ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

Sensitivity to parameter and data variations in dimensionality reduction techniques

Francisco J. García-Fernández[1,2], Michel Verleysen[2],
John A. Lee[3] and Ignacio Díaz[1] *

1- Univ. of Oviedo - Department of Electrical Engineering
Campus Viesques, 33204 Gijón, Spain

2- Univ. Catholique de Louvain - Machine Learning Group
ICTEAM/ELEN - Place du Levant, 3 1348 Louvain-la-Neuve, Belgium

3- Univ. Catholique de Louvain-Molecular Imaging, Radiotherapy and Oncology
IREC - Avenue Hippocrate 55 1200 Bruxelles, Belgium

Abstract. Dimensionality reduction techniques aim at representing high-dimensional data in a meaningful and lower-dimensional space, improving the human comprehension and interpretation of data. In recent years, newer nonlinear techniques have been proposed in order to address the limitation of linear techniques. This paper presents a study of the stability of some of these dimensionality reduction techniques, analyzing their behavior under changes in the parameters and the data. The performances of these techniques are investigated on artificial datasets. The paper presents these results by identifying the weaknesses of each technique, and suggests some data-processing tasks to improve the stability.

1 Introduction

In recent years, we have witnessed an unprecedented data revolution. Sensors and other data sources, such as social networks, e-government open data, etc., pervade our lives and a huge amount of data is available in many different fields, such as medical imaging, process control and text mining, just to mention a few. As a consequence, we obtain not only better but also bigger datasets. This requires to have suitable techniques for analyzing and visualizing all these data. Typically, these data are high-dimensional, so it is not possible to directly visualize them, in a two/three dimensional lattice. It is at this point that dimensionality reduction (DR) techniques play a key role: they make a transformation of these high-dimensional data into a meaningful visualization with a reduced dimensionality.

DR includes various techniques that are able to construct meaningful data representations in a space of given dimensionality. Linear DR is a well-known field with techniques such as principal component analysis (PCA) [1] or multidimensional scaling (MDS) [2, 3]. Unlike linear DR, nonlinear DR techniques

*This work has been financed by a grant under PCTI funds, by the Spanish Ministry of Science and Education and FEDER funds under grants DPI2009-13398-C02-01 and by the Council of Education and Science of the Government of Principado de Asturias, under the agreement with the University of Oviedo for the Plan of viability and conversion to International Excellence Campus. J.A.Lee is a Research Associate with the Belgian F.R.S.-FNRS (Fonds National de la Recherche Scientifique).

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

[4] have the ability to deal with complex manifolds, which is a typical fact with real-world datasets which are likely to lie on nonlinear manifolds. In this field, techniques appeared later, especially with nonlinear variants of multidimensional scaling [5] and neural approaches [6, 7]. However, in the beginning of this century, a renewed interest in DR techniques emerged and new nonlinear algorithms based on spectral techniques were proposed. Some of these techniques are *Isomap* [8], *Local Linear Embedding* (LLE)[9] and *Laplacian Eigenmaps* (LE)[10]. Also, non-convex techniques, such as *Stochastic Neighbor Embedding* (SNE) [11] and *t-Stochastic Distributed Neighbor Embedding* (t-SNE) [12], were proposed.

In this paper, we focus on the analyses of DR techniques attending to their stability under variations in the parameters and in data. Although some reviews and analysis have been published [4, 13, 14], some questions referring to the uncertainty of the resulting projections remain open, despite being of particular importance for a practical use of the methods. This paper contributes to answering this questions. Section 2 introduces the motivation and the aim of this work. Section 3 describes the experimental methodology, and Section 4 illustrates both the results and the conclusions arisen from the study.

2 Stability of dimensionality reduction techniques

In dimensionality reduction, we transform a set of N high-dimensional vectors, $\mathbf{X} = [\mathbf{x}_i]_{1 \leq i < N}$, into N low-dimensional vectors ($d << D$), $\mathbf{Y} = [\mathbf{y}_i]_{1 \leq i < N}$. Mathematically, a DR technique can be understood as an application $f : \mathbb{R}^D \rightarrow \mathbb{R}^d$, where $d < D$. The general idea of DR is to embed, in the visualization space, close –or similar– points next to each other, while keeping large distances among faraway –or dissimilar– points.

Several ways of constructing the embedding exist. A general approach is to preserve pairwise distances, with an appropriate metric [5, 7]. Another possibility is using a probability-based approach to obtain a pairwise similarity matrix [11, 12]. Also, another smart way to address this problem is using a graph-based model of data, whose edges exist depending on the known entries of the pairwise distance matrix. In this approach, the weights of the edges are determined according to the nature of the data. Finally, the embedding can be constructed retaining the global [8] or the local structure of data [9, 10].

If the stability of DR algorithms is analyzed attending to parameter and data variations, certain behaviors are expected. For instance, graph-based solutions have a major problem if the constructed graph is not completely connected –a typical situation with clustered datasets–, even if the complete pairwise distance matrix is known. In this case, they are not capable of reducing the complete dataset $\mathbf{X}$, whereas methods based on pairwise distances or similarity matrix are. Moreover, the behavior of various graph-based algorithms may differ. Local-based graphs have a larger dependency on small changes in the data points than global-based ones. This implies that if newer points are added to a dataset, the embedding should have more differences in local than in global coordinates with respect to the original embedding. If adding a moderate number of new points

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

heavily modifies the representation, this is considered as negative for visualization purposes, because there is no consistency in the resulting embeddings.

On the other hand, the problem of DR can be solved in different ways. Attending to whether the objective function contains local optima or not, techniques can be classified [13] into non-convex and convex, respectively. Most of the newer techniques, such as *Isomap*, *LLE* or *LE*, are part of the second group, whereas *SNE* or *t-SNE* are in the group of the non-convex techniques. Convex techniques imply an eigenvalue decomposition. This mechanism introduces an uncertainty in the embedding, that results in possible geometrical variations, such as mirroring, rotation and translation, for the same dataset. On the other hand, non-convex algorithm embeddings are usually quite difficult to compare, because they also experience geometrical variations and they are usually initialized with a set of random points and then iteratively computed.

Attending to visualization purposes, the stability of DR techniques is an important question. The more stable the visualizations obtained with DR technique under small changes are, the easier it is to obtain visual insights in datasets, which can result into an improved knowledge discovery process. The study made in this paper is intended to provide an initial approach to help in the choice of the most suitable algorithm attending to visualization features, such as geometrical variations or cluttering.

3 Experimental methodology

To evaluate the stability of DR techniques, we present four experiments. Several DR methods are used and compared according to each stability criterion. The experiments are carried out on four well-known synthetic datasets: *Swiss roll*, *helix*, *broken swiss roll* and *twin-peaks*–see second column in Figure 1.

In this analysis, we selected the following DR techniques: *PCA*, *Isomap*, *LLE*, *LE* and *t-SNE*. The parameter settings for the experiments are shown in Table 1.

Technique	**Parameters**	**Settings**
PCA	None	None
Isomap	k: number of neighbors	$5 < k < 20$
LLE, LE	k: number of neighbors	$5 < k < 45$
t-SNE	*Perplexity*: size of a soft K-ary neighborhood	$5 < P < 45$

Table 1: Parameter settings for the experiments.

A brief description of the experiments and their objectives is shown below. While experiments 1 and 2 analyze the techniques and the influence of the parameters, experiments 3 and 4 are oriented to usual situations when working with real-datasets.

Experiment 1. Using the same dataset ($N = 1000$ points), data points are introduced to the DR technique in a random order. By applying this experiment,

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

geometrical variations in the projections are expected to appear.

Experiment 2. The objective of the experiment is to analyze the stability of the DR algorithms under changes in their parameters. Thereby, using identical datasets (1000 points), we test the behavior of each technique analyzing the continuity of the projections with different values for the parameters: number of neighbors for convex techniques and *perplexity*, size of a soft K-ary neighborhood, for t-SNE.

Experiment 3. Starting with a dataset of 1000 points, we increase its size to 1100, 1500 and 2000 points and each DR technique is trained. The objective is to observe which of the selected DR techniques perform more stably with incrementally changing datasets.

Experiment 4. We introduce different sets of points over the same topological space –i.e. swiss roll manifold–, with the objective of studying which technique yields the more stable results.

In order to improve the behavior and the stability of these DR algorithms, we propose two simple, easily applicable and low computational pre- and post-processing methods.

1. For convex DR techniques, we propose the use of *Procrustes Analysis* [15], a least-squares orthogonal mapping for manifold alignment. This post-processing method aligns the shape of a new projection according to a previous one, obtaining a linear transformation that minimizes the sum of squared errors between points. We only focus on rotation and translation, as scaling can lead to cluttered visualizations. Since this algorithm makes a comparison point by point, it can only be used with datasets that share points, so it is not applied to Experiment 4.

2. For t-SNE, we apply a pre-processing methodology: we fix for each run identical initial conditions for the location of $[\mathbf{y}_i]$, drawn randomly from $N(0, 10^{-4}I)$. By doing that, the randomness of the initialization is avoided, while the order of presentation of points in the gradient descent remains random.

4 Results and Conclusions

Due to paper length constrains, only the most relevant results are shown in Figure 1 –the complete experiments can be consulted here[1].

In Exp. 1, the general behavior of the techniques is reasonably stable, excluding *t-SNE* due to the randomness in the initialization, whose performance is really improved with the pre-processing. Rotation and mirroring effects, which are the usual variations among different orders, are easily avoided using Procrustes Analysis for manifold alignment, as in the cases of *LLE*, *Isomap* and *LE*.

In Exp. 2, attending to the influence of the parameters, the behavior of *LLE* and *LE* is quite unstable. Both techniques tend to converge into a single point

[1] `http://gsdpi.dieecs.com/actividad/stability-of-dr-techniques/`

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

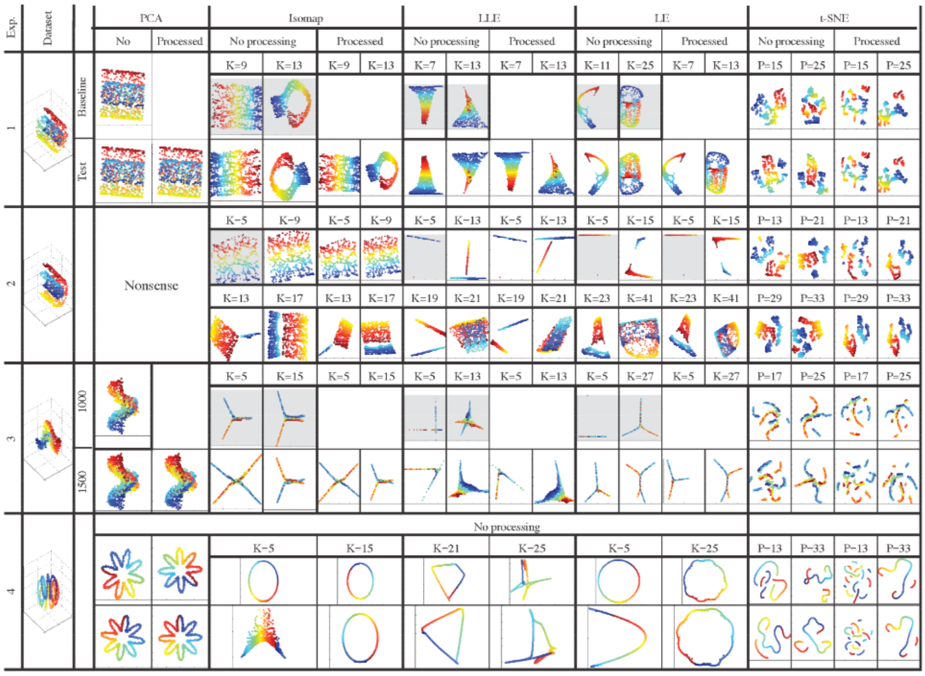

Fig. 1: Non-processed and processed experimental results are shown. Second column displays the datasets used. The experiments are presented by rows, while DR techniques are presented in columns 4-8. Shaded cells are the baselines for *Procrustes Analysis*. Notice that for Experiments 1, 3 and 4 the stability is considered for the same value of the parameter, while in Exp. 2 the stability is studied among all the values.

if the underlying manifold is not well determined –in Fig. 1, see k=5 in *LLE* and *LE*– and there is no continuity in the evolution of the projections when the number of neighbors is increased. This behavior does not appear with *Isomap*, *PCA* or *t-SNE*.

Working with incrementally changing datasets (Exp. 3) is a hard task for DR techniques, although it is important to notice that if the number of points in the dataset is large, the influence of the setup is lower. With a low number of points, it is possible that *Isomap*, *LE* and *LLE* couldn't obtain a fully connected graph, so the projection could experiment major changes when increasing the number of points. In PCA or *t-SNE*, that are not based in graph principles, the projections obtained remains quite stable.

In Exp. 4, we can see the behavior of graph-based techniques when the graph is not fully connected (see this in Isomap, *LE* and *LLE* for $k = 5$). In the case of *PCA* and *t-SNE*, the projections are reasonably stable, with only small geometric differences among the embeddings.

ESANN 2013 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 24-26 April 2013, i6doc.com publ., ISBN 978-2-87419-081-0. Available from http://www.i6doc.com/en/livre/?GCOI=28001100131010.

In general, local methods are more likely to be influenced by small changes in both data and parameter variations. LE and LLE tend to provide cluttered visualizations, whereas data points in t-SNE, *Isomap* and PCA are much more scattered. *t-SNE*, due to the nature of its gradient, tends to form small clusters in the embedding. It is interesting to emphasize that if the visualization of the whole dataset is a major requirement, graph-based techniques are not a good solution, due to the proper limitation in the construction of the graph. In contrast, both PCA and *t-SNE* are always capable of projecting the whole dataset, although the quality of the embedding is generally better in *t-SNE*, specially when working with non-linear manifolds. The performances of the pre-processing via manifold alignment are higher if the shapes are similar.

As future work, this analysis can be extended considering the behavior with real-world datasets, regarding the stability against outliers and studying alternatives to the pre- and post-processing approaches proposed.

References

[1] I.T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, 2002.

[2] G. Young and A.S. Householder. Discussion of a set of points in terms of their mutual distances. *Psychometrika*, 3(1):19–22, 1938.

[3] W. Torgerson. Multidimensional scaling: I. theory and method. *Psychometrika*, 17:401–419, 1952.

[4] J.A. Lee and M. Verleysen. *Nonlinear Dimensionality Reduction.* Information Science and Statistics. Springer, 2007.

[5] J.W. Sammon Jr. A nonlinear mapping for data structure analysis. *Computers, IEEE Transactions on*, C-18(5):401 – 409, may 1969.

[6] T. Kohonen, M. R. Schroeder, and T. S. Huang, editors. *Self-Organizing Maps.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 3rd edition, 2001.

[7] P. Demartines and J. Herault. Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. on Neural Networks*, 8(1):148 –154, January 1997.

[8] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, December 2000.

[9] S. T. Roweis and L. K. Saul. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science*, 290(5500):2323–2326, December 2000.

[10] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, June 2003. WOS:000182530600005.

[11] G. Hinton and S. T. Roweis. Stochastic neighbor embedding. *Advances in neural information processing systems*, 15:833–840, 2003.

[12] L. van der Maaten and G. Hinton. Visualizing Data using t-SNE. *Journal of machine learning research*, 9:2579–2605, November 2008.

[13] L. van der Maaten, E. Postma, and J. Van den Herik. Dimensionality reduction: A comparative review, 2009.

[14] J. Khoder, R. Younes, and F. Ouezdou. Stability of dimensionality reduction methods applied on artificial hyperspectral images. In *Computer Vision and Graphics*, volume 7594 of *Lecture Notes in Computer Science*, pages 465–474. Springer, 2012.

[15] D. G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2):87–99, 1989.