# Nonlinear Dimensionality Reduction for Visualization

Michel Verleysen[1] and John A. Lee[2]

[1] ICTEAM, Université catholique de Louvain, Belgium
michel.verleysen@uclouvain.be
http://perso.uclouvain.be/michel.verleysen/
[2] IREC, Université catholique de Louvain, Belgium
john.lee@uclouvain.be

**Abstract.** The visual interpretation of data is an essential step to guide any further processing or decision making. Dimensionality reduction (or manifold learning) tools may be used for visualization if the resulting dimension is constrained to be 2 or 3. The field of machine learning has developed numerous nonlinear dimensionality reduction tools in the last decades. However, the diversity of methods reflects the diversity of quality criteria used both for optimizing the algorithms, and for assessing their performances. In addition, these criteria are not always compatible with subjective visual quality. Finally, the dimensionality reduction methods themselves do not always possess computational properties that are compatible with interactive data visualization. This paper presents current and future developments to use dimensionality reduction methods for data visualization.

**Keywords:** visualization, dimensionality reduction, manifold learning.

## 1 Introduction

Data analysis has become an overwhelming discipline in many areas of our everyday life. Modern ways to acquire and to store information are responsible for a data deluge. Extracting relevant information from huge amounts of data is a challenge that is responsible for important and recent scientific developments in statistics and machine learning. Applications of such new data analysis tools range from (bio)medicine to consumer profiling, industrial and quality control, environmental monitoring, and many others.

A specific aspect of data analysis is visualization. Visual inspection of data is unavoidable in many practical situations. The main reason is that, despite the power of modern data analysis tools, few of them are really blind in the sense that they can be applied without any understanding of the data at hand: preliminary qualitative knowledge is needed, and visualization might help in this context for example in finding outliers, clusters, etc. Another reason, among many other ones, to visualize data is that non-experts are often difficult to convince about the benefits of mathematical tools, if they cannot see the results in the way they are used to see and to analyse them.

M. Lee et al. (Eds.): ICONIP 2013, Part I, LNCS 8226, pp. 617–622, 2013.

Visualization has been developed rather independently by two research communities. One one side, the machine learning community has developed dimensionality reduction (DR) methods that may be used for visualization if the resulting dimensionality is restricted to be 2 or 3. One the other side, the information visualization (IV) community has developed graphic ways of representing the information under the angle that is most usable by the user. Unfortunately, only few attempts exist to combine the features and advantages of both fields.

Dimensionality reduction is a generic term including manifold learning, nonlinear projection, etc. The goal of DR is to handle data that contain a high number of attributes (and therefore cannot be visualized easily), and to reduce them (through the optimization of mathematical information content criteria) to a lower-dimensional space, while preserving as much as possible the information content in the data. If the dimension of the latter space is 2 or 3, this provides an obvious way to visualize data. On the other hand, information visualization focuses on user-centric graphic objectives, and largely relies on controllability (the user decides which is the best way he needs for representing data) and interaction (the controllability is achieved through a user interface that responds almost immediately, making different views affordable in a single session).

Controllability and interaction are two concepts that are mostly absent from dimensionality reduction. Most DR methods rely on the algorithmic optimization of a predefined information criterion; although the results can be satisfactory on the point of view of information content preservation, they are usually not in terms of effective visualization. Problems such as the sensitivity to initial conditions, possible rotations and mirroring are common. More dramatically, the criterion to be optimized has to be predefined; adjusting the criterion to another balance between conflicting goals (see below for details) needs to run the algorithm again, which implies prohibitive computational load and simulation times.

## 2 Dimensionality Reduction: State-of-the-Art

Dimensionality reduction [1] has its roots in methods such as principal component analysis (PCA) [3]. PCA can be used to reduce the dimensionality of high-dimensional data; new features are generated by linear combinations of the original features, by optimizing a maximum variance/minimum loss of information criterion. If the resulting dimension is limited to 2, PCA provides an easy way to represent high-dimensional data; however, PCA only aims to preserve simple second-order statistics (directions of main variance) and can miss the important characteristics of more complicated data distributions.

Except for a few older methods like Sammon's mapping [4], most nonlinear extensions to PCA were frenetically developed in the 80s and 90s. Relaxing the linearity constraint has been found to open the way to new information preservation criteria, which tend to yield better low-dimensional data representations in practice. Methods such as Sammon's mapping and curvilinear component analysis (CCA) [5] result from a nonlinear view of PCA: while PCA tries to preserve all Euclidean distances between pairs of points in a data set (while projecting it

to a lower-dimensional space), Sammon's mapping and CCA emphasize preservation of small distances, which are usually the ones that are the most important for effective visualization. At the same time, methods such as Curvilinear distance analysis (CDA) [6] and Isomap [7] were developed, where the distances to be preserved are based on the data distribution itself, such as geodesic or graph distances. The graph distances allow a better representation of the important-to-preserve similarities and distances in the data; in some sense they act against the well-known unreliability of estimating data properties in high-dimensional spaces, called the curse of dimensionality [8].

The diversity of the many DR methods has revealed how difficult it can be to analyse relationships between different methods and their suitability for a particular analyst's needs. Part of the problem is that nonlinear DR has been done by optimizing relatively abstract criteria, and the relationships of the criteria to helping analysts in a meaningful task has not been clear. Recent analysis has made it clear that (at least) two conflicting goals exist in DR: in terms of a relaxed form of distance preservation called neighbourhood preservation, 1) two data items that are neighbours in the original space should remain neighbours in the projection space, and 2) two data items should be shown as neighbours in the projection space only if they are neighbours in the original space. Recently, both two goals have been shown to correspond to performance in an information retrieval task, visual retrieval of neighbours from the output display, as measured by the information retrieval measures precision and recall respectively. The conflict between the goals yields a natural trade-off between the precision and recall measures, and between visualizations that are good for one goal versus the other [9] [10]. For example, projecting a spherical surface distribution to a two-dimensional space results in flattening the sphere surface onto a circle if only goal 2 (recall) is optimized or cutting the surface open like an orange-peel world map if only goal 1 (precision) is optimized. This example illustrates the conflicting requirements in DR and visualization; neither result is obviously better than the other, the choice, or the compromise, should be guided by the users needs.

## 3 Visualization

Information visualization [2] has developed ways to visualize data in a user-centric way. IV relies on the adequation between the method and the cognitive goal of understanding data. Many information visualizations are interactive, reflecting the difficulty to represent data in a unique, undebatable way. Interaction also closes the loop with the user: interaction is based on cognition, therefore helps reflecting the users' needs.

Information visualization methods are largely based on extensive software that combine user goals, modern computer visualization features, and interaction. The representation principles behind the methods are usually quite simple (parallel coordinates, dendograms, trees, heatmaps, etc.), although recent information visualization often involves dimensionality reduction methods, such as principal component analysis and Sammon's mapping.

Machine-learning based dimensionality reduction and information visualization are complementary: the DR field has developed advanced mathematical criteria and ways to optimize them, while IV takes into account users' needs, cognitive aspects and computer resources. However combining the advantages from two fields requires an in-depth study of performance criteria and computational requirements.

## 4 Quality Criteria and Computational Requirements

Quality criteria exist to measure the performances of nonlinear dimensionality reduction methods [10]. Most of them yield a pair of values (trustworthiness and continuity, mean relative rank errors, etc.), which also reflects these dual or conflicting requirements. These measures can assess the compromise between the requirements for a given DR result (visualization), but so far the only way to influence the compromise is to change the criterion of the DR method. Changing the criterion yields two difficulties: A) the DR algorithm must be rerun, and since most algorithms take from tens of seconds to hours on standard computers, fast interaction with the user becomes impossible; B) the link between control parameters in the mathematical optimization goal and the behaviour of the DR algorithm is far from straightforward, especially when several control parameters are involved.

These limitations exist even in the most recent DR algorithms. For example, algorithms from the stochastic neighbour embedding (SNE) family [11] [12] have been shown to outperform distance-based methods in the last years, especially when the original space is high-dimensional. They optimize a divergence between distributions of distances or neighbours in both spaces, and can partially alleviate the curse of dimensionality by adjusting priors on the distributions, but the same basic difficulties remain: the need to rerun the algorithm when the criterion is modified, and the difficulty of controlling in an intuitive way the compromise between conflicting objectives.

On the other hand quality criteria when visualization is involved are far from the information content perspective brought by such trustworthiness and continuity pairs of measures. Performances in visualization are measured in a way that takes the cognitive process into account, thus involving the user. In this context it is much more difficult to define in advance the exact mathematical criterion to be optimized. Interaction is thus needed between the method and the user: the visualization is modified by the user, who can estimate in real-time the adequacy between the visual result and his expected goal.

Interaction necessitates speed: one cannot reasonably expect the user to wait for more than a few seconds between the request and the response. In DR methods, the quality criteria (or a proxy) are directly optimized to give the resulting projection. As most modern criteria are nonlinear, non-linear optimization is involved, with a number of parameters to optimize that is proportional to the number of data in the database. In most situations, depending on the application and on the DR method, this results in unaffordable computation times.

## 5 New Developments

In order to lead to effective and usable visualization methods, the modern tools in the dimensionality reduction field have to be adapted from several perspectives.

First, effective visualization necessitates parameters that may be controlled by the user, in order to take cognitive aspects into account and adapt the results of the algorithm to the user's needs. Most DR methods do contain parameters. For example, often one of them implements a compromise between the trustworthiness and the continuity of the projection. Another might control the influence of outliers, ... In theory it is thus possible to influence the visualization through user-controlled parameters. However, nothing indicates that the choice of these parameters, guided by algorithmic and mathematical considerations, is appropriate with regards to the cognitive control. There is thus a need for identifying the role of existing parameters and, if necessary, to change them to parameters closer to the user's needs.

Secondly the DR methods have to be rethought in the light of visualization needs. Most modern DR methods are shown to outperform competitors in specific settings, and according to specific quality criteria. But are these criteria the most appropriate ones when visualization is concerned? Is it reasonable possible to use them as a proxy of subjective, cognitive criteria? Conversely, would it be possible to express subjective criteria in a mathematical form and optimize them directly?

Third, the question of stability has to be investigated. DR methods result in *a* representation of the data. However what concerns visual perception, several equivalent projections could be thought of (for example rotations, scalings, ...). When the parameters of the DR methods are modified, even slightly, another optimum of the criterion to optimize can be found, leading to an almost equivalent but completely different projection. Such instability is of course undesirable in the context of visualization, and must be controlled at the algorithmic level.

Finally, the computational requirements should be seriously investigated, in the light of the possibility for user interaction. Fast algorithms have a clear advantage. When necessary, incremental methods could be developed: slight changes in the parameters of a method should not result in largely different representations. This "continuity" in the process could be exploited to reduce the computation time after user interaction.

## 6 Conclusion

The field of Machine Learning has generated a large number of dimensionality reduction methods. These methods can be used for the visualization of data, which is a fundamental step in exploratory data analysis. In parallel, the field of Information Visualization develops user-centric graphic ways to visualize data based on cognitive results. The complementarity of the approaches is a challenge for the future developments of dimensionality-based visualization methods: how to

incorporate user control, cognitive criteria, stability and computational requirements in DR methods are key questions opening new perspectives for research on dimensionality reduction.

## References

1. Lee, J.A., Verleysen, M.: Nonlinear dimensionality reduction. Springer (2007)
2. Keim, D.A., Mansmann, F., Schneidewind, J., Thomas, J., Ziegler, H.: Visual analytics: Scope and challenges. In: Simoff, S.J., Böhlen, M.H., Mazeika, A. (eds.) Visual Data Mining. LNCS, vol. 4404, pp. 76–90. Springer, Heidelberg (2008)
3. Jolliffe, I.T.: Principal Component Analysis. Springer-Verlag, New York, NY (1986)
4. Sammon, J.W.: A nonlinear mapping algorithm for data structure analysis. IEEE Transactions on Computers, CC-18(5), 401–409 (1969)
5. Demartines, P., Hérault, J.: Curvilinear component analysis: A self-organizing neural network for nonlinear mapping of data sets. IEEE Transactions on Neural Networks 8(1), 148–154 (1997)
6. Lee, J.A., Verleysen, M.: Curvilinear distance analysis versus isomap. Neurocomputing 57, 49–76 (2004)
7. Tenenbaum, J.B., de Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. Science 290(5500), 2319–2323 (2000)
8. Donoho, D.L.: High-Dimensional Data Analysis: The Curse and Blessings of Dimensionality. Lecture for the American Math. Society Math. Challenges of the 21st Century (2000)
9. Venna, J., Peltonen, J., Nybo, K., Aidos, H., Kaski, S.: Information retrieval perspective to nonlinear dimensionality reduction for data visualization. Journal of Machine Learning Research 11, 451–490 (2010)
10. Lee, J.A., Verleysen, M.: Quality assessment of dimensionality reduction: Rank-based criteria. Neurocomputing 72(7-9), 1431–1443 (2009)
11. Hinton, G., Roweis, S.T.: Stochastic neighbor embedding. In: Becker, S., Thrun, S., Obermayer, K. (eds.) Advances in Neural Information Processing Systems (NIPS 2002), vol. 15, pp. 833–840. MIT Press (2003)
12. van der Maaten, L., Hinton, G.: Visualizing data using t-SNE. Journal of Machine Learning Research 9, 2579–2605 (2008)