ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

How to project 'circular' manifolds using geodesic distances?

John Aldo Lee, Michel Verleysen*

Université catholique de Louvain
Place du Levant, 3, B-1348 Louvain-la-Neuve, Belgium
{lee,verleysen}@dice.ucl.ac.be

Abstract. Recent papers have clearly shown the advantage of using the geodesic distance instead of the Euclidean one in methods performing non-linear dimensionality reduction by means of distance preservation. This new metric greatly improves the performances of existing algorithms, especially when strongly crumpled manifolds have to be unfolded. Nevertheless, neither the Euclidean nor the geodesic distance address the issue of 'circular' manifolds like a cylinder or a torus. Such manifolds should ideally be torn before to be unfolded. This paper describes how this can be done in practice when using the geodesic distance.

1 Introduction

Non-linear dimensionality reduction, a.k.a. non-linear projection, aims at representing a given set of high-dimensional points in a lower-dimensional space. In that framework, it is commonly assumed that the high-dimensional data points lie on a low-dimensional smooth manifold. If the manifold is non-linear, then an intuitive way to reduce the embedding dimensionality consists in transforming the manifold in order to 'unroll' or 'unfold' it. Once unrolled, the manifold becomes linear and the dimensionality reduction is trivial.

A widely used paradigm to project a manifold is distance preservation. Indeed, if Euclidean inter-point distances are identical in both high- and low-dimensional embeddings, then the manifold topology remains unchanged in the low-dimensional space. Classical metric multidimensional scaling [2] (MDS) implements that idea in a straightforward way. However, distance-preservation fully works only for linear manifolds. If a manifold has to be transformed in a non-linear way, then some distances need to be either stretched or shrunk. As it is likely that short distances can be more easily preserved than longer ones, distance preservation may still work if distances are weighted. This is precisely

*M.V. works as a senior research associate of the Belgian FNRS.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

the way methods like Sammon's non-linear mapping [9] (NLM) or Hérault's Curvilinear Component Analysis [3] (CCA) work.

Another way to enhance the principle of distance preservation consists in replacing the Euclidean metric by the geodesic one. Briefly put, the geodesic metric measures lengths along the manifold, contrarily to the Euclidean distance which measures lengths along straight lines. It is intuitively clear (Fig. 1 left) that such a metric makes the distance preservation much easier, since geodesic distances do not change as much as the Euclidean ones when the manifold is unrolled. Geodesic versions of MDS and CCA have been published

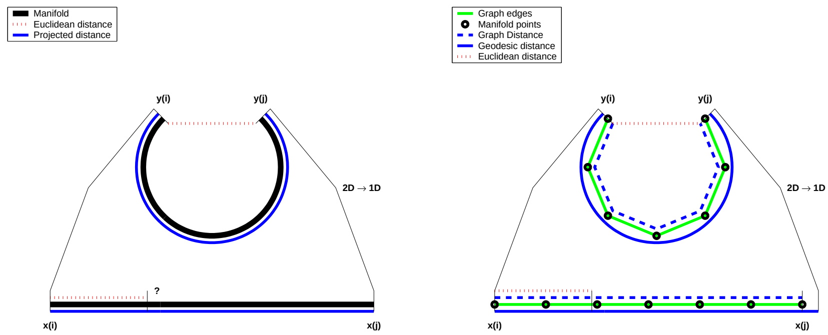

Figure 1 : Geodesic and graph distances for dimensionality reduction in the case of a 'c'-shaped curve: (left) geodesic distances can be better preserved than Euclidean ones, (right) graph distances approximate geodesic ones.

recently [10, 6] and outperform their Euclidean cousins in many cases. But in spite of many advantages, the geodesic distance leaves an open question: how to project 'circular' manifolds, i.e. manifolds with loops and holes?

This paper defines the geodesic distance (Section 2) and circular manifolds (Section 3), explains how to 'tear' them efficiently (Section 4) and shows a few experimental results (Section 5). Finally, Section 6 draws the conclusions and outlines perspectives for future work.

2 Geodesic and graph distances

Contrarily to a Euclidean distance, which only depends on the coordinates of two points in the embedding space, the geodesic distance also depends on the manifold which underlies the points. In the framework of dimensionality reduction by distance preservation, the geodesic distance allows projecting a wider class of manifolds than the Euclidean one. Indeed, a perfect isometry between Euclidean distances in the data and projection spaces only exists for the plane. On the other hand, an isometry between geodesic distances in the data space and Euclidean distances in the projection space exist for numerous manifolds. It has been shown experimentally that the use of the geodesic distance makes the distance preservation easier in many cases.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

Theoretically, the geodesic distance between two manifold points is computed as the shortest arc length between them. The word 'arc' refers to a one-dimensional smooth submanifold. The computation of the shortest arc length requires to minimize an integral over all possible arcs. Obviously, this cannot be achieved since only a few points of the manifold are known.

Consequently, geodesic distances are approximated in practice by computing graph distances [1]. Instead of computing a distance as the length of an arc, all the process is discretized (Fig. 1 right). Starting from the available points, the underlying manifold is replaced by a graph (Fig. 2), which is obtained by connecting neighboring points (e.g. the K closest neighbors or all points inside an ϵ-ball). Next, edges of the graph are labeled with their Euclidean length and Dijkstra's algorithm is run on the graph, in order to compute the shortest path for each pair of points. The 4th plot of Fig. 2 shows that the shortest path well approximates the arc of the geodesic distance.

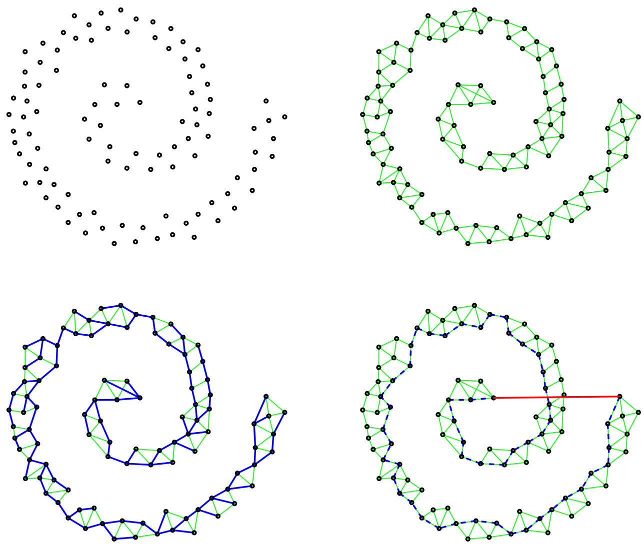

Figure 2 : Practical procedure to compute graph distances: (1st plot) a few manifold points are available, (2nd plot) each point becomes a graph vertex and is connected with its 3 closest neighbors in order to obtain a graph, (3rd plot) after labeling the graph edges with their length, Dijkstra's algorithm is run on the graph, with the central point of the spiral as source vertex, (4th plot) the Euclidean and graph distances between the same two points.

3 Circular manifolds

The word 'circular' may qualify manifolds having a hole. In topology, a hole in a manifold is a structure that prevents the manifold to be continuously shrunk to a single point. A hole is intuitively equivalent to a (cup) handle: a kind of

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

loop allowing one's hand to pick the manifold up. Manifolds may be classified according to their number of holes. For example, planes and spheres have no hole, but cylinders and tori have one. The number of holes is related to other topological concepts like the genus or the Betti number of the manifold. The latter gives the maximum number of cuts that can be made in a manifold without dividing it into several parts.

In the context of dimensionality reduction, circular manifolds raise a major difficulty. For example, in the case of a cylinder, it is known that the manifold is two-dimensional and that any piece of the cylinder is isometric to a subset of $\mathbb{R}^2$ using the geodesic distance. Nonetheless, most projection methods fail because the cylinder is circular: the ideal projection onto a plane requires to tear the cylinder, otherwise it cannot be unrolled. This leads to the following question: Is it possible to tear (or cut) a circular manifold before attempting to reduce its dimensionality? In a theoretical setting, the answer to that question would be very complex. In practice, the situation seems even worse, since only a few points of the manifold are known; but things becomes surprisingly simple when graph distances are used, and the answer to the above question is yes.

4 Tearing circular manifolds

Circular smooth manifolds may be characterized by the fact that they contain closed one-dimensional smooth submanifolds that cannot be continuously shrunk to a single point. For example, in a torus, such submanifolds are circles along the torus body section or around its 'donut' hole.

When the manifold is represented by a graph, non-shrinkable closed submanifolds somehow correspond to 'chordless' cycles. In a graph, a cycle is a path (i.e. a sequence of graph vertices such that an edge connects each pair of subsequent vertices), which is closed (i.e. an edge also connects the starting and ending points of the path). And by 'chordless' it is meant that there exists no path between any two points of the cycle which contains less edges than the two branches of the cycle. Using this definition, a graph having chordless cycles with more than three edges somehow corresponds to a circular manifold. Hence, tearing a circular manifold amounts to remove all chordless cycles longer than three edges in its associated graph. At first sight, this task seems untractable because chordless cycles must be identified in the graph. There is, however, a two-stage procedure which is far simpler: (i) remove *all* cycles in the graph and (ii) reintroduce all cycles which are small or not chordless.

The first stage (Fig. 3 center) is achieved very easily by computing a spanning tree of the graph. By definition, a tree in a connex graph is a connex subgraph with no cycles. Two kinds of spanning trees are minimum spanning trees (MST) and shortest path trees (SPT). They respectively result from Prim's algorithm [7] and Dijkstra's shortest path algorithm [4].

The second stage (Fig. 3 right) is performed by knowing that all edges left outside the spanning tree introduce one or several cycles in the tree. Edges outside the tree are considered one by one: if the smallest chordless cycle

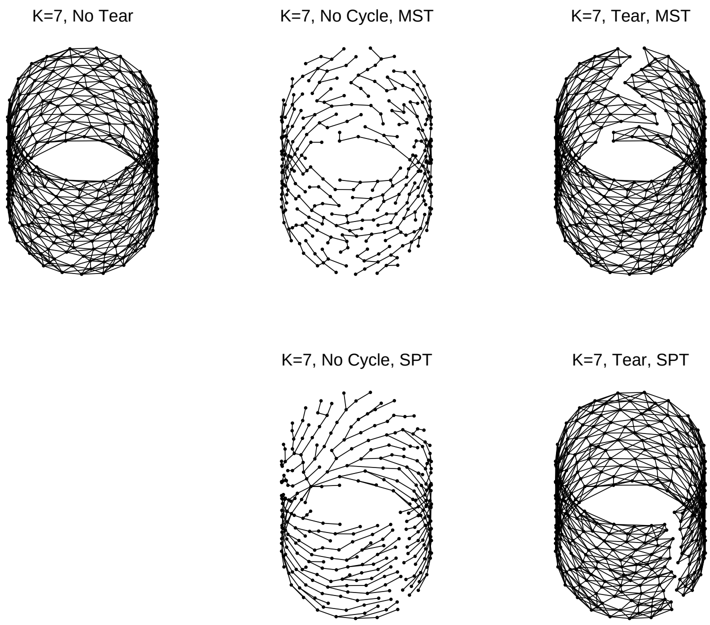

Figure 3 : The two-stage tearing procedure: (left) 300 points lying on a cylinder, connected with their 7 closest neighbors, (center) the minimum spanning tree (MST) and a shortest path tree (SPT) computed on the resulting graph, (right) the torn graph after reintroduction of all edges that do not generate chordless cycles with more than 4 edges.

introduced by an edge contains more than c edges, it remains outside, otherwise it is put back in the graph. The list of remaining edges is repeatedly swept until no change occurs anymore. Practically, the number of edges of the smallest cycle due to the addition of an edge can be computed with a breadth-first algorithm, which computes the straightest path between both vertices of the considered edge (i.e. the path with the minimum number of edges).

5 Experimental results

This section presents a few results of the proposed tearing technique on simple manifolds. Figure 4 shows the advantage of tearing a circular manifold (a cylinder) before projecting it. The data set contains 300 (noise-free) points of the cylinder. In order to build the graph, each point is connected with its 7 closest neighbors. Next, graph distances are computed using Dijkstra's algorithm and fed into Sammon's NLM, leading to the first projection. As can be seen, the cylinder is flattened or crushed instead of being unrolled. On the other hand, if all chordless cycles with more than 4 edges are removed, then the graph is torn and Sammon's NLM yields the desired projection.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

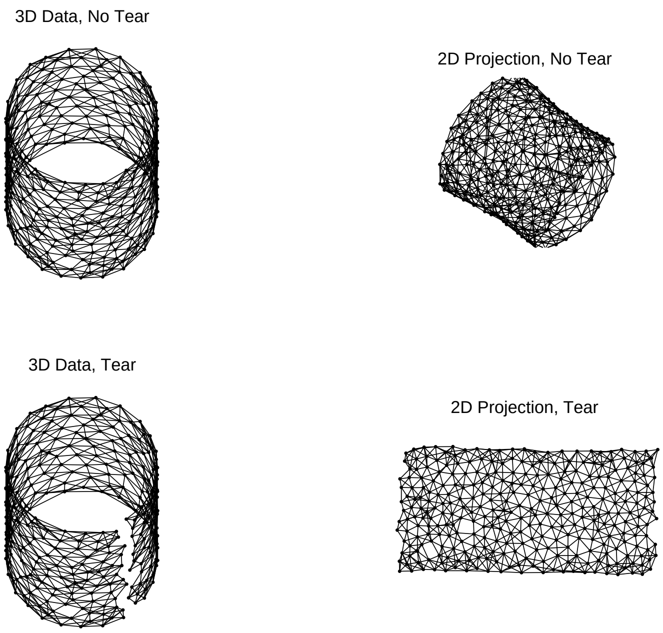

Figure 4 : Two-dimensional projection of the cylinder of Fig. 3 using Sammon's non-linear mapping with the graph distance: (top) the 300-vertex graph is projected without tearing any edges, (bottom) projection obtained when the same graph is torn (using a SPT).

Another example is the torus, shown in top of Fig. 5 . The data set contains 600 (noise-free) points. Each point is connected with its 6 closest neighbors. This time, the two-dimensional projection has been computed by CCA, which is the only distance-preserving method method able to tear manifolds. The projection on the left is obtained by runing CCA with Euclidean distances (the graph is shown for visualization purposes only). The projection on the right is computed by using the same method but with graph distances and after removing all chordless cycles with more than 4 edges. In this second setting, the manifold is torn beforehand and CCA converges faster.

As illustrated by Fig. 3, the shape of the tear depends on the chosen kind of spanning tree. Actually, the MST computed by Prim's algorithm has the property of being unique (for a connex graph with real edge labels that are all different). Moreover, the idea of a minimum-weight spanning tree is intuitively appealing: the tree includes the shortest edges, i.e. the 'strongest' ones. Experimentally, however, the MST yields irregular tears. On the other hand, a SPT computed by Dijkstra's algorithm is not unique (it strongly depends on the source vertex), but leads to neater cuts.

The removal of chordless cycles may be useful for non-circular manifolds too. Indeed, when computing the graph distance, it may happen that parasitic edges

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

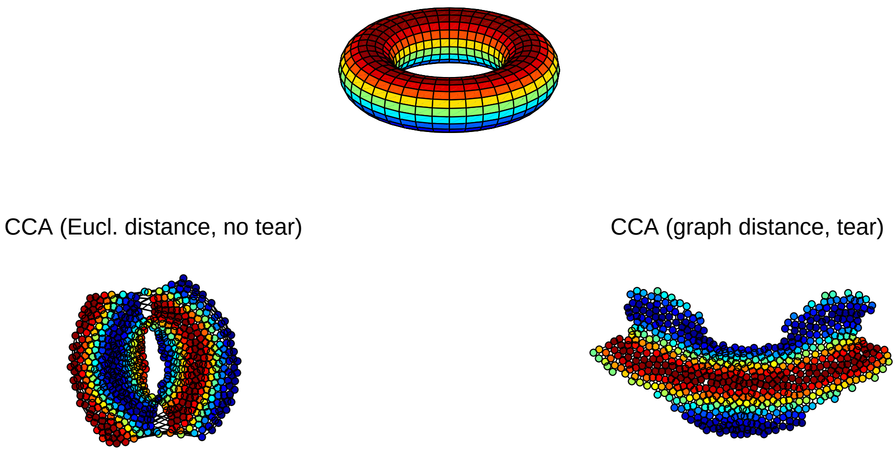

Figure 5 : Two-dimensional projection of a torus using CCA: (top) the torus manifold, (left) projection of 600 points obtained by CCA using the Euclidean distances, (right) projection of the same data set with same method but using the graph distances computed after removing all chordless cycles with more than 4 edges.

appear in the graph and jeopardize the good approximation of the geodesic distances. This typically happens when each point is connected to a too high number of neighbors (Fig. 6). With three neighbors only (1st plot), there are no parasitic edges but the approximation of the geodesic distances is poor. Approximations would be far better with 7 neighbors (2nd plot), but then how to remove the undesired edges? In most cases, edges are undesired when they accidentally generate chordless cycles. The tearing method described above allows removing them in an easy way. The two last plots in Fig. 6 show the torn graphs when using respectively an SPT and MST. Clearly, the SPT does not give the expected result. On the other hand, the appealing properties of the MST lead to a much better result.

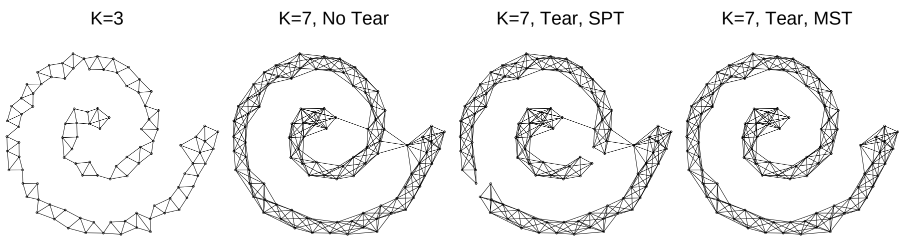

Figure 6 : Tearing undesired edges: (1st plot) some noisy points of a spiral-shaped manifold, each point is connected with its 3 closest neighbors, (2nd plot) the same points connected to their 7 closest neighbors, undesired edges appear in the graph, (3rd plot) the latter graph after removing the chordless cycles with a SPT, (4th plot) with a MST.

ESANN'2004 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 28-30 April 2004, d-side publi., ISBN 2-930307-04-8, pp. 223-230

6 Conclusion

This paper has presented an efficient technique to tear circular manifolds when they have to be projected by distance-preserving methods using the geodesic distance. The technique relies on the fact that those methods approximate geodesic distances by graph distances, i.e. shortest paths in a graph. In that case, tearing the manifold is equivalent to removing all chordless cycles in the corresponding graph. Instead of directly identifying and removing the chordless cycles, which would be a tedious task, all cycles are removed by computing a spanning tree and then edges are put back in the graph, one by one, if they do not generate a long chordless cycle.

The proposed technique may be useful in many methods using either geodesic distances or merely K-ary neighborhoods. Such methods are Isomap [10], Sammon's NLM [9], CDA [6], LLE [8], Isotop [5], etc.

Experimental results have shown that the technique can tear circular manifolds like a cylinder or a torus. Moreover, the same technique can efficiently remove undesired edges in the graph used to approximate the geodesic distances. This elegantly adresses one of the shortcomings of that metric.

Future work aims at comparing the proposed tearing technique with Curvilinear Component Analysis, that can project and tear the manifold at the same time.

References

[1] M. Bernstein, V. de Silva, J. C. Langford, and J. B. Tenenbaum. Graph approximations to geodesics on embedded manifolds. Technical report, Stanford University, Stanford, December 2000.

[2] T. F. Cox and M. A. A. Cox. *Multidimensional scaling.* London, 1995.

[3] P. Demartines and J. Hérault. Curvilinear Component Analysis: A self-organizing neural network for nonlinear mapping of data sets. *IEEE Transactions on Neural Networks*, 8(1):148–154, January 1997.

[4] E. W. Dijkstra. A note on two problems in connection with graphs. *Numerical Mathematics*, 1:269–271, 1959.

[5] J. A. Lee, C. Archambeau, and M. Verleysen. Locally Linear Embedding versus Isotop. In M. Verleysen, editor, *Proceedings of ESANN'2003, 11th European Symposium on Artificial Neural Networks*, pages 527–534. D-Side public., Bruges (Belgium), April 2003.

[6] J. A. Lee, A. Lendasse, and M. Verleysen. Curvilinear Distances Analysis versus Isomap. In M. Verleysen, editor, *Proceedings of ESANN'2002, 10th European Symposium on Artificial Neural Networks*, pages 13–20. D-Facto public., Bruges (Belgium), April 2002.

[7] R. C. Prim. Shortest connection networks and some generalisations. *Bell System Tech. J.*, 36:1389–1401, 1957.

[8] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

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

[10] 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.