EuroVis 2013 Workshop on Visual Analytics using Multidimensional Projections (2013)
M. Aupetit and L. van der Maaten (Editors)

Stability comparison of dimensionality reduction techniques attending to data and parameter variations

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

[1]University of Oviedo, Spain
[2]Université Catholique de Louvain, Belgium

Abstract

The analysis of the big volumes of data requires efficient and robust dimension reduction techniques to represent data into lower-dimensional spaces, which ease human understanding. This paper presents a study of the stability, robustness and performance of some of these dimension reduction algorithms with respect to algorithm and data parameters, which usually have a major influence in the resulting embeddings. This analysis includes the performance of a large panel of techniques on both artificial and real datasets, focusing on the geometrical variations experimented when changing different parameters. The results are presented by identifying the visual weaknesses of each technique, providing some suitable data-processing tasks to enhance the stability.

Categories and Subject Descriptors (according to ACM CCS): I.2.6 [Computing Methodologies]: Artificial Intelligence—Machine learning

1. Introduction

The technological evolution in recent years has resulted in an unprecedented generalization of data sources, which usually implies not only better, but also bigger datasets. Social networks and *open data* initiatives are clear examples of this new trend. For many economic sectors, these huge amounts of data include potential information, which usually is hidden and therefore it is necessary to be extracted. As a consequence, this requires having suitable techniques for analyzing and visualizing all these data. As these data are typically high-dimensional and so they can not be visualized directly in a two/three dimensional lattice, dimensionality reduction (DR) techniques play a key role, making a transformation of these data into a meaningful, visualizable and reduced-dimensional space.

Dimensionality reduction includes techniques that allow the user to obtain meaningful data representations of a given dimensionality, improving the process of comprehension and analysis of data. In this field, several techniques have been proposed –we only focus on unsupervised techniques. Principal Component Analysis (PCA) [Jol05] or Multidimensional Scaling (MDS) [Tor52, YH38] are well-known examples of linear DR techniques. Although linear techniques usually perform well, they fail when working with complex datasets, which lie on a nonlinear manifold. In these cases, nonlinear techniques performs better as they have the ability to deal with this kind of data. Nonlinear DR techniques [LV07] started to appear later, especially with nonlinear variants of multidimensional scaling [Sam69] and neural approaches [KSH01, DH97]. In recent years, the evolution in the DR field has focused on spectral techniques, such as Isomap [TSL00], Local Linear Embedding (LLE) [RS00], Laplacian Eigenmaps (LE) [BN03], and non-convex techniques, such as Stochastic Neighbor Embedding (SNE) [HR02] and t-Distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08]. These modern DR techniques are usually known as manifold learning algorithms [TDBET98].

This paper contributes to the study the stability of unsupervised DR techniques with respect to variations of their parameters and of the data. In contrast to [KYO12], where the authors present a study the stability by introducing a perturbation to data, this paper focuses on answering question referring to both data and parameters. The reminder of this paper is organized as follows. Section 2 introduces the motivation and the aim of this work. Section 3 describes the experimental methodology and a description of the data used. Section 4 illustrates the results. Finally Section 5 ends up with the conclusions.

DOI: 10.2312/PE.VAMP.VAMP2013.005-009

2. Stability of dimensionality reduction techniques

Dimensionality reduction transforms 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 $d < D$. Thereby, the main idea of DR is to keep far-away the points which are very dissimilar in the input space, while keeping close the ones that are near to each other in the original lattice.

In order to better understand the stability of DR techniques, it is necessary to take into account the way the embedding is constructed. One of the simplest approaches to data projection is to preserve pairwise distances, either using an appropriate metric [Sam69, DH97] or using a probability-based approach [HR02, vdMH08], obtaining a pairwise distance or dissimilarity matrix respectively. However, the previous approaches cannot be applied if the pairwise distance matrix has unknown elements. In this case, an alternative is to compute a graph model of the data, whose edges depend on the known elements of the pairwise distance matrix. Particularly, this method can also be applied when all the elements of the pairwise matrix distance are known, like in many manifold learning algorithms. Finally, the embedding can be obtained from this model whether retaining the global [TSL00] or the local structure of data [RS00, BN03].

If the stability of DR algorithms is analyzed attending to parameter and data variations, some behaviors are expected. For instance, graph-based solutions have a major drawback 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 matrices 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. If adding a moderate number of new points heavily modifies the representation, this is considered as negative for visualization purposes, because there is no continuity in the resulting embeddings.

Regarding the way the embedding is solved, two major alternatives exist [vdMPVdH09]. On the one hand, convex techniques –minimization of a convex cost function–, such as Isomap, LLE or LE, involve an eigenvalue decomposition. The mathematical procedure introduces indeterminacies in the embeddings, which can lead to irrelevant geometric transformations like mirroring, rotation and translation of the projection between different results. On the other hand, non-convex techniques –minimization of a non-convex cost function–, e.g. SNE or t-SNE, use the gradient descent algorithm in order to obtain the final projection. The problem with these algorithms comes from the randomness introduced in the process: the initialization is random for all non-convex techniques, and the way in which the data points are presented in each iteration is also random only in stochastic gradient descent algorithm, so it is difficult to obtain comparable projection under the same conditions.

From the visual analytics point of view, some behaviors are desired [War08]. Perception and cognition are important parts of the process of visual analytics and it is necessary to take them into account in order to select the most suitable technique for a visual analytics application. Thereby, some of the behaviors previously described about the performances of DR techniques are not good for this process.

Geometric variations –e.g. rotation, translation, ...– in the projection make the analysis for the user difficult. Internally, the human brain needs to revert these transformations in order to ease the comparison of projections, slowing the process. Thus, it is necessary that the DR techniques take into account this fact. These geometric transformations include not only the variations caused by the algorithm, but also the discontinuity when changing some data or algorithm parameters, such as the order of the data points or the neighborhood parameter respectively.

Apart from these, other requirements for DR techniques in the visual analytics field can be related to time computation. Particularly in the case of interactive applications, the time between the action of the user and change in the display must be the shortest possible, so if the algorithm is time-consuming, maybe it is not suitable for interactive purposes.

The study made in this paper intends to provide an initial approach to helping in the selection of a suitable algorithm that combines a good dimensionality reduction performance and good visualization features.

3. Experimental methodology

In this section, we describe the experiments and the methodology applied, as well as the processing tasks we propose for improving the stability of DR techniques.

3.1. Dimensionality Reduction Techniques analyzed

In order to evaluate the stability and robustness of the DR techniques, we propose the following experiments. Each experiment focuses on different desired features for the application of DR techniques in the visual analytic field. The experiments are carried out on four well-known synthetic datasets: S-curve, Swiss roll, helix and twin-peaks, and two natural datasets, *MNIST* [LBBHov] and *Olivetti faces* [SH94] (see Figure 1).

In this analysis, we select six DR techniques, which are enumerated below. The settings of the techniques for the experiments are shown in Table 1.

- Principal component analysis (PCA) [Jol05].
- Isomap [TSL00].
- Laplacian Eigenmaps (LE) [BN03].

- Locally Linear Embedding (LLE) [RS00].
- Stochastic Neighbor Embedding (SNE) [HR02].
- t-Distributed Stochastic Neighbor Embedding (t-SNE) [vdMH08].

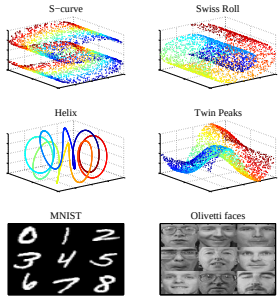

Figure 1: *Datasets used for the stability comparison.*

Technique	Parameters	Settings
PCA	None	None
Isomap LE LLE	k: number of neighbors	$4 < k < 40$
SNE t-SNE	*Perplexity* (P): size of a k-ary neighbor	$4 < P < 40$

Table 1: *Parameter settings for the experiments.*

3.2. Experiment description

A detailed description of each experiment applied to the data is shown below. On the one hand, Experiments 1 and 2 are related to each technique in order to study the influence of the parameters in the resulting embeddings. On the other hand, Experiments 3 and 4 are oriented to common scenarios when working with natural datasets.

Experiment 1. This experiment focuses on analysing the influence of the order in which the data points are introduced to the DR algorithm. Datasets of 1000 points are introduced in different random orders, with the aim of analyzing the geometric variations experimented by the techniques.

Experiment 2. For this experiment, the objective is to study the stability of the DR techniques under changes in their parameters (see Table 1). Thus we test the behavior of the DR techniques using identical datasets, analyzing their visual continuity in the resulting projections for a wide set of values for the parameters.

Experiment 3. We apply this experiment regarding the performance of DR techniques when working with incrementally changing datasets. Starting from a dataset of 800 points, we increase the number of points in several steps observing the transformations generated by each technique.

Experiment 4. In this case, the experiment aims at analyzing the variations observed when datasets from the same topological space, but with different points, are projected, in order to study which technique yields the more stable results.

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

1. In the case of convex techniques, we propose the use of *Procrustes Analysis* [Ken89]. This algorithm is a mathematical procedure in statistical shape analysis that allows one to analyze a set of shapes. Basically, this method computes the rotation matrix and the translation vector of each projection, according to a projection considered as the baseline.
   Since this algorithm makes a point-by-point comparison, it can only be used with datasets that share them, so it does not apply to Experiment 4.
2. For non-convex techniques, our approach focuses on controlling the initialization of the algorithms, fixing the initial conditions by fixing the random seed used to generate them, while using a stochastic gradient descent algorithm to obtain the optimal solution. This *semi-random* approach can control the initialization, while the random introduction of the points helps to avoid local minima.
   It is important to point out that our approach is a post-processing task in the case of convex techniques, while in non-convex algorithms, the method is a pre-processing.

4. Results

Due to paper length constrains, only the most relevant results are shown in Figure 2.

For Exp. 1 (Figure 2, top left), we show the results for Isomap, LLE, whose performance is similar to LE, and t-SNE, similar to SNE. 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. In the case of convex techniques, the approach proposed, based on Procrustes Analysis, is able of aligning the embedding in a suitable way, avoiding geometrical transformations.

Referring to the influence of the parameters (Figure 2, top righ), the behavior of LE –similar to LLE– is more unstable than Isomap or t-SNE. As it can bee seen in Figure 2, the continuity in the case of Isomap is better than LE, which also tends to obtain cluttered visualization. In the case of convex techniques, the semi-random approach gives more stability to the projections obtained than with a random initialization.

When working with incrementally changing datasets (Figure 2, bottom left), Isomap, LE and LLE are not always capable of obtaining a fully connected graph if the number of points is not large enough, while this problem does not appear in PCA or t-SNE. Concerning this experiment, the

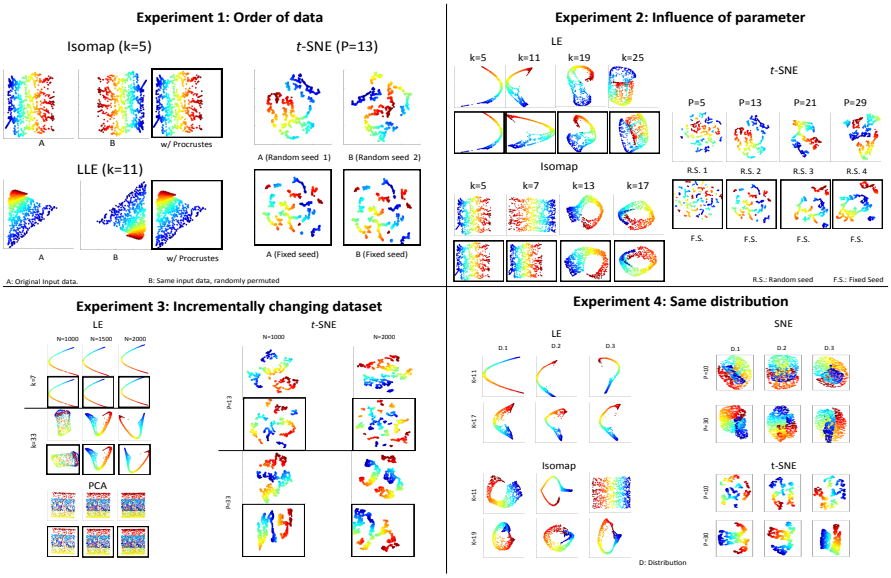

Figure 2: *Relevant results obtained from the experimental methodology proposed for the Swiss roll dataset. Notice that framed projections corresponds to processed results with Procrustes Analysis and the semi-random approach.*

post-processing in t-SNE improves the continuity between projections with the same perplexity. Also, it is important to emphasize that if the number of points in the dataset is large, the influence of the parameter is lower.

In Experiment 4 (Figure 2, bottom righ), the projection of graph-based techniques –Isomap, LE and LLE– may differ a lot depending on the data, while in the case of SNE and t-SNE –or PCA, which is not shown–, the projections are reasonably stable for the same parameter.

5. Conclusions and future work

In this paper, we present a study of the analysis of stability of unsupervised dimensionality reduction techniques under variations in data and in the parameters of the algorithms. The analysis is achieved through different experiments over several artificial and natural datasets.

As a general conclusion, local methods –LE and LLE, which retain local structure of data– are more likely to be influenced by small changes in both data and parameter variations. They also 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 point out that if the visualization of the whole dataset is a major requirement, graph-based techniques are not a good solution, as the construction of the graph can lead to not fully-connected graphs and so not all points will be embedded. On the other, PCA, *t*-SNE and SNE are not affected by this problem. Among them, the quality of the embedding is usually better in *t*-SNE and SNE, particularly when working with non-linear manifolds.

Our approaches to improving the stability of DR techniques obtain satisfying results. The Procrustes Analysis algorithm applied to convex techniques performs well in most of the cases, although it has a major dependence on the projection chosen as the baseline for the geometric transformation. In the case of non-convex techniques, the *semi-random* approach proposed makes a stronger control of the final shape, leading to more comparable projections.

As a future work, some possible directions are promising. Particularly, the application of the methodologies proposed in this paper to visual analytics tools in order to ease the knowledge discovery process by stabilizing the projections, as well as the extension of the study including other an out-of-sample comparison, supervised techniques and the definition of a metric of stability for DR techniques.

Acknowledgement

This work has been financed by the Spanish Ministry of Science and Education and FEDER funds under grants DPI2009-13398-C02-01 and by the Government of Asturias. J.A.Lee is a Research Associate with the Belgian F.R.S.-FNRS (Fonds National de la Recherche Scientifique).

References

[BN03] BELKIN M., NIYOGI P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation 15*, 6 (June 2003), 1373–1396. 1, 2

[DH97] DEMARTINES P., HERAULT J.: Curvilinear component analysis: a self-organizing neural network for nonlinear mapping of data sets. *IEEE Trans. on Neural Networks 8*, 1 (Jan. 1997), 148 –154. 1, 2

[HR02] HINTON G., ROWEIS S.: Stochastic neighbor embedding. *Advances in neural information processing systems 15* (2002), 833–840. 1, 2, 3

[Jol05] JOLLIFFE I.: *Principal component analysis*. Wiley Online Library, 2005. 1, 2

[Ken89] KENDALL D. G.: A survey of the statistical theory of shape. *Statistical Science 4*, 2 (1989), 87–99. 3

[KSH01] KOHONEN T., SCHROEDER M. R., HUANG T. S. (Eds.): *Self-Organizing Maps*, 3rd ed. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2001. 1

[KYO12] KHODER J., YOUNES R., OUEZDOU F. B.: Stability of dimensionality reduction methods applied on artificial hyperspectral images. In *Computer Vision and Graphics*, Bolc L., Tadeusiewicz R., Chmielewski L. J., Wojciechowski K., (Eds.), no. 7594 in Lecture Notes in Computer Science. Springer Berlin Heidelberg, Jan. 2012, pp. 465–474. 1

[LBBHov] LECUN Y., BOTTOU L., BENGIO Y., HAFFNER P.: Gradient-based learning applied to document recognition. *Proceedings of the IEEE 86*, 11 (Nov), 2278–2324. 2

[LV07] LEE J., VERLEYSEN M.: *Nonlinear dimensionality reduction*. Springer, 2007. 1

[RS00] ROWEIS S. T., SAUL L. K.: Nonlinear dimensionality reduction by locally linear embedding. *Science 290*, 5500 (Dec. 2000), 2323–2326. 1, 2, 3

[Sam69] SAMMON J.W. J.: A nonlinear mapping for data structure analysis. *Computers, IEEE Transactions on C-18*, 5 (may 1969), 401 – 409. 1, 2

[SH94] SAMARIA F. S., HARTER A. C.: Parameterisation of a stochastic model for human face identification. In *Applications of Computer Vision, Proceedings of the Second IEEE Workshop on* (1994), pp. 138–142. 2

[TDBET98] TOLLIS I. G., DI BATTISTA G., EADES P., TAMASSIA R.: *Graph Drawing: Algorithms for the Visualization of Graphs*. Prentice Hall, July 1998. 1

[Tor52] TORGERSON W.: Multidimensional scaling: I. theory and method. *Psychometrika 17*, 4 (1952), 401–419. 1

[TSL00] TENENBAUM J. B., SILVA V. D., LANGFORD J. C.: A global geometric framework for nonlinear dimensionality reduction. *Science 290*, 5500 (Dec. 2000), 2319–2323. 1, 2

[vdMH08] VAN DER MAATEN L., HINTON G.: Visualizing data using t-SNE. *Journal of Machine Learning Research 9* (Nov. 2008), 2579–2605. 1, 2, 3

[vdMPVdH09] VAN DER MAATEN L., POSTMA E., VAN DEN HERIK J.: Dimensionality reduction: A comparative review, 2009. 2

[War08] WARE C.: *Visual thinking for design*. Morgan Kaufmann Pub, 2008. 2

[YH38] YOUNG G., HOUSEHOLDER A.: Discussion of a set of points in terms of their mutual distances. *Psychometrika 3*, 1 (1938), 19–22. 1