Neurocomputing 112 (2013) 92–108

Contents lists available at SciVerse ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Type 1 and 2 mixtures of Kullback–Leibler divergences as cost functions in dimensionality reduction based on similarity preservation

John A. Lee [a,*,1], Emilie Renard [b], Guillaume Bernard [a], Pierre Dupont [b], Michel Verleysen [b]

[a] Université catholique de Louvain, Molecular Imaging, Radiotherapy, and Oncology—IREC, Avenue Hippocrate 55, B-1200 Bruxelles, Belgium
[b] Université catholique de Louvain, Machine Learning Group—ICTEAM, Place du Levant 3, B-1348 Louvain-la-Neuve, Belgium

ARTICLE INFO

Available online 7 March 2013

Keywords:
Dimensionality reduction
Manifold learning
Stochastic neighbor embedding
Divergence

ABSTRACT

Stochastic neighbor embedding (SNE) and its variants are methods of dimensionality reduction (DR) that involve normalized softmax similarities derived from pairwise distances. These methods try to reproduce in the low-dimensional embedding space the similarities observed in the high-dimensional data space. Their outstanding experimental results, compared to previous state-of-the-art methods, originate from their capability to foil the curse of dimensionality. Previous work has shown that this immunity stems partly from a property of shift invariance that allows appropriately normalized softmax similarities to mitigate the phenomenon of norm concentration. This paper investigates a complementary aspect, namely, the cost function that quantifies the mismatch between similarities computed in the high- and low-dimensional spaces. Stochastic neighbor embedding and its variant *t*-SNE rely on a single Kullback–Leibler divergence, whereas a weighted mixture of two dual KL divergences is used in neighborhood retrieval and visualization (NeRV). We propose in this paper a different mixture of KL divergences, which is a scaled version of the generalized Jensen–Shannon divergence. We show experimentally that this divergence produces embeddings that better preserve small *K*-ary neighborhoods, as compared to both the single KL divergence used in SNE and *t*-SNE and the mixture used in NeRV. These results allow us to conclude that future improvements in similarity-based DR will likely emerge from better definitions of the cost function.

1. Introduction

Dimensionality reduction (DR) aims at producing faithful and meaningful representations of high-dimensional data into a lower-dimensional space. The general intuition that drives DR is that close or similar data items should be represented near each other, whereas dissimilar ones should be represented far from each other. Through the history of DR, authors have formalized this idea of neighborhood preservation in various ways, using several models for the mapping or embedding of data from the high-dimensional (HD) space to the low-dimensional (LD) one. For instance, principal component analysis (PCA) [1–3] and classical multidimensional scaling (MDS) [4–6] rely on linear projections that maximize variance preservation and dot product preservation, respectively. Nonlinear variants of metric MDS [7] are based on (weighted) distance preservation: they build a low-dimensional embedding that reproduce as faithfully as possible the pairwise distances measured in the data space. These distances can be Euclidean or approximation of geodesic lengths [8–12]. The use of similarities in DR is quite recent and emerged with methods based on spectral optimization. Among many other examples, Laplacian eigenmaps [13], locally linear embedding [14], and diffusion maps [15] involve sparse matrices of similarities, also called affinity matrices. In spite of a sound theoretical framework, these methods fail to outperform older techniques in typical visualization tasks [16–18]. A possible explanation is that these methods can be reformulated into classical MDS achieved in an unknown feature space [19,20]. In this case, the definition of the similarities merely determines the implicit, arbitrary nonlinear mapping from the data space to the feature space [21,22].

Genuine similarity preservation appeared later with a technique called stochastic neighbor embedding (SNE) [23]. In contrast with spectral methods that directly convert the pairwise similarities defined in the HD space into inner products, SNE matches similarities that are computed both in the HD and LD spaces. To some extent, the set of normalized similarities between a given datum and all others can be seen as an a priori distribution of neighbors, which justifies the term 'stochastic' in the method name. Interest in the new paradigm developed in SNE grew

* Corresponding author. Tel.: +32 2 7649528.
E-mail address: john.lee@uclouvain.be (J.A. Lee).
[1] J.A.L. is a Research Associate with the Belgian National Fund of Scientific Research (F.R.S.-FNRS).

0925-2312/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.neucom.2012.12.036

significantly after the publication of variants such as Student t-distributed SNE (t-SNE) [18] and NeRV [24], standing for neighborhood retrieval and visualization. These variants led to breakthroughs in terms of DR quality, with outstanding experimental results [18,24]. Nevertheless, the reasons of this performance leap remain partly unknown. The role played by SNE's very specific similarities has been investigated in [25], where it was shown that similarities defined as softmax ratios with an appropriate normalization benefit from a shift invariance property. This normalization allows the similarities to alleviate the phenomenon of norm concentration [26], which has been identified as the main cause of the poor performance of DR techniques based on distance preservation [25]. In that perspective, SNE and its variants are among the seldom nonlinear DR methods that effectively defeat the curse of dimensionality [27,28]. Another approach is followed in [29,30], where shift invariance is gained in a MDS variant by maximizing the Pearson correlation between the (vectorized) distance matrices in the HD and LD spaces, instead of the norm of their difference.

This paper focuses on a complementary aspect of similarity-based DR, namely, the definition of pertinent cost functions [31]. In SNE and its t-distributed variant t-SNE, the cost function is a sum of Kullback–Leibler (KL) divergences. For each datum, a divergence measures the mismatch between an a priori distribution of its neighbors in the HD space and the corresponding distribution computed in the LD space. As the KL divergence is asymmetric with respect to the two distributions it compares, NeRV also involves the 'dual' KL divergence, where the two distributions are swapped. A metaparameter controls the weight of two dual divergences in the cost function. In an information retrieval perspective, it has been shown that this metaparameter allows NeRV to reach different tradeoffs between *precision* and *recall* [24]. According to the nomenclature developed in [32], NeRV entails a type 1 mixture of KL divergences, that is, a linear mixture of two dual divergences. For equal mixture weights, the resulting divergence is symmetric with respect to the two compared distributions. In this paper, we investigate a type 2 mixture of KL divergences [32], which involves a composite distribution and also a nonlinear mixture of two divergences. This second type of mixture is closely related to the generalized Jensen–Shannon divergence [33,34] and, like the type 1 mixture, it is also symmetric when both weights are equal to one half. Using a criterion of K-ary neighborhood preservation, we show experimentally that the type 2 mixture outperforms both the type 1 mixture and the usual non-blended divergence. A careful examination of the gradient of each mixture reveals some clues to justify this better behavior.

The rest of this paper is organized as follows. Section 2 describes the normalized similarities used in SNE and its variants to define a priori distributions of neighbors. Section 3 deals with the two different types of divergence mixtures that measure the mismatch between these distributions. Section 4 focuses on optimization issues and analyzes the gradient of the considered divergence mixtures. Section 5 presents and discusses the experimental results. Finally, Section 6 draws the conclusions and sketches some perspectives.

2. Shift-invariant softmax similarities

Let $\mathbf{\Xi}=[\boldsymbol{\xi}_i]_{1\le i\le N}$ denote a set of N points in some M-dimensional space. Similarly, let $\mathbf{X}=[\mathbf{x}_i]_{1\le i\le N}$ be its representation in a P-dimensional space, with $P\le M$. The Euclidean distances between the ith and jth points are given by $\delta_{ij}=\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|_2$ and $d_{ij}=\|\mathbf{x}_i-\mathbf{x}_j\|_2$ in the HD and LD spaces respectively. The term similarity generally refers to a quantity that decreases as the distance grows. In SNE, the similarities associated with δ_{ij} and d_{ij} are defined for $i\neq j$ by

$$\sigma_{ij}=\frac{\exp(-\gamma_{ij})}{\sum_{k,k\neq i}\exp(-\gamma_{ik})} \quad \text{and} \quad s_{ij}=\frac{\exp(-g_{ij})}{\sum_{k,k\neq i}\exp(-g_{ik})}, \tag{1}$$

where $\gamma_{ij}=\gamma(\delta_{ij}/\lambda_i)$ and $g_{ij}=g(d_{ij})$. Functions γ and g are both non-negative with a non-negative derivative. Parameter λ_i is a bandwidth that can be seen as a soft neighborhood radius. By convention, $\sigma_{ij}=s_{ij}=0$ if $i=j$.

In SNE and NeRV, the similarities are Gaussian, with γ and g being defined as

$$\gamma(u)=g(u)=u^2/2. \tag{2}$$

In t-SNE, the similarities in the LD space are defined in a different way than in the HD space, by using an unnormalized probability mass function of a Student t distribution with m degrees of freedom[2]

$$s_{ij}=\frac{(1+s_{ij}^2/m)^{-(m+1)/2}}{\sum_k(1+s_{ij}^2/m)^{-(m+1)/2}}. \tag{3}$$

This amounts to opting for

$$g(u)=\frac{m+1}{2}\ln(1+u^2/m). \tag{4}$$

in (1). The heavier tail of the Student t function, as compared to the Gaussian, induces an exponential transformation between the HD and LD distances [35]. The longer the distance is in the HD space, the stronger it is stretched in the LD space.

An important feature of similarities defined as softmax exponential ratios such as above is their normalization, that is, $\sum_j\sigma_{ij}=\sum_j s_{ij}=1$. Combined with positivity, it allows the similarities σ_{ij} and s_{ij} to be interpreted as a priori probabilities for $\boldsymbol{\xi}_j$ and $\mathbf{x}_j$ to be neighbors of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$, respectively. But more importantly, normalization implies scale invariance with respect to $\exp(-\gamma_{ij})$ in σ_{ij}, which in turn translates into shift invariance with respect to $\gamma(\delta_{ij}/\lambda_i)=\delta_{ij}^2/(2\lambda_i^2)$ [25]. Since null distances have a trivial distribution that differs from that of non-zero distances, they are excluded from the sum in the normalization denominators in (1). As a direct result, the shift applicable to δ_{ij}^2 can range from $-\min_{j,j\neq i}\delta_{ij}^2$ to ∞. The lower end of this interval ensures that the shifted distances remain positive. A negative shift close to this lower bound is particularly interesting to mitigate the phenomenon of norm concentration [26]. One manifestation of this phenomenon is the following: for a finite sample of points $\mathbf{\Xi}$, $\min_{j,j\neq i}\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|$ grows faster with M than $\max_j\|\boldsymbol{\xi}_i-\boldsymbol{\xi}_j\|$. In other words, the relative variance of a discrete distribution of Euclidean distances (namely, its variance divided by the square of its mean) vanishes when M tends to ∞. The changing shape of distance distributions, depending on the dimensionality, partly explains the failure of DR methods based on distance preservation. The distances in LD spaces are always and systematically 'too scattered' to match those observed in HD spaces. Invariance to shifts in similarities circumvents this problem [25].

3. Divergences to measure the similarity mismatch

Thanks to positivity and normalization, vectors $\boldsymbol{\sigma}_i=[\sigma_{ij}]_{1\le j\le N}$ and $\mathbf{s}_i=[s_{ij}]_{1\le j\le N}$ can be seen as discrete probability

[2] The exact definition of s_{ij} in [18] also entails a slightly different normalization of the similarity, with a sum in the denominator that runs over both indices instead of the second one only. In practice, doing so simplifies the gradient of t-SNE but has no significant effect on the method results. Our definitions of σ_{ij} and s_{ij} in (1) reproduce those of SNE in [23] and have the advantage of instantiating twice the very same template.

distributions. Therefore, divergences can be used to assess their mismatch.

In SNE, the Kullback–Leibler divergence of $\mathbf{s}_i$ with respect to $\boldsymbol{\sigma}_i$ is used. It is defined as

$$\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\sum_j \sigma_{ij}\ln(\sigma_{ij}/s_{ij}). \tag{5}$$

The cost function of SNE [23] can then be written as

$$E(\mathbf{X};\boldsymbol{\Xi},\boldsymbol{\lambda})=\sum_i \mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i), \tag{6}$$

where vector $\boldsymbol{\lambda}=[\lambda_i]_{1\le i\le N}$ includes all similarity bandwidths associated with each datum. An advantage of using the KL divergence is that it contains logarithms that cancels some of the exponential functions in the similarities, which eventually leads to a very simple analytical formulation of its gradient.

In NeRV, the cost function blends two dual KL divergences. The resulting mixture is also a divergence and can be written as

$$\mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=(1-\kappa)\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)+\kappa\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i), \tag{7}$$

where $0\le\kappa\le 1$ is parameter that controls the importance of both terms. The cost function of NeRV is then

$$E(\mathbf{X};\boldsymbol{\Xi},\boldsymbol{\lambda},\kappa)=\sum_i \mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i). \tag{8}$$

Divergence $\mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ used in NeRV can be rewritten into discrete Shannon entropies and cross-entropies, namely,

$$\mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=(1-\kappa)(H(\boldsymbol{\sigma}_i,\mathbf{s}_i)-H(\boldsymbol{\sigma}_i))+\kappa(H(\mathbf{s}_i,\boldsymbol{\sigma}_i)-H(\boldsymbol{\sigma}_i)), \tag{9}$$

where $H(\mathbf{u},\mathbf{v})=-\sum_i u_i\ln(v_i)$ and $H(\mathbf{u})=H(\mathbf{u},\mathbf{u})$. Due to the presence of cross-entropies, it can easily be seen that $\mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ ranges from 0 to ∞. For $\kappa=1/2$, $\mathrm{D}^{1/2}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ is symmetric, namely, $\mathrm{D}^{1/2}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\mathrm{D}^{1/2}_{\mathrm{KLt1}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)$. According to the nomenclature in [32], $\mathrm{D}^{1/2}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ is the type 1 symmetric generalization of the KL divergence.

Another way to combine two KL divergences is given by

$$\mathrm{D}^{\kappa}_{\mathrm{JS}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\kappa\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{z}_i)+(1-\kappa)\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\mathbf{z}_i), \tag{10}$$

$$=H(\mathbf{z}_i)-\kappa H(\boldsymbol{\sigma}_i)-(1-\kappa)H(\mathbf{s}_i), \tag{11}$$

where $\mathbf{z}_i=\kappa\boldsymbol{\sigma}_i+(1-\kappa)\mathbf{s}_i$ and $0\le\kappa\le 1$. This mixture is known as the generalized Jensen–Shannon divergence [33,34] (with

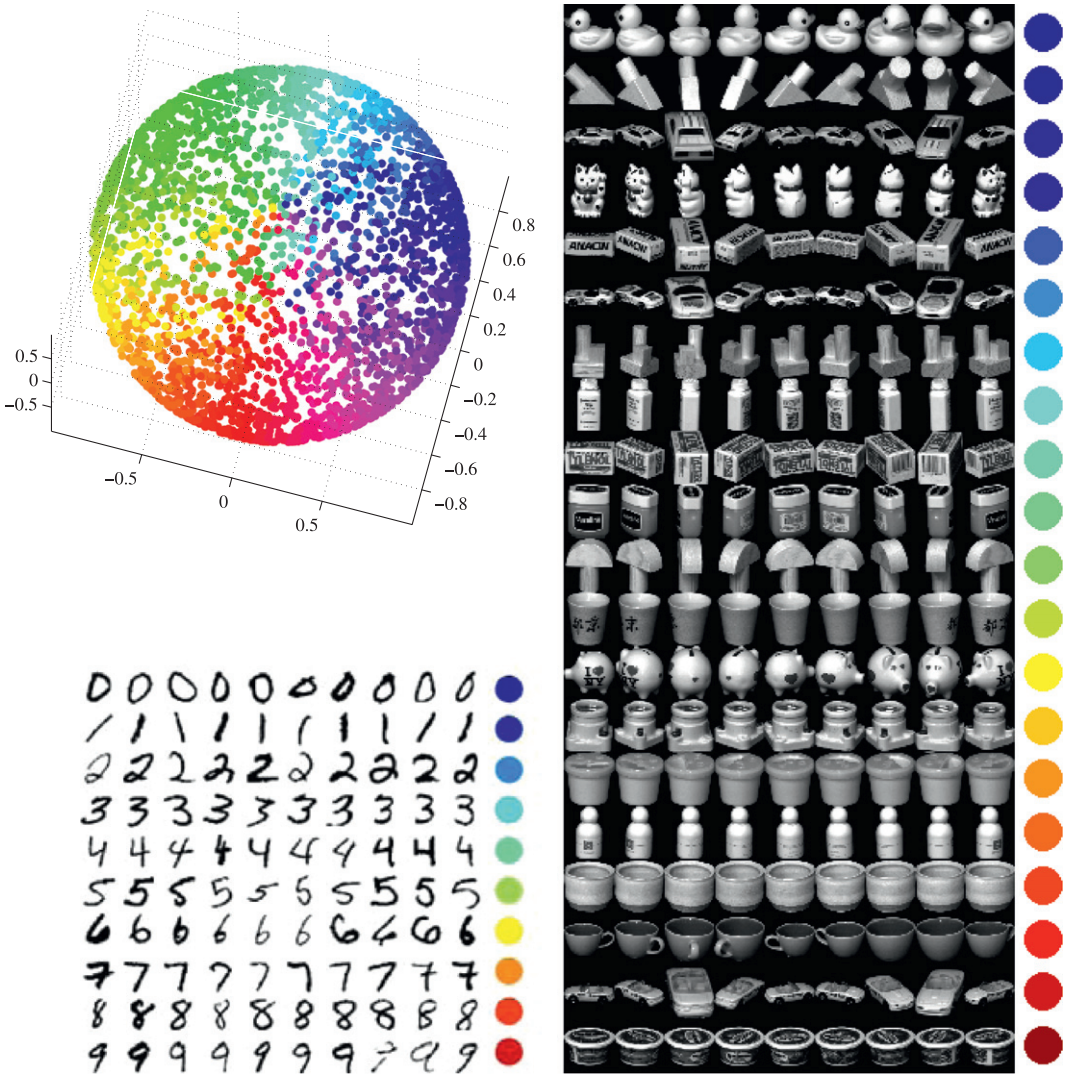

Fig. 1. The three data sets used in the experiments are the spherical shell (3 dimensions, 3000 points, upper left corner), the COIL-20 image bank (16384 dimensions, 1440 points, right column), and a random subset of the MNIST image bank (784 dimensions, 6000 points, lower left corner). The COIL-20 and MNIST images are just vectorized before dimensionality reduction. In all three cases, a two-dimensional embedding is sought.

only two distributions in this case). For $\kappa=1/2$, $\mathrm{D}_{\mathrm{JS}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ is symmetric and referred to as the type 2 symmetric generalization of the KL divergence [32]. There are only entropies and no cross-entropies in the JS divergence, which varies between 0 and $H([\kappa,1-\kappa]^T)$. When κ tends to 0 or 1, it can easily be seen that $\mathrm{D}_{\mathrm{JS}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ vanishes, even if $\boldsymbol{\sigma}_i\neq\mathbf{s}_i$. This contrasts with $\mathrm{D}_{\mathrm{KLt1}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$, for which we have $\mathrm{D}_{\mathrm{KLt1}}^{0}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ and $\mathrm{D}_{\mathrm{KLt1}}^{1}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)$. A similar behavior can be obtained with a slightly different definition of the type 2 mixture, namely,

$$\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\frac{1}{\kappa(1-\kappa)}\mathrm{D}_{\mathrm{JS}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\frac{1}{1-\kappa}\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{z}_i)+\frac{1}{\kappa}\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\mathbf{z}_i). \tag{12}$$

Using l'Hôpital's rule, the limits for κ close to 0 or 1 then become

$$\lim_{\kappa\to 0}\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)\quad\text{and}\quad\lim_{\kappa\to 1}\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)=\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i). \tag{13}$$

Hence, both the type 1 and 2 mixtures include the two dual KL divergences $\mathrm{D}_{\mathrm{KL}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ and $\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)$ as particular cases. The division of the JS divergence by $\kappa(1-\kappa)$ in the type 2 mixture also scales up its maximal value, which is moreover no longer bounded if κ is equal to 0 or 1.

To our best knowledge, $\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\mathbf{s}_i||\boldsymbol{\sigma}_i)$ has never been used as a substitute for the KL divergence in SNE. Formally, we have thus

$$E(\mathbf{X};\boldsymbol{\Xi},\boldsymbol{\lambda},\kappa)=\sum_i\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i). \tag{14}$$

Due to its direct relationship with the JS divergence, we coined this novel nonlinear DR method Jensen–Shannon embedding (JSE in short, to be pronounced like 'Jessie').

4. Optimization

The cost functions of SNE, *t*-SNE, NeRV, and JSE can be written in a generic form as

$$E(\mathbf{X};\boldsymbol{\Xi},\boldsymbol{\lambda},\kappa)=\sum_i B(\boldsymbol{\sigma}_i)+C(\mathbf{s}_i;\boldsymbol{\sigma}_i)=\sum_i\left(\sum_j b(\sigma_{ij})+\sum_j c(s_{ij};\sigma_{ij})\right), \tag{15}$$

where we distinguish the terms depending only on σ_{ij} from those depending on both s_{ij} and σ_{ij}. In SNE and *t*-SNE, we have $B(\boldsymbol{\sigma}_i)=H(\boldsymbol{\sigma}_i)$ and $C(\mathbf{s}_i;\boldsymbol{\sigma}_i)=H(\boldsymbol{\sigma}_i,\mathbf{s}_i)$. In NeRV, we have $B(\boldsymbol{\sigma}_i)=(1-\kappa)H(\boldsymbol{\sigma}_i)$ and $C(\mathbf{s}_i;\boldsymbol{\sigma}_i)=(1-\kappa)H(\boldsymbol{\sigma}_i,\mathbf{s}_i)+\kappa\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)$. In JSE, we have $B(\boldsymbol{\sigma}_i)=H(\boldsymbol{\sigma}_i)/(1-\kappa)$ and $C(\mathbf{s}_i;\boldsymbol{\sigma}_i)=H(\boldsymbol{\sigma}_i,\mathbf{z}_i)/(1-\kappa)+\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)/\kappa$.

For all methods, each term $B(\boldsymbol{\sigma}_i)$ in (15) is proportional to the Shannon entropy of $\boldsymbol{\sigma}_i$ and can be interpreted as a constant baseline, which the varying term $C(\mathbf{s}_i;\boldsymbol{\sigma}_i)$ is to be compared to. In this view, the magnitude of each baseline $B(\boldsymbol{\sigma}_i)$ somehow reflects how important is the embedding quality of each datum $\boldsymbol{\xi}_i$ relatively to the others. Being proportional to the entropy of $\boldsymbol{\sigma}_i$, the baseline can largely vary, depending on bandwidth λ_i and the given distribution of the distances δ_{ij} around $\boldsymbol{\xi}_i$. If $B(\boldsymbol{\sigma}_i)\ll B(\boldsymbol{\sigma}_j)$, then poor preservation of the neighborhood around $\boldsymbol{\xi}_i$ is less penalized than that of the neighborhood around $\boldsymbol{\xi}_j$. For these reasons, it is advised to individually adjust each bandwidth λ_i in order to equalize all baselines [23,18,24]. In practice, this goal can be reached by solving $B_0=B(\boldsymbol{\sigma}_i)$ with respect to λ_i, for all i, where B_0 is specified by the user. In [23,18,24], the user indirectly but conveniently fixes B_0 by specifying a perplexity value. The perplexity of $\boldsymbol{\sigma}_i$ is defined as $\exp(H(\boldsymbol{\sigma}_i))$ and can be thought of intuitively as the size of a soft K-ary neighborhood. For instance, a perplexity equal to 50 individually adjusts each bandwidth λ_i in the normalized Gaussian function centered on $\boldsymbol{\xi}_i$, in such a way that approximately 50 surrounding vectors $\boldsymbol{\xi}_j$ out of $N-1$ have a similarity value σ_{ij} that is not in the tail of the Gaussian function.

Section 4.1 describes an efficient procedure to equalize all baseline terms, whereas Section 4.2 deals with the subsequent minimization of the equalized cost function, in order to determine the optimal embedding $\mathbf{X}$.

4.1. Equalization of the baselines

In *t*-SNE [18], the method of solving $B(\boldsymbol{\sigma}_i)=B_0$ to determine λ_i is a dichotomous search (bisection method). Since $B(\boldsymbol{\sigma}_i)$ is differentiable with respect to λ_i, an alternative possibility is Newton's method. This iterative technique finds the closest root of any differentiable function $f(u)$ by applying repeatedly the update $u^{(t+1)}=u^{(t)}-f(u^{(t)})/f'(u^{(t)})$, where t denotes the current iteration. Both the dichotomous search and Newton's method converge quickly, but the former requires the solution to be bracketed whereas the latter needs a single initialization value, which is easier in practice.

In the particular case of the baseline equalization, we need the partial derivative of $B(\boldsymbol{\sigma}_i)$ with respect to λ_i. For the sake of simplicity, let us reparameterize $B(\boldsymbol{\sigma}_i)$ by defining the precision as

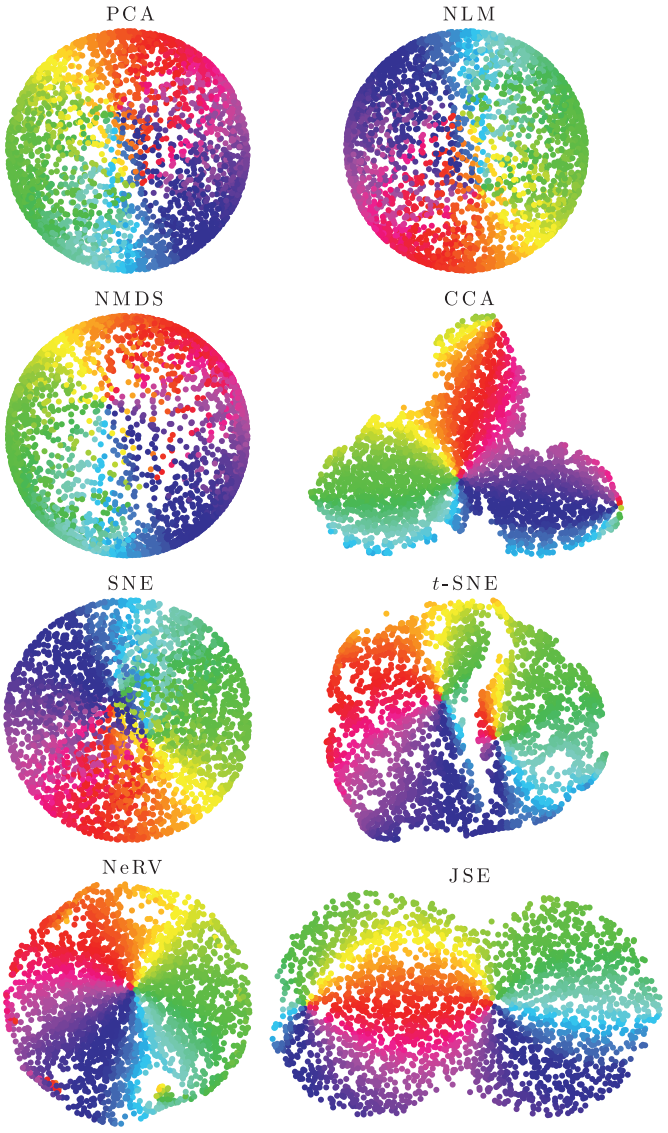

Fig. 2. Embeddings of the spherical shell with eight different DR methods.

$\pi_i = \lambda_i^{-2}$. Thanks to the chain rule, we can write

$$\frac{\partial B(\boldsymbol{\sigma}_i)}{\partial \pi_i} = \sum_j \frac{\partial b(\sigma_{ij})}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial \pi_i}. \tag{16}$$

The second factor in each term develops into

$$\frac{\partial \sigma_{ij}}{\partial \pi_i} = \sum_k \frac{\partial \sigma_{ij}}{\partial \gamma_{ik}} \frac{\partial \gamma_{ik}}{\partial \pi_i} = \sigma_{ij} \left(-\frac{\partial \gamma_{ij}}{\partial \pi_i} + \sum_k \sigma_{ik} \frac{\partial \gamma_{ik}}{\partial \pi_i} \right), \tag{17}$$

where $\gamma_{ij} = \gamma(\delta_{ij}/\lambda_i) = \pi_i \delta_{ij}^2/2$ and therefore

$$\frac{\partial \gamma_{ij}}{\partial \pi_i} = \delta_{ij}^2/2. \tag{18}$$

Eventually, the final update rule for the precision is

$$\pi_i^{(t+1)} = \pi_i^{(t)} + \frac{2(B_0 - B(\boldsymbol{\sigma}_i))}{\sum_j \sigma_{ij} \frac{\partial b(\sigma_{ij})}{\partial \sigma_{ij}} \left(-\delta_{ij}^2 + \sum_k \sigma_{ik} \delta_{ik}^2 \right)}. \tag{19}$$

To avoid negative precisions, an absolute value operator can be applied to the right-hand side of the update. Since $B(\boldsymbol{\sigma}_i) \propto H(\boldsymbol{\sigma}_i)$ for SNE and all its considered variants, we have $b(\sigma_{ij}) \propto \sigma_{ij} \ln(\sigma_{ij})$ and

$$\frac{\partial b(\sigma_{ij})}{\partial \sigma_{ij}} \propto \ln(\sigma_{ij}) + 1 \ . \tag{20}$$

For the initialization, we suggest $\pi_i^{(0)} = \delta_{ik}^{-2}$, where k is the index of the Kth shortest distance and K is the integer closest to the specified perplexity.

4.2. Optimization of the low-dimensional coordinates

Once the baselines are equalized and all bandwidths λ_i are determined, the next step is to minimize the cost function with respect to $\mathbf{X}$. Using the chain rule, the partial derivative with respect to the coordinates in the LD space can be written as

$$\frac{\partial E}{\partial \mathbf{x}_h} = \sum_i \frac{\partial C(\mathbf{s}_i; \boldsymbol{\sigma}_i)}{\partial \mathbf{x}_h} = \sum_{i,j,p,q} c'(s_{ij}; \sigma_{ij}) \frac{\partial s_{ij}}{\partial g_{pq}} \frac{\partial g_{pq}}{\partial \mathbf{x}_h}, \tag{21}$$

where $c'(u; v)$ denotes the derivative of $c(u; v)$ with respect to u.

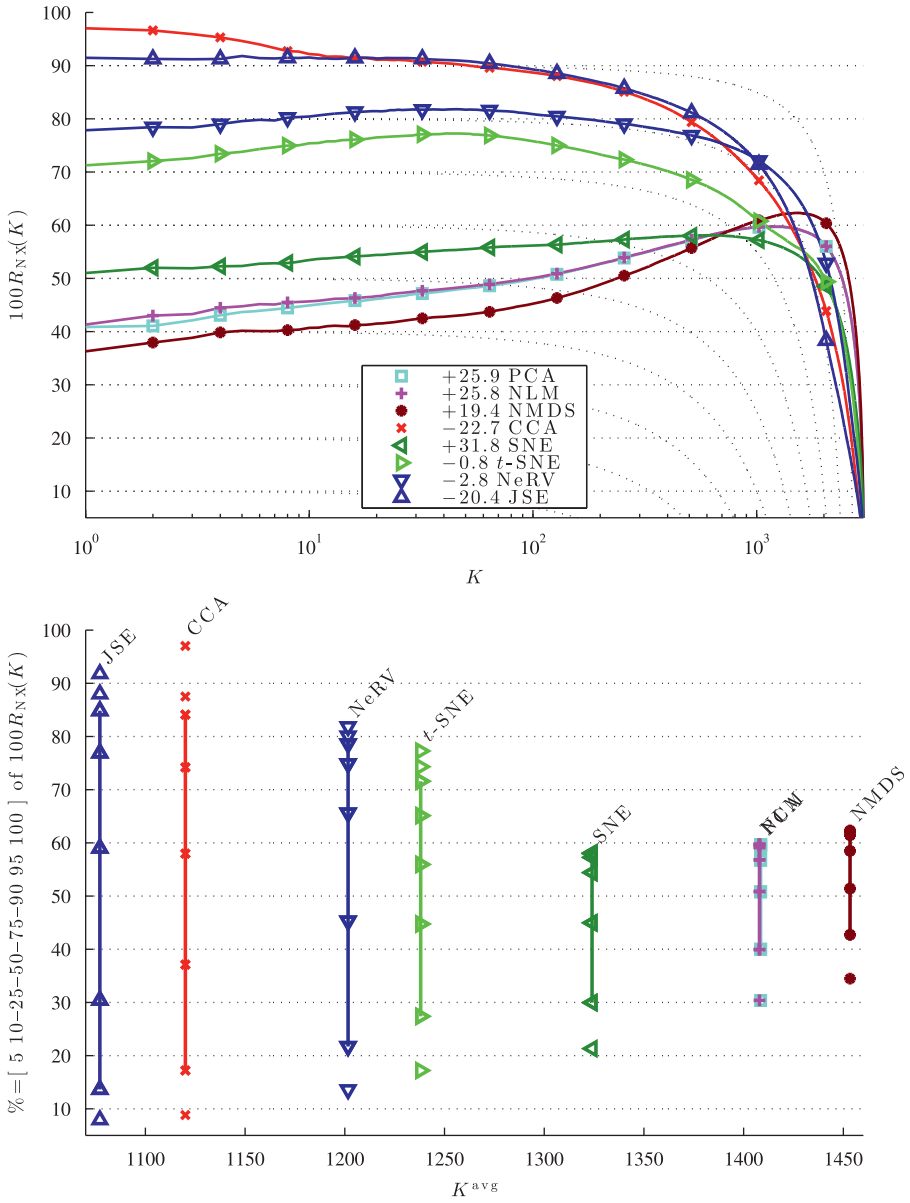

Fig. 3. Quantitative assessment of the spherical shell embeddings shown in Fig. 2. Each curve in the top diagram is associated with an embedding and reflects its improvement over a random embedding on various scales (0 means no improvement, 100 means perfection). The second diagram summarizes each curve by decorrelating pure performance (ordinate) and the average size of the best preserved K-ary neighborhoods (in abscissa).

The second factor in each term develops into

$$\frac{\partial s_{ij}}{\partial g_{pq}} = \begin{cases} \dfrac{\exp(-g_{pq})^2 - \exp(-g_{pq})\sum_k \exp(-g_{pk})}{\left(\sum_k \exp(-g_{pk})\right)^2} & \text{if } p=i\neq j=q \\ \dfrac{\exp(-g_{pq})\exp(-g_{pj})}{\left(\sum_k \exp(-g_{pk})\right)^2} & \text{if } p=i\neq j\neq q \\ 0 & \text{otherwise} \end{cases}$$
$$= \begin{cases} (s_{pq}-1)s_{pq} & \text{if } p=i\neq j=q \\ s_{pj}s_{pq} & \text{if } p=i\neq j\neq q \\ 0 & \text{otherwise} \end{cases}. \quad (22)$$

The third factors are given by

$$\frac{\partial g_{pq}}{\partial \mathbf{x}_h} = g'(d_{pq})\frac{\partial d_{pq}}{\partial \mathbf{x}_h}. \quad (23)$$

If the distances d_{pq} in the LD space are Euclidean, we can further write

$$\frac{\partial g_{pq}}{\partial \mathbf{x}_h} = \begin{cases} \dfrac{g'(d_{hq})}{d_{hq}}(\mathbf{x}_h - \mathbf{x}_q) & \text{if } p=h \\ \dfrac{g'(d_{ph})}{d_{ph}}(\mathbf{x}_h - \mathbf{x}_p) & \text{if } q=h \\ \mathbf{0} & \text{otherwise} \end{cases}. \quad (24)$$

Since $\partial s_{ij}/\partial g_{pq}$ is null if $p\neq i$ and $\partial g_{pq}/\partial \mathbf{x}_h$ is null if both i and q are different from h, we can simplify the partial derivative of the cost function into

$$\frac{\partial E}{\partial \mathbf{x}_h} = \sum_{j,q} c'(s_{hj};\sigma_{hj})\frac{\partial s_{hj}}{\partial g_{hq}}\frac{\partial g_{hq}}{\partial \mathbf{x}_h} + \sum_{i,j} c'(s_{ij};\sigma_{ij})\frac{\partial s_{ij}}{\partial g_{ih}}\frac{\partial g_{ih}}{\partial \mathbf{x}_h}, \quad (25)$$

$$\frac{\partial E}{\partial \mathbf{x}_h} = \sum_q \left(\sum_j c'(s_{hj};\sigma_{hj})\frac{\partial s_{hj}}{\partial g_{hq}}\right)\frac{\partial g_{hq}}{\partial \mathbf{x}_h} + \sum_i \left(\sum_j c'(s_{ij};\sigma_{ij})\frac{\partial s_{ij}}{\partial g_{ih}}\right)\frac{\partial g_{ih}}{\partial \mathbf{x}_h}. \quad (26)$$

After renaming h into i, i and q into j, and j into k, we can expand all partial derivatives and we obtain

$$\frac{\partial E}{\partial \mathbf{x}_i} = \sum_j (\mathbf{x}_i - \mathbf{x}_j)\frac{g'(d_{ij})}{d_{ij}} s_{ij}\left(-c'(s_{ij};\sigma_{ij}) + \sum_k s_{ik}c'(s_{ik};\sigma_{ik})\right)$$
$$+ \sum_j (\mathbf{x}_i - \mathbf{x}_j)\frac{g'(d_{ij})}{d_{ij}} s_{ji}\left(-c'(s_{ji};\sigma_{ji}) + \sum_k s_{jk}c'(s_{jk};\sigma_{jk})\right).$$

If we define

$$w_{ij} = s_{ij}\left(-c'(s_{ij};\sigma_{ij}) + \sum_{k=1}^{N} s_{ik}c'(s_{ik};\sigma_{ik})\right), \quad (27)$$

then we end up with

$$\frac{\partial E}{\partial \mathbf{x}_i} = \sum_{j=1}^{N}(w_{ij}+w_{ji})\frac{g'(d_{ij})}{d_{ij}}(\mathbf{x}_i - \mathbf{x}_j). \quad (28)$$

This partial derivative provides a search direction that can be plugged in many gradient-based optimization techniques. Most of them work in an iterative fashion, with several successive updates of an initial guess. Newton and quasi-Newton techniques being too demanding in terms of memory consumption, a simplified but still generic update can be written as

$$\mathbf{x}_i^{(t+1)} = \mathbf{x}_i^{(t)} - \mu_i^{(t)}\frac{\partial E}{\partial \mathbf{x}_i}, \quad (29)$$

where $\mu_i^{(t)}$ is a step size. In order to accelerate convergence, each step size should ideally be the quotient of an adaptive gain factor divided by the magnitude of the second derivative $\partial^2 E/\partial \mathbf{x}_i^2$. In the spirit of Newton's method, this amounts to a diagonal, positive semidefinite approximation of the Hessian matrix, like in Sammon's nonlinear mapping [7]. The magnitude of the second derivatives in the denominators contribute to a steep falloff of the cost function in the first iterations, whereas the adaptive numerators, initialized to one, are intended to compensate for the incomplete Hessian.

4.2.1. Type 1 mixture of KL divergences

NeRV minimizes $E(\mathbf{X};\Xi,\boldsymbol{\lambda},\kappa) = \sum_i \mathrm{D}^{\kappa}_{\mathrm{KLt1}}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$. This amounts to instantiating the generic cost function in (15) with the terms

$$c(s_{ij};\sigma_{ij}) = (\kappa-1)\sigma_{ij}\ln(s_{ij}) + \kappa s_{ij}\ln(s_{ij}/\sigma_{ij}), \quad (30)$$

whose derivative with respect to s_{ij} is

$$c'(s_{ij};\sigma_{ij}) = (\kappa-1)\sigma_{ij}/s_{ij} + \kappa(1+\ln(s_{ij}/\sigma_{ij})). \quad (31)$$

Therefore it yields

$$w_{ij} = (1-\kappa)(\sigma_{ij}-s_{ij}) + \kappa s_{ij}(\ln(\sigma_{ij}/s_{ij}) + \mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i||\boldsymbol{\sigma}_i)). \quad (32)$$

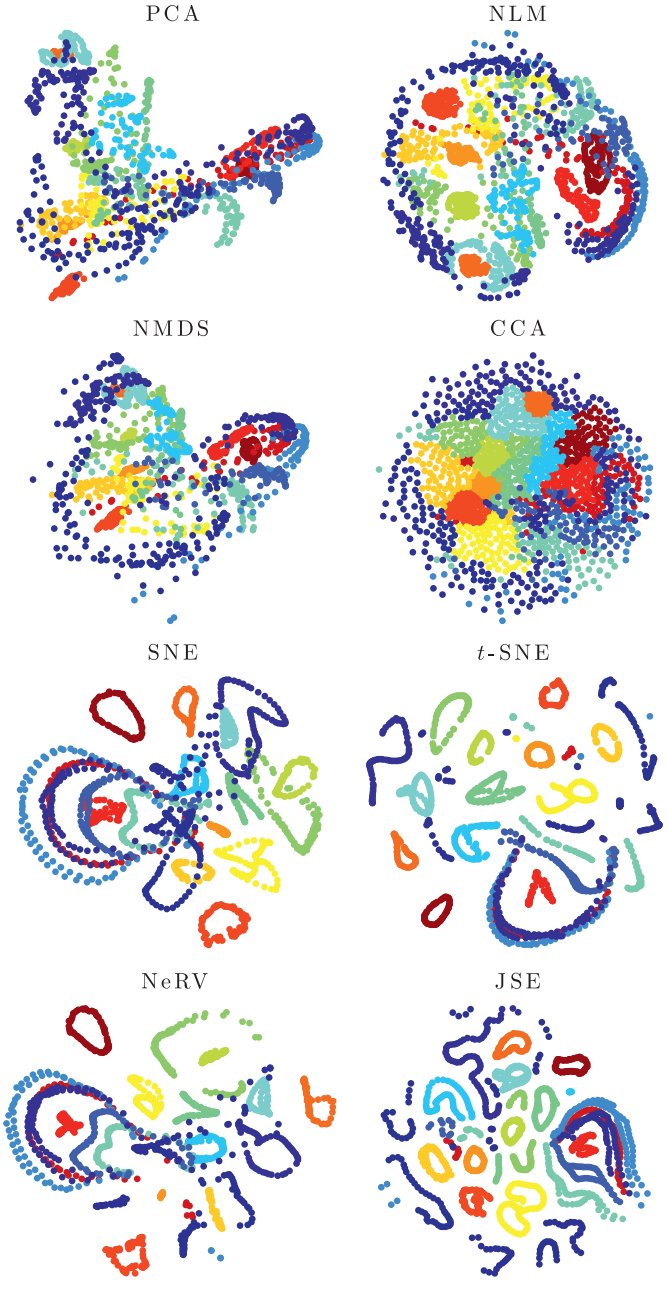

Fig. 4. Embeddings of the COIL-20 image bank with eight different DR methods.

The last formula is valid for NeRV, SNE, and t-SNE. In SNE and t-SNE, we have $\kappa=0$ and therefore $w_{ij}=\sigma_{ij}-s_{ij}$. In NeRV and SNE, we have

$$g(d_{ij})=d_{ij}^2/2 \quad \text{and} \quad \frac{g'(d_{ij})}{d_{ij}}=1. \tag{33}$$

In t-SNE, we have

$$g(u)=\frac{m+1}{2}\ln(1+u^2/m) \quad \text{and} \quad \frac{g'(d_{ij})}{d_{ij}}=\frac{m+1}{m+d_{ij}^2}. \tag{34}$$

4.2.2. Type 2 mixture of KL divergences

JSE minimizes $E(\mathbf{X};\boldsymbol{\Xi},\boldsymbol{\lambda},\kappa)=\sum_i \mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$. This amounts to instantiating the generic cost function in (15) with the terms

$$c(s_{ij};\sigma_{ij})=\frac{1}{\kappa-1}\sigma_{ij}\ln(z_{ij})+\frac{1}{\kappa}s_{ij}\ln(s_{ij}/z_{ij}), \tag{35}$$

where $z_{ij}=\kappa\sigma_{ij}+(1-\kappa)s_{ij}$. The derivative of $c(s_{ij};\sigma_{ij})$ with respect to s_{ij} is

$$c'(s_{ij};\sigma_{ij})=\frac{1}{\kappa}\ln(s_{ij}/z_{ij}), \tag{36}$$

which yields

$$w_{ij}=\frac{s_{ij}}{\kappa}\left(\ln(z_{ij}/s_{ij})+\mathrm{D}_{\mathrm{KL}}(s_i||z_i)\right). \tag{37}$$

Using l'Hôpital's rule, the limit for κ tending to zero is $w_{ij}=\sigma_{ij}-s_{ij}$, as expected. In JSE, $\mathrm{D}_{\mathrm{KLt2}}^{\kappa}(\boldsymbol{\sigma}_i||\mathbf{s}_i)$ is used in conjunction with $g(u)=u^2/2$. Therefore, $g'(d_{ij})/d_{ij}=1$. The next subsection justifies why t-distributed similarities are not necessary in JSE.

4.2.3. Interpretation as a force-directed layout

Some nonlinear DR methods and graph embedding techniques [36,37] stem from analogies with mechanical or electromagnetic systems. Each datum is then considered as a mass or a charged particle, which interacts with other data by means of springs or electromagnetic forces. These techniques proceed by

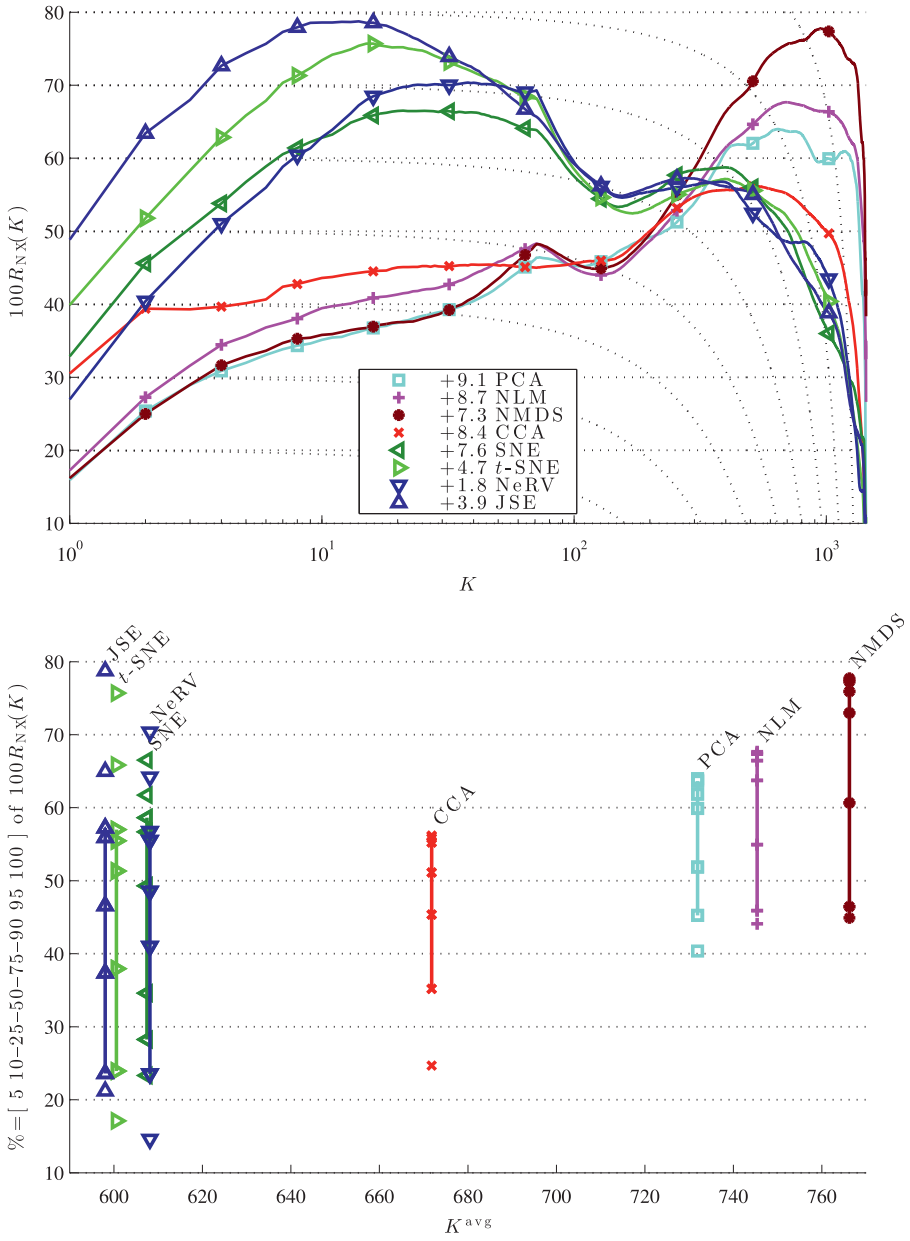

Fig. 5. Quantitative assessment of the COIL-20 embeddings shown in Fig. 4.

transposing the observed mechanical or electromagnetic system in a low-dimensional space, where they let it evolve and reach a state of equilibrium in which all attractive and repulsive forces cancel each other. The result is usually referred to as a force-directed layout. Within this framework, the cost function can be interpreted as free energy to be minimized. Similarly, the partial derivative in (28) can be seen to some extent as the sum of all forces applied to $\mathbf{x}_i$ by all other points $\mathbf{x}_j$, for $1 \le j \le N$. Unit vector $(\mathbf{x}_i - \mathbf{x}_j)/d_{ij}$ indicates the direction of each force, whereas $(w_{ij}+w_{ji})g'(d_{ij})$ gives its magnitude. For the sake of simplicity and without loss of generality, let us consider the relative magnitude $(w_{ij}+w_{ji})g'(d_{ij})/d_{ij}$ and its constituting asymmetric contributions denoted by $F_{ij} = w_{ij}g'(d_{ij})/d_{ij}$.

In the case of SNE, the terms of the relative magnitude turn out to be the similarity differences, namely, $F_{ij} = \sigma_{ij} - s_{ij}$. This means that the force between $\mathbf{x}_i$ and $\mathbf{x}_j$ vanishes only if $\sigma_{ij} = s_{ij}$ and $\sigma_{ji} = s_{ji}$. In contrast, in t-SNE, we have

$$F_{ij} = (\sigma_{ij} - s_{ij}) \frac{m+1}{(m+d_{ij})}. \tag{38}$$

The second factor modulates the force magnitude, which decreases when d_{ij} grows. For a large distance value, this provides an alternative possibility to nearly cancel the force.

In NeRV, the terms of the relative magnitude are given by

$$F_{ij} = (1-\kappa)(\sigma_{ij} - s_{ij}) + \kappa s_{ij}(\ln(\sigma_{ij}/s_{ij}) + \mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i \| \boldsymbol{\sigma}_i)). \tag{39}$$

In this case, the force vanishes only if $\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i \| \boldsymbol{\sigma}_i) = 0$ and $\mathrm{D}_{\mathrm{KL}}(\mathbf{s}_j \| \boldsymbol{\sigma}_j) = 0$. Notice however that the second term of F_{ij} is multiplied by s_{ij}, which decreases when d_{ij} grows, like the modulation factor in t-SNE. This means that NeRV is in an intermediate position between SNE and t-SNE, with a modulation factors that fully plays its role only if $\kappa = 1$.

In JSE, the terms of the relative magnitude are given by

$$F_{ij} = \frac{s_{ij}}{\kappa}(\ln(z_{ij}/s_{ij}) + \mathrm{D}_{\mathrm{KL}}(\mathbf{s}_i \| \mathbf{z}_i)). \tag{40}$$

Like in t-SNE, they are modulated with a factor that decreases when d_{ij} grows, except in the limit case $\kappa = 0$.

One may suppose that the presence of a modulation factor is a desirable feature in a nonlinear DR method. This conjecture has already been verified in the case of DR methods based on distance preservation [12,38]. At least, it explains why Demartines' curvilinear component analysis [39] outperforms Sammon's nonlinear mapping [7]. The presence of a modulation factor can also account for the superiority of t-SNE over SNE, which precisely lacks such a factor. With only a partial modulation, one can expect NeRV to perform somewhere in between, depending on the value of κ. Eventually, JSE benefits from a full modulation like t-SNE and yields comparable or even improved experimental results, such as detailed below.

5. Experiments and results

This section aims at evaluating experimentally the embedding performances of the two different types of divergence mixtures. For this purpose, several data sets are used (Section 5.1). Quality assessment is achieved with recent rank-based criteria (Section 5.2). The divergence mixtures are compared to each other and to four other standard NLDR methods (Section 5.3). Eventually, the results are presented and discussed (Section 5.4).

5.1. Data sets

The experiments involve three data sets. The first one contains 3000 points that are uniformly sampled from the surface of a sphere, as illustrated in the upper left corner of Fig. 1. In all representations of this spherical shell, the rainbow colors are constant along the longitudes and join together at the North and South poles. The goal is to re-embedded this locally two-dimensional manifold in a two-dimensional space, such as a planisphere. In this academic exercise, the main difficulty is essentially to succeed in cutting the manifold in the most appropriate way, which is not an easy task for all NLDR methods. Without any cut, the embedding from three to two dimensions entails a necessary superimposition of two hemispheres.

The second data set is the COIL-20 image bank [40]. It contains 72 gray-level images of 20 different objects, as shown in the right of Fig. 1. The 72 images correspond to 4-degree rotations around each object; nine of them are represented in Fig. 1. All images are square, with height and width of 128 pixels. The images are converted into 128^2-dimensional vectors that are fed into the

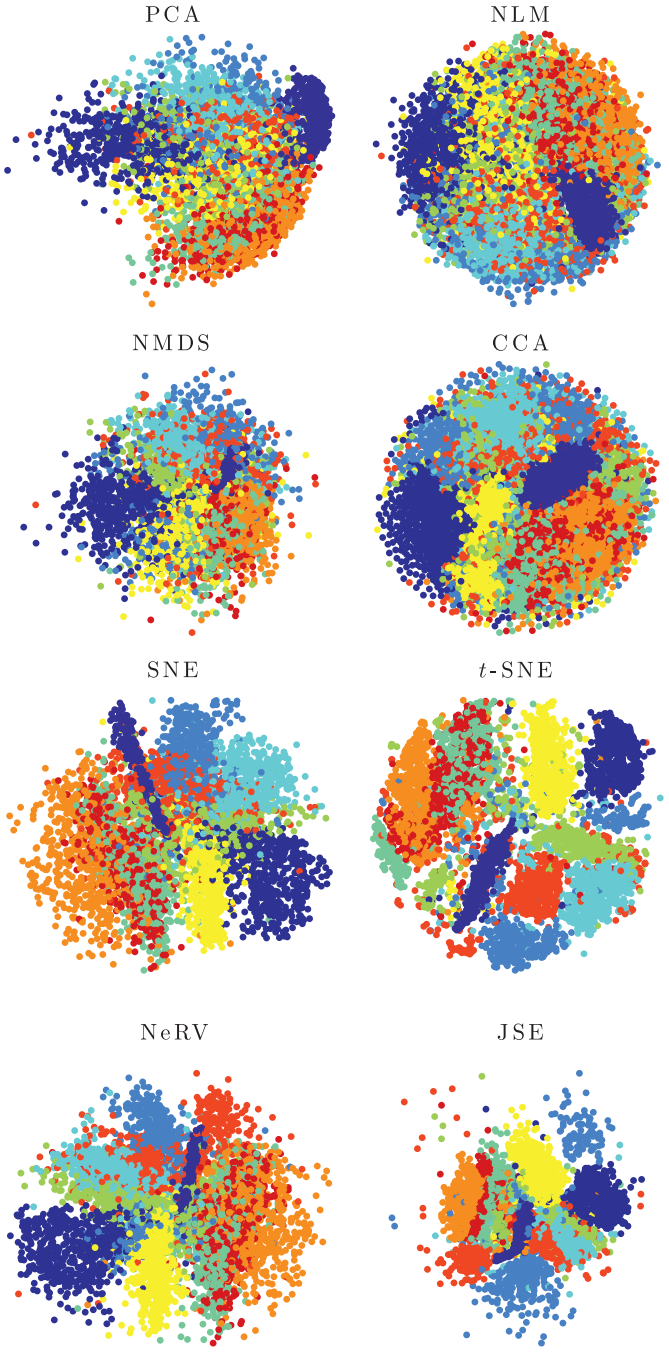

Fig. 6. Embeddings of the MNIST image banks with eight different DR methods.

various NLDR methods, without any further preprocessing (no feature selection nor any linear projection). The two most prominent characteristics of this data set are its very high dimensionality and the presence of 20 one-dimensional manifolds [31]. Each of them is topologically equivalent to a ring, the only degree of freedom being the rotation angle. The goal is to get a two-dimensional representation of the 1440 points.

The third data set is a random subset of the MNIST image bank [41]. It contains 6000 gray-level images of scanned handwritten digits (out of 60 000). There are about 600 images of each of the ten digits; a few of them are shown in the lower left corner of Fig. 1. Like in the previous data set, each 28-by-28 image is vectorized and fed into the various NLDR methods without any other processing, in order to get a two-dimensional representation. The main characteristics of this data set are its high dimensionality and the presence of ten clusters that might overlap, due to the resemblance between some 4 and 9 or some 3 and 8, for instance. The intrinsic dimensionality of each cluster is likely to be high, due to noise and the numerous degrees of freedom that affect hand writing.

5.2. Quality assessment

The quality criterion used to assess the various embeddings evaluates the preservation of K-ary neighborhoods [17]. The rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ in the HD space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \le k < j \le N)\}|$, where $|A|$ denotes the cardinality of set A. Similarly, the rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ in the LD space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \le k < j \le N)\}|$. The K-ary neighborhoods of $\boldsymbol{\xi}_i$ and $\mathbf{x}_i$ are the sets defined by $\nu_i^K = \{j : 1 \le \rho_{ij} \le K\}$ and $n_i^K = \{j : 1 \le r_{ij} \le K\}$, respectively. A first performance index can be written as

$$Q_{\mathrm{NX}}(K) = \sum_{i=1}^{N} \frac{|\nu_i^K \cap n_i^K|}{KN}. \tag{41}$$

This criterion varies between 0 and 1 and measures the average normalized agreement between corresponding K-ary neighborhoods in the HD and LD spaces. If the coranking matrix [17,42] is defined as $\mathbf{Q} = [q_{kl}]_{1 \le k,l \le N-1}$ with $q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|$, then we can

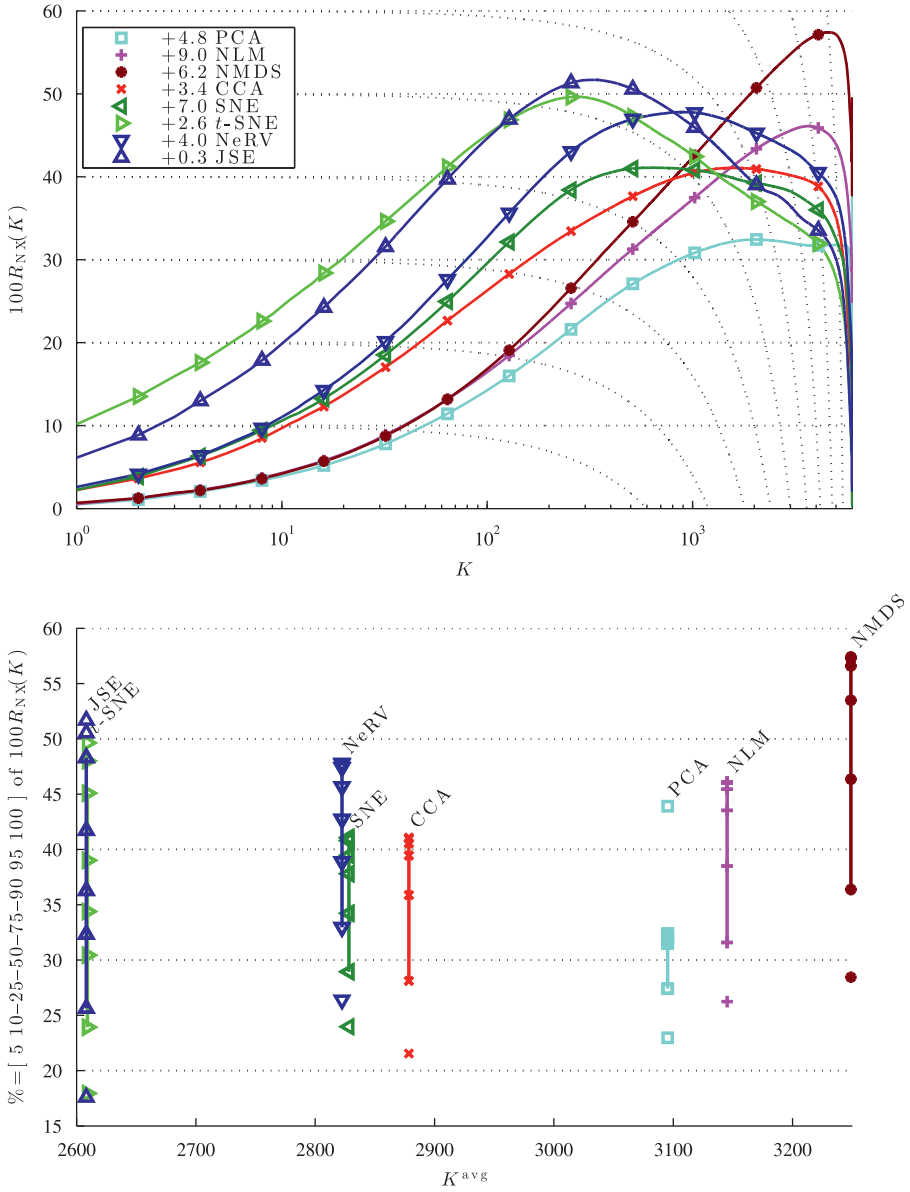

Fig. 7. Quantitative assessment of the MNIST embeddings shown in Fig. 6.

rewrite $Q_{NX}(K)$ as

$$Q_{NX}(K) = \sum_{1 \le k \le K} \sum_{1 \le l \le K} \frac{q_{kl}}{KN}. \quad (42)$$

Hence, $Q_{NX}(K)$ counts in the upper left K-by-K block of $\mathbf{Q}$ both the preserved ranks (on the main diagonal) and the within-neighborhood permutations (on each side of the diagonal). The coranking matrix also allows us to refine the quality assessment with a second criterion, defined as

$$B_{NX}(K) = \left(\sum_{1 \le k \le K} \sum_{1 \le l < k} \frac{q_{kl}}{KN} \right) - \left(\sum_{1 \le k \le K} \sum_{k < l \le K} \frac{q_{kl}}{KN} \right). \quad (43)$$

The sign of $B_{NX}(K)$ indicates whether the majority of rank errors is located above or below the main diagonal of the upper left K-by-K block. This criterion allows us to distinguish between two types of errors [17]: neighboring points in the HD space that are pulled away in the LD space (neighborhood extrusions, $B_{NX}(K) < 0$) and non-neighbors in the HD space that are erroneously represented close to each other in the LD space (neighborhood intrusions, $B_{NX}(K) > 0$). This distinction exists in other (pairs of) quality indices [43]. It stems from an analogy with false positives and false negatives, which also motivates the use of dual divergences in NeRV [24].

In each of the following experiments, we compare several embeddings. For each of them we report in a first diagram the curve given by

$$R_{NX}(K) = \frac{(N-1)Q_{NX}(K) - K}{N-1-K}, \quad (44)$$

for $1 \le K \le N-2$. A value of zero for this criterion corresponds to a random embedding ($Q_{NX}(K) \approx K/(N-1)$) [17], whereas 1 means perfect K-ary neighborhood agreement ($Q_{NX}(K) = 1$). In the legend of each curve, we also provide the average of $100B_{NX}(K)$, that is, $B_{NX}^{avg} = (100/(N-1)) \sum_{K=1}^{N-1} B_{NX}(K)$. It indicates which kind of neighborhood errors dominates in the embedding (neighborhood intrusions or extrusions). See Fig. 3 for an example. Notice the dotted isolevels for both $R_{NX}(K)$ (horizontal) and $Q_{NX}(K)$ (curved), as well as the logarithmic scale for the abscissa, allowing for an easier inspection of local neighborhoods.

In a second diagram, we try to provide a synthetic representation of each $R_{NX}(K)$ curve. Considering each curve as a set of points with coordinates $(K, R_{NX}(K))$, our goal is to visually separate the overall performance level, given by $R_{NX}(K)$ alone, from scale information, that is, how $R_{NX}(K)$ is distributed along the K-axis. In order to reflect the overall performance level, we compute the percentiles 5, 10, 25, 50, 75, 90, 95, and 100 of $\{R_{NX}(K)\}_{1 \le K \le N-2}$, whereas the weighted average

$$K^{avg} = \frac{\sum_{K=1}^{N-2} K R_{NX}(K)}{\sum_{K=1}^{N-2} R_{NX}(K)} \quad (45)$$

indicates where the gross mass of each $R_{NX}(K)$ curve is distributed along the K-axis. We plot these percentiles vertically at the abscissa indicated by K^{avg}. Hence, this second diagram allows a quick visual comparison of all curves (the higher the better; local preservation to the left, global preservation to the right).

5.3. Methods

Eight DR methods are compared in the experiments. The first and oldest one is principal component analysis (PCA) [3], which is equivalent to Torgerson–Gower classical metric multidimensional scaling (MDS) [4,5,44]. The linear projection along the principal directions is found by spectral decomposition of the covariance matrix (or the Gram matrix for classical MDS). The second method is Sammon's nonlinear mapping [7], which is a nonlinear variant of stress-based MDS. The embedding is computed by gradient descent, with a diagonal approximation of the Hessian matrix to accelerate convergence. The third method is Shepard–Kruskal nonmetric MDS (NMDS) [45,46], which combines gradient descent and isotonic regression in each iteration. The fourth method is yet another variant of nonlinear stress-based MDS, namely, curvilinear component analysis (CCA) [39]. It relies on a specific optimization technique, very close in spirit to stochastic gradient descent. All three nonlinear variants of MDS minimize the following generic stress function,

$$E(\mathbf{X}; \Xi) = \sum_{j \ne i} b_{ij} (\phi(\delta_{ij}) - d_{ij})^2, \quad (46)$$

where δ_{ij} and d_{ij} are the pairwise Euclidean distances in the HD and LD spaces, respectively. For Shepard–Kruskal MDS, weight b_{ij} is equal to 1 and the positive and monotonic function ϕ is determined by isotonic regression. For NLM and CCA, ϕ is the identity function. In Sammon's NLM, $b_{ij} = \delta_{ij}^{-1}$. In Demartines' CCA, $b_{ij} = H(\ell - d_{ij})$, where ℓ is a neighborhood radius and H is a step function. To avoid poor local minima, ℓ slowly decreases after each iteration in the specific optimization procedure of CCA. Quite clearly, $H(\ell - d_{ij})$ corresponds to a (binary) modulation factor such as discussed in Section 4.2.3.

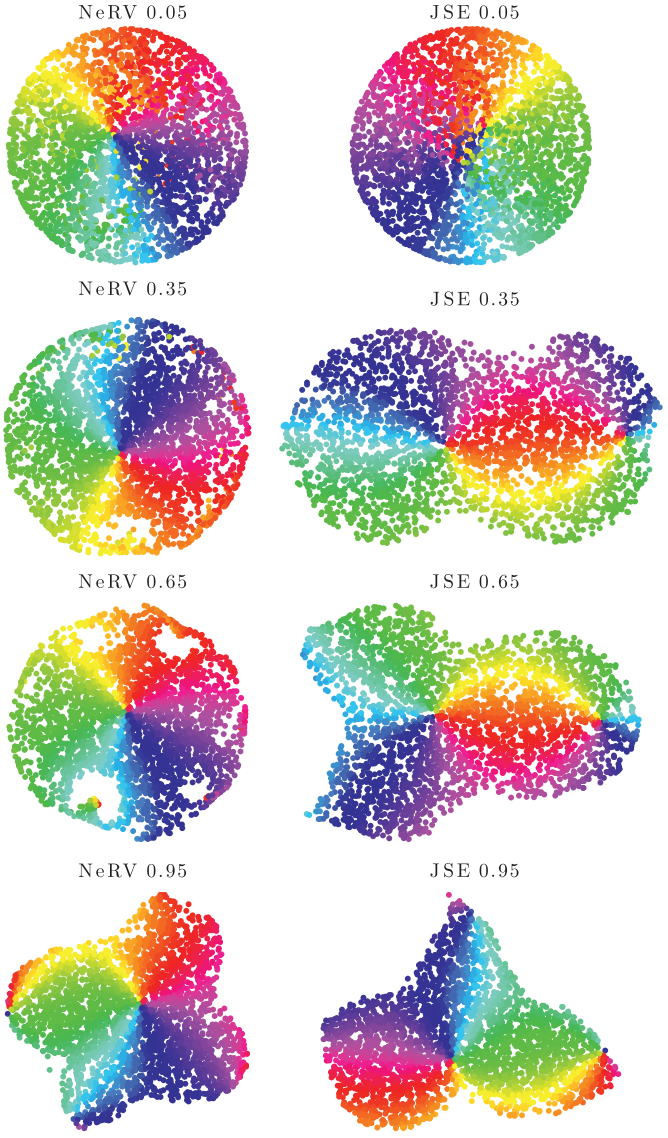

Fig. 8. Embeddings of the spherical shell with type 1 and 2 mixtures of KL divergence (NeRV and JSE, respectively), with a varying value of κ (mentioned after the method name).

The last four methods are based on similarity preservation. They are SNE [23], t-SNE [18], NeRV [24] (type 1 mixture of KL divergences), and JSE (type 2 mixture of KL divergences, our proposal). The cost functions of these four methods are minimized by gradient descent with a diagonal approximation of the Hessian matrix to accelerate convergence, like in Sammon's NLM. Information about the magnitude of the second derivatives significantly speeds up convergence and simplifies the initialization and update of adaptive gain factors in the gradient descent. Like in CCA, poor local minima are avoided by running first the method with a few higher perplexity values than the targeted one. The highest perplexity value is $N/2$, where N is the data set size, whereas its final value is $N/20=150$ for the spherical shell, $N/20=72$ for the COIL-20 images (72 is the cluster size), and $N/20=300$ for the MNIST digits (300 is half the cluster size, some cluster overlap being expected).

All eight methods are deterministic, in the sense that they always sort the input data set in a unique way and start with a PCA initialization of the embedding. Therefore, repeated runs even with shuffled data sets lead exactly to the same outcome.

Due to space and readability constraints, the experiments are limited to the eight methods mentioned above. Comparisons between other methods and t-SNE or NeRV can be found in [18,24].

5.4. Results

Two different experiments are reported and discussed. The first one is a comparison between the eight competing methods applied on the three data sets. In the case of the type 1 and 2 mixtures of divergences, κ is equal to one half. The divergences are thus symmetric with respect to $\boldsymbol{\sigma}_i$ and $\mathbf{s}_i$. The influence of balancing parameter κ is studied in a second experiment, where only the type 1 and 2 mixtures are compared.

5.4.1. All methods

Fig. 2 shows the 2D embeddings of the spherical shell with the eight considered methods.

A quick glance shows that the methods follow two very different strategies to embed the sphere. Four methods (PCA,

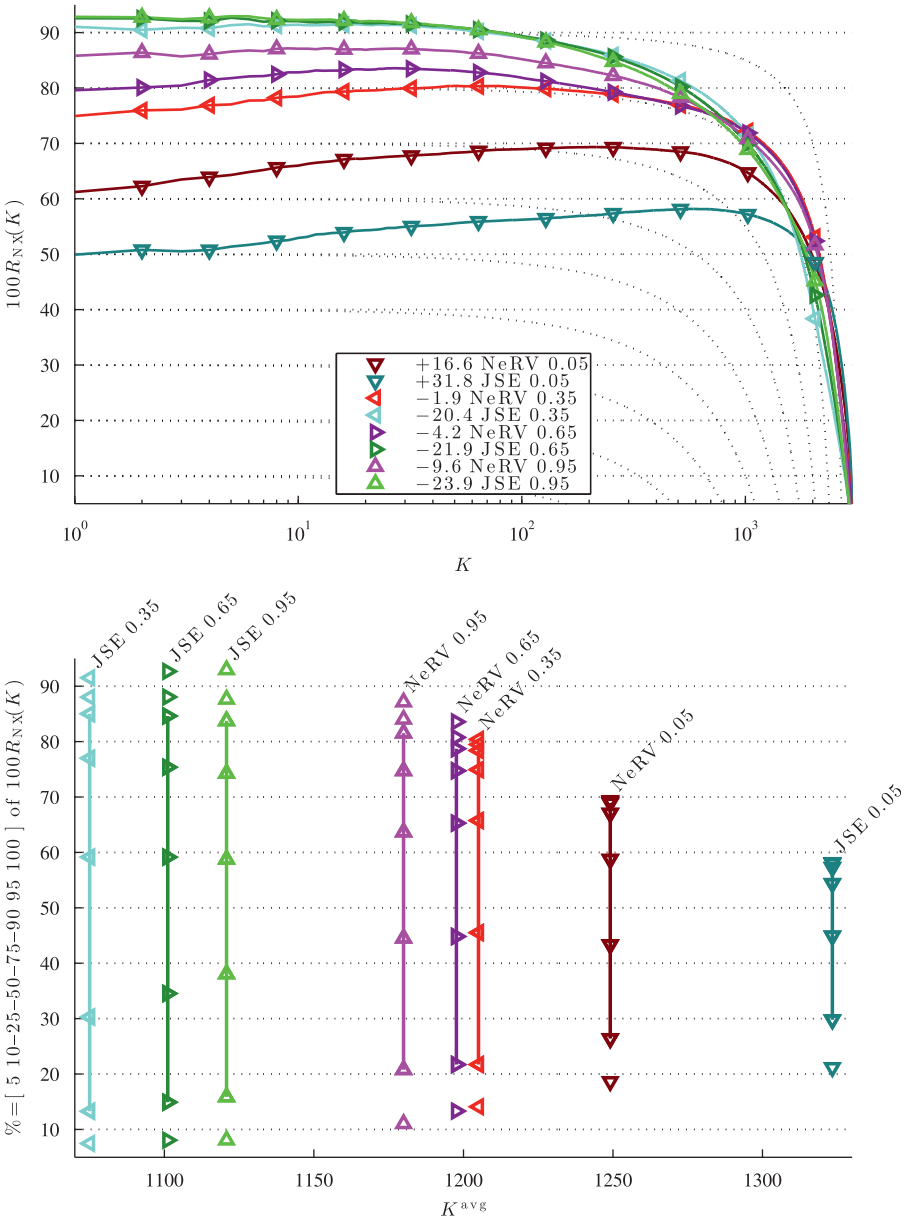

Fig. 9. Quantitative assessment of the spherical shell embeddings shown in Fig. 8.

NLM, NMDS, SNE) just squash the sphere onto the 2D plane, while the four others (CCA, t-SNE, NeRV, and JSE) cut the sphere and unfold it. Without any surprise, these last four methods are precisely those benefitting from a modulation factor in the force-directed interpretation developed in Section 4.2.3. Since any cut can be characterized by long, stretched distances in the LD space, the modulation factor is low or null and the patches that are torn apart undergo a very weak attraction. All terms of the gradient that would otherwise contribute to 'heal' the cut have a negligible magnitude compared to those ensuring the cohesion with closeby neighbors.

The result of PCA looks quite natural, since any linear projection of the sphere lead to the superimposition of two hemispheres. For the NLM, NMDS, and SNE, the impossibility to cut the manifold stems from the high penalty associated with any cut. In contrast, in CCA, t-SNE, NeRV, and JSE, this cost is much lower, thanks to their modulation factor.

The capability to tear a manifold comes however at a high price. It introduces many local minima in the cost function, since in the case of the sphere many different cuts can lead to almost equivalent results. In this respect, one sees that CCA and JSE produces the two most satisfying embeddings, with neat cuts and a good rendering of the uniform point density. This can be confirmed by looking at the quality curves in Fig. 3. Both CCA and JSE score very high for the preservation of small neighborhoods. They are closely followed by NeRV and then t-SNE. The curves corresponding to 'squashing' methods show a different shape. Because they basically embed two hemispheres on top of each other, these methods are totally unable to preserve small neighborhoods. On the other hand, they render pretty well the global, spherical structure of the shell. Thanks to isotonic regression, NMDS can modify the distances measured in the HD space and it achieves the best score for K larger than 1650.

The first diagram in Fig. 3 also indicates the value of $B_{\mathrm{NX}}^{\mathrm{avg}}$ for each method in the legend, in front of its name. The methods achieving the highest $R_{\mathrm{NX}}(K)$ values for small neighborhoods have a negative $B_{\mathrm{NX}}^{\mathrm{avg}}$, which reveals their propensity to tear and cut. Conversely, $B_{\mathrm{NX}}^{\mathrm{avg}}$ is positive for methods that squash the sphere and therefore have worse results in terms of small neighborhood preservation.

The COIL-20 image bank is a much more difficult problem than the spherical shell, due to the very high dimensionality of the vectorized images. As already shown in the literature [18,24,31], one can expect DR methods based on similarity preservation to largely outperform the older ones. Fig. 4 confirms this intuition. Only SNE, t-SNE, NeRV, and JSE yield embeddings that faithfully reveal the intrinsically unidimensional structure of all twenty manifolds. As can be seen in Fig. 5, JSE brings the best quantitative result for small-size neighborhoods; t-SNE, NeRV and SNE follow. The four other methods that do not rely on similarities fail to preserve small neighborhoods. Like with the sphere, they are instead quite good at representing the global arrangement of the data set, except CCA. In this respect, NMDS works the best, followed by the NLM and PCA. In contrast with the sphere, embedding the COIL-20 data set does not require any significant manifold cut or tear and all methods therefore yield a positive $B_{\mathrm{NX}}^{\mathrm{avg}}$ value. Nevertheless, the methods that best preserve small neighborhoods are those having the lowest $B_{\mathrm{NX}}^{\mathrm{avg}}$ values. The second diagram in Fig. 5 shows that CCA performs poorly on all scales, likely because it progressively overlook the global structure as its parameter ℓ decreases during the optimization process.

The MNIST and COIL-20 image banks share almost the same characteristics: a very high dimensionality and the presence of clusters. All four DR methods that do not rely on similarities yield highly cluttered representations of the ten clusters corresponding to the ten digits, as shown in Fig. 6. Separation between the various clusters is much better rendered with methods using similarities. The quality curves in Fig. 7 provide a quantitative assessment of the eight embeddings. Considering intra-cluster neighborhoods ($K \lesssim 600$), the ranking from best to worst is: JSE, t-SNE, NeRV, SNE, CCA, NMDS, the NLM, and PCA. As to larger neighborhood sizes ($K \approx 3000$), the list becomes: NMDS, the NLM, NeRV, CCA, SNE, JSE, t-SNE, and PCA. All values of $B_{\mathrm{NX}}^{\mathrm{avg}}$ are positive; the ranking from low to high: JSE, t-SNE, CCA, NeRV, PCA, NMDS, SNE, and NLM.

5.4.2. *Type 1 and 2 mixtures with varying κ*

In this second experiment, only two methods are considered: similarity preservation with cost functions defined as either

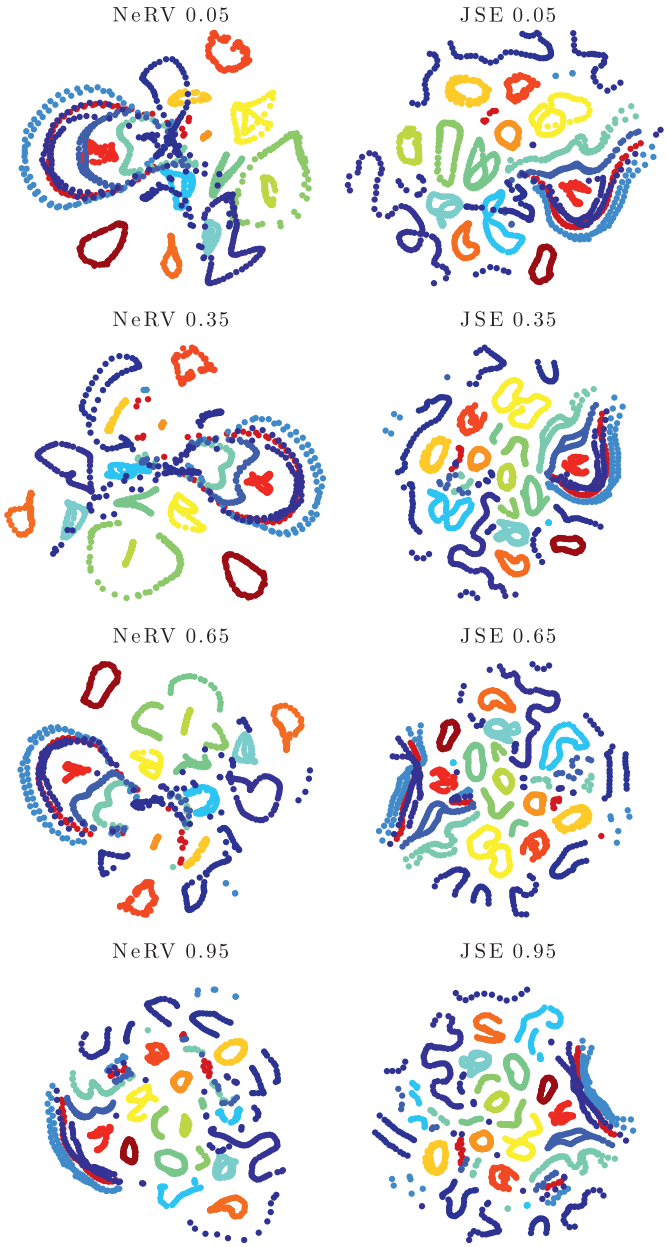

Fig. 10. Embeddings of the COIL-20 image bank with type 1 and 2 mixtures of KL divergence (NeRV and JSE, respectively), with a varying value of κ (mentioned after the method name).

type 1 or type 2 mixtures of KL divergences, namely, NeRV and JSE.

Fig. 8 shows the embeddings of the spherical shell for NeRV and JSE, with κ equal to 0.05, 0.35, 0.65, and 0.95. If $\kappa=0$, then both mixtures degenerate into the single KL divergence used in SNE. The corresponding embedding can be found in Fig. 2, as well as those of NeRV and JSE for $\kappa=0.5$. As can be intuitively expected, the lowest value of κ in this experiment ($\kappa=0.05$) yields embeddings that closely resemble that of SNE, for both mixture types. The sphere is squashed onto the 2D plane without any cut. Once κ grows, the modulated terms in the gradient of NeRV gain in importance and consequently tears and cuts progressively appear in the embeddings. The same effect can be observed with JSE, although it is much faster, since, unlike NeRV, JSE has no non-modulated terms in its gradient. Already at $\kappa=0.35$, JSE nicely unfolds the sphere without any twist. NeRV requires κ to be close to 0.95 to reach a similar result. Fig. 9 quantitatively confirms this analysis. For small neighborhood sizes, the curves of JSE for κ equal to 0.35, 0.65, and 0.95 are the highest and lie very close to each other. The corresponding curves for NeRV are somewhat lower and show an increased sensitivity to κ. Looking at the values of $B_{\mathrm{NX}}^{\mathrm{avg}}$ in the legend, this sensitivity is obvious: a higher κ translates into a lower $B_{\mathrm{NX}}^{\mathrm{avg}}$. The range of $B_{\mathrm{NX}}^{\mathrm{avg}}$ for NeRV is between 16.6 and -9.6, whereas JSE travels between 31.8 and -23.9.

In contrast with the spherical shell, embedding the COIL-20 image bank does not require important tears and cuts. The embeddings in Fig. 10 show little variations and a weak sensitivity to κ. For the highest values of κ, one can however see that the strings associated with the three cars are better unfolded and better separated from the two 'car-looking' boxes. Like with the

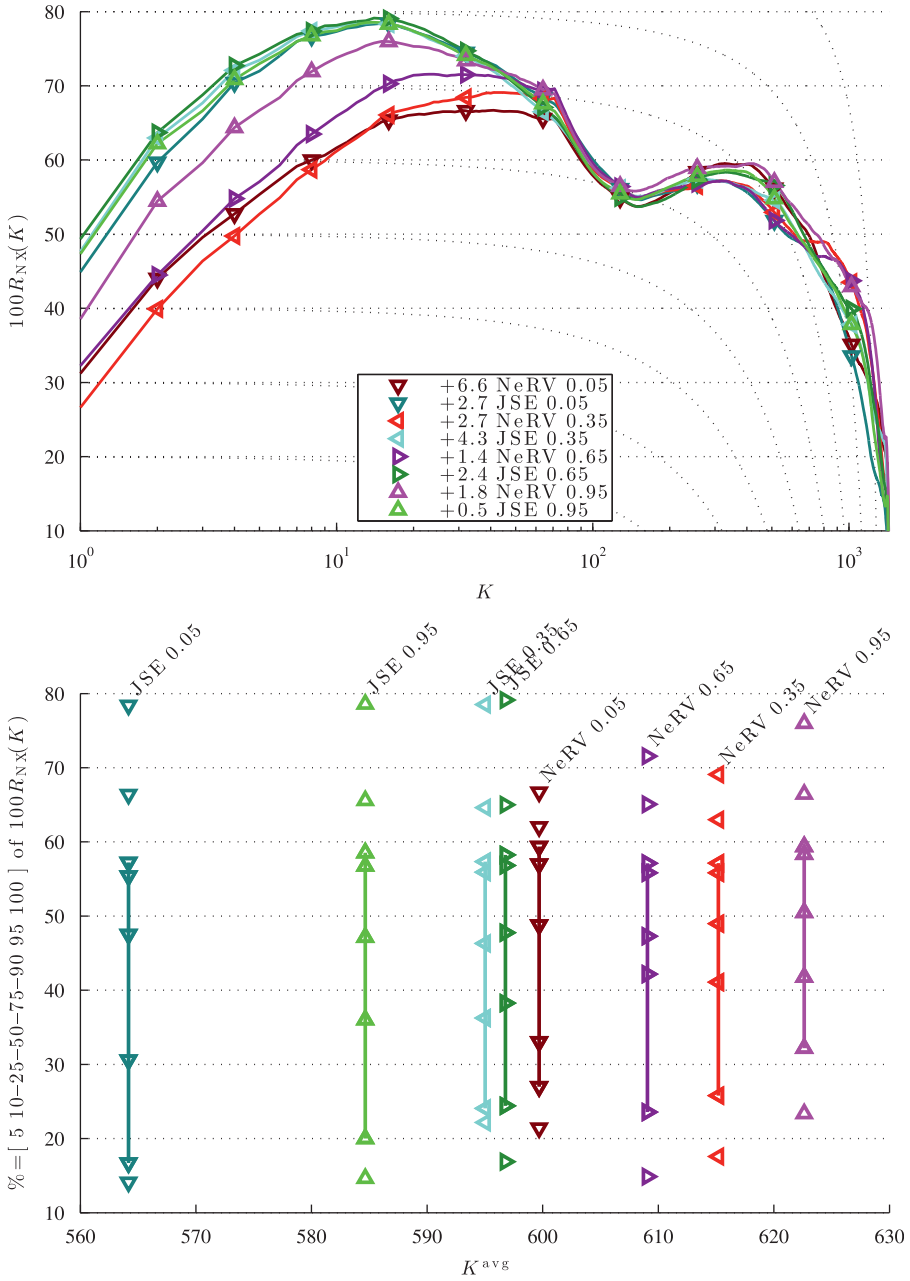

Fig. 11. Quantitative assessment of the COIL-20 embeddings shown in Fig. 10.

sphere, JSE attains this result with a lower value of κ and NeRV is slightly more sensitive to κ. On the quantitative side, in Fig. 11, NeRV and JSE barely differ from each other, although JSE systematically outperforms NeRV for neighborhoods smaller than the cluster size. Like with the sphere, κ directly influences the value of $B_{\mathrm{NX}}^{\mathrm{avg}}$. A larger κ reduces the risk of erroneous neighborhood intrusions, increases the propensity to tear them, and globally leads to higher $R_{\mathrm{NX}}(K)$ values for small K.

Changing κ in the case of the MNIST image bank leads to the embeddings shown in Fig. 12. When κ grows, NeRV better separates the ten clusters, whereas JSE achieves this result already with low values of κ. With both methods, the numerous outliers present in this data set (abnormally bold and deformed digits) are more faithfully represented. From a quantitative point of view and for neighborhood sizes smaller than the cluster size, JSE outperforms NeRV, which however gives a better rendering of the global inter-cluster arrangement. This is shown in Fig. 13, where the lower sensitivity to κ of JSE is also visible. Like with the other data sets, the larger κ is in NeRV, the lower $B_{\mathrm{NX}}^{\mathrm{avg}}$, and the higher $R_{\mathrm{NX}}(K)$ for small K.

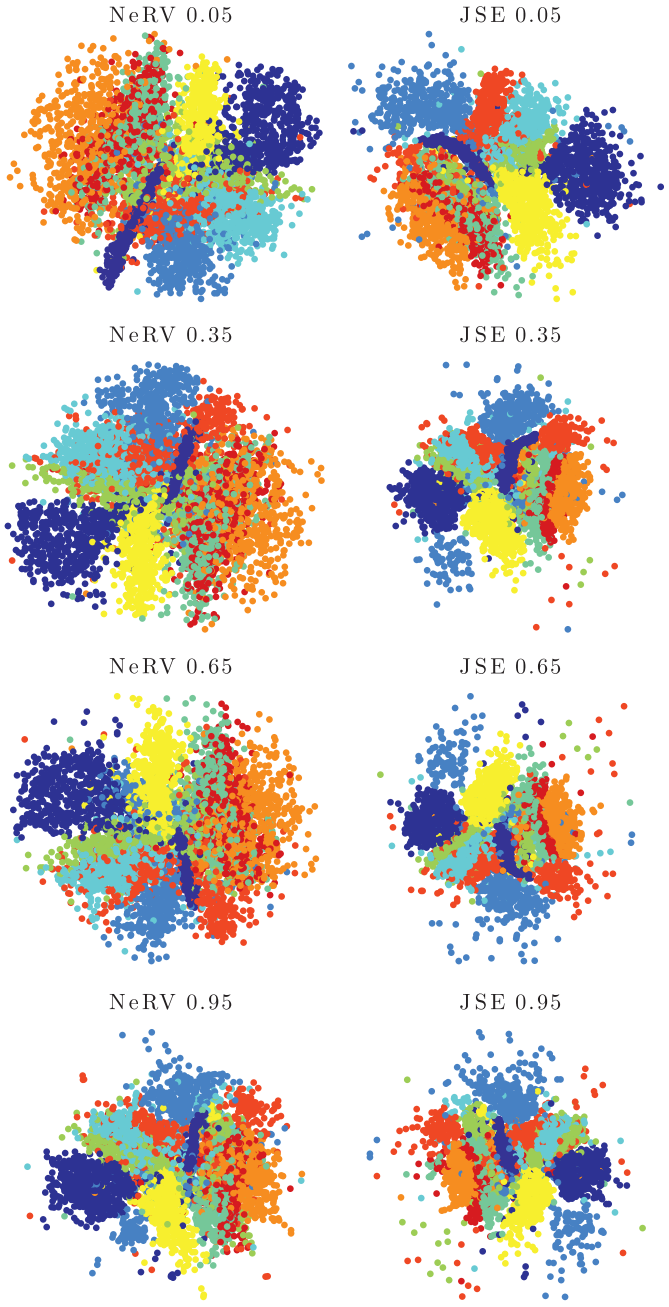

Fig. 12. Embeddings of the MNIST image bank with type 1 and 2 mixtures of KL divergence (NeRV and JSE, respectively), with a varying value of κ (mentioned after the method name).

5.5. General discussion

The experimental results reported above confirm the intuition developed in Section 4.2.3: the cost function that measures the similarity mismatch in SNE and its variants has a significant impact on the embedding quality. In particular, the cost function must have flexible penalties and the ability to 'cut some slack' when necessary. Putting equal effort in the preservation of all neighborhood relationships is hopeless and leads to mediocre results for all of them. On the other hand, sacrificing completely a few of these relationships, with for instance some manifold tears and cuts, allows the vast majority of them to be better preserved. In Section 4.2.3, the low penalty associated with tears and cuts takes a visible form when looking at the gradient of the various cost functions, which can be interpreted as the sum of attractive or repulsive forces applied to each pair of points in the embedding.

There are basically three strategies to modulate the gradient:

- t-SNE uses non-identical similarity definitions in the HD and LD spaces. This introduces an additional factor in each term of its gradient, compared to SNE. The main drawback of this approach is that it also implicitly induces an exponential transformation between the HD and LD distances. In other words, t-SNE cannot yield isometric embeddings of a linear manifold.
- NeRV blends two dual divergences, but only one of them leads to a modulated gradient. NeRV directly inherits the other divergence from SNE.
- JSE relies on a slightly more complicated mixtures of divergences, which leads to a fully modulated gradient like in t-SNE. Unlike the latter, however, JSE gains this advantage while keeping identical similarity definitions in the HD and LD spaces.

6. Conclusion and perspectives

Nonlinear DR methods based on similarity preservation occupy a more and more enviable place in the state of the art. Although the specific similarity definitions used in these methods is certainly one key of their success, thanks to their immunity to norm concentration in HD spaces, other aspects such as the cost function that measures the similarity mismatch cannot be overlooked. Switching from a simple asymmetric KL divergence to parameterized mixtures of KL divergences improves the DR results, which then become comparable to those of the best SNE variants around, like t-SNE. This shows that other approaches than the use of heavy-tailed similarities in the LD space work very well too, without having to cope with the inconsistency of non-identical similarity definitions in the HD and LD spaces. In the near future, we will extend the framework of type 1 and 2 mixtures to $\alpha\beta$-divergences [47]. This family includes generalized KL divergence as a special case, as well as the sum of squared differences and the Itakura–Saito divergence.

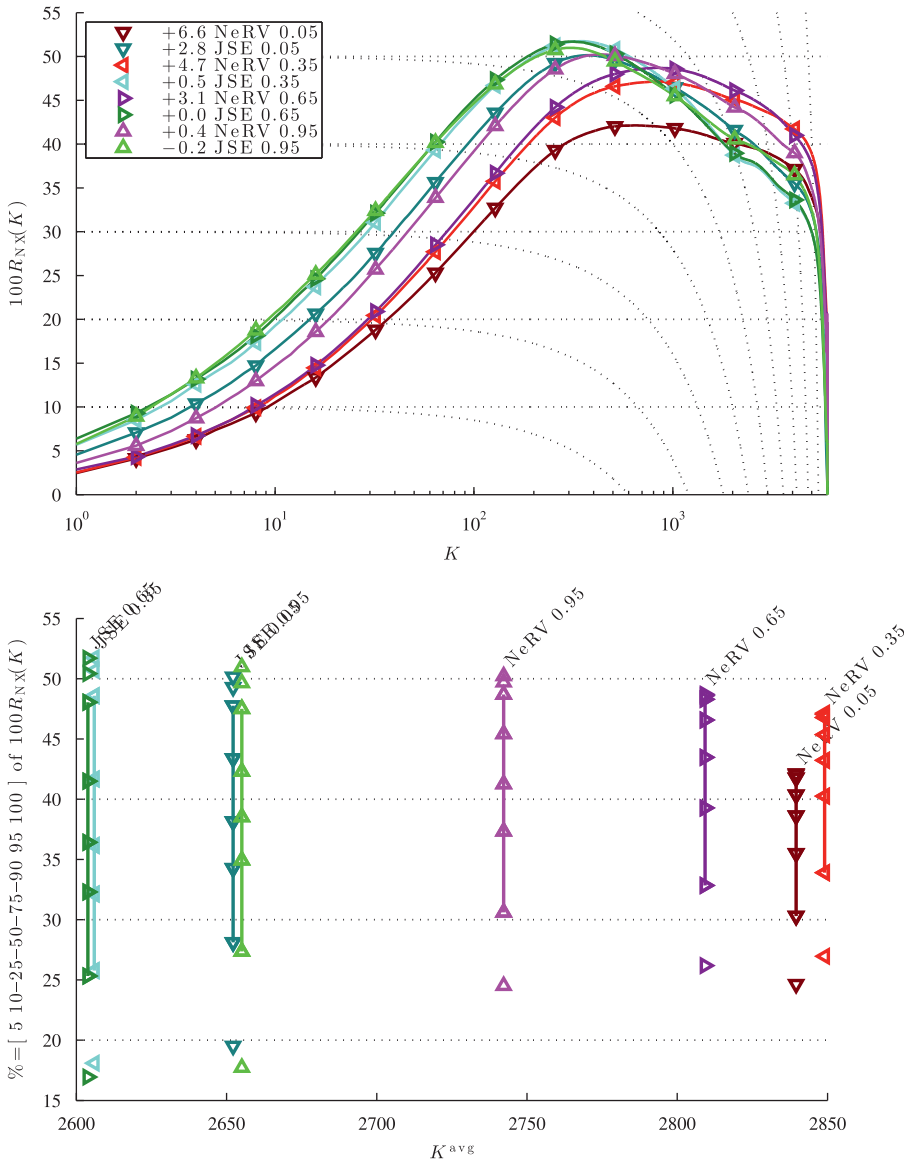

Fig. 13. Quantitative assessment of the MNIST embeddings shown in Fig. 12.

References

[1] K. Pearson, On lines and planes of closest fit to systems of points in space, Philos. Mag. 2 (1901) 559–572.

[2] H. Hotelling, Analysis of a complex of statistical variables into principal components, J. Educ. Psychol. 24 (1933) 417–441.

[3] I. Jolliffe, Principal Component Analysis, Springer-Verlag, New York, NY, 1986.

[4] G. Young, A. Householder, Discussion of a set of points in terms of their mutual distances, Psychometrika 3 (1938) 19–22.

[5] W. Torgerson, Multidimensional scaling, I: theory and method, Psychometrika 17 (1952) 401–419.

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

[7] J. Sammon, A nonlinear mapping algorithm for data structure analysis, IEEE Trans. Comput. CC-18 (5) (1969) 401–409.

[8] J. Tenenbaum, Mapping a manifold of perceptual observations, in: M. Jordan, M. Kearns, S. Solla (Eds.), Advances in Neural Information Processing Systems (NIPS) 1997, vol. 10, MIT Press, Cambridge, MA, 1998, pp. 682–688.

[9] J. Tenenbaum, V. de Silva, J. Langford, A global geometric framework for nonlinear dimensionality reduction, Science 290 (5500) (2000) 2319–2323.

[10] J. Lee, A. Lendasse, N. Donckers, M. Verleysen, A robust nonlinear projection method, in: M. Verleysen (Ed.), Proceedings of ESANN 2000, Eighth European Symposium on Artificial Neural Networks, D-Facto Publications, Bruges, Belgium, 2000, pp. 13–20.

[11] J. Lee, A. Lendasse, M. Verleysen, Curvilinear distances analysis versus isomap, in: M. Verleysen (Ed.), Proceedings of ESANN 2002, 10th European Symposium on Artificial Neural Networks, d-side, Bruges, Belgium, 2002, pp. 185–192.

[12] J. Lee, M. Verleysen, Curvilinear distance analysis versus isomap, Neurocomputing 57 (2004) 49–76.

[13] M. Belkin, P. Niyogi, Laplacian eigenmaps and spectral techniques for embedding and clustering, in: T. Dietterich, S. Becker, Z. Ghahramani (Eds.), Advances in Neural Information Processing Systems (NIPS 2001), vol. 14, MIT Press, 2002.

[14] S. Roweis, L. Saul, Nonlinear dimensionality reduction by locally linear embedding, Science 290 (5500) (2000) 2323–2326.

[15] B. Nadler, S. Lafon, R. Coifman, I. Kevrekidis, Diffusion maps spectral clustering and eigenfunction of Fokker–Planck operators, in: Y. Weiss, B. Schölkopf, J. Platt (Eds.), Advances in Neural Information Processing Systems (NIPS 2005), vol. 18, MIT Press, Cambridge, MA, 2006.

[16] J. Lee, M. Verleysen, Nonlinear Dimensionality Reduction, Springer, 2007.

[17] J. Lee, M. Verleysen, Quality assessment of dimensionality reduction: rank-based criteria, Neurocomputing 72 (7–9) (2009) 1431–1443.

[18] L. van der Maaten, G. Hinton, Visualizing data using t-SNE, J. Mach. Learn. Res. 9 (2008) 2579–2605.

[19] B. Schölkopf, A. Smola, K.-R. Müller, Nonlinear component analysis as a kernel eigenvalue problem, Neural Comput. 10 (1998) 1299–1319, also available as technical report 44 at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany, December 1996.

[20] J. Ham, D. Lee, S. Mika, B. Schölkopf, A kernel view of the dimensionality reduction of manifolds, in: 21th International Conference on Machine Learning (ICML-04), 2004, pp. 369–376, also available as technical report TR-102 at Max Planck Institute for Biological Cybernetics, Tübingen, Germany, 2003.

[21] M. Saerens, F. Fouss, L. Yen, P. Dupont, The principal components analysis of a graph, and its relationships to spectral clustering, in: Proceedings of the 15th European Conference on Machine Learning (ECML 2004), 2004, pp. 371–383.

[22] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, N. Le Roux, Spectral clustering and kernel PCA are learning eigenfunctions, Technical Report. 1239, Département d'Informatique et Recherche Opérationnelle, Université de Montréal, Montréal (Jul. 2003).
[23] G. Hinton, S. Roweis, Stochastic neighbor embedding, in: S. Becker, S. Thrun, K. Obermayer (Eds.), Advances in Neural Information Processing Systems (NIPS 2002), vol. 15, MIT Press, 2003, pp. 833–840.
[24] J. Venna, J. Peltonen, K. Nybo, H. Aidos, S. Kaski, Information retrieval perspective to nonlinear dimensionality reduction for data visualization, J. Mach. Learn. Res. 11 (2010) 451–490.
[25] J. Lee, M. Verleysen, Shift-invariant similarities circumvent distance concentration in stochastic neighbor embedding and variants, in: Proceedings of the International Conference on Computational Science (ICCS 2011), Singapore, 2011, pp. 538–547.
[26] D. François, V. Wertz, M. Verleysen, The concentration of fractional distances, IEEE Trans. Knowl. Data Eng. 19 (7) (2007) 873–886.
[27] R. Bellman, Adaptive Control Processes: A Guided Tour, Princeton University Press, Princeton, NJ, 1961.
[28] D. Donoho, High-Dimensional Data Analysis: The Curse and Blessings of Dimensionality, aide-mémoire for a lecture for the American Mathematical Society "Mathematical Challenges of the 21st Century", 2000.
[29] M. Strickert, S. Teichmann, N. Sreenivasulu, U. Seiffert, High-throughput multi-dimensional scaling (HiT-MDS) for cDNA-array expression data, in: W. Duch (Ed.), Proceedings of the ICANN 2005, Artificial Neural Networks: Biological Inspirations (Part I), Lecture Notes in Computer Science, vol. 3696, Springer, 2005, pp. 625–634.
[30] M. Strickert, N. Sreenivasulu, B. Usadel, U. Seiffert, Correlation-maximizing surrogate gene space for visual mining of gene expression patterns in developing barley endosperm tissue, BMC Bioinformatics 8 (2007) 165.
[31] K. Bunte, S. Haase, M. Biehl, T. Villmann, Stochastic neighbor embedding (SNE) for dimension reduction and visualization using arbitrary divergences, Neurocomputing 90 (2012) 23–45.
[32] A. Cichocki, S.-i. Amari, Families of alpha- beta- and gamma-divergences: flexible and robust measures of similarities, Entropy 12 (2010) 1532–1568.
[33] J. Burbea, C. Rao, On the convexity of some divergence measures based on entropy functions, IEEE Trans. Inf. Theory 28 (3) (1982) 489–495.
[34] J. Lin, Divergence measures based on the Shannon entropy, IEEE Trans. Inf. Theory 37 (1) (1991) 145–151.
[35] J. Lee, M. Verleysen, On the role and impact of the metaparameters in t-distributed stochastic neighbor embedding, in: Y. Lechevallier, G. Saporta (Eds.), Proceedings of the 19th COMPSTAT, Paris (France), 2010, pp. 337–348.
[36] G. Di Battista, P. Eades, R. Tamassia, I. Tollis, Algorithms for drawing graphs: an annotated bibliography, Technical Report. Brown University (Jun. 1994).
[37] G. Di Battista, P. Eades, R. Tamassia, I. Tollis, Graph Drawing: Algorithms for the Visualization of Graphs, Prentice-Hall, 1999.
[38] J. Venna, S. Kaski, Local multidimensional scaling, Neural Networks 19 (2006) 889–899.
[39] P. Demartines, J. Hérault, Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets, IEEE Trans. Neural Networks 8 (1) (1997) 148–154.
[40] S.A. Nene, S.K. Nayar, H. Murase, Columbia object image library (COIL-20), Technical Report. CUCS-005-96, Columbia University, 1996.
[41] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[42] J. Lee, M. Verleysen, Scale-independent quality criteria for dimensionality reduction, Pattern Recognition Lett. 31 (14) (2010) 2248–2257.
[43] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: an experimental study, in: G. Dorffner, H. Bischof, K. Hornik (Eds.), Proceedings of ICANN 2001, Springer, Berlin, 2001, pp. 485–491.
[44] J. Gower, Some distance properties of latent root and vector methods used in multivariate analysis, Biometrika 53 (1966) 325–338.
[45] R. Shepard, The analysis of proximities: multidimensional scaling with an unknown distance function (parts 1 and 2), Psychometrika 27 (1962) 125–140 219–249.
[46] J. Kruskal, Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis, Psychometrika 29 (1964) 1–28.
[47] A. Cichocki, S. Cruces, S.-i. Amari, Generalized alpha-beta divergences and their application to robust nonnegative matrix factorization, Entropy 13 (2011) 134–170.

John Aldo Lee was born in 1976 in Brussels, Belgium. He received the M.S. degree in Applied Sciences (Computer Engineering) in 1999 and the Ph.D. degree in Applied Sciences (Machine Learning) in 2003, both from the Université catholique de Louvain (UCL, Belgium). His main interests are dimensionality reduction, intrinsic dimensionality estimation, clustering, vector quantization, and various aspects of image processing. He is a member of the UCL Machine Learning Group and a Research Associate with the Belgian F.N.R.S. (Fonds National de la Recherche Scientifique). Together with Michel Verleysen, he wrote a monography entitled 'Nonlinear Dimensionality Reduction' published by Springer-Verlag in 2007. His current work aims at developing specific image enhancement techniques for positron emission tomography in the center of Molecular Imaging, Radiotherapy, and Oncology (MIRO).

Emilie Renard was born in 1989 in Belgium. She received the M.S. degree in Applied Mathematics in 2011 from the Université catholique de Louvain (UCL), Belgium. She is now a Ph.D. student in the UCL Machine Learning Group. Her research interests include data mining, supervised and unsupervised dimensionality reduction, as well as data visualization.

Guillaume Bernard was born in 1986 in Belgium. He received an M.S. in Computer Engineering in 2010 from the Université catholique de Louvain, Belgium. As a Ph.D. student in the Machine Learning Group, he works on image segmentation with machine learning methods in the center of Molecular Imaging, Radiotherapy, and Oncology (MIRO).

Pierre Dupont received an M.Sc. in Electrical Engineering in 1988 from the Université catholique de Louvain, Belgium, and a Ph.D. in Computer Science in 1996 from l'Ecole Nationale Supérieure des Télécommunications, Paris, France. From 1988 to 1991, he was a Research Staff Member at the Philips Research Laboratory Belgium. In 1992, he joined the France Telecom R&D Center, Lannion, France as a Research Staff Member. In parallel, he completed a Ph.D. in automatic induction of grammar from data, also known as grammatical inference. In 1996–1997, he was a post-doctoral researcher in the Speech Group of the Computer Science Department at Carnegie Mellon University, Pittsburgh, USA. In 1997–2001, he was Associate Professor in the Computer Science Department of Université Jean Monnet, Saint-Etienne, France. Since September 2001, Pierre Dupont is Professor at the Université catholique de Louvain. His current research interests include novel machine learning methods to tackle real problems arising in computational biology, bio-statistics and medical research, feature selection and dimensionality reduction, high-throughput data analysis and biomarker identification, statistical modeling of sequential processes, hidden Markov models and automata induction. He is also co-founder of DNAlytics, a UCL spin-off offering an analytical support to predictive and personalized medicine.

Michel Verleysen was born in 1965 in Belgium. He received the M.S. and Ph.D. degrees in Electrical Engineering from the Université catholique de Louvain (Belgium) in 1987 and 1992, respectively. He was an invited professor at the Swiss E.P.F.L. (Ecole Polytechnique Fédérale de Lausanne, Switzerland) in 1992, at the Université d'Evry Val d'Essonne (France) in 2001, and at the Université Paris I—Panthéon-Sorbonne from 2002 to 2011, respectively.

He is now Full Professor at the Université catholique de Louvain, and Honorary Research Director of the Belgian F.N.R.S. (National Fund of Scientific Research). He is editor-in-chief of the Neural Processing Letters journal (published by Springer), chairman of the annual ESANN conference (European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning), past associate editor of the IEEE Transactions on Neural Networks journal, and member of the editorial board and program committee of several journals and conferences on neural networks and learning. He was the

chairman of the IEEE Computational Intelligence Society Benelux chapter (2008–2010), and member of the executive board of the European Neural Networks Society (2005–2010). He is author or co-author of more than 250 scientific papers in international journals and books or communications to conferences with reviewing committees. He is the co-author of the scientific popularization book on artificial neural networks in the series «Que Sais-Je?» in French, and of the "Nonlinear Dimensionality Reduction" book published by Springer in 2007. His research interests include machine learning, feature selection, artificial neural networks, self-organization, time-series forecasting, nonlinear statistics, adaptive signal processing, and high dimensional data analysis.