3iA Côte d'Azur
Interdisciplinary Institute for Artificial Intelligence

Ph.D. research topic

- Title of the proposed topic: Multivariate topological data analysis for statistical machine learning
- Research axis of the 3IA: Axis 1 – Core elements of AI
- **Supervisor (name, affiliation, email): Jean-Daniel Boissonnat, DataShape team, Inria Sophia Antipolis, Jean-Daniel.Boissonnat@inria.fr**
- Co-supervisor (name, affiliation): Mathieu Carrière, DataShape team, Inria Sophia Antipolis
- The laboratory and/or research group: DataShape team, Inria Sophia Antipolis

---

Apply by sending an email directly to the supervisor.
The application will include:

- Letter of recommendation of the supervisor indicated above
- Curriculum vitæ.
- Motivation Letter.
- Academic transcripts of a master's degree(s) or equivalent.
- At least, one letter of recommendation.
- Internship report, if possible.

- Description of the topic:

Multivariate topological data analysis for statistical machine learning

Context

The huge democratization of machine learning and data science over the last years has permitted an unprecedented, open source access to a wide range of data sets coming from almost all fields of science and industry. However, many difficulties remain as it is very frequent that these data sets actually lie close to some hidden, lower-dimensional geometric structures such as manifolds. Hence, it has become critical to be able to efficiently unravel those intrinsic, geometric structures, which, if not detected or properly taken into account, can dramatically impede statistical machine learning models to perform well due to various reasons, the most famous of which being the so-called curse of dimensionality. There exists a large range of methods for taking geometric structures into account when doing data analysis (clustering, non-linear dimensionality reduction, manifold learning, etc.) but they remain limited by the

strong assumptions they usually require and their lack of ability to detect intrinsic properties more complicated than mere connectivity. On the other hand, Topological Data Analysis (TDA) [1, 2, 13] has gained a lot of attention in the last years, due to its wide applicability, its compact descriptors encoding the space's topological features with nice guarantees, and its methods for including these descriptors in further data analysis tasks. This led to drastic improvement, both theoretically and practically speaking, for a wide range of data science problems.

The main descriptors of TDA are computed by inferring the topological features of a space (connectivity, loops, cavities, etc.) from the variations of a scalar-valued continuous function called *filter.* This means that the topological features detected by TDA heavily depend on the filter function, which is usually chosen a priori. In particular, all features on which the filter is constant are actually missed, and fail to be recovered. Hence, TDA descriptors can be very uninformative if the filter function is chosen poorly. In order to cope with this issue, generalizations of such descriptors to *multivariate* filters, i.e., continuous filters that output several values instead of just one, has become one of the most active fields of TDA, and has already been able to propose several richer and more powerful descriptors [3, 4]. However, many steps are missing: as of today, there is no clear set of algorithms, theoretical guarantees, and statistical frameworks for multivariate TDA descriptors. This dramatically prevents multivariate TDA to be widely used in data science, despite the powerful and relevant topological features that it can detect.

An example application domain where multivariate TDA is critically needed is computational biology. Indeed, it is very often the case that several continuous, biological phenomenon happen at the same time, and create geometric and topological features inside data sets: for instance, data sets of single cells in genomics are often subject to batch effects (which create connected components), cell differentiation (which create branches) and cell cycle (which create loops), among other phenomenona. See for instance [5, 6]. In these data sets, using scalar filters is not enough in order to jointly capture all of these phenomena at once, and multivariate TDA already appeared as a potential candidate to succesfully retrieve those biological, geometric properties.

Goals and expected work:

The goal of this PhD is to develop a mathematically grounded set of methods for defining, computing and using multivariate topological descriptors for statistical machine learning, with applications in computational biology. This work will be organized in two parts, corresponding to the two main methods for computing topological features in TDA: topological persistence, and Mapper.

1. **Multivariate topological persistence and machine learning.** In this first axis, the candidate will work on building new descriptors based on topological persistence. The current descriptor, called the persistence diagram, encodes the topological variations of a scalar filter in a set of points in the Euclidean plane [1]. It enjoys statistical guarantees and several linearization methods for data science [7, 8, 9]. The goal of this

axis is to build on recent theoretical results about multivariate topology [10, 11] to come up with new descriptors that enjoy similar theoretical guarantees, and to develop methods for using these new descriptors in machine learning tasks. An emphasis will be put on quantitative immunofluorescence data sets, which are comprised of point clouds with multivariate filters, each point cloud representing individual cells, and each multivariate filter representing the cell type confidences. As a first step, the candidate will build on recent works that showcased the usefulness of topological approaches for this particular application [5, 11].

2. **Multivariate Mapper and statistical data analysis.** In this second axis, the candidate will work on Mapper graphs and complexes [4], which are the main data visualization tools of TDA, and which suffer from the same limitations than topological persistence since they are computed from scalar filters. More precisely, the main issue that currently prevents multivariate Mappers to be widely used in data science is their lack of statistical robustness. Starting with recent preliminary works [12], the candidate will work on developing a statistical framework for assessing stability of multivariate Mapper graphs and complexes. He will also focus on applying its results to data sets from genomics, in which it has already been shown that the Mapper visualization tool was useful for detecting biological events, such as continuous differentiation from stem cells to given cell types [6].

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 DataShape team at Inria Sophia Antipolis, which primarily works on geometric inference, TDA and their statistical and algorithmic aspects.
Most of the data sets will be generated and studied with Columbia University teams based in New-York, in collaboration with Raul Rabadan, Rami Vanguri and Andrew Blumberg. Moreover, the PhD work and results will potentially be included in the group's main open-source library Gudhi (https://gudhi.inria.fr/).

References

[1] Herbert Edelsbrunner and John Harer.
Computational topology: an introduction.
American Mathematical Society, 2010.

[2] Gunnar Carlsson.
Topology and data.
Bulletin of the American Mathematical Society, 46(2):255–308, 2009.

[3] Frédéric Chazal, Vin de Silva, Marc Glisse, and Steve Oudot.
The structure and stability of persistence modules.
Springer-Verlag, 2016.

[4] Gurjeet Singh, Facundo Mémoli, and Gunnar Carlsson.
Topological methods for the analysis of high dimensional data sets and 3D object recognition.
In 4th Eurographics Symposium on Point-Based Graphics (SPBG 2007), pages 91–100. The Eurographics Association, 2007.

[5] Andrew Aukerman, Mathieu Carrière, Chao Chen, Kevin Gardner, Raúl Rabadán, and Rami Vanguri.
Persistent homology based characterization of the breast cancer immune microenvironment: a feasibility study.
In 36th International Symposium on Computational Geometry (SoCG 2020), pages 11:1—-11:20. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2020.

[6] Abbas Rizvi, Pablo Cámara, Elena Kandror, Thomas Roberts, Ira Schieren, Tom Maniatis, and Raúl Rabadán.
Single-cell topological RNA-seq analysis reveals insights into cellular differentiation and development.
Nature Biotechnology, 35:551–560, 2017.

[7] Peter Bubenik.
Statistical topological data analysis using persistence landscapes.
Journal of Machine Learning Research, 16(3):77–102, 2015.

[8] Henry Adams, Tegan Emerson, Michael Kirby, Rachel Neville, Chris Peterson, Patrick Shipman, Sofya Chepushtanova, Eric Hanson, Francis Motta, and Lori Ziegelmeier.
Persistence images: a stable vector representation of persistent homology.
Journal of Machine Learning Research, 18(8):1–35, 2017.

[9] Mathieu Carrière, Marco Cuturi, and Steve Oudot.
Sliced Wasserstein kernel for persistence diagrams.
In 34th International Conference on Machine Learning (ICML 2017), volume 70, pages 664–673. JMLR.org, 2017.

[10] Jérémy Cochoy and Steve Oudot.
Decomposition of exact pfd persistence bimodules.
Discrete & Computational Geometry, pages 1–39, 2019.

[11] Mathieu Carrière and Andrew Blumberg.
Multiparameter persistence image for topological machine learning.
In Advances in Neural Information Processing Systems 33 (NeurIPS 2020), 2020.

[12] Mathieu Carrière and Bertrand Michel.
Statistical analysis of Mapper for stochastic and multivariate filters
In CoRR. ArXiv:1912.10742, 2019.

[13] Jean-Daniel Boissonnat, Frédéric Chazal, and Mariette Yvinec.
Geometric and Topological Inference.
Cambridge University Press, 2018.