Statistical and Learning Algorithms for Topologigal Data Analysis

PhD proposal

Duration: 3 years.

Topic: computational geometry learning, topological data analysis (TDA).

Institution and Lab:
INRIA Saclay, Geometrica team
1 rue Honoré d'Estienne d'Orves
Bâtiment Alan Turing
Campus de l'Ecole Polytechnique
91120 Palaiseau
FRANCE

Advisor: Frédéric Chazal
http://geometrica.saclay.inria.fr/team/Fred.Chazal

Contact: frederic.chazal@inria.fr

Context:

During the last decade, the wide availability of measurement devices and simulation tools has led to an explosion in the amount of available data in almost all domains of Science, industry, economy and even everyday life. Often these data come as point clouds embedded in Euclidean spaces, or as more general metric spaces, e.g. when data are just given as a matrix of pairwise distances between points which is often the case for sensor networks or social networks data. These points are usually not uniformly distributed in the embedding space but lie close to some lower-dimensional geometric structure (manifold or more general stratified space) which reflects important properties of the "systems" from which the data have been generated.

With the recent explosion in the amount and variety of available data, identifying, extracting and exploiting their underlying geometric structures has become a problem of fundamental importance for data analysis and statistical learning. On the theoretical side, the topological and geometric properties of data are of great help to analyze them and can be used for further learning or classification tasks. On the algorithmic and applied side, understanding the underlying geometric structure of data can help face the so-called curse of dimensionality phenomenon, and down the road lead to drastic improvements in the complexity of algorithms.

There exist various statistical and machine learning methods that intend to uncover the geometric structure of data, such as clustering, manifold learning and non linear dimensionality reduction, principal curves, sets estimation, to name a few. Most of them assume the underlying structure to have a very simple geometry — diffeomorphic to a disc or isometric to an open set of a Euclidean space. Furthermore the only topological information they seek for is connectivity.

On another hand, with the emergence of distance based approaches [7, 12, 5] and persistent topology [10, 2], geometric inference and computational topology have recently known an important development. New mathematically well-founded theories gave birth to the field of Topological Data Analysis [1] (TDA). So far the obtained results rely mostly on deterministic assumptions which are not satisfactory from a statistical viewpoint. As a consequence the corresponding methods remain exploratory. they do not benefit from a sound probabilistic framework and connot be easily used in a learning framework. Despite a few notable attempts to overcome this issue, the development

[1] See here for a recent talk about the state of TDA by G. Carlsson (Stanford): http://www.birs.ca/events/2012/5-day-workshops/12w5081/videos/watch/201210150903-Carlsson.mp4

of a statistical approach to Topological Data Analysis is still in its infancy and there is now a real need to combine computational topology and geometry, statistical and learning approaches to go beyond.

Goals and expected work: The goal of this PhD is to develop well founded statistical and learning methods and algorithms for Topological Data Analysis (TDA). The PhD student will particularly focus topological persistence [10, 2] that is a fundamental tools in TDA.

He will work on the design of various statistical frameworks and models for topological persistence. In topological data analysis, persistence diagrams are computed from filtered simplicial complexes built on top of finite data sets. The space of persistence diagrams being a metric space (endowed with the so-called bottleneck distance [9, 3]) they provide a "topological signature" of the data that can be used to compare different data sets. However, up to now, this comparison remains largely heuristic and rarely comes with statistical guarantees. The PhD work will be organized in two main parts:

1. **Statistical properties of persistence diagrams:** A first task will be to propose various statistical models that are relevant from the topological data analysis point of view and from which rigorous statistical results for persistence diagrams can be established. In a first step, building on recent results from [4] the statistical properties of persistence diagrams obtained from families of simplicial complexes built on top of point clouds sampled from a possibly unknown compact metric space endowed with a probability measure will be considered. Other important frameworks and models occurring in practice such as for example the statistics of persistence diagrams of distance-to-measure functions of empirical measures associated to point clouds sampled according to a given measure will also be considered. The objective of this part is to provide frameworks and results in which persistent diagrams can be used as well founded statistics and to design new TDA and geometric inference algorithms taking advantage of these statistics.

2. **Kernel-based learning algorithms for TDA:** Using the approach of [1] where persistent diagrams are represented as elements of an Hilbert space, we also intend to exploit our statistical results to design kernels carrying topological information that could be used with classical kernel-based methods in statistical learning. The objective of this part is to provide a set of mathematically well-founded statistical and machine learning algorithms that explicitly exploit the topological structure of data encoded in persistence diagrams. The relevance and efficiency of the designed tools will be tested on synthetic and real data sets coming from various applications areas. For example, building on [8] semi-supervised learning applications for 3D shapes databases based upon topological signatures will be explored.

Requested experience:
- a good mathematical background and some knowledge in computational geometry/topology and/or statistical learning. - Some notions of C/C++ or Python would also be welcome.

Related projects:
The PhD student will be a member of the Geometrica team in Saclay. Geometric inference, TDA and their statistical and algorithmic aspects are at the core of the research done by the Geometrica team in Saclay. The PhD work and results will contribute to the EU project CG-Learning (http://cglearning.eu/); to an INRIA associated team COMET between Geometrica, the Geometric computing group at Stanford and T. Dey's group at Ohio Univ (http://www.inria.fr/en/teams/comet); and, to a submitted ANR project entitled "TopData: Topological Data Analysis: Statistical Methods and Inference".

References

[1] P. Bubenik. *Statistical topology using persistence landscapes.* arXiv:1207.6437v1 [math.AT], 2012.

[2] G. Carlsson, A. Zomorodian, *Computing Persistent Homology*, Discrete & Computational Geometry, Volume 33 (2), pp 249-274, 2005.

[3] F. Chazal, V. de Silva, M. Glisse, S. Oudot, *The Structure and Stability of Persistence Modules*, arXiv:1207.3674, July 2012: http://arxiv.org/abs/1207.3674

[4] F. Chazal, V. de Silva, S. Oudot, *Persistence Stability for Geometric complexes*, arXiv:1207.3885, July 2012: http://arxiv.org/abs/1207.3885

[5] F. Chazal, D. Cohen-Steiner, Q. Mérigot, *Geometric Inference for Probability Measures*, Journal on Foundation of Computational Mathematics, 11, 6, 2011.

[6] F. Chazal, D. Cohen-Steiner, A. Lieutier, *A Sampling Theory for Compact Sets in Euclidean Space*, in Discrete & Computational Geometry, Vol 41, 3, 2009.

[7] F. Chazal, S. Oudot, *Towards Persistence-Based Reconstruction in Euclidean Spaces*, in proc. 2008 ACM Symposium of Computational Geometry, p.232-241

[8] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mï¿½moli, S. Oudot, Gromov-Hausdorff Stable Signatures for Shapes using Persistence, Computer Graphics Forum (proc. SGP 2009), pp. 1393-1403, 2009.

[9] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer, *Stability of persistence diagrams*, Discrete Comput. Geom., 37(1):103-120, 2007.

[10] H. Edelsbrunner and J. Harer. *Persistent homology — a survey.* Surveys on Discrete and Computational Geometry. Twenty Years Later, 257-282, eds. J. E. Goodman, J. Pach and R. Pollack, Contemporary Mathematics 453, Amer. Math. Soc., Providence, Rhode Island, 2008.

[11] H. Edelsbrunner and J. L. Harer. *Computational Topology. An Introduction*. Amer. Math. Soc., Providence, Rhode Island, 2010.

[12] F. Chazal, D. Cohen-Steiner, A. Lieutier, *A Sampling Theory for Compact Sets in Euclidean Space*, in Discrete & Computational Geometry, Vol 41, 3, 2009.