ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 465-470

Visualizing distortions in continuous projection techniques.

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

Abstract. The visualization of multi-dimensional data has been studied for a long time. Here we propose to visualize any distortion measure associated to a projected datum in continuous projection techniques, by coloring its corresponding Voronoï cell in the projection space. We apply this approach to detect where the high-dimensional manifold has been torn or glued during the projection. We experiment this technique with the Principal Component Analysis and the Curvilinear Component Analysis for different databases.

1 Introduction

Projection techniques are useful for the exploratory analysis of a set of multidimensional data. They project the data in a lower-dimensional space (usually two-dimensional) which helps the expert to apprehend visually their topology. We consider the "continuous" projection techniques such as the Principal Component Analysis (PCA) (Hotelling 1933) which projects linearly the data onto the principal axes of their distribution, and the Curvilinear Component Analysis (CCA) (Demartines et al. 1997), a variant of the Multi-Dimensional Scaling (MDS) (Torgerson 1952) and the Non Linear Mapping (NLM) (Sammon 1969) which aim at preserving pairwise distances between data from the input space to the projection space.

In general, these projection techniques lead to a loss of information through the stretching or the compression of some of the pairwise distances. Measuring and visualizing these distortions is crucial to the expert to detect whether the features observed in the projection space which are the basis of the exploratory data analysis, are faithful images of some features in the input space or artifacts of the projection technique. Therefore it is a way to help the expert to apprehend the topology of the data in the input space.

A large number of measures and associated visualization techniques have been proposed for the Self-Organizing Map (Kohonen 1988; Vesanto 1999), but very few for the other techniques (*e.g.* "dx-dy" (Demartines et al. 1997), "Trustworthiness" (Venna et al. 2001)). Here, we propose a technique to visualize any measure associated to a datum, and we apply this technique to detect where a manifold has been torn or glued during the projection with the PCA and the CCA.

2 Principle

We consider a set of data $\underline{x} = (x_1, \ldots, x_n)$ in the input manifold $E \subset \mathrm{R}^d$ euclidean, and the set of corresponding projections $\underline{y} = (y_1, \ldots, y_n)$ in the projection manifold $F \subset \mathrm{R}^2$ euclidean. A projection function $f : E \to F$ associates to each point x_i a projection $y_i = f(x_i)$. X with $X_{ij} = \|x_i - x_j\|$ and Y with $Y_{ij} = \|y_i - y_j\|$, are the euclidean distance matrices computed between the points of E and F respectively. We consider that a distortion measure $m_i \in [0, 1]$ is associated to each projection y_i.

We propose to give the white color for $m_i = 1$ and the black one for $m_i = 0$ following a linear grey scale, to the Voronoï cell V_i associated to y_i and defined as (Okabe et al. 1992):

$$\forall y_i \in \underline{y},\ V_i = \{v \in F \mid \forall y_j \in \underline{y}, (v - y_i)^2 \leq (v - y_j)^2\}$$

Using Voronoï cells makes easier the inference of the color (hence the measure) associated to a projection point, because any point in the space belongs by definition to the Voronoï cell of its closest projection point, so is given the color of this point. It also allows to cover the entire projection space, so there is no need to define a specific background color which could disturb the visual analysis. This idea has been proposed for Self-Organizing Maps (Vesanto 1999) with square or hexagonal Voronoï cells corresponding to different topologies of the maps, but it had not yet been experimented with continuous projection techniques.

3 Visualizing where the input manifold has been torn or glued

The "similarity coloring" has been proposed in (Kaski et al. 1998) and in a matrix form in (Rousset et al. 2001) for the SOM. It is based on the measure and the visualization as a color, of the normalized distance of any datum x_i to a selected datum x_s in the input space E. Following this principle, we propose a "proximity" measure m_i^{loc} associated to any point y_i and defined as:

$$\forall i \in \{1, \ldots, n\},\ m_i^{loc} = 1 - \frac{X_{is}}{\max_k(X_{ks})} \tag{1}$$

The data whose projections are associated to the same color in F are at the same distance to the selected datum in E. The brighter the color, the closer the distance. Thus if two projections y_i and y_s are far from (*close to*) each other in

F while the corresponding x_i and x_s are close to (*far from*) each other in E (*i.e.* their Voronoï cells have a similar (*different*) color but are separated by some darker cells (*are adjacent*)), then the manifold E has been torn (*glued*) between these two points during the projection f. Notice the measure is normalized so the black color is given in F to the farther point to the selected point in E, and the white color to the selected point itself.

To help selecting the points y_s likely to reveal such tearing or gluing, we propose to visualize for each projection point y_i the corresponding compression for the PCA: $M_i = \sum_j D^+_{ij}$ with $D_{ij} = X_{ij} - Y_{ij}$ and $D^+_{ij} = D_{ij}$ if $D_{ij} > 0$, 0 else, or the stretching for the CCA: $M_i = \sum_j D^-_{ij}$ with $D^-_{ij} = -D_{ij}$ if $D_{ij} < 0$, 0 else. The corresponding normalized measure m_i^{all} associated to a projection y_i is defined as:

$$m_i^{all} = \frac{M_i - \min_k(M_k)}{\max_k(M_k) - \min_k(M_k)}$$

The brighter the color associated to a projection, the higher the probability for the corresponding datum to be close to a tearing or gluing of the input manifold E.

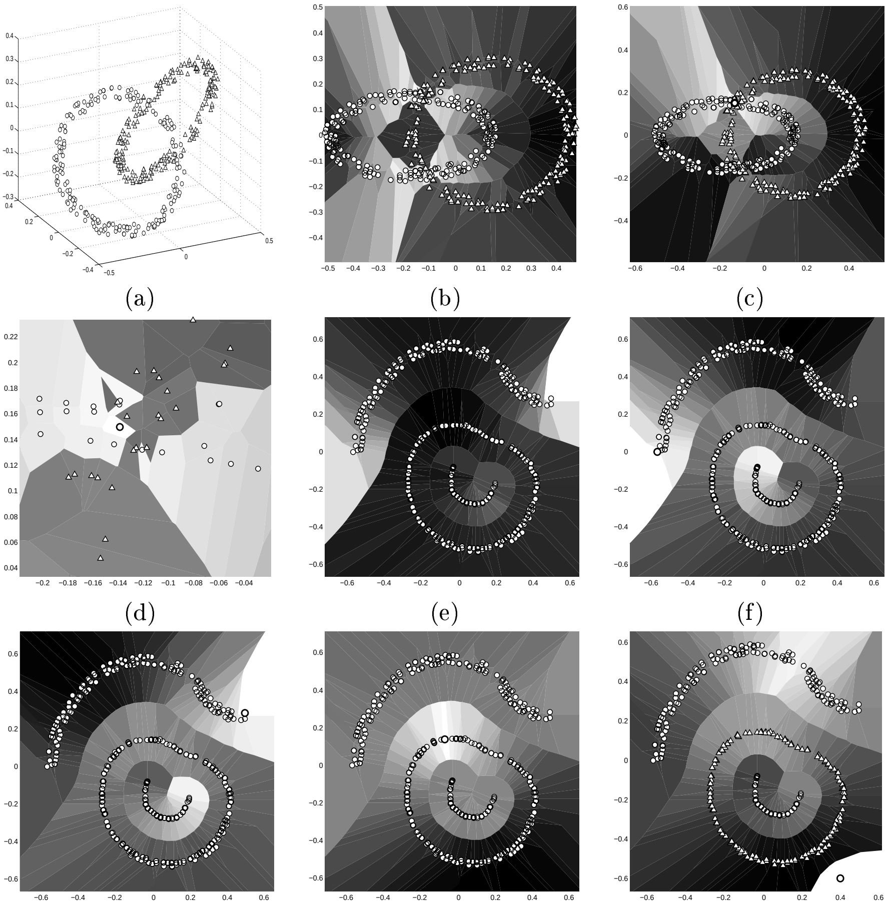
4 Experiments

We compare PCA and CCA onto two simple data sets with known topology: two interlaced rings on the figure 1 and a sphere on the figure 2. The results are interpreted in the caption of the figures.

5 Discussion

As far as the X and Y matrices have been computed, the computation load is very low ($O(n)$ to compute the distortion measures and $O(n\log(n))$ to compute the Voronoï diagram). For all the continuous projection techniques derived from MDS and NLM, both these matrices are necessarily computed to carry out the projection. For PCA, both matrices have to be computed from scratch which involves $O(dn^2)$ operations.

The expert which has selected some projection technique with some similarity measure to compute X and Y, is prepared to analyze visually the projection distribution according to this knowledge. The distortion measures and the visualization technique we propose only makes visible to the expert the proximity information which is contained in X and Y, but which may not appear through the projected distribution: projections which seem to be close to or far from each other in the projection space according to the similarity measure used in Y and known by the expert, are not necessarily so in the input space according to X. So it seems possible to use any metric or similarity measure to define X and Y with the distortion measure and visualization proposed, but this point needs further experiments.

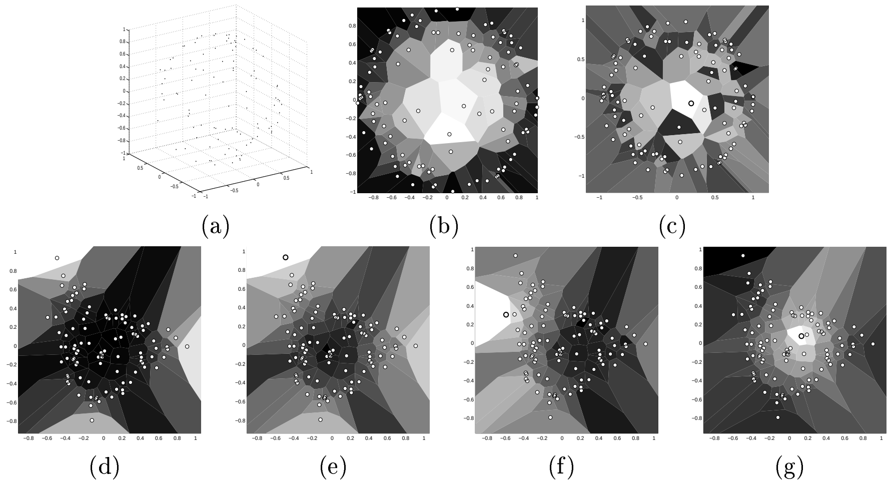


(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 1: **The PCA and CCA of two interlaced rings :** (a) Initial 3D data. (b) Compressions after the PCA appear close to the crossing of both rings. The left ring is the most distorted. (c) (magnified on (d)) Proximity measure associated to a point selected at a crossing (thick circle with white Voronoï cell): projections close to each other with adjacent Voronoï cells have very different colors, showing that both rings have been glued while they are normally separated in the input space. (e) Stretching after the CCA gather at the extreme areas of the top and the inside groups of projections. (f-g) Proximity measure: bright areas are in fact connected in the input space. (h) All the points associated to middle grey cells (like the top and the inside groups) are at the same distance from the selected point, and there is neither gluing nor tearing at this point. From the observations (f-g-h) it is possible to infer what the data look like in the input space: the top and the inside groups are topologically connected in the way shown in (f-g) (*i.e.* homeomorphic to a circle), and geometrically lie on a sphere with the selected point in (h) as the plausible center. (i) A badly projected point (”outlier”) after the optimization phase of the CCA is detected: the selected point in the bottom right corner should have been projected in the bright area of the top group.



(a) (b) (c)

(d) (e) (f) (g)

Figure 2: **The PCA and CCA of a sphere:** (a) Data on a sphere. (b) Compressions after the PCA gather at the center of the projected distribution. (c) The proximity measure shows that some points from the rear and the front of the sphere are close in the projection space, there is gluing of the input manifold (dark and bright adjacent cells around the selected point). (d) Stretching after the CCA gather at the vertices of the triangular projected distribution. (e-f) The proximity measure shows how the vertices and the edges of the triangle are in fact connected in the input space. (g) Neither tearing nor gluing are detected at the center of the projected distribution.

The combination of the distortion measure and the visualization process we propose, not only allows to detect outliers, or gluing and tearing of the input manifold, but also allows to reconstruct the topology of the data in the input space from the observation of their projection.

The technique only depends on X and Y so it applies in the case where only X is provided but $\underline{x}$ is not. In the case of incomplete X, the distortion measures proposed have to be adapted, and a specific color or texture should be used in the visualization process to account for the lack of some of the distances.

6 Conclusion

We proposed a simple way to visualize any distortion measure associated to a projection point in continuous projection techniques. This technique associated to a relevant distortion measure provides a very useful view of the quality of a projection. It allows to detect projection artifacts such as outliers, and to visualize distortions such as tearing or gluing, avoiding errors of interpretation and helping to reconstruct the topology of the data in the input space.

We intend to adapt this technique to visualize measures associated to pairs of projections. Additional experiments with higher dimensional data or other metrics should also be carried out to set definitely the benefits and the limits of this promising approach.

References

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

[Hotelling 1933] H. Hotelling, *Analysis of a complex of statistical variables into principal components*, Journal of Educational Psychology, 24:417:441, 498-520, 1933.

[Kaski et al. 1998] S. Kaski, T. Kohonen, J. Venna, *Visual Explorations in Finance using Self-Organizing Maps*, Chapter Tips for SOM Processing and Colorcoding of Maps. Springer-Verlag, London, 1998.

[Kohonen 1988] T. Kohonen, *Self-Organization and Associative Memory Formation*, Springer-Verlag, 1988.

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

[Rousset et al. 2001] P. Rousset, C. Guinot *Distance between Kohonen classes, visualization tool to use SOM in data set analysis and representation*, J. Mira and A. Prieto (Eds.): IWANN 2001, LNCS 2085, pp. 119-126, Springer-Verlag Berlin Heidelberg, 2001.

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

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

[Venna et al. 2001] J. Venna, S. Kaski, *Neighborhood preservation in nonlinear projection methods: an experimental study,* Artificial Neural Networks - ICANN 2001: International Conference on Artificial Neural Networks, Vienna, Austria, Proceedings , G. Dorffner, H. Bischof, K. Hornik (Eds.) Lecture Notes in Computer Science 2130,485:491, 2001.

[Vesanto 1999] J. Vesanto, *SOM-based data visualization methods*, In Intelligent Data Analysis, Volume 3, Number 2, Elsevier Science, pp. 111-126. IOS Press 1999.