approach has been introduced as a nonlinear DR method that uses Gaussian probability functions to measure the similarity among instances and then minimizes the Kullback Leibler (KL) divergence between the HD and LD distribution to find an LD representation. Indeed, this paradigm was well received after the publication of variants such as t-distributed SNE (t-SNE) [42], neighbourhood retrieval visualizer (NeRV) [66], Jensen-Shannon embedding (JSE) [39], and its multi-scale versions Ms.SNE, Ms.NeRV, Ms. JSE, Ms.t-SNE [11, 40]. In this sense, SNE-based approaches outperform traditional linear and nonlinear DR techniques for data visualization concerning the preservation of local neighborhoods around each datum [33, 53].

Moreover, some DR methods incorporate the samples' class membership, provided as labels per sample, during the embedding within an SNE-based framework [73]. A more recent publication proposes a modification of t-SNE (*cat*-SNE) that employs class labels to adjust the widths of the Gaussian neighborhoods around each datum instead of deriving those from a perplex-

ity set by the user and obtaining rather good results in neighborhood preservation assessment [12]. Still, the labels of all data samples are required. Unfortunately, in real-world applications, labeled data are scarce and hard to obtain, yielding to semi-supervised DR [61, 74], which is a field aiming at performing DR on partially labeled data sets. However, most semi-supervised DR methods do not perform neighborhood preservation assessment to verify their capability in data visualization because it has been prioritized the classification tasks [27, 28]. Here, we emphasize that the goals of both tasks are different and, in some applications, visualization can be misaligned with the classification; for instance, the boundary between two classes could be fuzzy, or unrecognized classes could be present by samples that can not be categorized. Notwithstanding, label information may be used to obtain a specific visualization of data, as the intra-class and inter-class data structure, to try to understand the latent variables of supervised models [4], even though the classification stage is not performed.

On the other hand, deep learning and autoencoder-based methods have become an essential alternatives for classification tasks from DR to process large-scale data-sets [21, 56]. In this aspect, some alternatives have used the objective function of t-SNE to regularize the optimization of the LD coordinates with variational autoencoders and convolutional neural networks to preserve local data structures [1, 4, 9, 17, 23]. This demonstrates that is pertinent to continue exploring the foundation of SNE-based methods to be used in all kind of applications.

This paper introduces a semi-supervised DR approach based on the t-SNE algorithm, termed semi-supervised t-SNE (SS.t-SNE), which gathers

accessible label information and input samples similarities in HD space to support data visualization of partially labeled samples. In particular, SS.t-SNE adjusts the Gaussian neighborhood's radius depending on whether the point is labeled or unlabeled in the HD space. For this purpose, we combine the Ms.t-SNE preservation of global and local neighborhoods for unlabeled data and the *cat*-SNE capability to cluster instances sharing the same label. Namely, the imposed SS.t-SNE data similarities for labeled points ignore the probability mass of unlabeled ones to establish a region size where the same class predominates without affecting the entropy-based neighborhood's distribution. Simultaneously, the unlabeled points work with the average of multiple Gaussian areas with a growing radius. As a result, the labeled points stretch with other instances of the same class, while the unlabeled samples support the data structure's codification. Testing is carried out on real-world data sets, for which we assess quality based on neighborhood preservation and recognition rates in LD spaces. From obtained results, the introduced SS.t-SNE proves to be a competitive method for visualization and classification tasks compared to other state-of-the-art techniques.

The remainder is organized as follows. Section 2 summarizes related work. Section 3 describes the theoretical background of t-SNE and its relevant variants, as well as introduces the proposed semi-supervised DR approach. Sections 4 and 5 describes experiments and results. Finally, Section 6 draws the conclusions and future work.

2. Related Work

To better understand semi-supervised DR methods, let us start with the big picture of semi-supervised learning. In this light, we deal with real-world applications where data mixes labeled and unlabeled samples. From a machine learning perspective, the semi-supervised issue can be cover from a cluster or manifold perspective [54, 61]. Tthe supervised clustering point of view assumes that if samples are in the same group, they are likely to be of the same class. On the other hand, we can assume that the HD data follows an LD manifold [52]. A manifold can be understood as a data set's intrinsic structure because of the variables' dependence. Thus, some geometrical shapes can be induced to represent an object to be recognized in the HD space [37]. In both cluster and manifold assumptions, the intuitive feature to be preserved in constructing an embedding space is the observed proximity among data points. The latter can be associated with neighborhoods formed around each sample. Cluster assumption looks for subsets of spatially grouped points inside a decision boundary; meanwhile, the manifold perspective concerns preserving the pairwise relationships between points, being necessary to impose a "proximity", i.e. distance criterion. Table 1 lists the assumption used in some DR methods contrasted against its purpose.

Overall, DR methods quantify proximity as spatial distances [19, 59, 64], graph distances [36, 63], and as metrics in reproducing kernel Hilbert space (RKHS) [60]. However, methods based on distance preservation fix a rigid structure in manifold assumption. Alternatively, other researchers treat the problem as a topological application, establishing a lattice and representing proximity through the connection between two vertices. For example, self-

Approach	Perspective	Purpose	Quality assessment	Label information
MDS [64]	Manifold	Visualization	Visual Inspection	Unsupervised
Sammon's mapping [59]	Manifold	Visualization	Visual Inspection	Unsupervised
LE [7]	Manifold	Clustering	Visual Inspection	Unsupervised
Isomap [63]	Manifold	Manifold learning	Visual Inspection	Unsupervised
LLE [58]	Manifold	Visualization	Visual Inspection	Unsupervised
$t-$SNE [42]	Manifold	Visualization	Visual Inspection	Unsupervised
JSE [39]	Manifold	Visualization	Rank criteria	Unsupervised
			Visual Inspection	
NeRV [66]	Manifold	Visualization	Mean rank-based precision	Unsupervised
			Mean rank-based recall	
SNeRV [66]	Manifold	Visualization	Mean rank-based precision	Supervised
			Mean rank-based recall	
			Classification error rates	
MRE [44]	Manifold	Feature extraction	Visual Inspection	Supervised
SSDR [69]	Cluster	Clustering	Classification accuracy	Semi-supervised
	and Manifold			
SELF	Cluster	Clustering	Average misclassification	Semi-supervised
[62]	and Manifold			
SLDA [67]	Cluster	Clustering	Recognition rate	Semi-supervised
ASSDR [45]	Manifold	Clustering	Classification accuracy	Semi-supervised
GCDR-LP [18]	Manifold	Clustering	Average accuracy	Unsupervised
VLR [51]	Cluster	Classification	Mean accuracy	Semi-supervised
	or Manifold			
SODA [49]	Cluster	Classification	Recognition rate	Semi-supervised
GLPP [27]	Manifold	Classification	Recognition rate	Semi-supervised
UMAP [43]	Manifold	Visualization	Embedding stability	Unsupervised [1]
			Visual Inspection	
Interactive t-SNE	Manifold	interactive labeling	Classification accuracy	Semi-supervised
[41]			Visual inspection	
SS EME [75]	Manifold	Classification	Recognition rate	Semi-supervised

Table 1: Survey of relevant dimensionality reduction methods.

organizing maps (SOM) [32] set a predefined lattice with centroids as vertices that mold the data distribution. Moreover, methods like LLE [58], and LE [7] set a lattice where the same data points are used as vertices. These two latest techniques belong to the family of spectral approaches because they use component decomposition and select the principal eigenvectors as a basis within the LD space. However, even though LLE and LE work with manifold assumptions, they do not overcome traditional DR methods in neighborhood preservation for visualization tasks, even though they have

shown to be helpful for clustering [38, 42].

Alternatively, SNE proposes to measure the proximity as a pairwise similarity in terms of probability distribution functions, bringing the desired flexibility in constructing the embedding space [26]. In more detail, a Gaussian kernel measures the similarities among points, and each sample's bandwidth (length-scale) is set to hold a given entropy within the neighborhood. Each input point is embedded, and another Gaussian kernel with unitary bandwidth measures the similarities among LD points. Finally, the coordinates of the LD points are optimized to minimize the KL divergence between HD and LD distributions. The outstanding results of SNE have been improved in different variations of the original version. We can identify variants in four aspects from a literature review: i) Setting the similarity measure. ii) Fixing the divergence function between HD and LD distributions. iii) Optimization stage enhancement. iv) Bandwidth computation algorithms. Table 2 shows relevant SNE-based approaches according to their variation perspective regarding unsupervised, semi-supervised, and supervised learning. We are interested in highlighting that the method is flexible enough to combine complementary variations simultaneously according to the application's requirements. In this sense, it is worth exploring new possibilities of this paradigm.

Variants concerning the similarity measure tackle the crowding problem with t-SNE changing the Gaussian kernel in LD for a Student-t distribution [42]. Multiple relational embedding proposes to combine multiple similarity matrices weighted by the knowledge provided by the user. SNeRV uses the learning metric as the distance between samples to increase the similarities

Approach	Variation	Purpose	Quality assessment	Label information
Multiple relational embeding [44]	Similarity measure	Use of side information	Visual inspection	Semi-supervised
Symmetric - SNE [42]	Similarity measure	Symmetric similarity	Visual inspection	Unsupervised
t-SNE [42]	Similarity measure	Crowding Problem	Visual inspection	Unsupervised
SNeRV [66]	Similarity measure	Local or global structure preservation	Mean rank-based precision	Supervised
Parametric t-SNE [22]	Similarity measure	Out-of-sample extension	Rank-criteria	Unsupervised Supervised
H-SNE [53]	Similarity measure	Accelerating	Nearest neighbor error	Unsupervised
Fast-DSNE [73]	Similarity measure	SNE for classification	Recognition rate	Supervised
tt-SNE [11]	Similarity measure	Perplexity free variant	Rank - criteria	Unsupervised
NeRV [66]	Divergence	Local or global structure preservation	Mean rank-based precision	Unsupervised
Arbitrary divergences [16]	Divergence	Extension for different divergences	Nearest-neighbor error Rank - criteria	Unsupervised
JSE [39]	Divergence	Local or global structure preservation	Rank - criteria	Unsupervised
KEDR [3]	Divergence	Information theoretic learning extension	Rank - criteria	Unsupervised
f-divergences [29]	Divergence	Latent structure exploration	Precision - Recall curves	Unsupervised
JSE [39]	Optimization stage	Avoiding Local minima	Rank - criteria	Unsupervised
Tree-based acceleration [65]	Optimization stage	Accelerating	Nearest-neighbor error Computation time	Unsupervised
Interactive t-SNE [41]	Optimization stage	Interactive labeling		Semi-supervised
Opt-SNE [8]	Optimization stage	Early exaggeration accelerating	Iteration number	Unsupervised
FIt-SNE [33]	Optimization stage	Accelerating	CPU time performance Nearest neighbor error	Unsupervised
Fast$t$-SNE [13]	Optimization stage	Accelerating	CPU time performance Rank - criteria	Unsupervised
Multi-scale SNE-based methods [40]	Bandwidth computation	Local and global structure preservation	Rank - criteria	Unsupervised
ca$t$-SNE [12]	Bandwidth computation	Inter and intra-class structure preservation	Rank - criteria	Supervised

Table 2: Survey of SNE-based variants for dimensionality reduction.

for neighbors of the same class [66]. Indeed, the unsupervised neighbor retrieval visualizer (NeRV) is presented as a second aspect variant, with the same spirit of JSE [39], KEDR [3], and others [16, 29]. A supervised t-SNE (S-tSNE) [25] uses a dissimilarity measure with a parameter that forces points of the same class to be but penalizes similarities between points of different classes. For the divergence function, modified mixtures are studied to introduce a trade-off that allows independently revealing local or global structures. The optimization stage has been studied in Opt-SNE [8], FIt-SNE [33], and variants of the Barnes-Hut algorithm [65], looking for strategies to accelerate LD space convergence. Likewise, the SeStSNE approach modifies the gradient of the loss function to follow a descending direction that favors

classes isolation according to label information (the original SeStSNE is published in a GitHub web page of Leland McInnes, but the method is used for interactive labeling by [41]). Then, concerning the bandwidth computation, most variants change the idea of a unique perplexity value to fix the soft-neighborhood size. Ms.SNE, Ms.JSE, Ms.NeRV [40], and Ms.t-SNE [11] propose a multi-scale paradigm aiming to preserve local and global structure of the original data in the same LD embedding. Meanwhile, *cat*-SNE employs label information to adjust the bandwidths of the Gaussian distribution to fix a majority class neighborhood for each datum, obtaining encouraging results in neighborhood preservation assessment [12]. Fast multi-scale neighbor embedding algorithms accelerate the multi-scale methods, subsampling the data set at different scales with vantage-point trees and applying the Barnes-Hut algorithm to evaluate the loss function and its gradient. This variant shows that the multi-scale approach is better at preserving the HD neighborhoods than the single-scale schemes, leading to high-quality LD embeddings in a reduced computation time [13].

Cluster assumption seems to be the preferred choice when label information is available. Table 1 shows that most supervised/semi-supervised DR methods are used in classification and clustering applications. Unexpectedly, It is unusual to find methods based on the manifold perspective devoted to data visualization tasks extended to a supervised version. Besides, the quality assessment of this last group of approaches is related to classification performance instead of neighborhood preservation. In turn, plenty of traditional unsupervised DR methods have been extended to work with label information. For example, the supervised principal components approach extracts

the features with the most substantial estimated correlation to the label and then applies PCA to the reduced data set [5]. A semi-supervised PCA version applies PCA to labeled data, then some of the unlabeled instances are mapped and classified with the projection matrix. The newly labeled data set is used again to re-train the model in a self-training sequence until some heuristic convergence criterion is satisfied [57]. Supervised PCA estimates a sequence of principal components with maximal dependence on the labels, including a kernelized version [6]. A discriminant version of Isomap, coined M-Isomap (marginal Isomap), incorporates a pairwise constraint to cope the neighborhood graph into "Cannot-link" and "Must-link" neighbors to guide the manifold learning [71]. Later, a semi-supervised extension (SSD-Isomap) complements the constraints of must-link and cannot-link with a new constraint of likely-link to depict the neighborhoods of data points without class labels [28]. Likewise, LLE for classification [20] and its semi-supervised variant [70] extend the traditional LLE with a manifold for each class and propose a new distance measure to impose dissimilarity depending on label information. Similarly, other methods use pairwise constraints to set the neighbors graph [45, 51, 69, 72]. Moreover, semi-supervised orthogonal discriminant analysis via label propagation (SODA) is a semi-supervised variant of LDA that propagates the label information from labeled to unlabeled data and calculates a projection matrix using orthogonal discriminant analysis [49]. Globality-locality preserving projection (GLPP) aims to maintain the manifold using the graph Laplacian of the dynamic and static factors of data [27]. Flexible manifold embedding (FME) uses label information and the manifold structure of the data in semi-supervised embedding [50]. Discriminative SNE

(DSNE) and Fast-DSNE (FDSNE) [73], divide the probability function using inter and intra-class similarities.

Most of the above methods have been developed for the classification task, but data visualization nor neighborhood preservation. Some exhibit 2D embeddings; nonetheless, when label information is incorporated, the principal goal seems to be obtaining LD spaces where classes are isolated, no matter that the data structure is no longer preserved. Remarkably, none of the above semi-supervised methods contemplate using neighborhood preservation metrics to assess the embedding quality, even though the approach could be based on the manifold assumption. Recently, uniform manifold approximation and projection (UMAP) have been developed. UMAP uses a Riemannian geometry and fuzzy simplicity to set the manifold with several nearest neighbors where the points are uniformly distributed over the manifold in a generalization of LE [43]. The authors report that UMAP outperforms t-SNE and other widely used methods for unsupervised DR in visualization tasks. However, Kobak and Linderman [31] argue that there is no evidence that UMAP has any advantage over t-SNE in terms of preserving global structure. The online user guide documentation for UMAP software package in Python includes an extension to incorporate label information by combining an additional metric defined on the label space. As S-tSNE, Supervised UMAP rewards and penalizes distances whether or not two points are of the same class. However, as t-SNE, UMAP is a single-scale DR method, where the scale is associated with the chosen numbers of nearest neighbors. Although the authors claim superiority regarding the runtime performance, the method is based on an initial sub-sampling stage followed

by the nearest neighbor method to generate the original manifold. This can be analogous to the accelerated variants of t-SNE based on the Barnes-HUT algorithm or as FIt-SNE. In consequence, the embedding will be different from run to run. Moreover, some experiments in [4] show that when UMAP uses label information, the classification accuracy or the visualization quality using the features from a good classification model might not be good enough.

In short, we detect a lack of semi-supervised DR methods to use label information in neighborhood preservation with a multi-scale paradigm for data visualization of the intra-class and the inter-class structure. The modular architecture of the SNE-based algorithms makes us believe that it is the best choice for a new semi-supervised framework because it can be easily scalable according to the application requirements.

3. Methods

Since t-SNE is the basis of the proposed method, it is described deeply below, including its multi-scale variant Ms.t-SNE, and its supervised version *cat*-SNE.

3.1. t-SNE and Relevant Variants Fundamentals

Let $\boldsymbol{\Xi} = \{\boldsymbol{\xi}_i \in \mathbb{R}^M\}_{i=1}^N$ be a high-dimensional set of N points and M variables, and let $\boldsymbol{X} = \{\boldsymbol{x}_i \in \mathbb{R}^P\}_{i=1}^N$ be its low-dimensional representation in P variables ($P \leq M$). The HD and LD distances between the i-th and j-th samples are noted as $\delta_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2$ and $d_{ij} = \|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2$, respectively; where $\|\cdot\|_2$ stands for the L2 norm and $i, j \in \{1, 2, \ldots, N\}$. The t-SNE algorithm computes the HD and LD pairwise similarities as follows:

$$\sigma_{ij} = \frac{\exp(-\pi_i \delta_{ij}^2/2)}{\sum_{k \neq i} \exp(-\pi_i \delta_{ik}^2/2)}, \quad s_{ij} = \frac{(1+d_{ij}^2)^{-1}}{\sum_{k \neq l}(1+d_{kl}^2)^{-1}}; \tag{1}$$

with $\sigma_{ii} = s_{ii} = 0$, $\pi_i \in \mathbb{R}^+$ is a precision parameter imposing the squared inverse radius of a soft neighborhood centered on $\boldsymbol{\xi}_i$. The latter is fixed based on a binary search or Newton's method aiming to reach a user-predefined perplexity value $K_\star \in \mathbb{R}^+$, such that $\log(K_\star) = -\sum_{j=1}^{N} \sigma_{ij} \log(\sigma_{ij})$ [39, 42]. Of note, t-SNE symmetrizes the HD similarities as $\sigma_{ij} = (\sigma_{ji} + \sigma_{ij})/2N$ to circumvent outliers issues. Then, the vectors $\boldsymbol{\sigma}_i = [\sigma_{ij}]$ and $\boldsymbol{s}_i = [s_{ij}]$ gather the non-parametric probability distribution estimation for sample i in the HD and LD spaces. Therefore, the expected discrepancy between $\boldsymbol{\sigma}_i$ and $\boldsymbol{s}_i$ along the studied samples reflects if $\boldsymbol{X}$ is an appropriate embedding of $\boldsymbol{\Xi}$. Concerning this, the t-SNE algorithm computes the LD points, minimizing the following Kullback-Leibler divergence-based loss between the HD and LD similarity distributions:

$$E(\boldsymbol{\Xi}, \boldsymbol{X} | \{\pi_i\}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \sigma_{ij} \log(\sigma_{ij}/s_{ij}). \tag{2}$$

The loss function in Eq. (2) can be optimized employing a gradient-descent-based approach to compute the LD point $\boldsymbol{x}_i \in \boldsymbol{X}$, yielding:

$$\frac{\partial E}{\partial \boldsymbol{x}_i} = 4 \sum_{j=1}^{N} (\sigma_{ij} - s_{ij})(\boldsymbol{x}_i - \boldsymbol{x}_j)(1 + ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2^2)^{-1}. \tag{3}$$

Further, the Multi-scale framework of t-SNE, termed Ms.t-SNE, deals with more than one neighborhood radius, combining several HD similarities with increasing perplexities to capture local and global features [11]. In this way, Ms.t-SNE's HD similarity function is defined as:

$$\tilde{\sigma}_{ij} = \frac{1}{H}\sum_{h=1}^{H}\sigma_{ij}^{h}, \tag{4}$$

where $\sigma_{ij}^{h} \in [0,1]$ is computed as in Eq. (1) fixing its precision parameter $\pi_i^h \in \mathbb{R}^+$ using an exponentially growing perplexity framework, e.g., $K_\star^h = 2^h$, with $h \in \{1, 2, \ldots, H\}$ and $H = \lfloor \log_2(N/2) \rceil$ ($\lfloor \cdot \rceil$ stands for the rounding operation).

In turn, the class-aware t-SNE enhancement, named *cat*-SNE, includes the supervised information to compute HD similarities from the set $\boldsymbol{\mathcal{C}} = \{c_i \in \mathbb{Z}\}_{i=1}^{N}$, where c_i represents the output label of the input instance $\boldsymbol{\xi}_i$. Namely, the *cat*-SNE approach builds soft-neighborhoods with dominance class fixing the precision parameter $\hat{\pi}_i \in \mathbb{R}^+$ as follows:

$$\hat{\pi}_i = \begin{cases} \arg\min\limits_{\pi_i^h \in \Omega_i} \quad \pi_i^h, & \Omega_i \neq \emptyset \\ \arg\max\limits_{\pi_i^h} \sum\limits_{j \neq i | c_j = c_i} \sigma_{ij}^h, & \Omega_i = \emptyset, \end{cases} \tag{5}$$

where Ω_i is a set containing the precision values π_i^h holding $\sum_{j \neq i | c_j = c_i} \sigma_{ij}^h > \theta$, σ_{ij}^h is computed using the precision parameter π_i^h, and $\theta \in [0.5, 1)$ controls the required conditional distribution of the i-th instance in class c_i. Note that the lower bound value for the θ parameter equals 0.5 to ensure the dominance of the class c_i. Besides, the upper bound $\theta = 1$ is not considered to avoid the total dominance of the studied class, which can excessively tear classes in the LD space. Then, the *cat*-SNE's precision $\hat{\pi}_i$ in Eq. (5) is employed to compute the HD similarities $\hat{\sigma}_{ij} \in [0,1]$ as in Eq. (1). Thereby, *cat*-SNE seeks for the largest neighborhood radius centered on $\boldsymbol{\xi}_i$, that ensures the class label's dominance c_i (ruled by θ). It is worth mentioning that both

Ms.t-SNE and *cat*-SNE employ the t-SNE loss in Eq. (2) to compute the LD points through gradient-descent-based optimization.

3.2. Semi-Supervised t-SNE (SS.t-SNE)

Let $\mathcal{L} = \{\boldsymbol{\xi}_i | c_i \neq \emptyset\}$ and $\mathcal{U} = \{\boldsymbol{\xi}_i | c_i = \emptyset\}$ be the supervised (labeled) and unsupervised (unlabeled) input-output sets, respectively; where $|\mathcal{L}|+|\mathcal{U}| = N$ ($|\cdot|$ stands for set cardinality). Besides, $\sigma_{ij}^{(\mathcal{L})}$ (or $\sigma_{ij}^{(\mathcal{U})}$) is the similarity for each point in $\mathcal{L}$ (or $\mathcal{U}$) and its neighbors of the entire data set. As above exposed for the *cat*-SNE method, soft-neighborhoods' construction with a dominant class favors the encoding of points within the same group. Even though several real-world problems pose a challenge when both labeled and unlabeled data must be embedded, we can take advantage of a semi-supervised formulation by adding multi-scale neighborhood preservation from the unlabeled data to the class-aware embedding. In this way, we want to compute the similarities $(\sigma_{ij}^{(\mathcal{L})}, \sigma_{ij}^{(\mathcal{U})})$ fixing a single precision $\pi_i^{(\mathcal{L})} \in \mathbb{R}^+$ for each $\boldsymbol{\xi}_i \in \mathcal{L}$, and a multiple precision $\pi_i^{h(\mathcal{U})} \in \mathbb{R}^+$ with $h = \{1, 2, \ldots, H\}$ for each $\boldsymbol{\xi}_i \in \mathcal{U}$. The single precision for labeled points will preserve the intra-class similarity as in *cat*-SNE, whilst the multi-scale precision on unlabeled samples will preserve the local and global structure of the complete data-set. Hence, we can write the HD pairwise similarities as follows:

$$\sigma_{ij}^{(\mathcal{V})} = \frac{\exp(-\pi_i^{(\mathcal{V})} \delta_{ij}^2/2)}{\sum_{\forall \boldsymbol{\xi}_k \in \{\mathcal{L} \cup \mathcal{U} | j \neq i\}} \exp(-\pi_i^{(\mathcal{V})} \delta_{ik}^2/2)}, \tag{6}$$

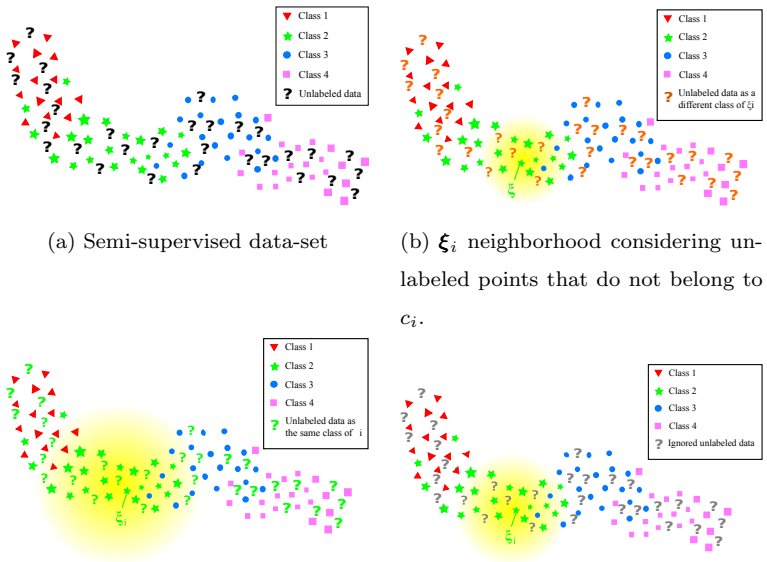
where $(\mathcal{V}) \in \{(\mathcal{L}), h(\mathcal{U})\}$. The labeled point's similarity $\sigma_{ij}^{(\mathcal{L})} \in [0, 1]$ is computed as in Eq. (6) using a precision parameter $\pi_i^{(\mathcal{L})}$. The unlabeled point's similarity, $\sigma_{ij}^{(\mathcal{U})} \in [0, 1]$, is computed as follows:

$$\sigma_{ij}^{(\mathcal{U})} = \frac{1}{H}\sum_{h=1}^{H}\sigma_{ij}^{h(\mathcal{U})}, \tag{7}$$

where $\sigma_{ij}^{h(\mathcal{U})} \in \mathbb{R}^{+}$ is computed as in Eq. (6) using a precision parameter $\pi_i^{h(\mathcal{U})}$.

As seen, $\sigma_{ij}^{(\mathcal{U})}$ does not rely on label information of the neighbors. However, $\sigma_{ij}^{(\mathcal{L})}$ does. Now we have to deal with unlabeled data in the neighborhood of $\boldsymbol{\xi}_i \in \mathcal{L}$. For the sake of clarity, Fig. 1 illustrates such a scenario. Fig. 1a exhibits a hypothetical partially labeled data-set with four classes. Fig. 1b, Fig. 1c, and Fig. 1d display three different approaches. The former, Fig. 1b, establishes that unlabeled data belong to a different class than c_i to use the ca$t$-SNE framework. Consequently, neighborhood sizes must be smaller to find the majority of neighbors of the same group, excluding unlabeled points that can be similar to $\boldsymbol{\xi}_i$ and bursting the classes in the LD space. Conversely, in Fig. 1c, the second approach labels these samples with the class c_i, producing the opposite effect than the above. Accordingly, ignoring the unlabeled data, as in Fig. 1d, can produce reliable neighborhood sizes as it does not need to impose a class. However, removing the unlabeled points when computing the single-precision of $\boldsymbol{\xi}_i \in \mathcal{L}$ is not a suitable alternative. These points must be included to calculate the HD similarities, altering the normalization factor and each neighborhood's entropy.

Here, we introduce a novel semi-supervised t-SNE approach, termed SS.t-SNE, by correctly removing the unlabeled point's probability mass instead of the points themselves as in Fig. 1d. This way, we couple unlabeled and labeled samples when computing the HD precisions. To this end, the non-parametric probability distribution estimation of sample $\boldsymbol{\xi}_i \in \mathcal{L}$ is rewritten



(a) Semi-supervised data-set

(b) $\boldsymbol{\xi}_i$ neighborhood considering unlabeled points that do not belong to c_i.

(c) $\boldsymbol{\xi}_i$ neighborhood considering unlabeled points that belong to c_i class.

(d) $\boldsymbol{\xi}_i$ neighborhood ignoring unlabeled points.

Figure 1: Handling unlabeled data in $\boldsymbol{\xi}_i \in \mathcal{L}$ neighborhood. Fig. 1a is a hypothetical data-set with four available classes and partially labeled instances; a question mark represents the unlabeled samples. Fig. 1b, Fig. 1c, Fig. 1d show three intuitive alternatives to handle unlabeled samples. The diffuse yellow spot in each figure represents the neighborhood size for each approach.

and constrained as:

$$\sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{L} | c_j = c_i, j \neq i\}} \sigma_{ij}^{(\mathcal{L})} + \sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{L} | c_j \neq c_i, j \neq i\}} \sigma_{ij}^{(\mathcal{L})} + \sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{U}\}} \sigma_{ij}^{(\mathcal{L})} = 1. \tag{8}$$

The first term on the left side in Eq. (8) corresponds to the sum of the HD similarities between $\boldsymbol{\xi}_i$ and all labeled points holding $c_j = c_i$, the second term is related to similarities against labeled instances with $c_j \neq c_i$, and the third term corresponds to the relationships between $\boldsymbol{\xi}_i$ and the unlabeled set $\mathcal{U}$. Next, we remove the total probability mass of unlabeled points by

subtracting the third term from both sides in Eq. (8). Therefore, to preserve the probability constraint, the following inequality must hold:

$$\sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{L} | c_j = c_i, j \neq i\}} \sigma_{ij}^{(\mathcal{L})} > \tilde{\theta} \left(1 - \sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{U}\}} \sigma_{ij}^{(\mathcal{L})} \right). \tag{9}$$

where $\tilde{\theta} \in [0.5, 1)$ allows tuning the portion of considered labeled samples when computing the HD precisions for points in $\mathcal{L}$. Note that the factor $\left(1 - \sum_{\forall \boldsymbol{\xi}_j \in \{\mathcal{U}\}} \sigma_{ij}^{(\mathcal{L})}\right)$ allows us to use the same range for the parameter $\tilde{\theta}$ as in *cat*-SNE, without having to be aware of the number of unlabeled samples in the data-set. Again, the lower bound $\tilde{\theta} = 0.5$ ensures the dominance of class c_i, and the upper bound $\tilde{\theta} = 1$ represents class unanimity in the neighborhood. As a consequence, Eq. (5) is expanded for semi-supervised data-sets as:

$$\pi_i^{(\mathcal{L})} = \begin{cases} \arg \min\limits_{\pi_i^{h(\mathcal{L})} \in \tilde{\Omega}_i} \pi_i^{h(\mathcal{L})}, & \tilde{\Omega}_i \neq \emptyset \\ \arg \max\limits_{\pi_i^{h(\mathcal{L})}} \sum\limits_{j \neq i | c_j = c_i} \sigma_{ij}^{h(\mathcal{L})}, & \tilde{\Omega}_i = \emptyset, \end{cases} \tag{10}$$

where $\tilde{\Omega}_i$ is a set containing the precision values $\pi_i^{h(\mathcal{L})}$ holding the condition exposed in Eq. (9). Here, $\pi_i^{h(\mathcal{L})}$ and $\pi_i^{h(\mathcal{U})}$, are computed as in Ms.*t*-SNE.

Next, we have all the HD pairwise similarities. As in the original *t*-SNE, we can achieve symmetric similarities to avoid problems with outliers. If we represent the similarities in a single variable as:

$$\tilde{\sigma}_{ij} = \begin{cases} \sigma_{ij}^{(\mathcal{L})}, & \boldsymbol{\xi}_i \in \mathcal{L} \\ \sigma_{ij}^{(\mathcal{U})}, & \boldsymbol{\xi}_i \in \mathcal{U}, \end{cases} \tag{11}$$

then, the symmetric similarity can be computed as $\sigma_{ij} = (\tilde{\sigma}_{ji} + \tilde{\sigma}_{ij})/2N$. It is noteworthy that the label information is not used in the LD similarities s_{ij}, so they can be computed as in Eq. (1). Note that the unsupervised Ms.t-SNE and the supervised *cat*-SNE become particular cases of SS.t-SNE. If label information is unavailable, the similarities are computed with Eq. (6) for the complete data set as in Ms.t-SNE. Likewise, when all points are labeled, the factor $\left(1 - \sum_{\forall \boldsymbol{\xi}_j \in \mathcal{U}} \sigma_{ij}\right)$ is equal to 1 in Eq. (9) as in *cat*-SNE.

Of note, SS.t-SNE only appraises label information to compute the HD similarities. Again, LD similarities remain as in the original t-SNE, which means that Eq. (2) can be used as a loss function, following Eq. (3) to perform a gradient descent-based optimization. However, as suggested by [40], Quasi-Newton techniques (as is L-BFGS) are considered an alternative to reach an optimal solution in fewer iterations, avoiding poor gradient scaling through second-order derivatives involvement. Although local minima affect both derivatives-based methods, we can take advantage of the multi-scale framework. The multi-scale optimization is performed as a cascade process in [40]. $\boldsymbol{X}$ is initialized with the first P principal components of $\boldsymbol{\Xi}$; then, the loss function is minimized with the highest perplexity $K_\star^H$. After convergence, the obtained value of $\mathbf{X}$ serves as the initialization for a subsequent minimization of the loss function with a smaller perplexity $K_\star^{H-1}$, and so on until reaching the smaller target perplexity for each point. Similar approaches have been used in [42, 66], suggesting that bigger perplexities smooth the high-frequency components of the loss function and decrease the probability of observing local minima. Additionally, even though the Newton method requires second derivative terms as the Hessian Matrix, L-BFGS uses an approximation of

it from the first derivative, avoiding additional computation time. Our SS.t-SNE algorithm can be summarized as in Algorithm 1.

Algorithm 1: SS.t-SNE.

Data: $\mathcal{D} = \boldsymbol{\Xi}, \boldsymbol{\mathcal{C}}$; $\tilde{\theta}$

Result: $\boldsymbol{X}$

1 Compute multi-scale precisions $\pi_i^{h(\mathcal{U})}$ of $\mathcal{U}$, and $\pi_i^{h(\mathcal{L})}$ of $\mathcal{L}$ for perplexities $K_{\star h} = 2^h$ with $1 \leq h \leq H$

2 Replace $\pi_i^{h(\mathcal{L})}$ in Eq. (10) for each ξ_i in $\mathcal{L}$, decreasing h from H until the condition is met to get the single-scale precision $\pi_i^{(\mathcal{L})}$.

3 For all ξ_i in $\mathcal{U}$, compute multi-scale similarities $\sigma_{ij}^{(\mathcal{U})}$ using $\pi_i^{h(\mathcal{U})}$ and Eq. (7). Then, apply the multi-scale optimization described in [40].

4 For all ξ_i in $\mathcal{L}$ and simultaneously with step 3, compute single-scale similarities $\sigma_{ij}^{(\mathcal{L})}$ using $\pi_i^{(\mathcal{L})}$ and Eq. (6). It is possible to perform multi-scale optimization as described in [40].

5 For all x_i in $\mathcal{U} \cup \mathcal{L}$, compute LD similarities s_{ij} using Eq. 1.

6 Optimize $\boldsymbol{X}$ using the gradient in Eq. (3) for the L-BFGS.

4. Experimental Set-up

Provided experiments aim to test three critical aspects of the proposed SS.t-SNE. First, we study the influence of hyperparameter $\tilde{\theta}$, within the range $[0.5, 1)$, and the percentage, from 1% to 99%, of unlabeled points to explore the impact of supervised information on the embeddings. Second, we study our proposal for neighborhood preservation in two-dimensional spaces compared to state-of-the-art semi-supervised DR methods. Finally, we compute

the classification rate as a function of the LD dimensions to test the DR support for separating classes under semi-supervised scenarios. Fig. 2 summarizes the three studied scenarios for SS.t-SNE experimental testing. First, complete data sets with all samples labeled are used. Then, a percentage of labels are randomly erased according to each experiment. Next, the DR method is provided with the data set, the hyperparameter $\tilde{\theta}$, and the dimensionality P of the LD space. Finally, the output $\boldsymbol{X}$ is compared with the original data to test the structure preservation performance. Additionally, the output is classified and compared with the original label information.

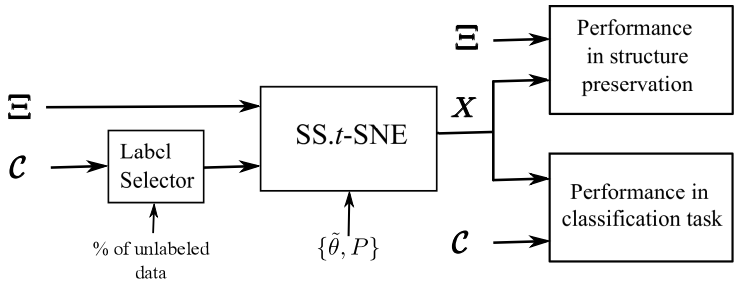


Figure 2: Experimental sketch to test our SS-t-SNE algorithm for data structure preservation and classification performance.

4.1. Data sets

To test our SS.t-SNE DR approach, we carried out experiments that involve both synthetic and real-world data sets (see Fig. 3), as follows:

- COIL20: 72 images (grayscale is used) sizing 128×128 pixels of 20 different objects corresponding to 4-degree rotations. ($M = 128^2, N = 1440$) [48].

– Fashion MNIST (fmnist): It is a random subsample of 1400 articles of Zalando. Each sample is a 28×28 grayscale image associated with a label from 10 classes. $(M = 28^2, N = 1400)$ [68]

– MNIST: This is a random subsample of 1400 scanned handwritten digits in 28×28 images. $(M = 28^2, N = 1400)$ [35].

– Swissroll: 3D synthetic data set of 500 samples following a nonlinear structure. We divide the samples into 5 groups simulating classes. $(M = 3, N = 500)$ [3]

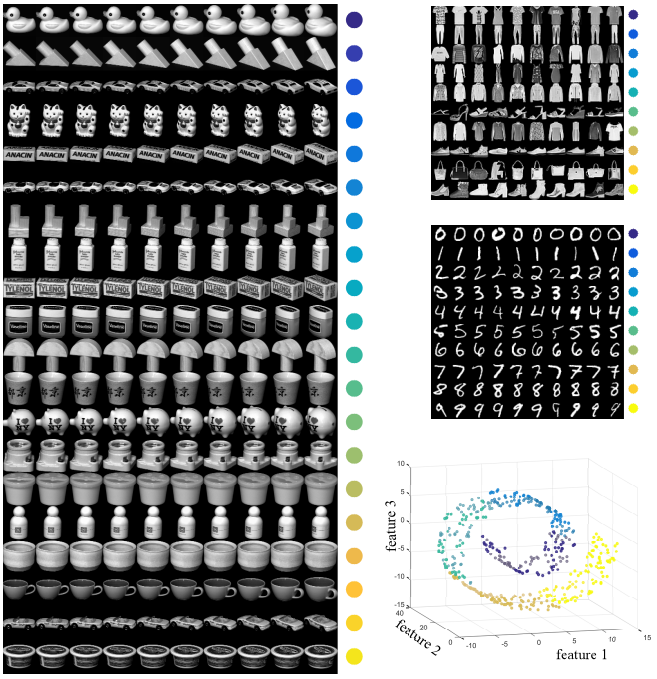


Figure 3: Data sets collection. COIL20, fmnist, MNIST, and Swissroll as explained in Section 4.1

4.2. Quality Assessment and Method Comparison

Neighborhood preservation effectiveness is evaluated using two criteria. First, unsupervised DR uses rank-based criteria to test the local and global structure preservation. We can set several K nearest neighbors to contemplate the intrusion and extrusion of points in a neighborhood for different scales [38]. Let ϑ_i^K and ζ_i^K denote the K nearest neighbor sets of $\boldsymbol{\xi}_i$ and $\boldsymbol{x}_i$ in the HD and LD spaces, respectively. In this context, the average normalized agreement between HD and LD set of neighborhoods can be written as:

$$Q_{NX}(K) = \frac{1}{KN} \sum_{i=1}^{N} |\vartheta_i^K \cap \zeta_i^K| \in [-1, 1]. \tag{12}$$

A random embedding corresponds to $Q_{NX}(K) \approx K/(N-1)$. In order to compare different neighborhood sizes independently, $Q_{NX}(K)$ is rescaled as:

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N-1-K}. \tag{13}$$

The $R_{NX}(K)$ is typically displayed with log-scale for K. Also, the area under the curve $AUC[R_{NX}(K)] = (\sum_{K=1}^{N-2} K^{-1})^{-1}(\sum_{K=1}^{N-2} R_{NX}(K)/K) \in [-1, 1]$ increases with the DR quality, quantified at all scales with an emphasis on small ones [40]. The same principle is adopted for the second criterion in the supervised framework known as kNN gain [12], considering the neighbors of the same class around each point. kNN gain is defined as:

$$G_{NN}(K) = \frac{1}{N} \sum_{i=1}^{N} \frac{|\{\boldsymbol{x}_j \in \zeta_i^K |\ c_i = c_j\}| - |\{\boldsymbol{\xi}_j \in \vartheta_i^K |\ c_i = c_j\}|}{K}. \tag{14}$$

Eq. (14) averages the gain (or loss, if negative) of neighbors of the same class around each point after DR. A global score summarizing the curve is

provided by its area $AUC[G_{NN}(K)] = (\sum_{K=1}^{N-2} K^{-1})^{-1}(\sum_{K=1}^{N-2} G_{NN}(K)/K) \in [-1, 1]$ [12].

In addition, a straightforward k-nearest neighbors classifier (kNNC) is applied to the LD space to evaluate the isolation of the classes in each embedding. Note that more sophisticated classifiers can be used, i.e., support vector machines, random forests, among others; however, the kNNC is preferred in our experiments to keep the notion of neighborhood preservation and to reveal the DR benefits for separating classes in LD spaces.

As a benchmark, manifold and clustering-based semi-supervised methods are used for comparison purposes. First, the SODA approach (clustering perspective) is a semi-supervised variant of the well-known LDA through a label propagation scheme. For the experiments, a covariance-based similarity matrix is computed. Besides, the regularization parameter μ used for SODA to avoid over-fitting is fixed as 0.1 [49]. Second, the GLPP algorithm (manifold assumption) is tested based on a Laplacian graph strategy [27]. Third, the UMAP technique (manifold assumption) is also tested as a generalization of the LE algorithm [43]. We use the semi-supervised version, publicly available[2], setting the required hyperparameters as recommended in the UMAP's user guide. It is worth mentioning that the SODA and GLPP Matlab implementations are also publicly available [3,4]

[2] `https://umap-learn.readthedocs.io/en/latest/`

[3] `https://sites.google.com/site/feipingnie/publications`

[4] `https://sites.google.com/site/siteshuang/home-1/publications`

4.3. SS.t-SNE Training Details

For all provided experiments regarding our SS.t-SNE approach, the PCA method is used to initialize the embedding coordinates $\boldsymbol{X}$. Algorithm 1 is employed, holding a line search with backtracking for the step size adjustment under the strong Wolfe conditions. The search direction is the product of the gradient with a diagonal estimate of the Hessian, which is polished by the limited-memory BFGS algorithm. The maximum number of iterations is set to 1000. As mentioned before, the optimization is multiscale. L-BFGS is run several times, starting with the highest perplexity value and then switching to a smaller perplexity until reaching the corresponding target. Our SS.t-SNE Matlab implementation is publicly available[5]. Moreover, concerning the classification experiments, which are carried out on the LD space computed by the tested semi-supervised DR approaches, we use an indirect evaluation proposed in [55, 66] to make an unbiased comparison of the performance of the methods by class prediction accuracy of the embedding spaces. In more detail, we use a $5-$fold cross-validation method. In each fold, we fix as testing set the unlabeled points (randomly erased as exposed in Section 4 according to the required percentage of unlabeled data) and we provide it during training. After the methods have computed their LD space, we classify the testing set by running a k-nearest neighbor classifier ($k = 4$) on the embedded data and evaluate the recognition rate of the methods.

[5] `https://github.com/wserna/Semi-supervised.t-SNE`

5. Results and Discussion

This section describes the outcomes of each experiment on the proposed SS.t-SNE method.

5.1. SS.t-SNE Semi-Supervised Capability and Hyperparameter Tuning

The first experiment aims to show the effect of the threshold $\tilde{\theta}$, as the portion of labeled samples of the same class inside a neighborhood, to compute the HD precision in our SS.t-SNE algorithm. Besides, we vary the percentage of unlabeled data from 1% to 99% to test the SS.t-SNE robustness. Fig. 4 evaluates a parameter array, where Y-axis and X-axis represent $\tilde{\theta}$ and the percentage of unlabeled data respectively. The colormap uses yellow for the higher and green for lower values of $AUC[R_{NX}(K)]$ and $AUC[G_{NN}(K)]$. In the first row of Fig. 4 we found a similar behavior for $AUC[R_{NX}(K)]$ in the four data sets. The higher the percentage of unlabeled data the higher the $AUC[R_{NX}(K)]$, whilst the parameter $\tilde{\theta}$ does not have a significant impact here. However, as we expected, the $AUC[R_{NX}(K)]$ decays when $\tilde{\theta}$ approaches one because the small bandwidth for class unanimity in a neighborhood bursts the local data structure.

As seen in the second row of Fig. 4, the kNN gain is influenced by $\tilde{\theta}$. As $\tilde{\theta}$ increases, SS.t-SNE requires smaller neighborhoods to be satisfied, improving the $AUC[G_{NN}(K)]$ value. Again we find that the $AUC[G_{NN}(K)]$ decays for $\tilde{\theta} = 0.99$. Note that there is not a single value for the percentage of unlabeled data to maximize the $AUC[R_{NX}]$ and $AUC[G_{NN}(K)]$ quality assessments simultaneously. Nonetheless, even from the results in all data sets, we can interpret the percentage of unlabeled data as a hyperparameter

to choose between intraclass or interclass neighborhood preservation. When label information is unavailable, SS.t-SNE takes advantage to preserve the data structure at multiple scales, getting high scores for $R_{NX}(K)$. Eventually, when we use a considerable amount of labeled instances, the bandwidths become single, allowing the method to improve the $G_{NN}(K)$ score.

Further, Fig. 5 presents a more detailed evaluation of the above exposed with the $R_{NX}(K)$ and $G_{NN}(K)$ curves for the COIL20 data set as a function of the K nearest neighbors (X-axis) for different values of $\tilde{\theta}$ and % of unlabeled points values as proposed in Section 4.2. The AUC is also computed and displayed in legends. Fig. 5a shows that the $R_{NX}(K)$ curves are not significantly altered for $\tilde{\theta}$ values between 0.5 and 0.9 for a 50% of unlabeled data, achieving a similar behavior along X-axis. In general, the loss of $AUC[R_{NX}]$ concerning $\tilde{\theta}$ is not representative compared with the gain of $AUC[G_{NN}(K)]$ reflected in Fig. 5b. In the same analysis, we fix $\tilde{\theta} = 0.9$ and explore the $R_{NX}(K)$ and $G_{NN}(K)$ curves for different % of unlabeled data. Again, Fig. 5c shows that the sacrificed score in $R_{NX}(K)$ is minimal as we use additional label information compared to the rising trend of the $G_{NN}(K)$ curves. (All the data sets yield similar, but we only present COIL20 to save space).

Then, the $\tilde{\theta}$ impact is also studied regarding the classification performance. The results are shown in Fig. 6 for COIL20, FMNIST, and MNIST data sets. As we expect, the recognition rate is better when label information is available and the hyperparameter $\tilde{\theta} \approx 0.9$. Regarding visual inspection, Fig. 7 shows the embedding space for the hyperparameter array of $\tilde{\theta}$ vs. % of unlabeled data. Here, we make evident how the unlabeled samples (Cross

mark) are attracted to the labeled samples (dot mark) of the same class and how this attraction becomes stronger as a greater portion of label information is provided.

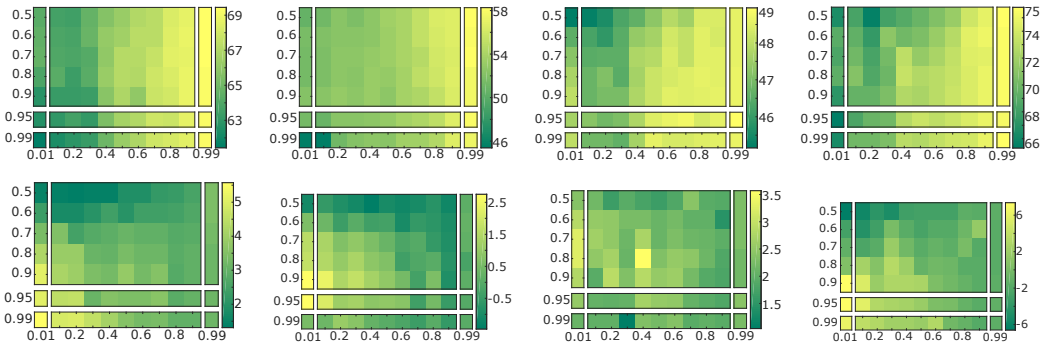
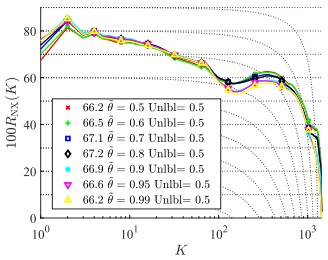


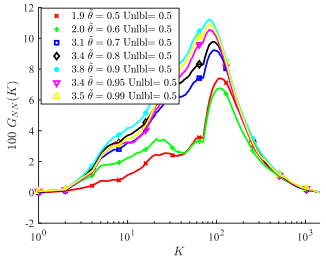
Figure 4: SS.t-SNE semi-supervised capability results. First column: COIL20, Second column: FMNIST, Third column: MNIST, Fourth column: SWISSROLL. The $AUC[R_{NX}(K)]$ (first row) and $AUC[G_{NN}(K)]$ (second row) are presented for different values of the $\tilde{\theta}$ hyperparameter (Y-axis, the portion of labeled data of the same class inside a neighborhood to compute the HD precision), vs. the percentage of unlabeled samples in the input set (X-axis).

5.2. Data structure preservation-based testing in 2D LD spaces

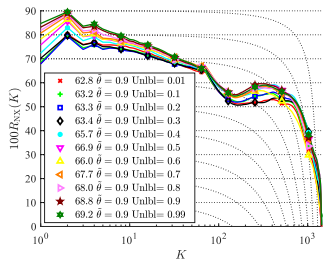
As studied in Section 5.1, a suitable $\tilde{\theta}$ value consists of 0.9 within our SS.t-SNE approach. Then, we compared our algorithm with state-of-the-art methods concerning neighborhood preservation-based assessments. Fig. 8 and 9 show the LD embeddings, and the neighborhood preservation curves for COIL20 and MNIST data sets fixing the LD space dimension as two, and the percentage of unlabeled data in 30%. As seen, SS.t-SNE improves the rank-based criteria and the kNN gain assessment for 2D LD spaces over the other semi-supervised methods, including UMAP. Besides, the SS.t-SNE-based em-



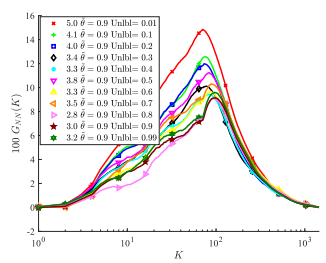
(a) COIL20 $R_{NX}(K)$ for 50% of unlabeled data



(b) COIL20 $G_{NN}(K)$ for 50% of unlabeled data



(c) COIL20 $R_{NX}(K)$ for $\tilde{\theta} = 0.9$

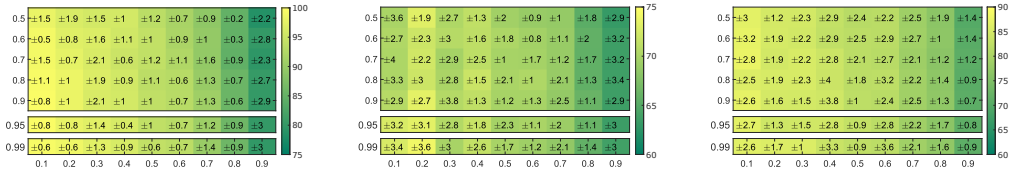


(d) COIL20 $G_{NN}(K)$ for $\tilde{\theta} = 0.9$

Figure 5: COIL20 illustrative results, $\tilde{\theta}$ vs. % of unlabeled samples (Unlbl). The $R_{NX}(K)$ and $G_{NN}(K)$ curves are shown. The AUC is also presented.

beddings show that unlabeled points (marked as crosses) feel attracted by their corresponding classes using the multi-scale structure information. The embeddings show how the unlabeled points, in most cases, are near to labeled points of the same class, preserving the intrinsic structure of each group. This tells us that the soft neighborhood size for labeled points computed with our method includes unlabeled points with a big chance of being of the same class.

The LD spaces generated with SODA show the isolation of some classes even better than GLPP. However, SODA is not efficient in preserving the structure, as seen in the R_{NX} and $G_{NN}(K)$ curves. On GLPP, the embed-



(a) COIL20 *Accuracy* (b) FMNIST *Accuracy* (c) MNIST *Accuracy*

Figure 6: kNNC-based ($k = 4$) classification accuracy concerning the $\tilde{\theta}$ value (Y-axis) and the percentage of unlabeled instances (X-axis). The embedding dimension is fixed as $P = 2$. The standard deviation of the $k-$fold is presented numerically for each parameter pair.

dings are a linear projection of the HD space, so this method is suitable to preserve the global but local structure as we see in the R_{NX} curves. Nevertheless, in complex data sets, linear projections increase the probability of superposing regions of points, producing negative values in the $G_{NN}(K)$ curves. The above is related to samples of other classes intruding into the original neighborhood. The UMAP algorithm preserves neighborhoods better than SODA and GLPP; even suitable isolation among classes is achieved. Still, we can see that SS.t-SNE separates the classes related to three kinds of cars in COIL20 data sets better than UMAP, conserving the interclass and intraclass neighborhoods. In addition, SS.t-SNE gets the highest scores for local and global structure preservation regarding the R_{NX} and $G_{NN}(K)$ curves.

5.3. Semi-Supervised DR-based Classification Testing

Figure 10 depicts the semi-supervised-based DR benefits regarding the classification performance, the LD space dimension, and the percentage of unlabeled instances. We apply the semi-supervised DR methods to generate

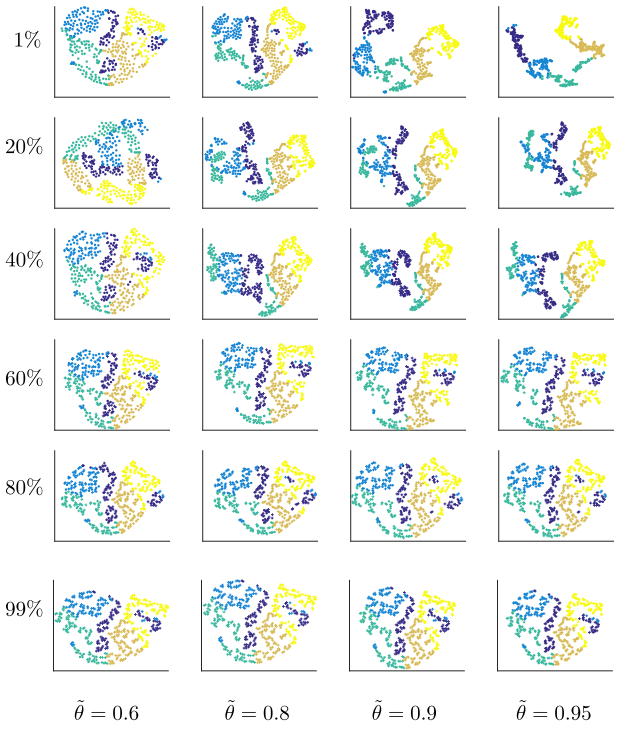
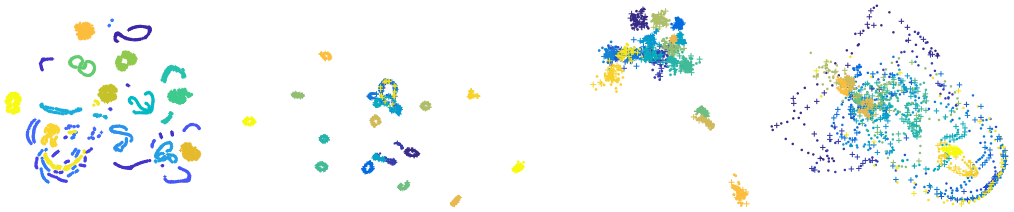


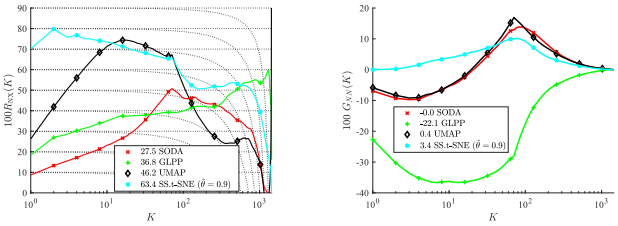
Figure 7: Visual Inspection results for the Swissroll data set. The X and Y axes represent $\tilde{\theta}$ and the % of unlabeled samples, respectively. Crosses and dots represent the unlabeled and labeled samples, respectively

LD spaces with $2 \leq P \leq 32$. Further, we apply a kNNC to classify the unlabeled samples fixing the number of nearest neighbors as 4. In this way, Y-axis and X-axis represent dimensionality P and the percentage of unlabeled data, respectively. The colormap represents the recognition rate, with high values of accuracy as yellow and low values as green. Here, SS.t-SNE presents better performance for low dimensions values ($P = 2$) and stays stable for the bigger ones. While the other methods must find the lower dimension in which the accuracy grows.

Results for SS.t-SNE show our new condition's efficiency to choose the



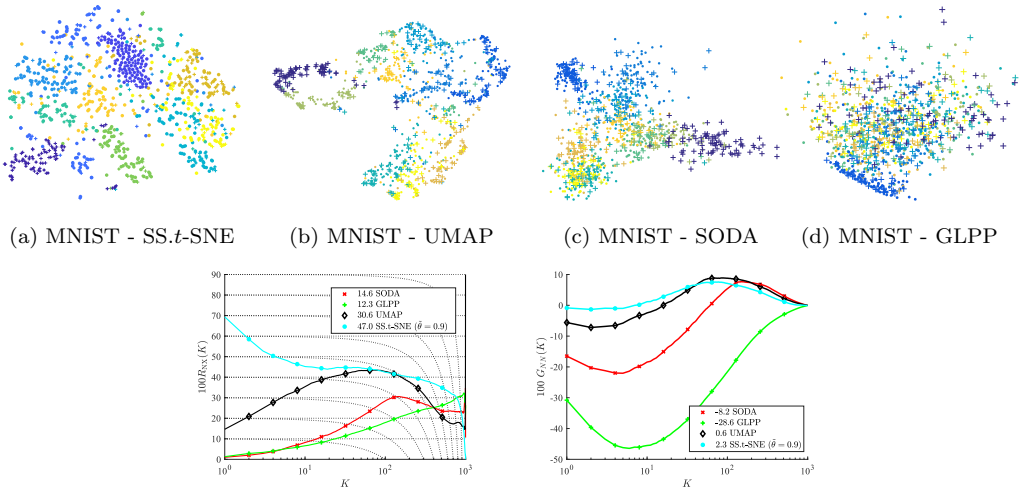
(a) COIL20 - SS.t-SNE (b) COIL20 - UMAP (c) COIL20 - SODA (d) COIL20 - GLPP



(e) COIL20 - $R_{NX}(K)$ curves (f) COIL20 - $G_{NN}(K)$ curves

Figure 8: COIL20 method comparison concerning the neighborhood preservation-based assessments with 30% of unlabeled data. From top to bottom: (a), (b), (c), and (d) show 2D embeddings from SS.t-SNE, UMAP, SODA, and GLPP approaches (crosses and dots represent the unlabeled and labeled samples, respectively); (e) and (f) present the $R_{NX}(K)$ and $G_{NN}(K)$ curves with the AUC in the legend.

Gaussian distribution's precision for labeled points without requiring difficult-to-tune parameters as the perplexity value. Notice that depending on the amount of unlabeled data, the class c_i may stop being the majority inside the neighborhood of $\boldsymbol{\xi}_i \in \mathcal{L}$. However, as the unlabeled data is being probabilistically ignored to compute the precision for labeled points, the intra-class structure information is well preserved. In fact, our experiments show that combining the single-scale forces of labeled points with the multi-scale forces of unlabeled ones is strong enough to attract unlabeled points of the corresponding class to the $\boldsymbol{\xi}_i$ neighborhood.



(a) MNIST - SS.t-SNE (b) MNIST - UMAP (c) MNIST - SODA (d) MNIST - GLPP

(e) MNIST - $R_{NX}(K)$ curves (f) MNIST - $G_{NN}(K)$ curves

Figure 9: MNIST method comparison concerning the neighborhood preservation-based assessments with 30% of unlabeled data. From top to bottom: (a), (b), (c) and (d) show 2D embeddings from SS.t-SNE, UMAP, SODA, and GLPP approaches; (e) and (f) present the $R_{NX}(K)$ and $G_{NN}(K)$ curves with the AUC in the legend.

6. Conclusions

A semi-supervised framework for dimensionality reduction based on similarity preservation is presented as a general approach to processing unsupervised, semi-supervised, and supervised data. The proposed framework is called SS.t-SNE and uses the class information to fit the bandwidths of the Gaussian neighborhoods of labeled data to reach a specified proportion of neighbors with the same class as the central point, ignoring the centers of unlabeled data but without excluding it from the probability density function. On the other hand, the local and global data structure is used to fit the bandwidths of the unlabeled samples in a multi-scale perspective. Though it is difficult to preserve all neighborhood relationships, the proposed

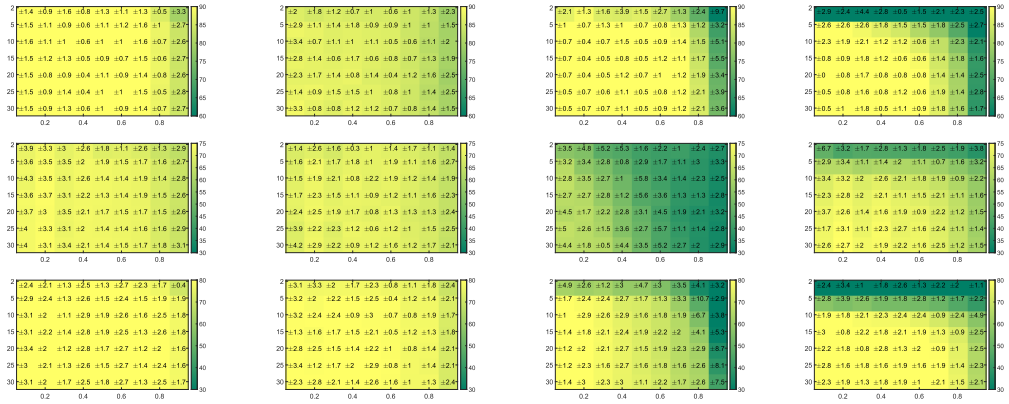


Figure 10: Classification results as a function of the LD space dimensionality P (Y-axis) and the percentage of unlabeled instances (X-axis). First row: COIL20, Second row: FMINST, Third row: MNIST. SS.t-SNE (first column), UMAP (second column), SODA (third column), and GLPP (fourth column) algorithms are studied. The standard deviation of the $k-$fold is presented numerically for each parameter pair.

method can decide what information to discard. When labels are available, the within-class structure is preserved. In the other case, our approach keeps a multi-scale structure. The experiments developed on four benchmark data sets reveal that SS.t-SNE outperforms other semi-supervised techniques for visualizing multidimensional data structure and isolating it by classes in low-dimensional spaces. Moreover, our approach uses the same cost function as in the original SNE method using Kullback-Leibler divergences because of its demonstrated capability of weighting the local structure.

Future work will include constructing an extension to map new samples from HD to LD space. Regarding this, we will explore matrix projection-based approaches from a reproducing kernel Hilbert space perspective. Also, we will couple our SS.t-SNE main ideas within a variational autoencoder framework to regularize the generation of the embedding space from large

semi-supervised data [23].

Acknowledgments

A. Alvarez thanks to the project: "Procesamiento de señales de electroencefalografía en interfaz cerebro-computador orientado a la detección de imaginación motora utilizando modelos de aprendizaje profundo y medidas de conectividad - (HERMES - 53223)", funded by Universidad Nacional de Colombia - Sede Manizales.

References

[1] Alemi, A., Poole, B., Fischer, I., Dillon, J., Saurous, R. & Murphy, K. Fixing a broken ELBO. *International Conference On Machine Learning*. pp. 159-168 (2018)

[2] Altman, N. & Krzywinski, M. The curse (s) of dimensionality. *Nat Methods*. **15**, 399-400 (2018)

[3] Álvarez, A., Lee, J., Verleysen, M. & Castellanos, G. Kernel-based dimensionality reduction using Renyi's α-entropy measures of similarity. *Neurocomputing*. **222** pp. 36-46 (2017)

[4] An, S., Hong, S. & Sun, J. ViVA: Semi-Supervised Visualization via Variational Autoencoders. *2020 IEEE International Conference On Data Mining (ICDM)*. pp. 22-31 (2020)

[5] Bair, E., Hastie, T., Paul, D. & Tibshirani, R. Prediction by supervised principal components. *Journal Of The American Statistical Association*. **101**, 119-137 (2006)

[6] Barshan, E., Ghodsi, A., Azimifar, Z. & Jahromi, M. Supervised principal component analysis: Visualization, classification and regression on subspaces and submanifolds. *Pattern Recognition*. **44**, 1357-1371 (2011)

[7] Belkin, M. & Niyogi, P. Laplacian eigenmaps and spectral techniques for embedding and clustering. *Advances In Neural Information Processing Systems*. pp. 585-591 (2001)

[8] Belkina, A., Ciccolella, C., Anno, R., Halpert, R., Spidlen, J. & Snyder-Cappione, J. Automated optimized parameters for T-distributed stochastic neighbor embedding improve visualization and analysis of large datasets. *Nature Communications*. **10**, 1-12 (2019)

[9] Benato, B., Gomes, J., Telea, A. & Falcão, A. Semi-supervised deep learning based on label propagation in a 2D embedded space. *ArXiv Preprint ArXiv:2008.00558*. **12702** pp. 371-388 (2020)

[10] Bishop, C. Pattern Recognition and Machine Learning. (springer,2006)

[11] Bodt, C., Mulders, D., Verleysen, M. & Lee, J. Perplexity-free t-SNE and twice Student tt-SNE. *ESANN*. pp. 123-128 (2018)

[12] Bodt, C., Mulders, D., López-Sánchez, D., Verleysen, M. & Lee, J. Class-aware t-SNE: cat-SNE. *ESANN*. pp. 409-414 (2019)

[13] Bodt, C., Mulders, D., Verleysen, M. & Lee, J. Fast Multiscale

Neighbor Embedding. *IEEE Transactions On Neural Networks And Learning Systems*. pp. 1-15 (2020)

[14] Borg, I. & Groenen, P. Modern multidimensional scaling: Theory and applications. *Journal Of Educational Measurement*. **40**, 277-280 (2003)

[15] Bunte, K., Biehl, M. & Hammer, B. A general framework for dimensionality-reducing data visualization mapping. *Neural Computation*. **24**, 771-804 (2012)

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

[17] Chien, J. & Hsu, C. Variational manifold learning for speaker recognition. *2017 IEEE International Conference On Acoustics, Speech And Signal Processing (ICASSP)*. pp. 4935-4939 (2017)

[18] Davidson, I. Knowledge Driven Dimension Reduction for Clustering.. *IJCAI*. pp. 1034-1039 (2009)

[19] Demartines, P. & Hérault, J. Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions On Neural Networks*. **8**, 148-154 (1997)

[20] De Ridder, D. & Duin, R. Locally linear embedding for classification. *Pattern Recognition Group, Dept. Of Imaging Science & Technology, Delft University Of Technology, Delft, The Netherlands, Tech. Rep. PH-2002-01*. pp. 1-12 (2002)

[21] Ghosh, T. & Kirby, M. Supervised dimensionality reduction and visualization using centroid-encoder. *ArXiv Preprint ArXiv:2002.11934*. pp. 20-1 (2020)

[22] Gisbrecht, A., Schulz, A. & Hammer, B. Parametric nonlinear dimensionality reduction using kernel t-SNE. *Neurocomputing*. **147** pp. 71-82 (2015)

[23] Graving, J. & Couzin, I. VAE-SNE: a deep generative model for simultaneous dimensionality reduction and clustering. *BioRxiv*. pp. 1 (2020)

[24] Gross, R. Face Databases. *Handbook Of Face Recognition*. (2005,2)

[25] Hajderanj, L., Weheliye, I. & Chen, D. A new supervised t-SNE with dissimilarity measure for effective data visualization and classification. *Proceedings Of The 2019 8th International Conference On Software And Information Engineering*. pp. 232-236 (2019)

[26] Hinton, G. & Roweis, S. Stochastic neighbor embedding. *Advances In Neural Information Processing Systems*. pp. 857-864 (2003)

[27] Huang, S., Elgammal, A., Huangfu, L., Yang, D. & Zhang, X. Globality-locality preserving projections for biometric data dimensionality reduction. *Proceedings Of The IEEE Conference On Computer Vision And Pattern Recognition Workshops*. pp. 15-20 (2014)

[28] Huang, R., Zhang, G. & Chen, J. Semi-supervised discriminant

Isomap with application to visualization, image retrieval and classification. *International Journal Of Machine Learning And Cybernetics*. **10**, 1269-1278 (2019)

[29] Im, D., Verma, N. & Branson, K. Stochastic Neighbor Embedding under f-divergences. *ArXiv Preprint ArXiv:1811.01247*. pp. 1 (2018)

[30] Jolliffe, I. Principal components in regression analysis. *Principal Component Analysis*. pp. 129-155 (1986)

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

[32] Kohonen, T. Self-organizing maps. (Springer Science & Business Media,2012)

[33] Linderman, G., Rachh, M., Hoskins, J., Steinerberger, S. & Kluger, Y. Fast interpolation-based t-SNE for improved visualization of single-cell RNA-seq data. *Nature Methods*. **16**, 243-245 (2019)

[34] Lawrence, N. & Hyvärinen, A. Probabilistic non-linear principal component analysis with Gaussian process latent variable models.. *Journal Of Machine Learning Research*. **6** (2005)

[35] LeCun, Y., Bottou, L., Bengio, Y., Haffner, P. & Others Gradient-based learning applied to document recognition. *Proceedings Of The IEEE*. **86**, 2278-2324 (1998)

[36] Lee, J., Lendasse, A., Verleysen, M. & Others Curvilinear distance analysis versus Isomap.. *ESANN*. **2** pp. 185-192 (2002)

[37] Lee, J. & Verleysen, M. Nonlinear dimensionality reduction. (Springer Science & Business Media,2007)

[38] Lee, J. & Verleysen, M. Quality assessment of dimensionality reduction: Rank-based criteria. *Neurocomputing*. **72**, 1431-1443 (2009)

[39] Lee, J., 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** pp. 92-108 (2013)

[40] Lee, J., Peluffo-Ordóñez, D. & Verleysen, M. Multi-scale similarities in stochastic neighbour embedding: Reducing dimensionality while preserving both local and global structure. *Neurocomputing*. **169** pp. 246-261 (2015)

[41] Luus, F., Khan, N. & Akhalwaya, I. Interactive Supervision with t-SNE. *Proceedings Of The 10th International Conference On Knowledge Capture*. pp. 85-92 (2019)

[42] Maaten, L. & Hinton, G. Visualizing data using t-SNE. *Journal Of Machine Learning Research*. **9**, 2579-2605 (2008)

[43] McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *ArXiv Preprint ArXiv:1802.03426*. pp. 1 (2018)

[44] Memisevic, R. & Hinton, G. Multiple Relational Embedding.. *NIPS*. pp. 913-920 (2004)

[45] Meng, M., Wei, J., Wang, J., Ma, Q. & Wang, X. Adaptive semi-supervised dimensionality reduction based on pairwise constraints weighting and graph optimizing. *International Journal Of Machine Learning And Cybernetics*. **8**, 793-805 (2017)

[46] Murphy, K. Machine Learning: A Probabilistic Perspective. (MIT press,2012)

[47] Nadler, B., Lafon, S., Kevrekidis, I. & Coifman, R. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. *Advances In Neural Information Processing Systems*. pp. 955-962 (2006)

[48] Nene, S., Nayar, S., Murase, H. & Others Columbia object image library (coil-20). *Technical Report CUCS-005-96*. pp. 1 (1996)

[49] Nie, F., Xiang, S., Jia, Y. & Zhang, C. Semi-supervised orthogonal discriminant analysis via label propagation. *Pattern Recognition*. **42**, 2615-2627 (2009)

[50] Nie, F., Xu, D., Tsang, I. & Zhang, C. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. *IEEE Transactions On Image Processing*. **19**, 1921-1932 (2010)

[51] Nie, F., Xu, D., Li, X. & Xiang, S. Semisupervised dimensionality reduction and classification through virtual label regression.

IEEE Transactions On Systems, Man, And Cybernetics, Part B (Cybernetics). **41**, 675-685 (2010)

[52] Olivier, C., Bernhard, S. & Alexander, Z. Semi-supervised learning. *IEEE Transactions On Neural Networks.* **20** pp. 1-8 (2006)

[53] Pezzotti, N., Höllt, T., Lelieveldt, B., Eisemann, E. & Vilanova, A. Hierarchical stochastic neighbor embedding. *Computer Graphics Forum.* **35-3** pp. 21-30 (2016)

[54] Pourbahrami, S., Balafar, M., Khanli, L. & Kakarash, Z. A survey of neighborhood construction algorithms for clustering and classifying data points. *Computer Science Review.* **38** pp. 100315 (2020)

[55] Peltonen, J., Aidos, H. & Kaski, S. Supervised nonlinear dimensionality reduction by neighbor retrieval. *2009 IEEE International Conference On Acoustics, Speech And Signal Processing.* pp. 1809-1812 (2009)

[56] Ramamurthy, M., Robinson, Y., Vimal, S. & Suresh, A. Auto encoder based dimensionality reduction and classification using convolutional neural networks for hyperspectral images. *Microprocessors And Microsystems.* **79** pp. 103280 (2020)

[57] Roli, F. & Marcialis, G. Semi-supervised PCA-based face recognition using self-training. *Joint IAPR International Workshops On Statistical Techniques In Pattern Recognition (SPR) And Structural And Syntactic Pattern Recognition (SSPR).* pp. 560-568 (2006)

[58] Roweis, S. & Saul, L. Nonlinear dimensionality reduction by locally linear embedding. *Science*. **290**, 2323-2326 (2000)

[59] Sammon, J. A nonlinear mapping for data structure analysis. *IEEE Transactions On Computers*. **100**, 401-409 (1969)

[60] Schölkopf, B., Smola, A. & Müller, K. Kernel principal component analysis. *International Conference On Artificial Neural Networks*. pp. 583-588 (1997)

[61] Sheikhpour, R., Sarram, M., Gharaghani, S. & Chahooki, M. A survey on semi-supervised feature selection methods. *Pattern Recognition*. **64** pp. 141-158 (2017)

[62] Sugiyama, M., Idé, T., Nakajima, S. & Sese, J. Semi-supervised local Fisher discriminant analysis for dimensionality reduction. *Machine Learning*. **78**, 35-61 (2010)

[63] Tenenbaum, J., De Silva, V. & Langford, J. A global geometric framework for nonlinear dimensionality reduction. *Science*. **290**, 2319-2323 (2000)

[64] Torgerson, W. Multidimensional scaling: I. Theory and method. *Psychometrika*. **17**, 401-419 (1952)

[65] Van Der Maaten, L. Accelerating t-SNE using tree-based algorithms. *The Journal Of Machine Learning Research*. **15**, 3221-3245 (2014)

[66] 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**, 451-490 (2010)

[67] Wang, S., Lu, J., Gu, X., Du, H. & Yang, J. Semi-supervised linear discriminant analysis for dimension reduction and classification. *Pattern Recognition*. **57** pp. 179-189 (2016)

[68] Xiao, H., Rasul, K. & Vollgraf, R. Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. *ArXiv Preprint ArXiv:1708.07747*. (2017)

[69] Zhang, D., Zhou, Z. & Chen, S. Semi-supervised dimensionality reduction. *Proceedings Of The 2007 SIAM International Conference On Data Mining*. pp. 629-634 (2007)

[70] Zhang, S. & Chau, K. Dimension reduction using semi-supervised locally linear embedding for plant leaf classification. *International Conference On Intelligent Computing*. pp. 948-955 (2009)

[71] Zhang, Z., Chow, T. & Zhao, M. M-Isomap: Orthogonal constrained marginal isomap for nonlinear dimensionality reduction. *IEEE Transactions On Cybernetics*. **43**, 180-191 (2012)

[72] Zhang, Y., Zhang, Z., Qin, J., Zhang, L., Li, B. & Li, F. Semi-supervised local multi-manifold Isomap by linear embedding for feature extraction. *Pattern Recognition*. **76** pp. 662-678 (2018)

[73] Zheng, J., Qiu, H., Xu, X., Wang, W. & Huang, Q. Fast Discriminative Stochastic Neighbor Embedding Analysis. *Computational And Mathematical Methods In Medicine*. **2013** (2013)

[74] Zhu, T., Pimentel, M., Clifford, G. & Clifton, D. Unsupervised bayesian inference to fuse biosignal sensory estimates for personalizing care. *IEEE Journal Of Biomedical And Health Informatics*. **23**, 47-58 (2018)

[75] Zhu, R., Dornaika, F. & Ruichek, Y. Semi-supervised elastic manifold embedding with deep learning architecture. *Pattern Recognition*. **107** pp. 107425 (2020)

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: