Geometry, Topology and Statistics in Data Sciences

Geometry and Statistics in Data Sciences
10-14 October 2022 - IHP, Paris

Monday

Stefan Sommer
Diffusions means in geometric statistics.

Analysis and statistics of shape variation and, more generally, manifold valued data can be formulated probabilistically with geodesic distances between shapes exchanged with (-log)likelihoods. This leads to new statistics and estimation algorithms. One example is the notion of diffusion mean. In the talk, I will discuss the motivation behind and construction of diffusion means and discuss properties of the mean, including reduced smeariness when estimating diffusion variance together with the mean. This happens both in the isotropic setting with trivial covariance, and in the anisotropic setting where variance is fitted in all directions. I will connect this to most probable paths to data and algorithms for computing diffusion means, particularly bridge sampling algorithms. Finally, we will discuss ways of sampling the diffusion mean directly by conditioning on the diagonal of product manifolds, thereby avoiding the computationally expensive iterative optimization that is often applied for computing means on manifolds.

Nina Miolane
Geomstats: a Python package for Geometric Machine Learning.

We introduce Geomstats, an open-source Python package for computations and statistics on nonlinear manifolds that appear in machine learning applications, such as: hyperbolic spaces, spaces of symmetric positive definite matrices, Lie groups of transformations, and many more. We provide object-oriented and extensively unit-tested implementations. Manifolds come equipped with families of Riemannian metrics with associated exponential and logarithmic maps, geodesics, and parallel transport. Statistics and learning algorithms provide methods for estimation, regression, classification, clustering, and dimension reduction on manifolds. All associated operations provide support for different execution backends — namely NumPy, Autograd, PyTorch, and TensorFlow. This talk presents the package, compares it with related libraries, and provides relevant examples.

We show that Geomstats provides reliable building blocks to both foster research in differential geometry and statistics and democratize the use of (Riemannian) geometry in statistics and machine learning. The source code is freely available under the MIT license at https://github.com/geomstats/geomstats.

Yusu Wang
Weisfeiler-Lehman Meets Gromov-Wasserstein .

The Weisfeiler-Lehman (WL) test is a classical procedure for graph isomorphism testing. The WL test has also been widely used both for designing graph kernels and for analyzing graph neural networks. In this talk, I will describe the so-called Weisfeiler-Lehman (WL) distance we recently introduced, which is a new notion of distance between labeled measure Markov chains (LMMCs), of which labeled graphs are special cases. The WL distance extends the WL test (in the sense that the former is positive if and only if the WL test can distinguish the two involved graphs) while at the same time it is polynomial time computable. It is also more discriminating than the distance between graphs used for defining the Wasserstein Weisfeiler-Lehman graph kernel. Inspired by the structure of the WL distance we identify a neural network architecture on LMMCs which turns out to be universal w.r.t. continuous functions defined on the space of all LMMCs (which includes all graphs) endowed with the WL distance. Furthermore, the WL distance turns out to be stable w.r.t. a natural variant of the Gromov-Wasserstein (GW) distance for comparing metric Markov chains that we identify. Hence, the WL distance can also be construed as a polynomial time lower bound for the GW distance which is in general NP-hard to compute. This is joint work with Samantha Chen, Sunhyuk Lim, Facundo Memoli and Zhengchao Wan.

Tuesday

Martin Bauer
Elastic Shape Analysis of Surfaces.

The past decades have seen tremendous advances in imaging techniques, which have led to a significant growth in the quantity and complexity of data in fields such as biomedical imaging, neuroscience and medicine. Naturally, this prompted the emergence of new mathematical and algorithmic approaches for the analysis of such data, which led to the emergence and growth of fields such as geometric shape analysis and topological data analysis. Infinite dimensional Riemannian geometry has proven to be a powerful tool to deal with the challenges that arise in this context. In my talk i will give a short introduction to the general concept of infinite dimensional Riemannian geometry, where i will discuss several of the striking phenomena that might arise in this situation. I will then focus on reparametrization invariant structures on spaces of immersions and, in particular, I will introduce the class of Sobolev metrics on spaces of curves and surfaces. For this class of Riemannian metrics I will discuss the local and global well-posedness of the geodesic equations and properties of the geodesic distance. Finally, to show how we can use this setup in practice, I will discuss the numerical implementation of a statistical framework based on such metrics.

Eric Klassen
The Square Root Normal Field and Unbalanced Optimal Transport.

The Square Root Normal Field (SRNF) is a distance function on shape spaces of surfaces in $\mathbb{R}^3$. Unbalanced Optimal Transport (UOT) is a variant of Optimal Transport in which mass is allowed to expand and contract as it is transported from one point to another. In this talk (joint work of Bauer, Hartman and Klassen) we discuss an unexpected relation between the SRNF distance for oriented surfaces in $\mathbb{R}^3$ and UOT for Borel measures on S^2.

Steve Oudot
Optimization in topological data analysis.

This talk will give an overview of the line of work on optimization for topological data analysis, from the initial attempts at differentiating the persistent homology operator, to the recent adaptations of stochastic gradient descent and gradient sampling.

Break

Omer Bobrowski
Universality in Random Persistence Diagrams.

One of the most elusive challenges within the area of topological data analysis is understanding the distribution of persistence diagrams. Despite much effort, this is still largely an open problem. In this talk we will present a series of conjectures regarding the behavior of persistence diagrams arising from random point-clouds. We claim that, viewed in the right way, persistence values obey a universal probability law, that depends on neither the underlying space nor the original distribution of the point-cloud. We back these conjectures with an exhaustive set of experiments, including both simulated and real data. We will also discuss some heuristic explanations for the possible sources of this phenomenon. Finally, we will demonstrate the power of these conjectures by proposing a new hypothesis testing framework for computing significance values for individual features within persistence diagrams.
This is joint work with Primoz Skraba (QMUL).

Kathryn Hess
Morse-theoretic signal compression and reconstruction.

In this lecture I will present work of three of my PhD students, Stefania Ebli, Celia Hacker, and Kelly Maggs, on cellular signal processing. In the usual paradigm, the signals on a simplicial or chain complex are processed using the combinatorial Laplacian and the resultant Hodge decomposition. On the other hand, discrete Morse theory has been widely used to speed up computations, by reducing the size of complexes while preserving their global topological properties. Ebli, Hacker, and Maggs have developed an approach to signal compression and reconstruction on chain complexes that leverages the tools of algebraic discrete Morse theory,, which provides a method to reduce and reconstruct a based chain complex together with a set of signals on its cells via deformation retracts, preserving as much as possible the global topological structure of both the complex and the signals. It turns out that any deformation retract of real degreewise finite-dimensional based chain complexes is equivalent to a Morse matching. Moreover, in the case of certain interesting Morse matchings, the reconstruction error is trivial, except on one specific component of the Hodge decomposition. Finally, the authors developed and implemented an algorithm to compute Morse matchings with minimal reconstruction error, of which I will show explicit examples.

Wednesday

Johannes Krebs

On the law of the iterated logarithm and Bahadur representation in stochastic geometry.

We study the law of the iterated logarithm and a related strong invariance principle for certain functionals in stochastic geometry. The underlying point process is either a homogeneous Poisson process or a binomial process. Moreover, requiring the functional to be a sum of so-called stabilizing score functionals enables us to derive a Bahadur representation for sample quantiles. The scores are obtained from a homogeneous Poisson process. We also study local fluctuations of the corresponding empirical distribution function and apply the results to trimmed and Winsorized means of the scores. As potential applications, we think of well-known functionals defined on the k-nearest neighbors graph and important functionals in topological data analysis such as the Euler characteristic and persistent Betti numbers as well as statistics defined on Poisson-Voronoi tessellations.

Katharine Turner

The Extended Persistent Homology Transform for Manifolds with Boundary.

The Persistent Homology Transform (PHT) is a topological transform which can be use to quantify the difference between subsets of Euclidean space. To each unit vector the transform assigns the persistence module of the height function over that shape with respect to that direction. The PHT is injective on piecewise-linear subsets of Euclidean space, and it has been demonstrably useful in diverse applications. One shortcoming is that shapes with different essential homology (i.e., Betti numbers) have an infinite distance between them.

The theory of extended persistence for Morse functions on a manifold was developed by Cohen-Steiner, Edelsbrunner and Harer in 2009 to quantify the support of the essential homology classes. By using extended persistence modules of height functions over a shape, we obtain the extended persistent homology transform (XPHT) which provides a finite distance between shapes even when they have different Betti numbers.

I will discuss how the XPHT of a manifold with boundary can be deduced from the XPHT of the boundary which allows for efficient calculation. James Morgan has implemented the required algorithms for 2-dimensional binary images as a forthcoming R-package. Work is also with Vanessa Robins.

Heather Harrington

Shape of data in biology.

TBA

Break

Frédéric Barbaresco
Symplectic Foliation Model of Information Geometry for Statistics and Learning on Lie Groups.

We present a new symplectic model of Information Geometry [1,2] based on Jean-Marie Souriau's Lie Groups Thermodynamics [3,4]. Souriau model was initially described in chapter IV "Statistical Mechanics" of his book "Structure of dynamical systems" published in 1969. This model gives a purely geometric characterization of Entropy, which appears as an invariant Casimir function in coadjoint representation, characterized by Poisson cohomology. Souriau has proved that we can associate a symplectic manifold to coadjoint orbits of a Lie group by the KKS 2-form (Kirillov, Kostant, Souriau 2-form) in the affine case (affine model of coadjoint operator equivariance via Souriau's cocycle) [5], that we have identified with Koszul-Fisher metric from Information Geometry. Souriau established the generalized Gibbs density covariant under the action of the Lie group. The dual space of the Lie algebra foliates into coadjoint orbits that are also the Entropy level sets that could be interpreted in the framework of Thermodynamics by the fact that dynamics on these symplectic leaves are non-dissipative, whereas transversal dynamics, given by Poisson transverse structure, are dissipative. We will finally introduce Gaussian distribution on the space of Symmetric Positive Definite (SPD) matrices, through Souriau's covariant Gibbs density by considering this space as the pure imaginary axis of the homogeneous Siegel upper half space where Sp(2n,R)/U(n) acts transitively. We will also consider Gibbs density for Siegel Disk where SU(n,n)/S(U(n)xU(n)) acts transitively. Gauss density of SPD matrices is then computed through Souriau's moment map and coadjoint orbits. Souriau's Lie Groups Thermodynamics model will be further explored in European COST network CaLISTA [6] and European HORIZON-MSCA project CaLIGOLA [7].

Victor Patrangenaru
Geometry, Topology and Statistics on Object Spaces.

The birth certificate of Statistics, as a mathematical science, is the Central Limit Theorem (CLT), showing that under general assumptions, the distribution of the standardized sample mean follows a standard normal distribution. In this talk one discusses the extension of the CLT to a random object on a smooth object space, and on a stratified space in general, by detailing the case of a random point on a tree, and in particular on a spider. As an application of such ideas, one investigates two possible origins of SARS-CoV-2, using an RNA analysis on rooted tree spaces.

Thursday

Nicolas Charon
Registration of shape graphs with partial matching constraints.

This talk will discuss an extension of the elastic curve registration framework to a general class of geometric objects which we call (weighted) shape graphs, allowing in particular the comparison and matching of 1D geometric data that are partially observed or that exhibit certain topological inconsistencies. Specifically, we generalize the class of second-order invariant Sobolev metrics on the space of unparametrized curves to weighted shape graphs by modelling such objects as varifolds (i.e. directional measures) and combining geometric deformations with a transformation process on the varifold weights. This leads us to introduce a new class of variational problems, show the existence of solutions and derive a specific numerical scheme to tackle the corresponding discrete optimization problems.

Irène Kaltenmark
Curves and surfaces. Partial matching in the space of varifolds.

The matching of analogous shapes is a central problem in computational anatomy. However, inter-individual variability, pathological anomalies or acquisition methods sometimes challenge the assumption of global homology between shapes. In this talk, I will present an asymmetric data attachment term characterizing the inclusion of one shape in another. This term is based on projection on the nearest neighbor with respect to the metrics of varifold spaces. Varifolds are representations of geometric objects, including curves and surfaces. Their specificity is to take into account the tangent spaces of these objects and to be robust to the choice of parametrization. This new data attachment term extends the scope of application of the pre-existing methods of matching by large diffeomorphic deformations (LDDMM). The partial registration is indeed induced by a diffeomorphic deformation of the source shape. The anatomical (topological) characteristics of this shape are thus preserved. This is a joint work with Pierre-Louis Antonsanti and Joan Glaunès.

Herbert Edelsbrunner
Chromatic Delaunay mosaics for chromatic point data.

The chromatic Delaunay mosaic of s+1 finite sets in d dimensions is an (s+d)-dimensional Delaunay mosaic that represents the individual sets as well as their interactions. For example, it contains a (non-standard) dual of the oay of the Voronoi tessellations of any subset of the s+1 colors. We prove bounds on the size of the chromatic Delaunay mosaic, in the worst and average case, and suggest how to use image, kernel, and cokernel persistence to get stable diagrams describing the interaction of the points of different colors.
Acknowledgements. This is incomplete and ongoing joint work with Ranita Biswas, Sebastiano Cultrera, Ondrej Draganov, and Morteza Saghafian, all at IST Austria.

Break

Claire Brecheteau

Approximating data with a union of ellipsoids and clustering.

I will introduce a surrogate for the distance function to the support of a distribution, which sublevel sets are unions of balls or of ellipsoids. I will expose different results, including rates of convergence for the approximadomtion of these surrogates with their empirical versions, built from pointclouds. I will explain how to use such estimators to cluster data with a geometric structure. The results have been published in the papers [1,2], and are still in progress.
[1] C. Brécheteau. Robust anisotropic power-functions-based filtrations for clustering. In 36th International Symposium on Computational Geometry (SoCG 2020), vol. 164, 2020.
[2] C. Brécheteau, C. Levrard. A k-points-based distance for robust geometric inference. Bernoulli, 26(4), 2020.

Dominique Attali

Reconstructing manifolds by weighted ℓ_1-norm minimization.

In this talk, we focus on one particular instance of the shape reconstruction problem, in which the shape we wish to reconstruct is an orientable smooth d-manifold embedded in $\mathbb{R}^N$. We reformulate the problem of searching for a triangulation as a convex minimization problem, whose objective function is a weighted ℓ_1-norm. I will then present the result in [1] which says that, under appropriate conditions, the solution of our minimization problem is indeed a triangulation of the manifold and that this triangulation coincides with a variant of the tangential Delaunay complex.

This is a joint work with André Lieutier.

References

[1] Dominique Attali and André Lieutier. Delaunay-like triangulation of smooth orientable sub-manifolds by least ℓ_1-norm minimization. In **38th International Symposium on Computational Geometry (SoCG'22)**, Berlin, Germany, June 7-10, 2022. http://www.gipsa-lab.grenoble-inp.fr/ dominique.attali/Publications/2022-socg-submission-optim.pdf[download].

Friday

Barbara Gris

Defining Data-Driven Deformation Models.

Studying shapes through large deformations allows to define a metric on a space of shapes from a metric on a space of deformations. When the set of considered deformations is not rel-evant to the observed data, the geodesic paths for this metric can be deceiving from a model-ling point of view. To overcome this issue, the notion of deformation module allows to incorpo-rate prior coming from the data in the set of considered deformations and the metric. I will pre-sent this framework, as well as the IMODAL library which enables to perform registration through such structured deformations. This Python library is modular: adapted priors can be easily defined by the user, several priors can be combined into a global one and various types of data can be considered such as curves, meshes or images. This is a joint work with Benjamin Charlier, Leander Lacroix and Alain Trouvé.

Laurent Younes

Stochastic Gradient Descent for Large-Scale LDDMM.

Large deformation diffeomorphic metric mapping (LDDMM) is a registration/comparison method that aligns two shapes or images using geodesics in the group of diffeomorphisms. Lagrangian implementations of the method, which are well adapted to compare, for example, discrete curves or triangulated surfaces, have a complexity proportional to the square of the number of points composing the compared objects, which can become intractable for large-scale problems. In this talk, we explore a stochastic gradient descent strategy, involving a variation of the standard reduction leading to Lagrangian LDDMM, that allow for cost-reducing randomization. We provide preliminary numerical highlights and experimental results for this method.

Stephen Preston

Isometric immersions and the waving of flags.

A physical flag can be modeled geometrically as an isometric immersion of a rectangle into space, with one edge fixed along the flagpole. Its motion, in the absence of gravity and wind, can be modeled as a geodesic in the space of all isometric immersions, where the Riemannian metric is inherited from the kinetic energy on the much larger space of all immersions. In this talk I will show how generically such an isometric immersion can be described completely by the curve describing the top or bottom edge, which gives a global version of a classical local result in differential geometry. Using this, I will show how to derive the geodesic equation, which turns out to be a highly nonlinear, nonlocal coupled system of two wave equations in one space variable, with tension determined by solving

an ODE system. The new model has the potential to describe motion of cloth with much fewer variables than the traditional method of strongly constraining three functions of two space variables.
This is joint work with Martin Bauer and Jakob Moeller-Andersen.