Dimensionality Reduction Techniques for Data Exploration

Flora S. Tsai and Kap Luk Chan
Nanyang Technological University
School of Electrical & Electronic Engineering
Singapore
fst1@columbia.edu

***Abstract*— Data exploration, or the search for features in data that may indicate deeper relationships among variables, relies heavily on visual methods because of the power of the human eye to detect structures. However, for large data sets with many variables and dimensions, the number of dimensions of the data can be reduced by applying dimensionality reduction techniques. This paper reviews current linear and nonlinear dimensionality reduction techniques. The nonlinear dimensionality reduction techniques which deal with finding a lower dimensional embedding of a nonlinear manifold can be classified under manifold learning algorithms. For basic types of nonlinear manifolds, experiments were performed on some of the current dimensionality reduction techniques. The nonlinear dimensionality reduction techniques generally do not perform well in the presence of noise, as seen from the results. When faced with a larger amount of noise, one of the algorithms was not able to converge to a solution. Thus, in order to apply nonlinear dimensionality reduction techniques effectively, the neighborhood, the density, and noise levels need to be taken into account.**

I. Introduction

Data exploration, also known as exploratory data analysis, is the search for features in data that may indicate deeper relationships among variables. Normally, data exploration relies heavily on visual methods because of the power of the human eye to detect structures. However, for large data sets with many variables and dimensions, it may be necessary to reduce the number of dimensions of the data by applying dimensionality reduction techniques. Dimensionality reduction is the search for a small set of features to describe a large set of observed dimensions. Dimensionality reduction is useful in visualizing data, discovering a compact representation, and decreasing computational processing time. In addition, reducing the number of dimensions can separate the important features or variables from the less important ones, thus providing additional insight into the nature of the data that may otherwise be left undiscovered.

When analyzing large data of multiple dimensions, it may be necessary to perform dimensionality reduction or projection techniques to transform the data into a smaller, more manageable set. By reducing the data set, we hope to uncover hidden structure that aids in the understanding as well as visualization of the data. Dimensionality reduction techniques such as Principal Component Analysis (PCA) [9] and Multidimensional Scaling (MDS) [2], [3], [8] have existed for quite some time, but most are capable only of handling data that is inherently linear in nature. Recently, some unsupervised nonlinear techniques for dimensionality reduction such as Locally Linear Embedding (LLE) [10] and Isometric Feature Mapping (Isomap) [12], which are primarily based on MDS, have achieved remarkable results for data with a certain type of topological manifold, such as the Swiss Roll. However, LLE and Isomap both fail in other types of nonlinear data, such as a sphere or torus [11]. In addition, the nonlinear techniques tend to be extremely sensitive to noise. This paper examines the reasons why certain techniques do well on some types but not other classes of data by performing a detailed analysis of existing techniques on various classes of nonlinear data sets. By understanding the reasons of failure, we hope to achieve success in developing new techniques that are suitable for a broader range of data topologies and which are more robust to varying levels of noise.

II. Linear Dimensionality Reduction Techniques

If the transformation to a lower-dimensional space is a linear combination of the original variables, then this is called linear dimensionality reduction. In feature extraction, all available variables are used and the data is transformed using a linear transformation to a reduced dimension space. The aim is to replace the original variables by a smaller set of underlying variables. The techniques covered here are also referred to as techniques of exploratory data analysis, geometric methods, or methods of ordination, where no prior assumption is made about the existence of groups or clusters in the data. Geometric methods are sometimes further categorized as being variable-directed when they are primarily concerned with relationships between variables, or individual-directed when they are primarily concerned with relationships between individuals [15].

A. Principal Component Analysis (PCA)

PCA, also known as Karhunen-Love transform, is a very established method of dimensionality reduction introduced by Pearson in 1901 [9]. The purpose of PCA is to derive new variables (in decreasing order of importance) that are linear combinations of the original variables and are uncorrelated. Geometrically, PCA can be described as a rotation of the axes

1–4244–0983–7/07/$25.00 © 2007 IEEE

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 15:50:01 UTC from IEEE Xplore. Restrictions apply.

of the original coordinate system to a new set of orthogonal axes that are ordered in terms of the amount of variation of the original data they account for [15].

One of the reasons for performing PCA is to find a smaller group of underlying variables that describe the data. The hope is that the first few components will account for most of the variation in the original data. PCA is a variable-directed technique, making no assumptions about the existence of groupings within the data, and is thus considered an unsupervised feature extraction technique [15].

PCA projects n-dimensional data onto a lower d-dimensional subspace in a way that minimizes the sum-squared error, or (equivalently) maximizes the variance, or (equivalently) gives uncorrelated projected distributions [5].

Since PCA is a linear transformation method, it is simple to compute and is guaranteed to work. It is useful in reducing dimensionality and finding new, more informative, uncorrelated features. However, since PCA is a linear dimensionality reduction technique, it may not be able to accurately represent nonlinear data. Also, one does not know how many principal components to keep, although as a general rule, the number may be chosen such that the variance of the components is roughly 90-95% of the original variance. In addition, PCA may not lead to an interesting viewpoint for data clustering because it is not good at discriminating data.

B. Multidimensional Scaling (MDS)

Multidimensional scaling (MDS) is a general approach which achieves a lower dimensional representation of data, while trying to preserve the distances between the data points [7]. This class of methods is sometimes called *distance methods*. The distance can be represented as either a similarity or dissimilarity measure. We can think of squeezing a high-dimensional point cloud into a small number of dimensions (2 or 3) while preserving as well as possible the interpoint distances [14].

MDS is equivalent to PCA when the distances are Euclidean. There are various MDS methods, differing in the types of metrics used as well as the calculations performed. However, all of the methods are governed by a set of similar principles. The starting point for MDS is the determination of the "spatial distance model" [3]. In order to determine the proximities, the following notations are used:

Let Δ and D, NxN matrices, represent the collection of objects, indexed by i and j, where the proximity or data value connecting object i with object j is represented by δ_{ij} [8], and the distances between pairs of points x_i and x_j be represented by d_{ij}, as shown in Equations (1) and (2):

$$\mathbf{\Delta} = \begin{bmatrix} \delta_{11} & \delta_{12} & \dots & \delta_{1N} \\ \delta_{21} & \delta_{22} & \dots & \delta_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ \delta_{N1} & \delta_{N2} & \dots & \delta_{NN} \end{bmatrix} \tag{1}$$

$$\mathbf{D} = \begin{bmatrix} d_{11} & d_{12} & \dots & d_{1N} \\ d_{21} & d_{22} & \dots & d_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ d_{N1} & d_{N2} & \dots & d_{NN} \end{bmatrix} \tag{2}$$

The aim of MDS is to find a configuration such that the distances d_{ij} match, as well as possible, the dissimilarities δ_{ij} [2].

The variations of MDS come in the differences in functions used to transform the dissimilarities. Classical Metric Multidimensional Scaling is a basic form of MDS, in which the distances between points in the results d_{ij} are as close as possible to the dissimilarities δ_{ij}, measured in Euclidean distances. This is also sometimes referred to as *principal coordinate analysis*, which is also equivalent to PCA [14]. MDS methods such as *Classical MDS* are called metric methods because the relationship between δ_{ij} and d_{ij} depend on the numerical or metric properties of the dissimilarities. Nonmetric MDS refers to those methods where the relationship between δ_{ij} and d_{ij} depend on the rank ordering of the dissimilarities [8].

In the original metric MDS [13] the distances between the data items have been given, and a configuration of points that would give rise to the distances is sought. Often a linear projection onto a subspace obtained with PCA is used. The key idea of the method, to approximate the original set of distances with distances corresponding to a configuration of points in a Euclidean space can, however, also be used for constructing a nonlinear projection method.

If each item x_i is represented with a lower-dimensional, say, two-dimensional data vector x'_i, then the goal of the projection is to optimize the representations so that the distances between the items in the two-dimensional space will be as close to the original distances as possible. If the distance between x_i and x_j is denoted by d_{ij} and the distance between x'_i and x'_i in the two-dimensional space by d'_{ij}, the metric MDS tries to approximate d_{ij} by d'_{ij}. If a square-error cost is used, the objective function to be minimized can be written as:

$$E_M = \sum_{i \neq j} [(d_{ij} - d'_{ij}]^2 \tag{3}$$

Whereas PCA takes a set of points in $\mathbb{R}^n$ and gives a mapping of the points in $\mathbb{R}^d$, MDS takes a matrix of pairwise distances and gives a mapping into $\mathbb{R}^d$. Some advantages of MDS are that it is relatively simple to implement, very useful for visualization, and thus, able to uncover hidden structure in the data. Some drawbacks of using MDS include the difficulty in selecting the appropriate dimension of the map, difficulty in representing small distances corresponding to the local structure, and unlike PCA, cannot obtain an (N-1)-dimensional map out of an N-dimensional one by dropping a coordinate [1].

III. Nonlinear Dimensionality Reduction Techniques

As seen in the previous section, PCA is useful in identifying significant coordinates and linear correlations in high dimen-

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 15:50:01 UTC from IEEE Xplore. Restrictions apply.

sional data. PCA as well as classical MDS are unsuitable if the data set contains nonlinear relationships among the variables. General MDS techniques are appropriate when the data is highly nonmetric or sparse. If the original high-dimensional data set contains nonlinear relationships, then nonlinear dimensionality reduction techniques may be more appropriate. In recent years, there has been much interest in the development of nonlinear dimensionality reduction techniques for data lying on a high-dimensional manifold, a topological space which is locally Euclidean. These methods, also known as *manifold learning* algorithms, are generally are based on the MDS approach, whereby a lower dimensional representation of data is achieved while preserving the original distances between the data points. However, there are some slight modifications and assumptions for these manifold learning techniques that make it different than MDS. A more detailed review of manifold learning is contained in the following section.

Some recently developed methods (LLE and Isomap) rely on applying linear techniques on a set of local neighborhoods, which are assumed to be locally linear in nature. As such, they fall into the category of *local linear* dimensionality reduction techniques. Although these local linear techniques perform well for particular classes of manifolds, they fail to achieve results with other classes of manifolds, to be discussed in detail in the following section on manifold classification.

A. Isometric Feature Mapping (Isomap)

Isomap[12] is a nonlinear dimensionality reduction technique that uses MDS techniques with geodesic interpoint distances instead of Euclidean distances. Geodesic distances represent the shortest paths along the curved surface of the manifold (a subspace of $\mathbb{R}^n$, measured as if the surface were flat)[6]. Unlike the linear techniques, Isomap can discover the nonlinear degrees of freedom that underlie complex natural observations [12].

Isomap deals with finite data sets of points in $\mathbb{R}^n$ which are assumed to lie on a smooth submanifold M_d of low dimension $d < n$. The algorithm attempts to recover M given only the data points. Isomap estimates the unknown geodesic distance in M between data points in terms of the graph distance with respect to some graph G constructed on the data points.

Isomap algorithm consists of three basic steps:

1) Determine which points are neighbors on the manifold M, based on the distances between pairs of points in the input space.
2) Estimate the geodesic distances between all pairs of points on the manifold M by computing their shortest path distances in the graph G.
3) Apply MDS to matrix of graph distances, constructing an embedding of the data in a d-dimensional Euclidean space Y that best preserves the manifold's estimated intrinsic geometry [12].

For two arbitrary points on a nonlinear manifold, their Euclidean distance in the high-dimensional input space may not accurately reflect their intrinsic similarity, as measured by geodesic distance along the low-dimensional manifold. Isomap is a very useful noniterative, polynomial-time algorithm for nonlinear dimensionality reduction if the data is severely nonlinear. Isomap is a special case of MDS, which is a special case of PCA. Isomap is able to compute a globally optimal solution, and for a certain class of data manifolds (Swiss roll), is guaranteed to converge asymptotically to the true structure[12]. However, Isomap may not easily handle more complex domains such as non-trivial curvature or topology.

B. Locally Linear Embedding (LLE)

LLE is a nonlinear dimensionality reduction technique that computes low-dimensional, neighborhood preserving embeddings of high-dimensional inputs. Unlike Isomap, LLE eliminates the need to estimate pairwise distances between widely separated data points and recovers global nonlinear structure from locally linear fits [10]. LLE assumes that the manifold is linear when viewed locally.

The LLE algorithm is summarized as follows:

1) Determine which points are neighbors on the manifold M, based on the distances between pairs of points in the input space (same as Isomap).
2) Compute the weights W_{ij} that best reconstruct each data point x_i from its neighbors.
3) Compute the vectors y_i that are best reconstructed by the weights W_{ij} [10].

Compared to Isomap, LLE is more efficient. However, LLE finds an embedding that only preserves the local structure, is not guaranteed to asymptotically converge, and may introduce unpredictable distortions. Both Isomap and LLE algorithms are unlikely to work well for manifolds like a sphere or a torus, require dense data points on the manifold for good estimation, and are strongly dependent on a good local neighborhood for success.

C. Hessian Locally Linear Embedding (HLLE)

Other variations of LLE, such as Hessian Eigenmaps [4], have been developed that combines LLE with Laplacian Eigenmaps [1]. Hessian Eigenmap modifies the Laplacian Eigenmap framework, where they substitute a quadratic form based on the Hessian in place of one based on the Lapacian [4].

HLLE is a method for recovering the underlying parametrization of scattered data (m_i) lying on a manifold M embedded in high-dimensional Euclidean space. The manifold M, viewed as a Riemannian submanifold of the ambient Euclidean space $\mathbb{R}^n$, is locally isometric to an open, connected subset Θ of Euclidean space $\mathbb{R}^d$. Because Θ does not need to be convex, HLLE is able to hander a wider class of situations than the original ISOMAP [4], such as data with a central square removed.

The computational demands of LLE algorithms are very different than those of the Isomap distance-processing step. LLE and HLLE are both capable of handling large N problems, because initial computations are performed only on smaller neighborhoods, whereas Isomap has to compute a full

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 15:50:01 UTC from IEEE Xplore. Restrictions apply.

matrix of graph distances for the initial distance-processing step. However, both LLE and HLLE are more sensitive to the dimensionality of the data space, n, because they must estimate a local tangent space at each point. Although an orthogonalization step was introduced in HLLE that makes the local fits more robust to pathological neighborhoods than LLE, HLLE still requires effectively a numerical second differencing at each point that can be very noisy at low sampling density [4].

D. Local Tangent Space Alignment (LTSA)

Based on a set of unorganized data points sampled with noise from a parameterized manifold, the local geometry of the manifold is learned by constructing an approximation for the tangent space at each data point, and those tangent spaces are then aligned to give the global coordinates of the data points with respect to the underlying manifold [16].

The LTSA algorithm, when compared with LLE, is less sensitive to choice of k neighborhoods than LLE. LTSA is a method for nonlinear dimensionality reduction that constructs approximations of tangent spaces in order to represent local geometry of the manifold and the global alignment of the tangent spaces to obtain the global coordinate system [16]. In general, LTSA is less sensitive to the choice of k neighborhoods, as compared to LLE.

Although LTSA as well as the other nonlinear dimensionality reduction algorithms are able to handle nonlinearities in data, they are generally not as robust as the linear dimensionality reduction techniques, and not able to handle certain types of nonlinear manifolds. The next section provides a more detailed discussion of manifold learning.

IV. Experiments and Results

A. Results of Dimensionality Reduction for Isometric Manifolds

Figure 1 shows the results obtained using classical MDS, LLE, Isomap, HLLE, and LTSA on an isometric nonlinear manifold, in the shape of an S-Curve, with 1000 data points. The MDS embedding appears to not be as good as the other embeddings because MDS is not able to take nonlinearities into account, and the Euclidean distance may not be as good a distance measure if the data lies on a nonlinear manifold.

However, if the number of data points, N, is set very small, then the results for MDS appears less distorted compared to the nonlinear dimensionality reduction algorithms. See Figure 2 for the results of using the dimensionality reduction algorithms on the S-Curve manifold, this time with only 200 data points. We can conclude from this that MDS may be better than the other algorithms for sparse data sets, but is not able to accurately capture the nonlinearities if the original data lies on a nonlinear manifold. Although LLE, Isomap, HLLE, and LTSA are all able to reduce the dimension of the nonlinear isometric manifold, they require the data points to be relatively dense.

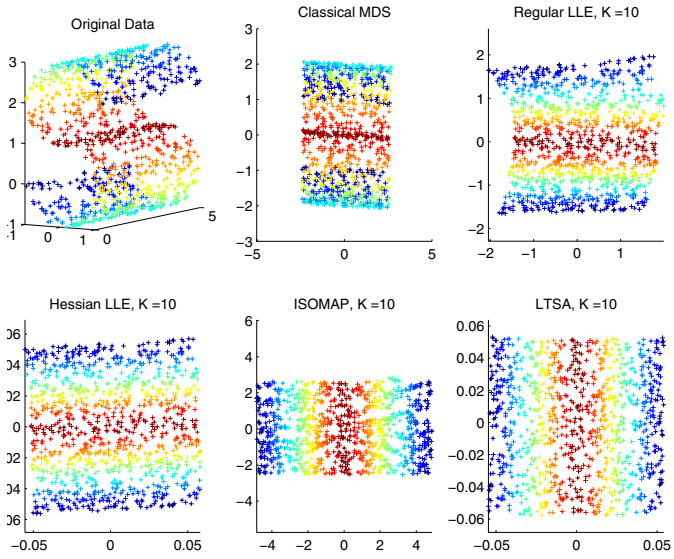

Fig. 1. Results of reducing dimension on an Isometric manifold, N=1000.

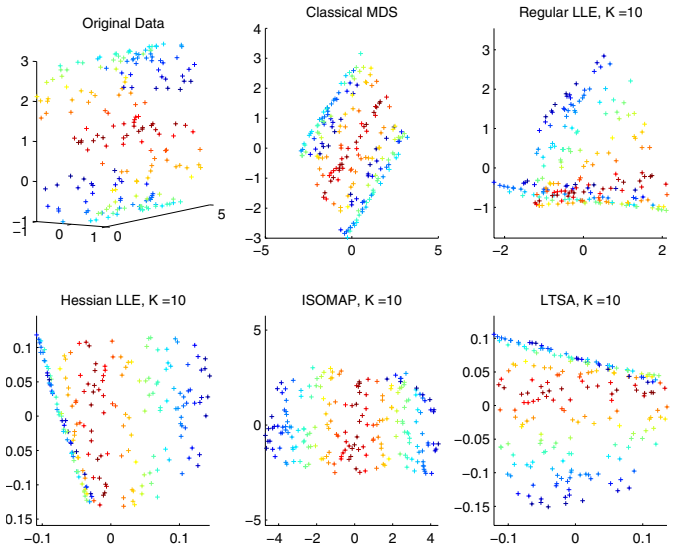

Fig. 2. Results of reducing dimension on an Isometric manifold, $N = 200$.

B. Results for Nonlinear Manifolds in Presence of Noise

To see the performance of the various algorithms in the presence of noise, experiments were conducted on adding Gaussian noise to a nonlinear manifold, in the shape of a Swiss Roll. Figure 3 show the results of the HLLE in the presence of noise. HLLE was not able to converge to a solution, when faced with a larger amount of noise.

From the results, we can conclude that locally linear techniques may look good in theory, but exhibit many problems in practice. Firstly, these techniques are highly dependent on the choice of neighborhood, which there are no clear rules for selecting. Secondly, because there are no clear rules in selecting an appropriate neighborhood, many trial runs may need to be conducted before optimal results are obtained.

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 15:50:01 UTC from IEEE Xplore. Restrictions apply.

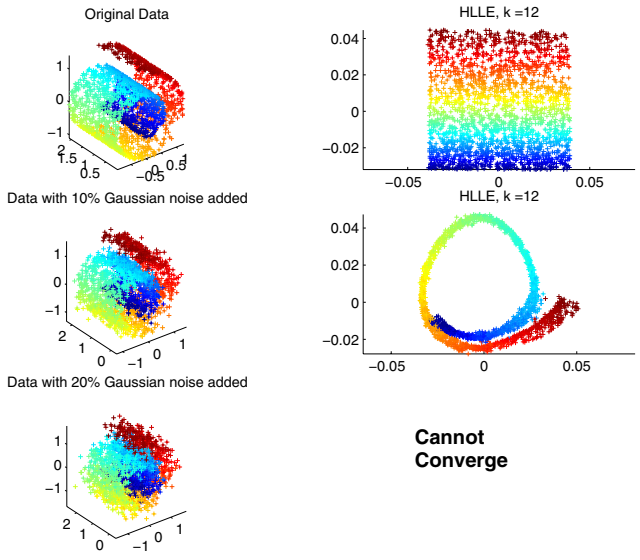

Fig. 3. Results on adding noise to Swiss Roll and applying HLLE.

Thirdly, identification of optimal results cause problems when the number of dimensions is high (more than three). Fourthly, for high noise levels, the techniques will fail to obtain good results. Lastly, because of the problems highlighted above, it is conceivable that one can go through many trial runs without obtaining good results, and still be unsure as to the reason why optimal results cannot be obtained. Is it because not enough trials are conducted, or because the nature of the data set is not suitable for the locally linear technique? Therefore, although these locally linear techniques can achieve remarkable results with certain classes of manifolds, in practice, they pose many more problems than they solve.

V. Conclusion

The applicability of dimensionality reduction techniques for data exploration has been evaluated in this work. A summary of some current linear and nonlinear dimensionality reduction techniques has been presented. PCA is useful in identifying significant coordinates and linear correlations in high dimensional data. PCA as well as classical MDS are unsuitable if the data set contains nonlinear relationships among the variables. General MDS techniques are appropriate when the data is highly nonmetric or sparse. If the original high-dimensional data set contains nonlinear relationships, then nonlinear dimensionality reduction techniques are more appropriate.

For basic types of nonlinear manifolds, an evaluation was performed on various dimensionality reduction techniques. Because the nonlinear dimensionality reduction techniques depend on a good neighborhood, the results when using different neighborhoods are significantly different. A good neighborhood will depend on the particular characteristics of the data set. If the neighborhood is too small, the global structure may not be captured effectively. If the neighborhood is too big, the nonlinearities of the data set may not be mapped appropriately. Because a "good" neighborhood is dependant on the data set, it is difficult to generalize an acceptable value that will work under all circumstances.

In addition, the algorithms require that the manifold consists of well-sampled data points, otherwise the nonlinearities in the manifold structure may not be effectively captured. Thus, the algorithms may not work properly if the data set is very sparse, as seen from the results. Classical MDS techniques are not able to fully capture the nonlinearities in the data; however, MDS algorithms generally performs better than the nonlinear algorithms when faced with a sparse data set.

The techniques generally do not perform well in the presence of noise, as seen from the results. When faced with a larger amount of noise, HLLE algorithm was not able to converge to a solution. Thus, in order to apply nonlinear dimensionality reduction techniques effectively, the neighborhood, the density, and noise levels need to be taken into account.

The techniques described fall under the general category of local linear techniques, which apply MDS-based techniques on a set of local neighborboods. This assumption of local linearity is valid when the original data set constitutes a high dimensional manifold, which appears in many applications where there is a time-varying component, such as a sequence of images in a video or microarray gene expression data taken under time-varying conditions. Thus, future work can focus on evaluating the algorithms on different types of nonlinear data.

References

[1] M. Belkin, P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15, no. 2, pp. 1373-1396, 2003.
[2] T. Cox, M. Cox, *Multidimensional Scaling*. Second Edition, New York: Chapman & Hall, 2001.
[3] M. Davison, *Multidimensional Scaling*. Florida: Krieger Publishing Company, 1992.
[4] D. Donoho and C. Grimes, "Hessian eigenmaps: Locally linear embedding techniques for high-dimensional data," *Proc. Natl Acad. Sci. USA*, 2003.
[5] R. Duda, P. Hart, D. Stork, *Pattern Classification*. New York: John Wiley & Sons, 2001.
[6] D. Gering, "Linear and Nonlinear Data Dimensionality Reduction," *Area Exam Report*, MIT, April 2002.
[7] D. Hand, H. Mannila, P. Smyth, *Principles of Data Mining*. Massachusetts: MIT Press, 2001.
[8] J. Kruskal, M. Wish, *Multidimensional Scaling*. London: Sage Publications, 1978.
[9] K. Pearson, "On lines and planes of closest fit to systems of points in space," *Phil. Mag., Ser. B*, vol. 2, no. 11, pp. 559-572, 1901.
[10] S. Roweis, L. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, Dec. 2000.
[11] L. Saul, S. Roweis, "Think Globally, Fit Locally: Unsupervised Learning of Low Dimensional Manifolds," *Journal of Machine Learning Research*, vol. 4, pp. 119-155, 2003.
[12] J. Tenenbaum, V. de Silva, J. Langford, "A Global Geometric Framework for Nonlinear Dimensionality Reduction," *Science*, vol. 290, pp. 2319-2323, Dec. 2000.
[13] W.S. Torgerson, "Multidimensional scaling: I. Theory and method," *Psychometrika*, vol. 17, pp. 401-419, 1952.
[14] W.N. Venables, B.D. Ripley, *Modern Applied Statistics with S*, Fourth Edition, Springer, New York, 2002.
[15] A. Webb, *Statistical Pattern Recognition*, Second Edition, John Wiley, London, 2002.
[16] Z. Zhang and H. Zha, "Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment," *SIAM Journal on Scientific Computing*, 26(1), 313–318, 2005.

Authorized licensed use limited to: b-on: UNIVERSIDADE DE AVEIRO. Downloaded on March 14,2023 at 15:50:01 UTC from IEEE Xplore. Restrictions apply.