Unsupervised Dimensionality Reduction: Overview and Recent Advances

John A. Lee and Michel Verleysen

***Abstract*— Unsupervised dimensionality reduction aims at representing high-dimensional data in lower-dimensional spaces in a faithful way. Dimensionality reduction can be used for compression or denoising purposes, but data visualization remains one its most prominent applications. This paper attempts to give a broad overview of the domain. Past develoments are briefly introduced and pinned up on the time line of the last eleven decades. Next, the principles and techniques involved in the major methods are described. A taxonomy of the methods is suggested, taking into account various properties. Finally, the issue of quality assessment is briefly dealt with.**

I. Introduction

The interpretation of high-dimensional data remains a difficult task, mainly because human vision is not used to deal with spaces whose dimensionality is higher than three. Part of this inability stems from the curse of dimensionality, a convenient expression that encompasses all weird and unexpected properties of high-dimensional spaces. If visualization is difficult in high-dimensional space, perhaps an (almost) equivalent representation in a lower-dimensional space could improve the readability of data. This is precisely the idea that underlies the field of dimensionality reduction (DR). This domain includes various techniques that are able to construct meaningful data representations in a space of given dimensionality. Beside visualization, other applications of DR are for instance data compression and denoising. Dimensionality reduction can also preprocess data, with the hope that a simplified representation can accelerate any subsequent processing or improve its outcome.

Linear DR is well known, with techniques such as principal component analysis [27] and classical metric multidimensional scaling [79], [62]. On the other hand, nonlinear dimensionality reduction [37] (NLDR) emerged later, with nonlinear variants of multidimensional scaling [57], [32], [59], such as Sammon's nonlinear mapping [52]. Research in NLDR is multidisciplinary and follows many approaches, ranging from artificial neural networks [29], [31], [44], [15], [42] to spectral techniques [55], [61], [48], [2], [19], [73], [76]. If linear DR assumes that data are distributed within or near a linear subspace, NLDR necessitates more complex models. The most generic extension consists in assuming that data are sampled from a smooth manifold. For this reason, modern NLDR is sometimes referred to as manifold learning [53], [73]. Under this hypothesis, one seeks to re-embed the manifold in a space of the lowest possible dimensionality, without modifying its topological properties. In practice, smooth manifolds are difficult to conciliate with the discrete nature of data. In constrast, graph structures have proven to be very useful and tight connections between NLDR and graph embedding [18] exist. Another usual hypothesis is to assume that data are distributed in clusters. Dimensionality reduction methods that emphasize clusters are often closely related to spectral clustering [5], [51], [43], [11].

Obviously, DR has to provide a low-dimensional representation that is meaningful in some sense. Regardless of the model (manifold, graph, clusters), the general idea of DR is to represent similar data items next to each other, while maintaining large distances between those that are dissimilar. In practice, the goal of DR is to preserve as well as possible simple properties such as soft or hard neighborhoods [29], [33], [75], similarities [23], [64], or ranks [57], [32], [45]. An even simpler and very popular solution is to preserve pairwise distances [57], [32], [52], [15], [16], [35], [61], [73]. This approach works indifferently with data that consist of coordinates or pairwise distances. If not all distances are specified, then the problem can elegantly be modeled using a graph, in which edges are present for known entries of the pairwise distance matrix. The edge weights can be binary- or real-valued, depending on the nature of the data. Some NLDR techniques exploit the sparsity of such graphs [61], [35], [48], [1], [73] even if all pairwise distances are available. This allows focusing on small neighborhoods [48], [1], [73] or to approximate geodesic distances [61], [6], [35] with weighted shortest paths. It illustrates the close relationship between NLDR and graph embedding.

This paper is organized as follows. Section II follows the timeline of past developments over more than a century. Section III presents the state of the art, in the form of short technical descriptions of major methods, starting from classical ones to recent discoveries. Next, Section IV attempts to categorize the methods according to various criteria. The important issue of quality assessment is dealt with in Section V. Finally, Section VI draws the conclusions and sketches some perspectives for the near future.

II. Historical background

The analysis of high-dimensional data is certainly not a new concern. The first major breakthrough occurred more than a hundred years ago, in 1901, with the first publications about principal component analysis [46], [26] (PCA), also

J.A.L. (john.lee@uclouvain.be) is with IREC Institute of the Université catholique de Louvain (UCL). M.V. is with the ICTEAM Institute of the UCL and with the SAMM laboratory, Université Paris I Panthéon-Sorbonne, France. J.A.L. and M.V. are with the UCL Machine Learning Group (MLG), Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium. J.A.L. is a Research Associate with the Belgian National Fund of Scientific Research (F.R.S.-FNRS).

978-1-4244-8126-2/10/$26.00 ©2010 IEEE

known as the Karhunen-Loève transform [28], [41]. PCA can be understood in several ways. From a statistical standpoint, it decorrelates variables and allows the selection of those that bear most of the data variance. As an optimization technique, PCA performs a total-least-squares (TLS) linear regression. PCA is also the first spectral DR method, as it projects data along the leading eigenvectors of the sample covariance matrix. Starting from the late 1930s until the early 1950s, several psychometrists extended PCA into classical metric multidimensional scaling [79], [62] (MDS). They gave an alternative way to compute the projections along the principal components, starting from either a Gram matrix or a matrix of pairwise Euclidean distances, instead of the sample covariance.

The 1960s saw the advent of nonmetric MDS [57], [32], in which spectral techniques were replaced with more generic optimization procedures, in order to deal with more complicated cost functions called ‘strain’ or ‘stress’. Nonmetric MDS was the first nonlinear DR method and extended the principle of distance preservation to the use of non-Euclidean metrics. Among several variants, Sammon’s nonlinear mapping (NLM) [52] is still very popular. Sammon’s NLM is a noticible milestone, as it put forward the idea that distance preservation should give more weight to short distances than to long ones. In the late 1980s and early 1990s, with the exception of principal curves [21], most developments in DR were inspired by brain studies (see e.g. [71]) and the boom of artificial neural networks [25], [50], [7]. The most emblematic and popular method in this stream is undoubtedly Kohonon’s self-organizing feature map [29], [47], [30] (SOM), a hybrid method mixing NLDR and vector quantization. Auto-association by means of deep feed-forward networks with a so-called bottleneck layer [31], [44], [63], [17] was an elegant way to achieve NLDR by minimizing TLS error such as in PCA. However, the difficulty to train deep networks with backpropagation prevented their immediate adoption. Artificial neural networks also influenced new developments in topographic mapping [8], [58] and in distance preservation, such as variants of Sammon’s NLM [42], [13] and curvilinear component analysis [15], [16], [22]. The latter method innovated by its ability to tear manifold when needed, which can be very handy to represent manifolds that are spherical or with loops.

The seminal work of Schölkopf et al. [56] about kernel PCA in 1996 set the trend for the next ten years, with a regain of interest in spectral DR methods [54]. Their idea was to apply the kernel trick to classical metric MDS, in such a way that principal components are computed in a so-called feature space. The application of the kernel leads to a Gram matrix in a space of nonlinearly mapped coordinates, without having to define explicitly the transformation. In a memorable issue of Science in December 2000, Isomap [61] was published as another extension of classical metric MDS, in which Euclidean distances where replaced with geodesic distances approximated by the length of shortest paths in a K-nearest-neighbors graph. Stress-based variants of MDS, such as CCA and NLM, utilize the same metric [34], [35]. In the very same issue of Science appeared the first spectral DR method that exploits the (nontrivial) trailing eigenvectors of a sparse matrix, namely locally linear embedding [48] (LLE). LLE minimizes a sum of local reconstruction errors: each datum is approximated by a linear combination of its neighbors in the high-dimensional space and the obtained coefficients are then used to compute its low-dimensional coordinates. Many spectral methods were published in the early and mid 2000s, such as Laplacian eigenmaps [1], Hessian LLE [19], coordination of local models [49], [69], [60], [10], [70], maximum variance embedding [72], [73], and diffusion maps [43], to cite only a few of them.

For a couple of years, there has been a resurgence of soft-computing in the field of NLDR. Although spectral methods come with the guarantee of finding the global optimum of their associated cost function, other more generic optimization procedures such as gradient descent can deal with a broader range of functions. State-of-the-art nonspectral methods are for instance local MDS [66] and auto-associative networks [24] with improved learning techniques especially tailored for so-called deep architectures [3]. Another very promising line of investigation is the preservation of similarities, instead of dissimilarities (that is, distances). Stochastic neighbor embedding [23] and its variants [64], [68] raised the interest towards similarity-based methods. In constrast to distances, similarities naturally emphasize local neighborhood relationships, as they rapidly decrease when data items are far from each other. With a proper choice of the similarity function, this new class of methods proves to be efficient and close to the natural criteria used to assess the quality of DR methods.

III. Dimensionality reduction methods

Let us denote $\Xi = [\boldsymbol{\xi}_i]_{1\leq i\leq N}$ the data set in the original space of representation. The goal of dimensionality reduction methods is to represent the data in a lower-dimensional space, by keeping some of the original properties of the data. Often, it is assumed that the $\boldsymbol{\xi}_i$ lie on a manifold (possibly corrupted by noise), in which case the goal is to preserve some properties of the manifold. The numerous dimensionality reduction methods differ by the properties of the data or the manifold they try to preserve. In the following, the data in the low-dimensional space will be denoted by $\mathbf{X} = [\mathbf{x}_i]_{1\leq i\leq N}$. The set $\mathbf{X}$ is the set of free parameters of the method, which are to be optimized.

A. Variance preservation

Principal component analysis [46], [26], [27] is a linear method, which aims at preserving a maximal fraction of the data set variance. Defining the covariance matrices in the original and low-dimensional spaces by $\mathbf{C}_{\Xi\Xi} = \frac{1}{N}(\Xi - \frac{1}{N}\Xi\mathbf{1}\mathbf{1}^T)^T(\Xi - \frac{1}{N}\Xi\mathbf{1}\mathbf{1}^T)$ and $\mathbf{C}_{\mathbf{XX}} = \frac{1}{N}(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}\mathbf{1}^T)^T(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}\mathbf{1}^T)$ respectively, PCA looks for the global optimum of

$$\min_{\mathbf{X}} \|\mathbf{C}_{\Xi\Xi} - \mathbf{C}_{\mathbf{XX}}\|_2^2 \ , \qquad (1)$$

where $\|\mathbf{A}\|_2 = \sqrt{\operatorname{tr}(\mathbf{A}^T\mathbf{A})}$ denotes the Frobenius norm. The linear data transformation combines translations and a rotation, and can thus be written as $\mathbf{X} = \mathbf{V}(\boldsymbol{\Xi} - \frac{1}{N}\boldsymbol{\Xi}\mathbf{1}\mathbf{1}^T)$. Given a target dimensionality P, the global minimum of (1) is attained when the columns of $\mathbf{V}$ correspond to the leading P eigenvectors of $\mathbf{C}_{\boldsymbol{\Xi}\boldsymbol{\Xi}}$.

B. From inner product preservation to distance preservation

Classical metric multidimensional scaling [79], [62] is the dual method to PCA: instead of involving the sample covariance matrix, it utilizes the Gram matrices of inner products $\mathbf{G}_{\boldsymbol{\Xi}\boldsymbol{\Xi}} = (\boldsymbol{\Xi} - \frac{1}{N}\boldsymbol{\Xi}\mathbf{1}\mathbf{1}^T)^T(\boldsymbol{\Xi} - \frac{1}{N}\boldsymbol{\Xi}\mathbf{1}\mathbf{1}^T)$ and $\mathbf{G}_{\mathbf{X}\mathbf{X}} = (\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}\mathbf{1}^T)^T(\mathbf{X} - \frac{1}{N}\mathbf{X}\mathbf{1}\mathbf{1}^T)$. MDS uses a spectral decomposition to find the global optimum of

$$\min_{\mathbf{X}} \|\mathbf{G}_{\boldsymbol{\Xi}\boldsymbol{\Xi}} - \mathbf{G}_{\mathbf{X}\mathbf{X}}\|_2^2 \ . \tag{2}$$

The low-dimensional coordinates in $\mathbf{X}$ consists of the leading P eigenvectors of $\mathbf{G}_{\boldsymbol{\Xi}\boldsymbol{\Xi}}$, after transposition and scaling by the square root of their associated eigenvalues. It can be shown that PCA and classical metric MDS are equivalent, in the sense that they lead to the same value of $\mathbf{X}$. In contrast to PCA, classical metric MDS works indifferently with coordinates, inner products, or pairwise Euclidean distances, thanks to double centering [62]. If $\boldsymbol{\Delta} = [\delta_{ij}]_{1\leq i,j\leq N}$ denotes the matrix of squared Euclidean distances, left and right multiplications with centering matrix $\mathbf{H} = \mathbf{I} - \frac{1}{N}\mathbf{1}^T\mathbf{1}$ lead to the centered Gram matrix:

$$\begin{aligned}-\frac{1}{2}\mathbf{H}\boldsymbol{\Delta}\mathbf{H} &= -\frac{1}{2}\mathbf{H}(\operatorname{diag}(\boldsymbol{\Xi})^T\mathbf{1} - 2\boldsymbol{\Xi}^T\boldsymbol{\Xi} + \mathbf{1}^T\operatorname{diag}(\boldsymbol{\Xi}))\mathbf{H}\\ &= -\frac{1}{2}\mathbf{H}(-2\boldsymbol{\Xi}^T\boldsymbol{\Xi})\mathbf{H}\\ &= (\boldsymbol{\Xi} - \frac{1}{N}\boldsymbol{\Xi}\mathbf{1}\mathbf{1}^T)^T(\boldsymbol{\Xi} - \frac{1}{N}\boldsymbol{\Xi}\mathbf{1}\mathbf{1}^T) = \mathbf{G}_{\boldsymbol{\Xi}\boldsymbol{\Xi}} \ .\end{aligned}$$

Starting from the second equality shows that double centering can also be applied to $\boldsymbol{\Xi}^T\boldsymbol{\Xi}$.

The formulation of MDS in terms of distances instead of inner products allows a much more intuitive geometrical interpretation. At the expense of replacing the spectral decomposition in classical metric MDS with more general optimization tools such as gradient descent, the cost function that formalizes distance preservation can be refined in many ways. For example, the minimization of Torgerson's strain function $\|\mathbf{G}_{\boldsymbol{\Xi}\boldsymbol{\Xi}} - \mathbf{G}_{\mathbf{X}\mathbf{X}}\|_2^2$ can be replaced with

$$\min_{\mathbf{X}} \sum_{i<j} w_{ij}(\delta_{ij}^2 - d_{ij}^2)^2 \ , \tag{3}$$

where the distances are denoted by $\delta_{ij} = \|\boldsymbol{\xi}_i - \boldsymbol{\xi}_j\|_2$ and $d_{ij} = \|\mathbf{x}_i - \mathbf{x}_j\|_2$. The multiplication of each term of the sum with a weighting factor w_{ij} gives an appreciable flexibility. For instance, one can favor the preservation of small distances, following the intuition that local neighborhood relationships convey more information than loose, remote connections. From a manifold standpoint, neglecting large distances is supposed to allow the data cloud to unfold and to make its embedding easier in a low-dimensional space. The importance given to small distance can also be reinforced by using non-squared distances, which leads to the minimization of the so-called stress function:

$$\min_{\mathbf{X}} \sum_{i<j} w_{ij}(\delta_{ij} - d_{ij})^2 \ . \tag{4}$$

The stress function is the basis of many nonlinear variants of classical metric MDS, such as Sammon's NLM [52]. In this method, w_{ij} is defined to be equal to $1/\delta_{ij}$ and the stress is minimized by a pseudo-Newton optimization procedure. Curvilinear component analysis [16] follows a similar approach, with the noticible difference that $w_{ij} = f(d_{ij}/\sigma)$, where $f : \mathbb{R}^+ \to \mathbb{R}^+$ is a decreasing function of its argument and σ is a neighborhood width. Although at first glance it looks very similar to Sammon's NLM, CCA shows a completely different behavior, due to the dependence of the weights upon the distances in the *low*-dimensional space. This pecularity gives CCA the ability to tear manifolds, which is very handy to embed spherical manifolds or to break and unfold loops in manifolds [36].

More fundamentally, NLDR can be prone to two types of errors: tearing errors (or extrusions) occur when close neighbors are represented far from each other, whereas flattening errors (or intrusions) indicate that remote data items erroneously become close neighbors. (See Section V for more details.) Within this framework, NLM can be shown to tolerate flattening errors, whereas CCA is a variant of MDS that allows tearing errors. These antagonist behaviors are combined in hybrid methods such as Venna's local multidimensional scaling [67], [66], where the weights in the stress function are given by $w_{ij} = \lambda f(d_{ij}/\sigma) + (1-\lambda)f(\delta_{ij}/\sigma)$. Parameter λ controls the balance between the two types of errors.

Another breakthrough in NLDR has been to replace the standard Euclidean norms and distances by other metrics. The most famous example is undoubtedly Isomap [61], which amounts to applying classical metric MDS to a matrix of pairwise geodesic distances. If data are assumed to be sampled from a manifold, geodesic distances are actually measured along the underlying manifold. This metric facilitates the preservation of distances for manifolds that are isometric to a subset of some Euclidean space (the length of straight lines drawn on a sheet of paper is invariant, regardless of the sheet curvature). In practice, geodesic distances are approximated by computing the length of shortest paths in a Euclidean graph corresponding to K-ary neighborhoods or ϵ-balls [6]. Geodesic distances have been used in Sammon's mapping [77] as well as in CCA [35].

In contrast to geodesic versions of NLM and CCA that involve gradient descents, Isomap keeps using a spectral decomposition to compute the embedding coordinates, just like classical metric MDS. This ensures that Isomap finds the global optimum of its cost function. On the other hand, the approximation of geodesic distances leads after double centering to a Gram matrix that is not guaranteed to be positive semidefinite. This issue is addressed in maximum variance unfolding (MVU) [72], [74], [73]. The idea of MVU is to unfold the data cloud by preserving the distances between neighboring points and maximizing all other longer distances. Knowing that geodesic distances are larger or equal to distances as the crow flies, MVU goes one step further than Isomap. Formally, MVU solves

$$\max_{\mathbf{X}} \sum_{i<j} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \ , \tag{5}$$

subject to $\mathbf{1}^T\mathbf{X} = \mathbf{0}$ and $\|\mathbf{x}_i - \mathbf{x}_j\|_2 = \delta_{ij}$ if $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are neighbors. In practice, MVU converts the Euclidean distances into the corresponding inner products and then modifies the resulting Gram matrix by means of semidefinite programming before applying classical metric MDS on it.

C. Distance preservation in feature spaces

Although they rely on classical metric MDS, methods such as Isomap and MVU achieve a nonlinear transformation of the data coordinates. This ability stems from their use of a modified Gram matrix. A pioneering method in this direction is kernel PCA. The idea is to 'kernelize' classical metric MDS, by means of the so-called kernel trick. Mercer kernels are symmetric functions of two arguments that can be rewritten in the form of an inner product. For any Mercer's kernel$k$, the theory allows us to write $k(\boldsymbol{\xi}_i, \boldsymbol{\xi}_i) = \langle \boldsymbol{\Phi}(\boldsymbol{\xi}_i), \boldsymbol{\Phi}(\boldsymbol{\xi}_j)\rangle$, where $\langle\cdot,\cdot\rangle$ denotes an inner product in a so-called feature space. The beauty of the kernel trick is that k implicitly defines the mapping function $\boldsymbol{\Phi}$. In other words, there is no need to map the data into the feature space and to explicitly compute the inner products. Instead, the Gram matrix in the feature space can be built directly by applying the kernel k on pairs of vectors drawn from data set $\boldsymbol{\Xi}$. The main concern about kernel PCA is the choice of an appropriate kernel, which at same time fulfills Mercer's theorem conditions and proves to be good at reducing the data dimensionality. In this respect, methods like Isomap and MVU often perform better than kernel PCA, because the kernel is determined in a data-driven way, either by computing geodesic distances or thanks to semidefinite programming. The kernel that leads to the modified Gram matrix is not known in analytical form but it is geometrically relevant.

Laplacian eigenmaps [2] is another DR method that can be seen as working in a feature space. The idea of Laplacian eigenmaps is to minimize small distances, while constraining the data covariance. Formally, Laplacian eigenmaps use a spectral decomposition to solve

$$\min_{\mathbf{X}} \sum_{i<j} w_{ij} \|\mathbf{x}_i - \mathbf{x}_j\|_2^2 \ , \tag{6}$$

subject to $\mathbf{1}^T\mathbf{X} = \mathbf{0}$, $\mathbf{C}_{\mathbf{XX}} = \mathbf{I}$, and $w_{ij} > 0$ if and only if $\boldsymbol{\xi}_i$ and $\boldsymbol{\xi}_j$ are neighbors. (K-ary neighborhoods or ϵ-balls can be used such as in Isomap.) The low-dimensional coordinates in $\mathbf{X}$ are given by the scaled and transposed trailing nontrivial eigenvectors of the Laplacian matrix, defined as $\mathbf{L} = \mathbf{D} - \mathbf{W}$, where $\mathbf{W} = [w_{ij}]_{1\leq i,j \leq N}$ and $\mathbf{D}$ is diagonal with $d_{ii} = \sum_j w_{ij}$. Several authors have shown that Laplacian eigenmaps amount to applying classical metric MDS to commute-time distances [78], that is, to distances related to random walks in a graph. In that sense, Laplacian eigenmaps can be understood in the same way as kernel PCA, that is, as a method of preserving inner products in a feature space. In the case of Laplacian eigenmaps, the latter is equipped with the inner product associated with commute-time distances.

To some extent, auto-associative feed-forward networks [31], [44], [63], [17], [24] can also be thought of as involving a feature space. Feed-forward networks are universal function approximator. In the case of NLDR, they are used in a very specific configuration, called auto-association, in which one tries to reproduce as output the data that are presented as input. This learning process is not trivial because the network has a deep and bottleneck[1] architecture, that is, it consists of many hidden layers, and the middle layer comprises fewer neurons than the input and output layers. Formally, if $f : \mathbb{R}^D \to \mathbb{R}^D$ denotes the network, NLDR can be achieved if we assume that $f = h \circ g$, where $g : \mathbb{R}^D \to \mathbb{R}^P$, $h : \mathbb{R}^P \to \mathbb{R}^D$, and $P < D$. Hence, the deep network f results from stacking two shallow ones, g and h. Computing $\mathbf{x}_i = g(\boldsymbol{\xi}_i)$ reduces the data dimensionality, whereas $\hat{\boldsymbol{\xi}}_i = h(\mathbf{x}_i)$ maps it back to its initial high-dimensional space. The learning process identifies the parameters of f that minimize the TLS error $\sum_i \|\boldsymbol{\xi}_i - \hat{\boldsymbol{\xi}}_i\|_2^2$. This cannot easily be achieved with backpropagation alone [50], due to the deep architecture of the network, but new promising learning techniques are being developed [24]. Notice that if g and h are linear, the auto-associative network reduces to PCA and does nothing more than fitting a linear subspace through the data cloud.

D. Self-organization

Self-organization finds its inspiration in the brain architecture [71]: the organization of sensory maps in the cortex reflects that of the corresponding sensing organ, in what is known as a topographic map. Neighboring points in the primary visual cortex, for example, correspond to neighboring points in the retina. Such an organization conforms to the intuition of DR: close data items should remain near to each other in the low-dimensional representation. Topographic mapping was popularized in the field of DR by Kohonen's self-organizing feature maps (SOMs). In contrast to the majority of other DR methods, SOMs do not build an embedding. Instead, low-dimensional coordinates are pre-established in $\mathbf{G} = [\mathbf{g}_k]_{1\leq k \leq Q}$. Most of the time, the vectors in $\mathbf{G}$ are located at the nodes of two-dimensional rectangular grid. The goal of the SOM is then to deform the grid in the high-dimensional data space, so that it fits through the data cloud. For this purpose, each vector $\mathbf{g}_k$ is associated with high-dimensional coordinates in the data space, denoted by $\boldsymbol{\gamma}_k$. In order to ensure the topographic consistence, the SOM considers each datum $\boldsymbol{\xi}_i$ succesively and moves all grid nodes according to

$$\boldsymbol{\gamma}_k \leftarrow \boldsymbol{\gamma}_k + \alpha K_\sigma(d(\mathbf{g}_k, \mathbf{g}_l))(\boldsymbol{\xi}_i - \boldsymbol{\gamma}_k) \ , \tag{7}$$

where $0 \leq \alpha \leq 1$ is a learning rate, K_σ is a positive and decreasing function of its argument (a Gaussian function,

[1] The term hourglass would actually describe better this kind of network.

for instance), $d(\cdot,\cdot)$ is a distance in the grid space, and $l = \arg\min_k \|\boldsymbol{\xi}_i - \boldsymbol{\gamma}_k\|_2$.

A pecularity of SOMs is that they accomplish a kind of vector quantization [20] in addition to DR, as the number of grid nodes is often much smaller than the data size. The fact that the low-dimensional coordinates of the grid node are imposed also means that the geometrical structure of data must be visualized with colored grid nodes or other artifacts. These issues are addressed in reversed SOM models [33], [75], in which vector quantization is no longer mandatory, the grid is built in the high-dimensional data space, and the update rule (7) is applied on low-dimensional coordinates.

E. Similarity preservation

Some recent publications put forward similarity preservation as an improvement over distance preservation. Whereas a pairwise dissimilarity such as a distance equals zero for identical items, a similarity is usually defined as a decreasing function of the distance. In the context of dimensionality reduction, the use of similarities is increasingly perceived as more consistent with the intuition that local properties such as K-ary neighborhoods should be preserved prior to global properties. This idea already guides many weighting schemes that are used in distance-preserving methods such as MDS, Sammon's mapping, CCA, and their variants. By using similarities, the dominating terms in a cost function are naturally associated with small distances. For instance, let us define normalized pairwise similarites with

$$\pi_{ij} = \frac{\gamma(\delta_{ij}^2)}{\sum_{k<l}\gamma(\delta_{kl}^2)} \quad \text{and} \quad p_{ij} = \frac{g(d_{ij}^2)}{\sum_{k<l} g(d_{kl}^2)} \ , \qquad (8)$$

where γ and g are positive and decreasing functions of their arguments.

Using normalized similarities makes it possible to consider them as probability densities. Stochastic Neighbor Embedding (SNE) [23] exploits this property in its cost function, which involves a Kullback-Leibler divergence between the normalized similarities in the high- and low-dimensional spaces. The KL divergence can be written as

$$D(\mathbf{X};\Xi) = \sum_{i<j} \pi_{ij}\log(\pi_{ij}/p_{ij}) \qquad (9)$$

and can be minimized by gradient descent. The formula of the partial derivative with respect to the low-dimensional coordinates turns out to be surprisingly concise and elegant:

$$\frac{\partial D(\mathbf{X};\Xi)}{\partial \mathbf{x}_i} = \sum_j (\pi_{ij} - p_{ij}) \frac{g'(d_{ij}^2)}{g(d_{ij}^2)} (\mathbf{x}_i - \mathbf{x}_j) \ . \qquad (10)$$

It also shows that the gradient is negligible for large distances, that is, for small similarities, provided $g'(d_{ij}) \leq g(d_{ij})$. Recent papers investigate the choice of the similarity functions [64] and the definition of the cost function [68]. As the KL divergence is not symmetric, the authors of [68] consider a weighted combination of two divergences, based on the same principle as their distance preserving method in [67], [66]. In particular, this allows them to cast their method within the framework of statistical information retrieval.

Notice that although previous DR methods such as kernel PCA, Laplacian eigenmaps, or LLE also involve similarities, they do not preserve them in a straightforward way as SNE and its variants do. Instead, one can consider their action as follows. First, they convert similarities into inner products in a feature space (such as those associated with commute-time distances in the case of Laplacian eigenmaps). Next, they discard all feature space dimensions that lead to moderate distortions the pairwise inner products. Experimental results [64] are in favor of true similarity preservation.

IV. Taxonomy of methods

The large variety of methods in the field of dimensionality reduction naturally raises the question of their classification. They can be gathered into several categories, which correspond to different conceptual ideas, assumptions of their underlying model, or algorithmic properties.

A well-known frontier is the one that separates linear methods from nonlinear ones. For instance, the models of PCA and classical metric MDS both assume that data are distributed on (or near to, because of noise) a linear subspace. These methods do not perform optimally if the data to be processed are sampled from a nonlinear manifold such as the popular Swiss roll. The majority of recent manifold learning methods can deal with this case.

Dimensionality reduction methods can also be classified according to their paradigm, which can be inner product preservation, distance preservation, similarity preservation, rank preservation, auto-association, or topological mapping, among many other possibilities. As to distance-preserving methods, one can refine the categories by considering Euclidean distances, geodesic ones, or random walks in a graph, also known as commute-time distances.

Considering that DR amounts to a total least square regression problem, all DR methods involve some kind of optimization. DR methods can thus be categorized according to the various optimization techniques they rely on. An important distinction is the one that separates spectral methods [54] from those that utilize more generic optimization schemes, such as (stochastic) gradient descent. Classical spectral methods are PCA and metric multidimensional scaling, which are both linear. Most nonlinear spectral methods result from applying classical metric MDS to nonlinearly transformed data, in a so-called feature space, following the idea developed in kernel PCA. Methods such as Isomap, MVU, LLE, and Laplacian eigenmaps can all be cast within this framework. Notice however that a further distinction can be drawn between dense and sparse spectral methods. The former (KPCA and Isomap) utilize the leading eigenvectors of a dense Gram matrix, whereas the latter involve the nontrivial trailing eigenvectors of a sparse positive semidefinite matrix, whose entries can often be considered as similarities. These two formulations are actually dual [76] and all methods reduce to MDS applied in a feature space. For example, the pseudo-inversion of the graph Laplacian matrix shows that Laplacian eigenmaps turn out to be MDS applied on commute time distances. Spectral methods own the appealing advantage of yielding the global optimum of their associated cost function. This nice property comes at the expense of less flexibility in the design of the cost function. Not all cost functions can be expressed in the form of an eigenproblem and those that can are not necessarily pertinent. Sparse spectral methods can also suffer from numerical problems in the computation of the trailing eigenvectors.

Another distinction can be drawn between parametric and nonparametric methods. For instance, PCA and classical metric MDS both rely on parametric models. The main advantage of having a parametric model is the ability to reduce the dimensionality of new data that were not included in the learning set. This is obvious for PCA, by projecting new data along the principal component, whereas Nyström formula [4] is used in the case of MDS. In contrast, many nonlinear methods are nonparametric. Sammon's NLM, CCA, SOMs, and (t-)SNE, to cite only a few, fall in this category. These methods cannot process new data in a straightforward way. For some nonlinear spectral methods, such as kernel PCA, Isomap, LLE, and Laplacian eigenmaps, Nyström formula can be applied [4], since they consists in applying classical metric MDS on a Gram matrix computed in a feature space. (MVU is a noticible exception, as the computation of the Gram matrix is not direct and involves itself an optimization step.) Auto-associative networks and generative topographic mapping [9] (GTM, a generalization of SOMs) are parametric.

Dimensionality reduction can be hard or soft [12], depending on the ratio of dimensionalities before and after reduction. Simple methods such as PCA or classical metric MDS can process very high-dimensional data and project them to very low-dimensional space, even below the intrinsic dimensionality of data. Most nonlinear methods are less robust, due to their higher model complexity, and often fail to converge if the target dimensionality is lower than the data intrinsic dimensionality. A noticible exception is t-SNE, which seems to be insensitive to the curse of dimensionality, unlike most other methods.

A few DR methods rely on vector quantization. The most emblematic one is undoubtedly the SOM, though some others have followed the same idea [15], [35], [33]. Vector quantization reduces the number of data items to be processed and is therefore dual or complementary to DR. Though it decreases the computational demand, it also implies that not all data items are represented in the low-dimensional space. An intermediate solution consists in replacing genuine vector quantization with the use of *landmarks* [14], [64]. Instead of computing distances or similarities between pairs of data items, they are measured from one datum to one landmark. This requires less storage and less computation, while still providing low-dimensional coordinates for each datum.

The literature often mentions a distinction between global and local DR methods [14]. Although these qualifiers are used for all kinds of methods, they are usually associated with dense and sparse spectral methods. Because these two type of methods are dual [76], any local (i.e. sparse) method becomes global (i.e. dense) once it is considered in an appropriate feature space.

V. Quality assessment

An important and yet not fully addressed issue of DR is quality assessment (QA). Until recently, QA has been overlooked and most of the effort has been devoted to designing new (NL)DR methods. Among the obvious ways to assess quality, the connection between DR and a total least squares regression problem suggests that the quadratic reconstruction error is an optimal criterion. However, it requires to re-embed the low-dimensional data back to the initial high-dimensional space. Only a few parametric methods such as PCA and auto-associative networks can do that.

Another possiblity is to choose a popular DR cost function, such as Sammon's NLM stress, and compute its value with any embedding. In addition to being unfair, this methodology raises several questions. Does the stress function faithfully translate the intuition of a good embedding? Is distance preservation a meaningful criterion? Clearly not, as we know that in order to embed manifolds with complicated shapes, distances should ideally be stretched and shrunk. On the other hand, the reason to be of Sammon's stress is differentiability: it can easily be optimized. In QA, this constraint disappears, as the quality measure merely needs to be evaluated. Therefore, the principle of DR —embed close neighbors next to each other and maintain large distances between faraway data items— is more faithfully rendered by rank preservation. This idea is investigated in [65], [68], [37] and a uniform framework for all rank-based criteria is suggested in [38], [39]. Connections between some criteria and statistical information retrieval are developed in [68].

In practice, the rank of $\boldsymbol{\xi}_j$ with respect to $\boldsymbol{\xi}_i$ in the high-dimensional space is written as $\rho_{ij} = |\{k : \delta_{ik} < \delta_{ij} \text{ or } (\delta_{ik} = \delta_{ij} \text{ and } 1 \leq k < j \leq N)\}|$, where $|A|$ denotes the cardinality of set A. Similarly, the rank of $\mathbf{x}_j$ with respect to $\mathbf{x}_i$ in the low-dimensional space is $r_{ij} = |\{k : d_{ik} < d_{ij} \text{ or } (d_{ik} = d_{ij} \text{ and } 1 \leq k < j \leq N)\}|$. The *co-ranking matrix* [39] can then be defined as $\mathbf{Q} = [q_{kl}]_{1 \leq k,l \leq N-1}$ with $q_{kl} = |\{(i,j) : \rho_{ij} = k \text{ and } r_{ij} = l\}|$. The co-ranking matrix is the joint histogram of the ranks and is actually a sum of N permutation matrices of size $N-1$. With an appropriate gray scale, the co-ranking matrix can also be displayed and interpreted in a similar way as a Shepard diagram [57]. Historically, this scatterplot has often been used to assess results of multidimensional scaling and related methods [16]; it shows the distances δ_{ij} with respect to the corresponding distances d_{ij}, for all pairs (i,j), with $i \neq j$. With the co-ranking matrix, distance preservation errors reduce to rank preservation errors, given by $\rho_{ij} - r_{ij}$. As in the case of distances, the error sign is important, in order to avoid considering the types of errors on the same footing. This explains why many quality measures, such as the trustworthiness and continuity [65] and the mean relative rank errors [37] come by pairs. They separately account for positive and negative rank errors (called intrusions and extrusions, respectively, or flattening and tearing errors).

The generic form of rank-based quality measures consists of weighted sums over the first K rows and columns of $\mathbf{Q}$, that is, $Q_1(K) = \sum_{k=1}^{K} \sum_{l=1}^{N} w_{kl} q_{kl}$ and $Q_2(K) = \sum_{k=1}^{K} \sum_{l=1}^{N} w_{lk} q_{lk}$. The weighting schemes range from very simple ideas (e.g. $w_{kl} \in \{0, 1/KN\}$) to more complicated ones, which can raise normalization issues [38]. Notice also that all of those criteria remain dependent on some scale parameter, generally given by K [40].

VI. Conclusions and perspectives

Dimensionality reduction is a boiling hot research topic. In the last decades, revolutionary ideas have reshaped the domain, leading to a wide range of methods pursuing similar goals. The multitude of methods, each one coming with its own advantages and drawbacks, makes comparisons rather difficult. This motivated the recent works around rank-based quality measures. The underlying idea is to evaluate the methods by criteria that are as close as possible to our intuition of how DR should ideally work. This raises the fundamental question of what we are really looking for when using DR methods. Is the application-driven objective close from neighbor preservation, distance preservation, or any other principle DR methods and quality measures are built upon? Naturally a good idea is first to adapt the evaluation criteria to the application-driven objectives. Another good idea is to try developing DR methods that optimize directly the selected quality criterion, or an approximation of it. Recent research has mostly been conducted in the opposite way: first designing a method, next evaluating it, and eventually trying to see whether it fits specific application goals. Though it is certain that modern DR techniques are extremely powerful and can adapt to many situations, the next challenge is certainly to reverse the design-evaluation-application process and make it closer to the application needs. As to visualization by DR, this includes bridging the gap between DR techniques developed in the context of machine learning, and advanced data visualization techniques.

References

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

[2] ——, "Laplacian eigenmaps for dimensionality reduction and data representation," *Neural Computation*, vol. 15, no. 6, pp. 1373–1396, Jun. 2003.

[3] Y. Bengio, "Learning deep architectures for AI," *Foundations and Trends in Machine Learning archive*, vol. 2, no. 1, pp. 1–127, 2009.

[4] Y. Bengio, J.-F. Paiement, P. Vincent, O. Delalleau, N. Le Roux, and M. Ouimet, "Out-of-sample extensions for LLE, isomap, MDS, eigenmaps, and spectral clustering," in *Advances in Neural Information Processing Systems (NIPS 2003)*, S. Thrun, L. Saul, and B. Schölkopf, Eds. Cambridge, MA: MIT Press, 2004, vol. 16.

[5] Y. Bengio, P. Vincent, J.-F. Paiement, O. Delalleau, M. Ouimet, and N. Le Roux, "Spectral clustering and kernel PCA are learning eigenfunctions," Département d'Informatique et Recherche Opérationnelle, Université de Montréal, Montréal, Tech. Rep. 1239, Jul. 2003.

[6] M. Bernstein, V. de Silva, J. Langford, and J. Tenenbaum, "Graph approximations to geodesics on embedded manifolds," Stanford University, Palo Alto, CA, Tech. Rep., Dec. 2000.

[7] C. Bishop, *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995.

[8] C. Bishop, M. Svensén, and C. Williams, "GTM: A principled alternative to the self-organizing map," in *Advances in Neural Information Processing Systems (NIPS 1996)*, M. Mozer, M. Jordan, and T. Petsche, Eds. Cambridge, MA: MIT Press, 1997, vol. 9, pp. 354–360.

[9] C. Bishop, M. Svensén, and K. Williams, "GTM: A principled alternative to the self-organizing map," *Neural Computation*, vol. 10, no. 1, pp. 215–234, 1998.

[10] M. Brand, "Charting a manifold," in *Advances in Neural Information Processing Systems (NIPS 2002)*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, vol. 15.

[11] M. Brand and K. Huang, "A unifying theorem for spectral embedding and clustering," in *Proceedings of International Workshop on Artificial Intelligence and Statistics (AISTATS'03)*, C. Bishop and B. Frey, Eds., Jan. 2003.

[12] M. Carreira-Perpiñán, "A review of dimension reduction techniques," University of Sheffield, Sheffield, Tech. Rep., Jan. 1997.

[13] D. de Ridder and R. Duin, "Sammon's mapping using neural networks: A comparison," *Pattern Recognition Letters*, vol. 18, no. 11–13, pp. 1307–1316, 1997.

[14] V. de Silva and J. Tenenbaum, "Global versus local methods in nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems 15*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, pp. 705–712.

[15] P. Demartines and J. Hérault, "Vector quantization and projection neural network," ser. Lecture Notes in Computer Science, A. Prieto, J. Mira, and J. Cabestany, Eds. New York: Springer-Verlag, 1993, vol. 686, pp. 328–333.

[16] ——, "Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets," *IEEE Transactions on Neural Networks*, vol. 8, no. 1, pp. 148–154, Jan. 1997.

[17] D. DeMers and G. Cottrell, "Nonlinear dimensionality reduction," in *Advances in Neural Information Processing Systems (NIPS 1992)*, D. Hanson, J. Cowan, and L. Giles, Eds. San Mateo, CA: Morgan Kaufmann, 1993, vol. 5, pp. 580–587.

[18] G. Di Battista, P. Eades, R. Tamassia, and I. Tollis, *Graph drawing: Algorithms for the visualization of graphs*. Prentice-Hall, 1999.

[19] D. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," in *Proceedings of the National Academy of Arts and Sciences*, vol. 100, 2003, pp. 5591–5596.

[20] A. Gersho and R. Gray, *Vector Quantization and Signal Processing*. Boston: Kluwer Academic Publisher, 1992.

[21] T. Hastie and W. Stuetzle, "Principal curves," *Journal of the American Statistical Association*, vol. 84, no. 406, pp. 502–516, 1989.

[22] J. Hérault, C. Jaussions-Picaud, and A. Guérin-Dugué, "Curvilinear component analysis for high dimensional data representation: I. Theoretical aspects and practical use in the presence of noise," in *Proceedings of IWANN'99*, J. Mira and J. Sánchez, Eds. Alicante, Spain: Springer, Jun. 1999, vol. II, pp. 635–644.

[23] G. Hinton and S. Roweis, "Stochastic neighbor embedding," in *Advances in Neural Information Processing Systems (NIPS 2002)*, S. Becker, S. Thrun, and K. Obermayer, Eds. MIT Press, 2003, vol. 15, pp. 833–840.

[24] G. Hinton and R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, Jul. 2006.

[25] J. Hopfield, "Neural networks and physical systems with emergent collective computational abilities," in *Proc. Natl. Acad. Sci. USA 79*, 1982, pp. 2554–2558.

[26] H. Hotelling, "Analysis of a complex of statistical variables into principal components," *Journal of Educational Psychology*, vol. 24, pp. 417–441, 1933.

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

[28] K. Karhunen, "Zur Spektraltheorie stochastischer Prozesse," *Ann. Acad. Sci. Fennicae*, vol. 34, 1946.

[29] T. Kohonen, "Self-organization of topologically correct feature maps," *Biological Cybernetics*, vol. 43, pp. 59–69, 1982.

[30] ——, *Self-Organizing Maps*, 2nd ed. Heidelberg: Springer, 1995.

[31] M. Kramer, "Nonlinear principal component analysis using autoassociative neural networks," *AIChE Journal*, vol. 37, no. 2, pp. 233–243, 1991.

[32] J. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, pp. 1–28, 1964.

[33] J. Lee, C. Archambeau, and M. Verleysen, "Locally linear embedding versus Isotop," in *Proceedings of ESANN 2003, 11th European Symposium on Artificial Neural Networks*, M. Verleysen, Ed. Bruges, Belgium: d-side, Apr. 2003, pp. 527–534.

[34] J. Lee, A. Lendasse, N. Donckers, and M. Verleysen, "A robust nonlinear projection method," in *Proceedings of ESANN 2000, 8th European Symposium on Artificial Neural Networks*, M. Verleysen, Ed. Bruges, Belgium: D-Facto public., Apr. 2000, pp. 13–20.

[35] J. Lee and M. Verleysen, "Curvilinear distance analysis versus isomap," *Neurocomputing*, vol. 57, pp. 49–76, Mar. 2004.

[36] ——, "Nonlinear dimensionality reduction of data manifolds with essential loops," *Neurocomputing*, vol. 67, pp. 29–53, 2005.

[37] ——, *Nonlinear dimensionality reduction*. Springer, 2007.

[38] ——, "Quality assessment of nonlinear dimensionality reduction based on k-ary neighborhoods," in *JMLR Workshop and Conference Proceedings (New challenges for feature selection in data mining and knowledge discovery)*, Y. Saeys, H. Liu, I. Inza, L. Wehenkel, and Y. Van de Peer, Eds., Sep. 2008, vol. 4, pp. 21–35.

[39] ——, "Quality assessment of dimensionality reduction: Rank-based criteria," *Neurocomputing*, vol. 72, no. 7–9, pp. 1431–1443, 2009.

[40] ——, "Scale-independent quality criteria for dimensionality reduction," *Pattern Recognition Letters*, 2010, in press.

[41] M. Loève, "Fonctions aléatoire du second ordre," in *Processus stochastiques et mouvement Brownien*, P. Lévy, Ed. Paris: Gauthier-Villars, 1948, p. 299.

[42] J. Mao and A. Jain, "Artificial neural networks for feature extraction and multivariate data projection," *IEEE Transactions on Neural Networks*, vol. 6, no. 2, pp. 296–317, Mar. 1995.

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

[44] E. Oja, "Data compression, feature extraction, and autoassociation in feedforward neural networks," in *Artificial Neural Networks*, T. Kohonen, K. Mäkisara, O. Simula, and J. Kangas, Eds. North-Holland: Elsevier Science Publishers, B.V., 1991, vol. 1, pp. 737–745.

[45] V. Onclinx, J. A. Lee, V. Wertz, and M. Verleysen, "Dimensionality reduction by rank preservation," in *Proceedings of IJCNN 2010*, Barcelona, Spain, 2010, in press.

[46] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Philosophical Magazine*, vol. 2, pp. 559–572, 1901.

[47] H. Ritter, T. Martinetz, and K. Schulten, *Neural Computation and Self-Organizing Maps*. Reading, MA: Addison-Wesley, 1992.

[48] S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *Science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[49] S. Roweis, L. Saul, and G. Hinton, "Global coordination of local linear models," in *Advances in Neural Information Processing Systems (NIPS 2001)*, T. Dietterich, S. Becker, and Z. Ghahramani, Eds. Cambridge, MA: MIT Press, 2002, vol. 14.

[50] D. Rumelhart, G. Hinton, and R. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, pp. 533–536, 1986.

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

[52] J. Sammon, "A nonlinear mapping algorithm for data structure analysis," *IEEE Transactions on Computers*, vol. CC-18, no. 5, pp. 401–409, 1969.

[53] L. Saul and S. Roweis, "Think globally, fit locally: Unsupervised learning of nonlinear manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119–155, Jun. 2003.

[54] L. Saul, K. Weinberger, J. Ham, F. Sha, and D. Lee, "Spectral methods for dimensionality reduction," in *Semisupervised Learning*, O. Chapelle, B. Schoelkopf, and A. Zien, Eds. MIT Press, 2006.

[55] B. Schölkopf, A. Smola, and K.-R. Müller, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998.

[56] ——, "Nonlinear component analysis as a kernel eigenvalue problem," *Neural Computation*, vol. 10, pp. 1299–1319, 1998, also available as technical report 44 at the Max Planck Institute for Biological Cybernetics, Tübingen, Germany, December 1996.

[57] R. Shepard, "The analysis of proximities: Multidimensional scaling with an unknown distance function (parts 1 and 2)," *Psychometrika*, vol. 27, pp. 125–140, 219–249, 1962.

[58] J. Svensén, "GTM: The generative topographic mapping," Ph.D. dissertation, Aston University, Aston, UK, Apr. 1998.

[59] Y. Takane, F. Young, and J. de Leeuw, "Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features," *Psychometrika*, vol. 42, pp. 7–67, 1977.

[60] Y. Teh and S. Roweis, "Automatic alignment of hidden representations," in *Advances in Neural Information Processing Systems (NIPS 2002)*, S. Becker, S. Thrun, and K. Obermayer, Eds. Cambridge, MA: MIT Press, 2003, vol. 15.

[61] J. Tenenbaum, V. de Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, Dec. 2000.

[62] W. Torgerson, "Multidimensional scaling, I: Theory and method," *Psychometrika*, vol. 17, pp. 401–419, 1952.

[63] S. Usui, S. Nakauchi, and M. Nakano, "Internal colour representation acquired by a five-layer neural network," in *Artificial Neural Networks*, T. Kohonen, K. Makisara, O. Simula, and J. Kangas, Eds. North-Holland: Elsevier Science Publishers, B.V., 1991.

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

[65] J. Venna and S. Kaski, "Neighborhood preservation in nonlinear projection methods: An experimental study," in *Proceedings of ICANN 2001*, G. Dorffner, H. Bischof, and K. Hornik, Eds. Berlin: Springer, 2001, pp. 485–491.

[66] ——, "Local multidimensional scaling," *Neural Networks*, vol. 19, pp. 889–899, 2006.

[67] ——, "Visualizing gene interaction graphs with local multidimensional scaling," in *Proceedings of ESANN 2006, 14th European Symposium on Artificial Neural Networks*, M. Verleysen, Ed. Bruges, Belgium: d-side, Apr. 2006, pp. 557–562.

[68] J. Venna, J. Peltonen, K. Nybo, H. Aidos, and S. Kaski, "Information retrieval perspective to nonlinear dimensionality reduction for data visualization," *Journal of Machine Learning Research*, vol. 11, pp. 451–490, 2010.

[69] J. Verbeek, N. Vlassis, and B. Kröse, "Coordinating mixtures of probabilistic principal component analyzers," Computer Science Institute, University of Amsterdam, Amsterdam, Tech. Rep. IAS-UVA-02-01, Feb. 2002.

[70] ——, "Self-organizing mixture models," *Neurocomputing*, vol. 63, pp. 99–123, 2005.

[71] C. von der Malsburg, "Self-organization of orientation sensitive cells in the striate cortex," *Kybernetik*, vol. 14, pp. 85–100, 1973.

[72] K. Weinberger and L. Saul, "Unsupervised learning of image manifolds by semidefinite programming," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR-04)*, vol. 2, Washington, DC, 2004, pp. 988–995.

[73] ——, "Unsupervised learning of image manifolds by semidefinite programming," *International Journal of Computer Vision*, vol. 70, no. 1, pp. 77–90, 2006.

[74] K. Weinberger, F. Sha, and L. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proceedings of the Twenty-First International Conference on Machine Learning (ICML-04)*, Banff, Canada, 2004, pp. 839–846.

[75] A. Wismüller, "The exploration machine - a novel method for data visualization," in *Lecture Notes in Computer Science. Advances in Self-Organizing Maps*, 2009, pp. 344–352.

[76] L. Xiao, J. Sun, and S. Boyd, "A duality view of spectral methods for dimensionality reduction," in *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburg, PA, 2006, pp. 1041–1048.

[77] L. Yang, "Sammon's nonlinear mapping using geodesic distances," in *Proc. 17th International Conference on Pattern Recognition (ICPR'04)*, 2004, vol. 2.

[78] L. Yen, D. Vanvyve, F. Wouters, F. Fouss, M. Verleysen, and M. Saerens, "Clustering using a random-walk based distance measure," in *Proceedings of ESANN 2005, 13th European Symposium on Artificial Neural Networks*, M. Verleysen, Ed. Bruges, Belgium: d-side, Apr. 2005, pp. 317–324.

[79] G. Young and A. Householder, "Discussion of a set of points in terms of their mutual distances," *Psychometrika*, vol. 3, pp. 19–22, 1938.