Approximation of Geodesic Distances for Geometric Data Analysis

Proposition de stage - Deuxième année de Master

Duration: 5/6 months.

Author of the subject: Frédéric Chazal - INRIA Futurs Saclay

Location : INRIA Futurs at Orsay (south of Paris) inside the GEOMETRICA team.

For further information: frederic.chazal@inria.fr

Subject: Due to the improvements of measurement devices and data storage tools, the available data about complex shapes or complex systems are growing very fast. These data being often represented as point clouds in high dimensional spaces there is a considerable interest in analyzing and processing data in such spaces. Although these point clouds usually live in high dimensional spaces, one often expects them to be located around an unknown, possibly non linear, low dimensional shape. Since a few years, there is a growing interest in developping methods and algorithms to infer topological and geometric characteristics of that shape from the data. It is motivated by the hope that such information will help to better understand the underlying complex systems from which the data are generated, to improve some existing learning algorithms and to develop new efficient learning algorithms.

In machine learning and data analysis, many algorithms make use of the geometry of the data to process it. Most often, their efficiency is closely related to the choice of a distance between the points of the data set. In practical situations, the distances between the points of a data set are often evaluated using the euclidean distance in the ambient space. Unfortunately, these distances may not reflect correctly the metric properties of the underlying shape, in particular when it is highly non-linear (see figure 1). A better adapted but harder to compute distance is the intrinsic (or geodesic) distance. Roughly, the intrinsic distance on a shape $K \subset \mathbf{R}^d$ between two points $a, b \in K$ is the infimum of the lengths of the paths contained in K joining a to b (this distance is set to $+\infty$ when no such path exist). Given a point cloud sampled around an unknown shape $K \subset \mathbf{R}^d$, approximating the intrinsic distance on K appears to be an important challenge in many settings. The goal of this work is to adress this problem. The first part of the work will consist in a review of the existing methods (and their applications in manifold learning) when the underlying shape is a smooth submanifold (curve, surface,...) of $\mathbf{R}^d$. Some existing algorithms will be implemented and tested. The second part of the work is more exploratory: we will consider the case where the underlying shape is a more general non smooth shape. Several approaches will be considered and experimented both from the theoretical point of view (convergence properties to the approximated distance) and from the algorithmic point of view (implementation of the adopted methods).

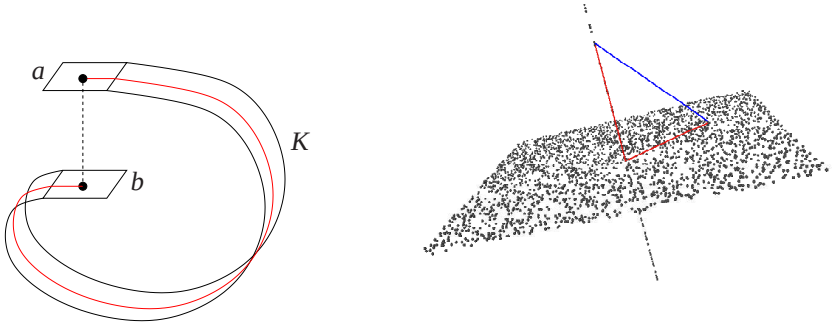

Figure 1: Simple examples showing that the intrinsic distance on a shape K may be very different from the euclidean one.

Required knowledge and background: Knowledge of C/C++. A background in computational geometry or manifold learning (non linear dimensionality reduction,...).

References

[1] M. Bernstein, V. de Silva, J.C. Langford, J.B. Tenenbaum, *Graph approximations to geodesics on embedded manifolds*, technical report.

[2] F. Chazal, D. Cohen-Steiner, A. Lieutier, *A Sampling Theory for Compact Sets in Euclidean Space*, Proceedings of the 22nd ACM Symposium on Computational Geometry 2006.

bibitemGW03 J. Giesen and U. Wagner, *Shape Dimension and Intrinsic Metric from Samples of Manifolds with High Co-dimension.* Proceedings of the 19th Annual ACM Symposium on Computational Geometry (SoCG), (2003) 329-337.

[3] J.B. Tenenbaum, V. de Silva, J. Langford, *A global geometric framework for nonlinear dimensionality reduction*, Science 290: 22 December 2000, pp. 2319-2323.