Visualizing the trustworthiness of a projection

Michaël Aupetit

CEA - Département
Analyse Surveillance Environnement
BP 12, 91680, Bruyères-Le-Châtel, France
michael.aupetit at cea.fr

Abstract. The visualization of continuous multi-dimensional data based on their projection in a 2-dimensional space is a way to detect visually interesting patterns, as far as the projection provides a faithful image of the original data. We propose to visualize directly in the projection space, how much the neighborhood has been preserved or not during the projection. We color the Voronoï cells associated with the segments of the Delaunay graph of the projections, according to their stretching or compression. We experiment these techniques with the Principal Component Analysis and the Curvilinear Component Analysis applied to different databases.

1 Introduction

1.1 Exploratory analysis by projecting data

The exploratory analysis of a set of multi-dimensional data is essential for the expert to apprehend and process ever growing databases. In this work, we consider the case of data expressed as vector of coordinates in some D-dimensional Euclidean space to which we refer as the original space in the sequel. Among other techniques [2], projecting the data onto a 2-dimensional space is a way for the expert to apprehend visually their topology in the original space, e.g. their connectedness, the number of clusters and eventually their shape or the local intrinsic dimension... Here we focus on the continuous projection techniques which associate to each original datum an image in the 2-dimensional Euclidean projection space through either a linear or a nonlinear projection. The former class of projections contains the axis parallel projection, the Principal Component Analysis (PCA) [4] and the classical linear Multi-Dimensional Scaling (MDS) [8]. The latter class of non-linear projections encompasses the Sammon's Non Linear Mapping (NLM) [7] which aims at preserving pairwise distances between data from the original space to the projection space focusing on the small distances in the original space, or the Curvilinear Component Analysis (CCA) [3], a variant of the NLM which focuses on the short distances in the projection space instead.

1.2 Visualizing distortions

As first highlighted by Venna and Kaski [9], it is very important for the expert to know whether nearby data in the projection space are actually nearby in the original space or not. A projection is told "trustworthy" in this case. Venna and Kaski proposed to measure the "trustworthiness" to perform more relevant comparisons between projection techniques. However, this measure is a sum over the trustworthiness measured at each point, providing a number which characterizes globally the projection. Here we would like to propose a way to "visualize" the trustworthiness between some pairs of points.

Most of the efforts in visualizing distortions have been carried on the Self-Organizing Maps [6, 10], but very few on the continuous projection techniques [1, 11]. In our previous work [1], we proposed to visualize compression, stretching and proximity in the original space by coloring Voronoï cells of each projected datum. The proximity measure is particularly interesting because it shows all the original pairwise distances associated with a selected datum in the projection space. However, it does not give an overview of the trustworthiness of all the projected data at one glance.

To complete this work, we propose to visualize at once all pairwise distortions associated with neighboring data in the projection space, by coloring some regions attached to these neighboring pairs of data. To avoid overlapping of the colored regions, we propose to consider the pairs of data whose Voronoï cells are adjacent in the projection space, so those projected data which are connected by an edge of their Delaunay graph.

2 Framework

We consider a (N, N) dissimilarity matrix X obtained by computing Euclidean distances X_{ij} between any pairs (x_i, x_j) of data $\underline{x} = (x_1, \ldots, x_N)$ in an original D-dimensional space $E = \mathbb{R}^D$, and the set of corresponding projections $\underline{y} = (y_1, \ldots, y_N)$ in the projection space $F = \mathbb{R}^2$ Euclidean with the distance matrix Y.

The Voronoï cell V_i of the point y_i is defined as [5]:$\forall y_i \in \underline{y},\ V_i = \{v \in F \mid \forall y_j \in \underline{y}, (v - y_i)^2 \leq (v - y_j)^2\}$.

These cells cover the projection plane. We then consider such pairs of projections for which the Voronoï cells are adjacent, so that the set of edges which connect such pairs $\underline{L} = \{\{i, j\} \in (1, \ldots, N)^2 | i \neq j, V_i \cap V_j \neq \emptyset\}$ is the set of edges of the Delaunay graph of the projections $\underline{y}$ in the projection space [5]. The "segment-Voronoï " cell, which is the Voronoï cell of the line segment joining y_i and y_j, is defined as: $\forall\{i, j\} \in \underline{L}, V_{ij} = \{v \in F \mid \forall k, l \in \underline{L}, d_{ij}(v) \leq d_{kl}(v)\}$ where $d_{ij}(v) = \min_{\alpha\in[0,1]} (v - (\alpha y_i + (1 - \alpha)y_j))^2$.

We consider that a measure $m_{ij} \in [-1, 1]$ is associated with a pair $\{y_i, y_j\}$ of Delaunay neighbors in the projection space. We propose to use a colormap going from black (-1) to white (+1) passing through grey (0) to color the corresponding segment-Voronoï cell. We define a measure of stretching and compression for the segments m_{ij} as

$$\forall\{i, j\} \in \underline{L}, m_{ij} = \frac{Y_{ij} - X_{ij}}{\max_{\{k,l\}\in\underline{L}}(\mathrm{abs}(Y_{kl} - X_{kl}))}$$

This measure is relative to the maximum distortion measured between pairs of points and their projections. It allows visualizing directly in the projection space, whether the original pairwise distance X_{ij} between Delaunay neighboring data $\{y_i, y_j\}$ in the projection space has been preserved (average grey), stretched (black) or compressed (white) after the projection. This gives an idea about how much the visualized closeness can be trusted upon.

3 Experiments

We experiment this method onto the same data sets used in our previous work [1] (except for the parallel planes) allowing to compare both visualization techniques. We show in which way the proposed visualization method is relevant to qualify the trustworthiness of the projections.

We use PCA and CCA onto three simple data sets with known topology: two interlaced rings in $\mathbb{R}^3$ in figure 1, a sphere in $\mathbb{R}^3$ in figure 2, and two parallel planes in $\mathbb{R}^3$ in figure 3. The results are discussed in the caption of the figures. Notice that no stretching occur during PCA.

4 Discussion

It appears that the analyst should pay attention to the areas where nearby cells show strong contrasting colors. It also appears that this technique is suitable to visualize whether pairs of nearby projections are images of effectively nearby original points, but it is less relevant to deal with pairs of far apart projections.

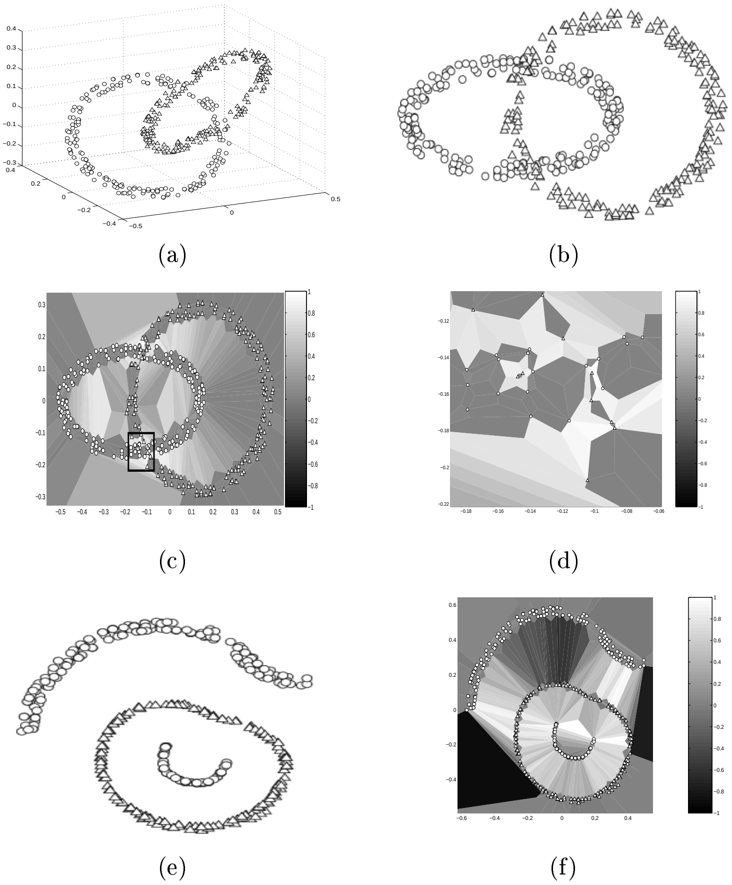
The fact it is a relative distortion measure makes it sensitive to noise, showing strong distortions where in fact the projection is not as bad. An absolute distortion measure should be used in parallel to distinguish between noise and true artifacts.

5 Conclusion

We proposed a method to evaluate visually the trustworthiness of a continuous projection. It is based on the coloring of segment-Voronoï cells associated with Delaunay neighboring pairs of projected data. The color is set according to the relative stretching or compression which occured on each segment during the projection. The real originality of our approach lies in the use of these segment-Voronoï cells making visible the pairwise distortions in the place where they just occur, helping the analyst to interprete the projections.

This allows detecting rapidly gluing of manifolds. However, it does not allow evaluating where tearing occured as can do the proximity visualization we proposed in our previous work [1].

An interesting idea to explore has been proposed by one of the anonymous reviewers, using the ranks of the data points instead of distances, as in the Trustworthiness measure [9]. This method is complementary to other existing visualization methods of distortions measures. We plan to develop this approach for 3-dimensional projections.

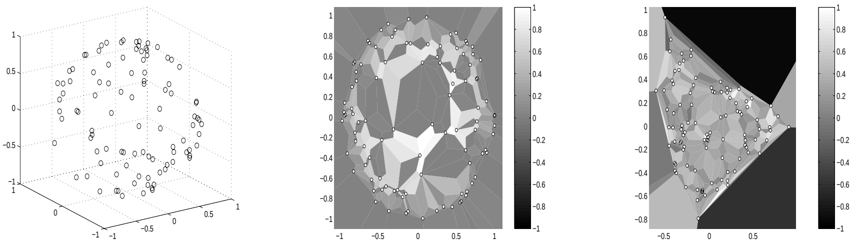


(a) (b)

(c) (d)

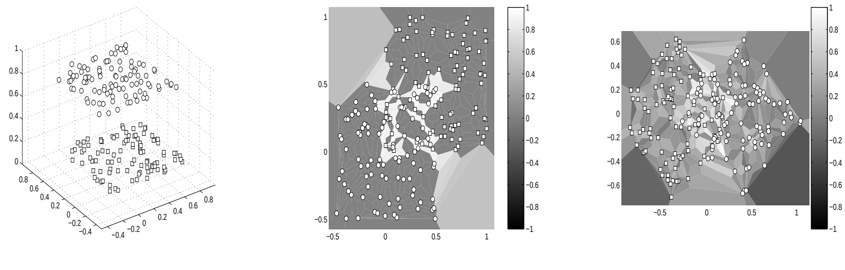
(e) (f)

Fig. 1: **Projection of two interlaced rings :** (a) Original 3D data whose topology and geometry is supposed to be unknown but is to be partially recovered through the projections. Data projected by the PCA (b) and the CCA (e) as it appears usually (The markers allow distinguishing to which ring belong the original data). Shall we trust what we see? (c) Segment-distortions after PCA. The left ring is the most distorted. The distortions occur across the rings (bright cells) but not along them (average grey cells) (d) A zoom at the bottom crossing (box in (c)) shows a gluing of manifolds: projections close to each other have adjacent segment-Voronoï cells with very different colors, showing that both rings have been glued while they are disconnected in the original space. This allows flushing out a topological distortion. (f) Data projected by the CCA show strong compression (bright cells) across the rings and some stretching (dark cells) too. One of the ring has been splitted but this visualization technique does not show tearing of manifolds, so we cannot say seeing the projections, whether the original data are grouped into two or three connected sets in the original space. However, we know at least they are separated into at most three connected sets because of the trust we got in the connectedness of the three different parts (pieces of rings) through the average grey color (no distortion) of the Voronoï cells we can see along these structures.



(a) (b) (c)

Fig. 2: **Projection of a sphere:** (a) Original data on a sphere. (b) Projected data using PCA. Most of the compressions (bright and dark cells) concentrate at the center of the projected distribution where back and front points originally far from each other are projected nearby, showing a gluing of the original manifold. Distances along the convex hull of the projections are well preserved (average grey cells). (c) Data projected using CCA as a triangular shape. Strong stretching (dark cells) occured along the right-hand sides of the triangle showing that tearing of the original sphere certainly occured here. However, this is not visible on the left-hand side of the triangle, while tearing occured there too. The distorsion we propose being a measure relative to the maximum distortion, it is sensitive to noise giving too much emphasis to the compressions and stretching occuring in the middle part of the projection while small distortions in this place are unavoidable and not as important for the analyst.



(a) (b) (c)

Fig. 3: **Projection of two parallel planes:** (a) Original data. (b) Projected data using PCA. Strong compressions occur in the middle showing a gluing of the originally disconnected manifolds. Thus, if classes are given, their apparent overlapping is likely to be an artifact of the projection and should not be taken for granted. (c) After CCA, compressions concentrate where both planes overlap, and stretching is partly visible on the border of the distribution of projected data. PCA provides a more uniform coloring than CCA due to the linearity of the projection, which eases the visual detection of the problematic area. The analyst should pay attention to the areas with strong contrasting colors.

References

[1] M. Aupetit, Visualizing distortions in continuous projection techniques. In M. Verleysen, editor, *proceedings of the* 12^{th} *European Symposium on Artificial Neural Networks* (ESANN 2004), d-side pub., pages 465-470, April 28-30, Bruges (Belgium), 2004.

[2] M. Aupetit, T. Catz, High-dimensional labeled data analysis with topology representing graphs. *Neurocomputing*, 63:139-169, Elsevier, 2005.

[3] P. Demartines, J. Hérault, Curvilinear Component Analysis: a Self-Organising Neural Network for Non-Linear Mapping of Data Sets. *IEEE Trans. on Neural Networks*, 8(1):148-154, IEEE, 1997.

[4] I.T. Jollife, *Principal Component Analysis*. Springer Verlag, New-York, 1986.

[5] A. Okabe, B. Boots, K. Sugihara, *Spatial tessellations: concepts and applications of Voronoï diagrams*. John Wiley, Chichester, 1992.

[6] P. Rousset, C. Guinot, Distance between Kohonen classes, visualization tool to use SOM in data set analysis and representation. In J. Mira and A. Prieto, editors, proceedings of the *International Workshop on Artificial Neural Networks* (IWANN 2001), Lecture Notes in Computer Science 2085, pages 119-126, Springer-Verlag Berlin, Heidelberg, 2001.

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

[8] W.S. Torgerson, Multidimensional scaling I - Theory and methods, *Psychometrica*, 17:401-419, 1952.

[9] J. Venna, S. Kaski, Neighborhood preservation in nonlinear projection methods: an experimental study. In G. Dorffner, H. Bischof, K. Hornik, editors, proceedings of the *International Conference on Artificial Neural Networks* (ICANN 2001), Vienna, Austria, Lecture Notes in Computer Science 2130, pages 485-491, Springer-Verlag, 2001.

[10] J. Vesanto, SOM-based data visualization methods, *Intelligent Data Analysis*, 3(2):111-126, Elsevier, IOS Press 1999.

[11] J. Warnking, A. Guerin-Duguet, A. Chéhikian, S.Olympieff, M. Dojat, C. Segebarth, Retinotopical mapping of visual areas using fMRI and a fast cortical flattening algorithm, *Neuroimage*, 11(5):S646, Elsevier, 2000.