CoMeT – Computation, Metric and Topology for Data Analysis

Scientific Project (2011 - 2013)

The wide availability of measurement devices and simulation tools has led to an explosion in the amount of available data, both in academia and in industry. Often this data is in the form of samples from some underlying geometric space or entity. Before such data can be effectively exploited, it needs to be processed so its underlying structures can be identified, extracted, and analyzed. In low dimensions, effective reconstruction techniques exist that can provide faithful approximations of the underlying structures from samples. Further processing makes it possible to study their topological and geometric properties. In high dimensions, however, the data often suffers from significant defects, including sparsity, noise, and outliers, violating sampling conditions required by extant methods. The problem is further compounded by the rapid growth in complexity of the data structures used for reconstruction as the dimensionality of the data increases, making them intractable in high dimensions.

To face these challenges, researchers have proposed other algorithmic tools that can uncover some of the properties of the structures underlying the data without full reconstruction. Dimensionality reduction techniques have been pretty successful in this vein: they can infer the intrinsic dimensionality of the data, as well as provide structure-preserving mappings of the data into lower-dimensional spaces when such mappings exist. Topological methods, including the ones developed by this team in an earlier project, have also been a success: they can infer the topological structure of the data without the need for an explicit reconstruction, even (and especially) in cases where no lower-dimensional embedding exists.

The amount of information provided by these techniques is lower than that given by an explicit reconstruction, and of course there is now a growing need to go beyond mere dimensionality reduction and topological inference. On the other hand, explicit reconstruction is still out of reach and may not be desirable anyway, in light of the previous discussion. With the current proposal we aim to find the sweet spot between these extremes. We aim to extract enough structure so that we can get a higher-level informative understanding of the data and of the *spaces* they originate from, both topologically and geometrically — but without the computational cost or immense sampling density that a reconstruction might require. For instance, in the context of shape matching and classification, we aim at finding mappings from spaces of shapes into structure-revealing, possibly higher-dimensional, spaces of signatures, and also at understanding precisely which structures are preserved through these mappings and which are not. Taking advantage of the respective expertise of the three groups participating in this project, we intend to address the following questions, gathered into three complementary research themes:

1. **Stable signatures for shapes:** define signatures for compact metric spaces, and compare these with the ones obtained from finite samplings. Understand how much information on the spaces is carried by the signatures: compare the metric on the *space of shapes* to the metric on the *space of signatures*. Signatures can be global or local, depending on the neighborhood size chosen. Devise more local descriptors that are stable with respect to perturbations of the base point, and exploit these descriptors in the context of partial shape matching.

2. **Geometric inference for measures:** considering point clouds as empirical measures rather than as compact sets, exploit the concept of *distance to a measure* [7] to devise effective algorithms for various data analysis problems, such as clustering or topological inference in the presence of noise and outliers. How much of the algorithmics for compact sets carries through?

3. **Geometry-aware topology calculation:** In relation with theme 1, study the stability of the manifold Laplace operator under topological and statistical noise. In topological inference, compute not only the topological type of the space underlying the data, but also a relevant set of generators that enable to *localize* homology classes for further processing (e.g. manifold unfolding or parametrization).

We believe these questions constitute the next step towards getting a higher-level understanding of geometric data, and down the road towards exploiting such data in practical settings. We will now detail their associated challenges, as well as the approaches we are planning to use.

1 Stable signatures for shapes

• Using the concept of *topological persistence* [11] to design signatures for shapes is a relatively new idea, and so far the existence and stability of such signatures are proven only for finite metric spaces [2]. The bottom line is to build a filtered simplicial complex on top of the input point cloud, and to use the topological structure of this filtration (encoded as a planar diagram called a *persistence diagram*) as a signature for the point cloud. We will try to generalize this construction for compact metric spaces, whether discrete or continuous. The main challenge will be to extend the construction of the filtered complex to infinite spaces, and to prove the stability of its topological structure with respect to perturbations of the space in the Gromov-Hausdorff distance. We are particularly interested in the generalization of the so-called Vietoris-Rips filtration, for which contributions by Haussman [13] and Latschev [15] may help.
• An important issue is how much information about the underlying shape can be recovered from descriptors. Previous work like [2] only provides lower bounds on a shape distance (e.g. the Gromov-Hausdorff distance) based on descriptor distance. It would be desirable to go the other way — to upper bound shape similarity based on descriptor similarity. A possible approach would be to enrich the pool of filtrations used to derive persistence-based signatures, and to show that with a sufficiently large family of filtrations we can guarantee that different shapes within a given class do get different signatures.
• Persistence-based signatures are scale-dependent by nature: a scaling of the sampled shape implies a scaling of the resulting filtration and persistence diagram. Can we find the scaling factor minimizing the distance between parsistence diagrams automatically? This would be useful for doing scale-oblivious shape classification. We have also studied multi-scale signatures based on heat diffusion on manifolds and Laplace-Beltrami operator [16], but their properties are still not well understood.
• The difficulty of matching two shapes or spaces is intimately tied to matching a shape to itself — shapes with many natural self maps (symmetries) can be difficult to match because of the ambiguities symmetries create (reflected in duplicate descriptors, etc.). It may be interesting to define some sort of *condition number* for a shape, which would capture the intrinsic difficulty of characteristic or matching against that shape.
• Finally, the signatures of [2] are global, not local. To make them more local, one can build one filtration per data point, whose persistence diagram will serve as local signature. What is then a relevant choice of filtration for a given data point? Given two points p, q, how can their signatures be compared at different scales? How can we detect up to which scale they match?

This would give an indication of how far the neighborhoods around these points look alike, and thus open the door to applications in partial shape matching.

2 Geometric inference for measures

A new paradigm for point cloud data analysis has 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 [3]. This distance can easily be computed pointwise in the case of a point cloud (simply average the squared distances to the k nearest neighbors), but its sublevel-sets, which carry the geometric information about the measure, remain hard to compute or approximate. A big challenge now is to find efficient algorithms in arbitrary dimensions to compute or approximate the topological structure of the sublevel-sets of the distance to a measure, in the same spirit as what was done in the recent years for distances to compact sets. Such algorithms would naturally find applications in topological inference in the presence of significant noise and outliers, but also in other less obvious contexts such as stable clustering. The current bottleneck is that there exist no equivalents of the union of balls and alpha-shape in the case of the distance to a measure. Our first goal will be to work out such equivalents. To start with, we will focus on medium dimensions and use a variant of the mesh-based inference algorithm [14] to approximate the sublevel-sets of the distance to a measure and get an idea of their topological structure.

3 Geometry-aware topology calculation

• Suppose we are given a scalar function already encoding some geometric information of interest over a point cloud data. Such a function could, for example, be one of the descriptors mentioned earlier. From this function, how can we reconstruct simple but meaningful topological structures, such as the contour tree or Reeb graph, with guarantees? This is along the line of recent work by the Geometrica and Geometric Computing groups [4, 5]. We may also take advantage of recent contributions by the Computational Geometry group on constructing Reeb graphs efficiently from a simplicial complex [12].
• Recent work to estimate shortest H_1 homology generators from point clouds [10] uses results by Chazal and Oudot [6] and also injects geometric constraints such as the shortest total length into the topology estimation process. However, the corresponding algorithm is still rather slow. How to improve its time complexity?
• How do we compute homology generators for the k-homology groups with $k > 1$? It is known that *shortest* such generators are NP-hard to compute or even approximate within any constant factor [8]. What other kinds of meaningful geometric information can we inject into the topological quantities in order to simplify the problem?
• Can we develop the theory and algorithms for a Laplace-type operator for stratified spaces from discrete samples, either point clouds or mesh structure? Currently, one can either use the manifold Laplacian [1] if it is known that the underlying space is a manifold, or alternatively, if there is no knowledge about the underlying space, one can construct a graph and use the (weighted) graph Laplace operator. What if the input space is in-between, e.g. a stratified space? This is more complex than a manifold, but has more structure than a general graph. How to define a Laplace-type operator, and what properties can one derive for this operator?
• The stability properties of Laplace operators [9] need to be further addressed, in particular in the context of small topological noise and in the context of statistical noise. This issue is related to Theme 2, and the tools developed therein may help here.

References

[1] M. Belkin, J. Sun, and Y. Wang. Constructing laplace operator from point clouds in $\mathbb{R}^d$. In *SODA '09: Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1031–1040, Philadelphia, PA, USA, 2009. Society for Industrial and Applied Mathematics.

[2] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. *Computer Graphics Forum (proc. SGP 2009)*, pages 1393–1403, 2009.

[3] F. Chazal, D. Cohen-Steiner, and Q. Mérigot. Geometric inference for measures based on distance functions. Research Report 6930, INRIA, May 2009.

[4] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Analysis of scalar fields over point cloud data. In *Proc. 20th ACM-SIAM Sympos. Discrete Algorithms*, pages 1021–1030, 2009.

[5] F. Chazal, L. J. Guibas, S. Y. Oudot, and P. Skraba. Persistence-based clustering in Riemannian manifolds. Research Report 6968, INRIA, June 2009.

[6] F. Chazal and S. Y. Oudot. Towards persistence-based reconstruction in Euclidean spaces. In *Proc. 24th ACM Sympos. Comput. Geom.*, pages 232–241, 2008.

[7] Frédéric Chazal, David Cohen-Steiner, and Quentin Mérigot. Geometric inference for measures based on distance functions. Research Report (version 2 - June 23 2010) 6930, INRIA, 2009.

[8] C. Chen and D. Freedman. Hardness results for homology localization. In *Proc. ACM-SIAM Sympos. Discrete Algorithms*, pages 1594–1604, 2010.

[9] T. K. Dey, P. Ranjan, and Y. Wang. Convergence, stability, and discrete approximation of laplace spectra. In *ACM-SIAM Symposium on Discrete Algorithms*, pages 650–663, 2010.

[10] T. K. Dey, J. Sun, and Y. Wang. Approximating loops in a shortest homology basis from point data. In *Proc. Annual Symposium on Computational Geometry*, pages 166–175, 2010.

[11] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete Comput. Geom.*, 28:511–533, 2002.

[12] W. Harvey, Y. Wang, and R. Wenger. A randomized O(m log m) time algorithm for computing reeb graphs of arbitrary simplicial complexes. In *Proc. Annual Symposium on Computational Geometry*, pages 267–276, 2010.

[13] J.-C. Hausmann. On the vietoris-rips complexes and a cohomology theory for metric spaces. *Prospects in topology (Princeton, NJ, 1994), Ann. of Math. Stud.*, 138:175–188, 1995.

[14] B. Hudson, G. L. Miller, S. Y Oudot, and D. R. Sheehy. Topological inference via meshing. In *Proc. Annual Symposium on Computational Geometry*, pages 277–286, 2010.

[15] J. Latschev. Vietoris-rips complexes of metric spaces near a closed Riemannian manifold. *Archiv der Mathematik*, 77(6):522–528, 2001.

[16] Jian Sun, Maks Ovsjanikov, and Leonidas Guibas. A concise and provably informative multi-scale signature based on heat diffusion. In *Eurographics Symposium on Geometry Processing (SGP)*, 2009.