Topological and Geometric Inference from Measures

Ph.D. Proposal

Advisors: Frédéric Chazal (DR) and Steve Oudot (CR) – Geometrica group, INRIA Saclay Île-de-France

Location : Parc Club Orsay Université, 4 rue Jacques Monod, 91893 Orsay Cedex.

Contact: `frederic.chazal@inria.fr` and `steve.oudot@inria.fr`

Context. The wide availability of measurement devices and simulation tools has led to an explosion in the amount of available geometric data, both in academia and in industry. Often this data is in the form of point clouds sampled from some unknown geometric spaces or entities. Before such data can be effectively exploited, its underlying structures must be identified, extracted, and analyzed. In the last decade or so, Computational Topology was a major contributor to the understanding of geometric structures in point cloud data, with the emergence of new theories like topological persistence or distances to compact sets. These theories rely on the assumption that the data lies on or close to some geometric structures and that the level of noise remains controlled. A new paradigm emerged recently, where point clouds are no longer treated as mere compact sets but rather as empirical measures. A notion of distance to such measures has been defined and shown to be stable with respect to perturbations of the measure, thus making it possible to deal with data sets corrupted with high-amplitude noise [1], as illustrated in Figure 1. Although this distance can easily be computed pointwise in the case of a point cloud (simply average the squared distances to the k nearest neighbors), its sublevel-sets, which carry the interesting geometric information about the measure, remain hard to compute or approximate.

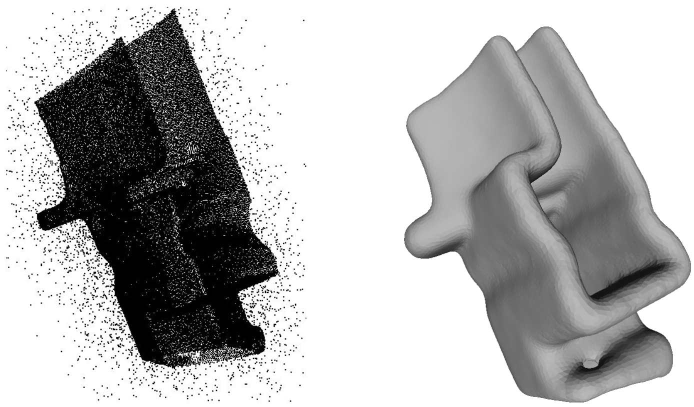

Figure 1: *Left: a 3-dimensional point cloud sampled from the surface of a mechanical part, to which* 10% *more points sampled uniformly at random from an enclosing box have been added. Right: a level-set of the distance function to the empirical measure based on the observations. This isosurface successfully recovers the topology of the mechanical part.*

Objectives. The current challenge is to turn the theory of distances to measures into effective algorithms for point cloud data analysis in arbitrary dimensions, in the same spirit as what has been done in the past for distances to compact sets [2, 3]. Such algorithms would find many applications in data analysis in the presence of significant noise and outliers.

On the mathematical front, the main task of the Ph.D. candidate will be to work out equivalents of the so-called *Čech filtrations* in the context of distances to measures. To put it simply, this means reinterpreting the sublevel-sets of distances to measures as unions of balls, from which Čech complexes are derived through nerve constructions. Another important question is to study the potential relationships that exist between the topology of these unions of balls and the ones of certain subcomplexes of the Delaunay or regular triangulations of the input point cloud, which can then be used in their stead. Such relationships were obtained in the classical setting through deformation retractions [4], and they proved a valuable tool for the analysis of many topological data analysis methods.

On the algorithmic front, the main task will be to devise tractable algorithms for estimating the topological and geometric properties of distances to measures, including their persistence diagrams. Although algebraic constructions based on Čech complexes can be naturally turned into theoretical algorithms, the latter are mostly useless in practice: not only do the combinatorial complexities of their data structures explode with the dimensionality of the data, but the geometric predicates involved in their construction lead to algorithmic problems that suffer from the *curse of dimensionality*. It is then necessary to use approximation, and in particular to show that simpler structures based on distance comparisons, like the *Vietoris-Rips* or *witness complexes*, lead to more tractable algorithms while keeping similar algebraic properties. Relating the properties of these simple structures to the ones of the former, more complex structures, will require to exploit the latest advances in topological persistence theory, in particular on the stability of persistence diagrams [5].

On the applications front, the data structures and algorithms designed by the Ph.D. candidate will be implemented and tested against real-life data. They will be integrated into the CGAL-HD library developped within the European project CG-Learning [6] for better dissemination. In addition, potential applications of this work will be considered, such as homology inference in the presence of significant noise and outliers, or geometric analysis of high-dimensional and potentially corrupted data.

References

[1] F. Chazal, D. Cohen-Steiner, Q. Mérigot, *Geometric Inference for Measures based on Distance Functions*, INRIA Research Report RR-6930.

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

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

[4] H. Edelsbrunner. *The union of balls and its dual shape*, in Proc. 9th Annual ACM Symposium on Computational Geometry, pages 218–231, 1993.

[5] F. Chazal, D. Cohen-Steiner, M. Glisse, L. J. Guibas, S. Y. Oudot. *Proximity of Persistence Modules and their Diagrams.* Proc. 25th Annual ACM Symposium on Computational Geometry, pages 237-246, 2009.

[6] *Computational Geometric Learning*, Specific Targeted Research Project (STREP) funded by the FET unit of the European Commission under contract No. 255827. `http://cgl.uni-jena.de/Home/WebHome`