Mathematisches Forschungsinstitut Oberwolfach

Report No. 3/2018

DOI: 10.4171/OWR/2018/3

Statistics for Data with Geometric Structure

Organised by
Aasa Feragen, Copenhagen
Thomas Hotz, Ilmenau
Stephan Huckemann, Göttingen
Ezra Miller, Durham

21 January – 27 January 2018

Abstract. Statistics for data with geometric structure is an active and diverse topic of research. Applications include manifold spaces in directional data or symmetric positive definite matrices and some shape representations. But in some cases, more involved metric spaces like stratified spaces play a crucial role in different ways. On the one hand, phylogenetic trees are represented as points in a stratified data space, whereas branching trees, for example of veins, are data objects, whose stratified structure is of essential importance. For the latter case, one important tool is persistent homology, which is currently a very active area of research. As data sets become not only larger but also more complex, the need for theoretical and methodological progress in dealing with data on non-Euclidean spaces or data objects with nontrivial geometric structure is growing. A number of fundamental results have been achieved recently and the development of new methods for refined, more informative data representation is ongoing. Two complimentary approaches are pursued: on the one hand developing sophisticated new parameters to describe the data, like persistent homology, and on the other hand achieving simpler representations in terms of given parameters, like dimension reduction. Some foundational works in stochastic process theory on manifolds open the doors to this field and stochastic analysis on manifolds, thus enabling a well-founded treatment of non-Euclidean dynamic data. The results presented in the workshop by leading experts in the field are great accomplishments of collaboration between mathematicians from statistics, geometry and topology and the open problems which were discussed show the need for an expansion of this interdisciplinary effort, which could also tie in more closely with computer science.

Mathematics Subject Classification (2010): 62xx, 53xx, 60xx, 65xx, 14xx.

Introduction by the Organisers

The workshop *Statistics for Data with Geometric Structure*, organized by Aasa Feragen (Copenhagen), Thomas Hotz (Ilmenau), Stephan Huckemann (Göttingen) and Ezra Miller (Durham) had 48 participants from many countries around the world. In particular, 14 of the 17 participants in the mini-workshop *Asymptotic Statistics on Stratified Spaces* held in 2014 at the MFO took part in this workshop. The interdisciplinary nature of the subject matter was reflected in the very diverse mathematical backgrounds of the speakers.

In the past years, data with geometric structure play an increasingly important role in statistics and lead to a surge in the application of geometric and topological concepts in statistical data analysis. Two major classes of approaches are pursued in this field. The first approach seeks to represent geometric objects as points in a non-Euclidean data space, while the second approach seeks to extract the major features of the geometric object to achieve a refined representation, not necessarily in a non-Euclidean space.

In the first approach, the spaces need not even be manifolds, but can be stratified spaces, in which case means can have non-standard properties, called stickiness [2] and repulsiveness. These are especially relevant for phylogenetic tree spaces which are used in population genetics. Calculation of geodesics [4] and analogues to principal components [3] is very challenging in these spaces.

On the other hand, measures on spaces with positive curvature can exhibit lower rates of asymptotic convergence of the sample mean, called smeariness [5, 6]. Such spaces of positive curvature are the principal object of concern in directional statistics and many shape representations.

For many spaces, refined methods have been developed, for example for dimension reduction and also some generic asymptotic results were achieved, see e.g. [7, 8, 9]. Furthermore, many specific difficulties for various data representations have been described and partly solved [12, 13].

A very important field, which is currently emerging, is the theory of stochastic processes and stochastic analysis on manifolds. Recent important foundational work has been done by Sommer and Joshi [11], whose collaboration was fostered by the mini-workshop *Asymptotic Statistics on Stratified Spaces*. The development of new models, the underlying computational theory, as well as computational tools are a milestone towards an effective treatment of stochastic processes on manifolds.

The second approach to data with geometric structure seeks to extract the major features of the geometric object to achieve a refined representation. A major technique to this effect is persistent homology (for an introduction and historical overview, see [1]), which is increasingly used in image and shape analysis. In this class of methods, scale-space-like transformations are used to represent complicated geometric objects in terms of topological properties. For example, separation of clusters and sizes of holes in a data set can be quantified in terms of persistence diagrams.

For every data set, the construction scheme of the persistence diagram must be reconsidered. In many applications, level sets of (possibly multivariate) functions

are considered; in some cases, objects are sliced in different angles to create a whole ensemble of persistence diagrams [14] and also several independent parameters, leading to higher dimensional persistence diagrams are considered (see the contribution by E. Miller).

Furthermore, the parameters of interest to extract from the persistence diagram must be determined for every application specifically. This can range from a reduction to simple scalar summary statistics, over curves [15] to sophisticated analysis applying tropical geometry [16].

The workshop provided an overview over the very diverse subject of statistics for data with geometric structure and a number of different ways to approach the subject. It initiated lively discussions concerning several topics, which were further discussed in five focus groups.

Acknowledgement: The MFO and the workshop organizers would like to thank the National Science Foundation for supporting the participation of junior researchers in the workshop by the grant DMS-1641185, "US Junior Oberwolfach Fellows". Moreover, the MFO and the workshop organizers would like to thank the Simons Foundation for supporting Ruriko Yoshida in the "Simons Visiting Professors" program at the MFO.

References

[1] H. Edelsbrunner and J. Harer, *Persistent Homology – a Survey*, in: Surveys on Discrete and Computational Geometry: Twenty Years Later (2008).

[2] T. Hotz, S. Huckemann, H. Le, J.S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer, *Sticky central limit theorems on open books*, Ann. Appl. Probab. **23** (2013), 2238–2258.

[3] T.M.W. Nye, X. Tang, G. Weyenberg, and R. Yoshida, *Principal component analysis and the locus of the Fréchet mean in tree space*, Biometrika **104** (2017), 901–922.

[4] M. Owen, and J.S. Provan, *A fast algorithm for computing geodesic distances in tree space*, IEEE/ACM Trans. Comput. Biol. Bioinf. **8** (2011), 2–13.

[5] T. Hotz and S. Huckemann, *Intrinsic means on the circle: Uniqueness, locus and asymptotics*, Annals of the Institute of Statistical Mathematics **67** (1) (2015), 177–193.

[6] B. Eltzner and S. Huckemann, *A Smeary Central Limit Theorem for Manifolds with Application to High Dimensional Spheres*, arXiv:1801.06581

[7] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–100, January 2010.

[8] Sungkyu Jung, Ian L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, September 2012.

[9] X. Pennec, *Barycentric Subspace Analysis on Manifolds*, To appear in Annals of Statistics, Institute of Mathematical Statistics. `https://arxiv.org/abs/1607.02833v2`, Oct 2017.

[10] Stefan Sommer. An Infinitesimal Probabilistic Model for Principal Component Analysis of Manifold Valued Data. *arXiv:1801.10341 [cs, math, stat]*, January 2018. arXiv: 1801.10341.

[11] S. Sommer, A. Arnaudon, L. Kuhnel, S. Joshi. Bridge Simulation and Metric Estimation on Landmark Manifolds. arXiv:1705.10943 [cs.CV]

[12] L. Devilliers, S. Allasonnire, A. Trouv and X. Pennec., *Template estimation in computational anatomy: Frchet means in top and quotient spaces are not consistent.* SIAM Journal of Imaging Science, 10(3):1139-1169. (2017).

[13] N. Miolane, S. Holmes, X. Pennec., *Template shape computation: correcting an asymptotic bias.* SIAM Journal of Imaging Science, 10(2):808-844, (2017).

[14] K. Turner S. Mukherjee D. M. Boyer . Persistent homology transform for modeling shapes and surfaces. *Information and Inference: A Journal of the IMA*, 3, 4:310–344, (2014)
[15] L. Crawford, A. Monod, A. X. Chen, S. Mukherjee, R. Rabadán. Functional Data Analysis using a Topological Summary Statistic: the Smooth Euler Characteristic Transform. arXiv:1611.06818 [stat.AP] (2017).
[16] A. Monod, S. Kališnik, J.Á. Patiño-Galindo, L. Crawford. Tropical Sufficient Statistics for Persistent Homology. arXiv:1709.02647 [math.ST] (2017).

Workshop: Statistics for Data with Geometric Structure

Table of Contents

Abstracts

Object Oriented Data Analysis: Principal Nested Submanifolds

J. Steve Marron

Object Oriented Data Analysis is the statistical analysis of populations of complex objects.This is seen to be particularly relevant in the Big Data era, where it is argued that an even larger scale challenge is Complex Data.Data objects with a geometric structure constitute a particularly active current research area.This is illustrated using a number of examples where data objects naturally lie in manifolds and spaces with a manifold stratification.An overview of the area is given, with careful attention to vectors of angles, i.e. data objects that naturally lie on torus spaces. Prinicpal Nested Submanifolds, which are generalizations of flags, are proposed to provide new analogues of Principal Component Analysis for data visualization.Open problems as to how to weight components in a simultaneous fitting procedure are discussed.

Introduction to Manifold Statistics

Sarang Joshi
(joint work with P. Thomas Fletcher)

1. Introduction

Over the last decade there has been intense interest in developing statistical methods for the analysis of manifold valued data. In this talk I will give a brief overview of some the methods we have developed [4, 3, 5]. One of the first application of Manifold Statistics as been the analysis of directional data [10]. In the analysis of two dimensional directional data the natural model space for the data is the unit circle. For three dimensional directional data analysis the natural data space is the unit sphere in three space. Both of these data spaces are examples of smooth Riemannian Manifolds. Another important application of Manifold Statistics has been the analysis of shape [1], in particular the configuration of N labeled landmark configurations modulo orientation and scale. This was first studied by Kendall [9] and is referred to as Kendall Shape Space. In this talk I will not go in to details of any particular application but rather outline the general concepts of methods for statistical analysis of manifold valued data.

2. Basic Statistics on Manifolds

2.1. **Point Estimation.** Two fundamental statistical concepts of characterizing the spread of set of data points are the **sample variance** and **mean absolute deviation.** Both concepts have a natural definition for a collection of points in an abstract metric space. The sample variance around the mean is the sum of normalized square distances: $\sigma^2 = \frac{1}{N}\sum_i d^2(\mu, x_i)$. The mean absolute deviation is similarly

defined as the average of the distances to the median m: $D_{med} = \frac{1}{N}\sum_i d(m, x_i)$.

Point Estimation of the Mean. Given a collection of data objects that are elements of an abstract Riemannian manifold, a natural statistical question is the point estimation of the mean. The concept of Fréchet mean is to define the "average" as the point on the Riemannian manifold as the minimizer of the sum of squared geodesic distances from the mean to all the data points, or the minimum variance estimate. The existence and uniqueness of the Fréchet mean is not guaranteed in general and depends on the completeness and sectional curvature properties of the metric [8]. By using weighted squared geodesic distances, one can use this concept to define the notion of interpolation and filtering of an abstract manifold valued data set. We have used this effectively on the space of positive definite matrices to define filtering of DTI data sets [3].

A stable gradient descent algorithm for computing the Fréchet mean consists of 1) initializing the estimate as one of the data points; 2) computing the geodesic distances between the current estimate and all the data points, i.e., solve the geodesic boundary value problem; and 3) updating the estimate of the mean by shooting in the direction of the average of the initial velocities of the geodesics computed previously, i.e., solving the geodesic initial value problem.

Point Estimation of the Median. Similar to the Fréchet mean, the Fréchet median is defined as the minimizer of the sum of absolute geodesic distance or the mean absolute deviation and is also the generalization of the Fermat-Weber problem. In [5] we used this to define a robust statistical estimation of the anatomical atlas and extended the notion of median filtering. Analogous to the gradient descent algorithm above, one can use the Weiszfeld's algorithm, which also requires completeness properties of the Riemannian metric.

2.2. **Regression Analysis.** Regression analysis is the study of the relationship between measured data and descriptive variables. As with most statistical techniques, regression analyses can be broadly divided into two classes: parametric and nonparametric. The most widely used parametric regression methods for data having a linear vector space structure are linear and polynomial regression, wherein a linear or polynomial function is fit in a least-squares fashion to observed data. Such methods are the staple of modern data analysis. The most common nonparametric regression approaches are kernel-based methods and spline-smoothing approaches, which provide great flexibility in the class of regression functions.

Geodesic Regression and Polynomial Regression. Recently, [2, 7] have each independently developed a form of geodesic regression that generalizes the notion of linear regression to Riemannian manifolds. In Hinkle-Fletcher-Joshi [6] geodesic regression was further generalized to polynomial regression for manifold valued data. The basic construction is to model manifold-valued random variable Y as

$$Y = \exp(\gamma(t), \epsilon) \ ,$$

where $\gamma(t)$ is a Riemannian polynomial of integer order k and exp is the Riemannian exponential map. Analogous to polynomials in a vector space, Riemannian

polynomials are defined as curves having the zero k^{th} order covariant derivative, i.e.,

$$\left(\nabla_{\dot\gamma(t)}\right)^k \dot\gamma(t) = 0 \ ,$$

where $\dot\gamma(t) = \frac{d}{dt}\gamma(t)$. As with regular polynomials Riemannian polynomials are fully determined by initial conditions at $t = 0$. Given observed data $x_i \in M$ at times t_i, the minimum variance $k^{th}-$ order polynomial regression is defined the minimization of the objective function

$$E\left(\gamma(0), v_1(0), \cdots, v_k(0)\right) = \frac{1}{N}\sum_{i=1}^{N} d^2\left(\gamma(t_i), x_i\right) \ ,$$

where $\gamma(0)$ is the initial point and $v_j(0), j = 1, \cdots, k$ are the initial conditions and parameters of the model. The energy function defined above is minimized using adjoint optimization.

References

[1] I. L. Dryden and K. Mardia. *Statistical Shape Analysis*. John Wiley & Son, 1998.

[2] P. T. Fletcher. Geodesic regression and the theory of least squares on riemannian manifolds. *International journal of computer vision*, 105(2):171–185, 2013.

[3] P. T. Fletcher and S. Joshi. Riemannian geometry for the statistical analysis of diffusion tensor data. *Signal Processing*, 87(2):250–262, 2007.

[4] P. T. Fletcher, C. Lu, S. M. Pizer, and S. Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *IEEE transactions on medical imaging*, 23(8):995–1005, 2004.

[5] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009.

[6] J. Hinkle, P. T. Fletcher, and S. Joshi. Intrinsic polynomials for regression on riemannian manifolds. *Journal of Mathematical Imaging and Vision*, 50(1-2):32–52, 2014.

[7] Y. Hong, N. Singh, R. Kwitt, and M. Niethammer. Time-warped geodesic regression. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 105–112. Springer International Publishing, 2014.

[8] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on pure and applied mathematics*, 30(5):509–541, 1977.

[9] D. G. Kendall. Shape manifolds, procrustean metrics and complex projective spaces. *Bulletin of London Mathematical Society*, 16:81–121, 1984.

[10] K. V. Mardia and P. E. Jupp. *Directional statistics*, volume 494. John Wiley & Sons, 2009.

Statistics with data on stratified spaces

Tom M. W. Nye

Conventional statistical methods typically rely on the data lying in a vector space. This assumption is fundamental in standard methods such as linear regression and principal component analysis, but also underlies results such as the central limit theorem. If the data instead lie in a smooth Riemannian manifold, much statistical methodology can be transferred to the new setting. However, some important applications give rise to data lying in so-called *manifold-stratified* spaces. Informally, a manifold-stratified space consists of a set of manifolds with boundary M_i,

$i = 1, 2, \ldots$, each equipped with a metric, together with a set of rules for gluing the manifolds together isometrically at their boundaries. Examples include simplicial complexes, cubical complexes (in which every cell is a unit Euclidean cube), orthant spaces (in which every cell is a copy of $\mathbb{R}^d_{\geq 0}$), and certain quotient spaces.

Example: A k-spider consists of k copies of $\mathbb{R}_{\geq 0}$, each equipped with the standard metric and glued together at the shared origin. An *open book* is the product of $\mathbb{R}^d$ with a k-spider. The 3-spider parametrizes the set of rooted leaf-labelled trees with three leaves, in which the single internal edge in each tree has a positive weight: given leaf labels $\{A, B, C\}$, there are three bifurcating labelled shapes $((A, B), C)$, $((C, A), B)$, $((B, C), A)$, together with the tree with no internal edges (A, B, C) corresponding to the origin of the spider. Each leg of the 3-spider corresponds to a different bifurcating shape, and the position along each leg determines the weight assigned to the internal edge on each tree.

Estimators on spiders and open books have unexpected properties which contrast to the usual properties on Euclidean vector spaces. The Fréchet mean (or barycenter) of a sample from a distribution on a 3-spider has a tendancy to 'stick' to the origin, with the estimate remaining at the origin despite small perturbations to the data. This stickiness phenomenon is due to the underlying non-positive curvature of the space, and a central limit theorem incorporating stickiness has been proved on the open book [2].

The 3-spider is a special case of a more general space trees, known as Billera-Holmes-Vogtmann (BHV) tree space [1]. The Billera-Holmes-Vogtmann tree space $\mathcal{T}_N$ is an orthant space which parametrizes of edge-weighted rooted trees for which the leaves are bijectively labelled $\{1, \ldots, N\}$. The Euclidean metric on each orthant extends globally and BHV tree space is non-positively curved. Owen and Provan [4] established a $O(N^4)$ algorithm for computing geodesics in $\mathcal{T}_N$. These ingredients enable practical statistics to be carried out, such as computation of Fréchet means and construction of principal geodesics. Recent work concerns construction of principal surfaces as barycentric subspaces of $\mathcal{T}_N$ [3].

There are many remaining challenges in this area. Analysis in BHV tree space relies critically on the non-positive curvature property, and there is a lack of results for spaces for which this property does not hold: for example much less is known about spaces of unlabelled trees, or spaces of trees with different numbers of leaves. General results about the effect of curvature on the asymptotics of estimators are beginning to be established. To date, most estimators studied are non-parametric and based on least-squares constructions. Recent work has started to consider parametric distributions constructed as the transition kernels of stochastic processes on tree space. This opens up an alternative approach to developing statistical methodology on these non-standard spaces.

References

[1] L.J. Billera, S.P. Holmes, and K. Vogtmann, *Geometry of the space of phylogenetic trees.* Adv. Appl. Math. **27** (2001), 733–767.

[2] T. Hotz, S. Huckemann, H. Le, J.S. Marron, J. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer, *Sticky central limit theorems on open books*, Ann. Appl. Probab. **23** (2013), 2238–2258.

[3] T.M.W. Nye, X. Tang, G. Weyenberg, and R. Yoshida, *Principal component analysis and the locus of the Fréchet mean in tree space*, Biometrika **104** (2017), 901–922.

[4] M. Owen, and J.S. Provan, *A fast algorithm for computing geodesic distances in tree space*, IEEE/ACM Trans. Comput. Biol. Bioinf. **8** (2011), 2–13.

Persistent Topology and Stochastic Geometry

Herbert Edelsbrunner

(joint work with many, mentioned as coauthors of listed papers)

Historical remarks. The idea of persistent homology was motivated by looking at protein structures, each represented by the family of alpha shapes we get by letting the radius of the atom balls go from zero to infinity. With the tool implemented by Ernst Mücke [4] and enhanced with Betti numbers by Jose Delfinado [2], we computed the number of tunnels in a cell membrane protein and noticed that it is not equal to one, as it should be, for any value of the radius. This prompted the question whether there is enough information in the sequence of homology groups to identify the visually important one tunnel from the mess of many. The answer was given a few years later in [3] with the introduction of persistent homology.

Definition of persistent homology. In a nutshell, *persistent homology* maps a sequence of spaces connected by inclusions (a *filtration*) to a sequence of homology groups connected by homomorphisms. For example, we may have a function on a topological space, $f\colon \mathbb{X} \to \mathbb{R}$, and we consider the filtration of its sublevel sets: $f^{-1}(-\infty, r]$. Using a field for the coefficients, the corresponding sequence of homology groups are vector spaces connected by linear maps. Homology classes are *born* in this sequence and *die* in this sequence, so we can record the classes by intervals or, equivalently, by points in two dimensions, where we record the birth on the horizontal coordinate axis and the death on the vertical coordinate axis. The resulting multi-set of point is commonly referred to as the *persistence diagram* of the filtration.

Stability. An important property of persistence is its stability. More precisely, consider two functions on a topological space, $f, g\colon \mathbb{X} \to \mathbb{R}$, and their respective persistence diagrams. We define the *bottleneck distance* between these diagrams as the length of the longest edge in a perfect matching, in which we choose the matching that minimizes this length and we are free to add points from the diagonal (where birth equals death) to either diagram if wish. The theorem, originally proved in [1], states that the bottleneck distance between the two diagrams is bounded from above by the L_∞-norm of $f - g$. Importantly, there are almost no

assumptions necessary, except that f and g be *tame*, which means that they both have only finitely many homological critical values and the homology groups of the sublevel sets have finite ranks.

Stochastic geometry. The use of persistence diagrams in statistical analyses of data begs the question of the expected diagram of noise, which we formalize as a stationary Poisson point process, $X \subseteq \mathbb{R}^d$. We are not able to answer this question in mathematical detail, but we have been able to shed light on the expected number of critical and non-critical simplices in the Delaunay mosaic of X. To define these notions, let $f\colon D(X) \to \mathbb{R}$ map every simplex of the Delaunay mosaic to the radius of the smallest ball whose bounding sphere passes through the vertices of the simplex, and whose interior does not contain any of the points of X. Assuming X is in general position, which happens with probability 1, the difference between two contiguous sublevel sets of f is an interval in the face lattice of $D(X)$. If this interval consists of a single simplex, then we call this a *critical simplex*; all simplices in intervals of size two or larger are called *non-critical simplices*. For example, in $\mathbb{R}^2$ every acute triangle is critical, and every obtuse triangle is non-critical (it occurs together with its longest edge). Incidentally, half the triangles are expected to be acute and half to be obtuse. The stochastic analysis in [5] gives precise statements about the expected number of critical and non-critical simplices of any type and of radius at most some given threshold. For infinite radius, this gives the expected number of simplices in the Delaunay mosaic, which were studied in [6].

References

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

[2] C.J.A. Delfinado and H. Edelsbrunner, An incremental algorithm for Betti numbers of simplicial complexes on the 3-sphere, *Comput. Aided Geom. Design* **12** (1995), 771–784.

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

[4] H. Edelsbrunner and E.P. Mücke, Three-dimensional alpha shapes, *ACM Trans. Graphics* **13** (1994), 43–72.

[5] H. Edelsbrunner, A. Nikitenko and M. Reitzner, Expected sizes of Poisson–Delaunay mosaics and their discrete Morse functions, *Adv. Appl. Probab.* **49** (2017), to appear.

[6] R.E. Miles, On the homogeneous planar Poisson point process, *Math. Biosci.* **6** (1970), 85–127.

Workflows for data science with geometric structure

Franz J. Király

Data and models with inherent geometric structure - for example directions, rotations, trees, graphs arising as observations or model parameters - are some of the most frequently found non-standard features in practical data analysis problems.

Despite this high practical relevance and ever-increasing demand on the data science market, the field is suffering from a usability crisis caused by the lack of

available toolsets and coding environments flexible enough to specify analyses and modelling primitives in a simple, user-accessible language.

The disconnection from end-users and high market pressures even appear to be have caused a "backspill" from commercial providers of geometric data science solutions into the community, exploiting the community's theory-oriented mindset to acquire academic credibility for unvalidated data science solutions.

While the talk was intended to provide solutions for the first issue, in consequence it led to quite heated discussions about the second, and the philosophical foundations of the scientific method in general - hence this extended abstract will discuss both.

1. Part I: a data *science* problem

Working with geometric data is inextricably linked to real world applications in which these occur. From a scientific perspective, a central question of method development is which methods or modelling strategies work. As there is no one approach that is valid or useful for all questions and all datasets, this is always in relation to the modelling task and the data at hand. One frequently heard claim made at the workshop was "method X is a great idea" - but to be able to make this claim, good scientific practice necessitates the following:

- A well-defined, testable scientific question, including clear statement of task, endpoint, and hypothesis assessed. A frequent mistake is stating a method, but not what problem it is supposed to solve - but without doing so, no testable claims are made.
- A state-of-art study design, including necessary comparisons against baselines and the gold standard for the task. A frequent mistake is a study on irrelevant data or a comparison which is unsound, or unfair, e.g., not to baselines but to worse methods.
- A clean quantitative evaluation, optimally including a significance and effect size for the main conclusions. A frequent mistake is providing effect size but not significance or vice versa.

The reader may also find helpful to keep in mind the parallel field of evidence based medicine, where questions such as "does homeopathy provide an effective treatment of (a certain type of) bowel cancer", or "are CT-scans a useful diagnostic procedure for chest infections?" arise, in parallel to questions such as "are topological persistence diagrams useful in predictive modelling of (a certain type of) tabular data?", or "is manifold-based PCA an useful exploratory tool for genomic data?".

One type of method discussed at particular length was methods that summarize geometric data - in a number of cases, it was unclear which problem they should solve: exploratory visualization? Extracting features? Supervised prediction? Or something else? - lacking a testable hypothesis. In medicine, for example, this would be similar to not stating which disease the treatment is supposed to cure.

Also in line with the evidence based medicine parallel, it was also interesting to observe how a number of phenomena known in the context of pseudo-medicine were emerging in a context of (pseudo-)data science:

- Denying the epistemiological basis of the scientific method: "one cannot prove anything anyway since everything can be falsified" - ignoring that it is testability and strength of evidence, rather than (mathematical?) "proof" which is at the scientific method's heart.
- Attempts to leave the burden of proof with the critic rather than with the proponent: "but can you show that it does *not* work?"
- Vague claims about application studies that may not exist: "This is being used widely, for example by hospitals, physicists, and government agencies such as the NSA!" (Which? Many! But where exactly? Let's take the discussion off-line!)
- Conflicts of interest where scientists are directly or indirectly benefitting from a company marketing and selling a potentially problematic methodology, but fail to declare this as a conflict of interest when claiming miraculous properties.

Like many data scientific areas in the times of the data science revolution, the field of statistics for geometric data is currently going through a crisis of scientific transparency and reproducibility - answers need to be found quickly, and much can be learnt from the transition of medicine to evidence based medicine - not only regarding technical content, but also regarding social and political dynamics, as well as effective implementation of community standards.

2. Part II: a reproducibility problem

Part of good data science practice is ensuring reproducibility and transparency. On the technical side, a general requirement for this are open dissemination and quality code design - which, as secondary beneficial effects, enables end users with geometric data problems to easily make use of relevant methodology, and facilitates the setting up of validatory studies.

While "open science" is largely consensus in the geometric data science community, a solid codebase that would allow easy use of the most popular methods does not exist. The talk suggested to jointly design a workflow interface which implements a workflow API for:

(i) Representing and storing data which may include structured and geometric data types such as shape, direction, trees. For example, a tabular dataset of patients where for each patient, demographics, an image, and a collection of shapes is recorded.

(ii) The most important modelling tasks involving geometry. These fall in two broad categories: (A) models *for* geometric data, including: (A.1) feature transformation and feature extraction for geometric data; (A.2) exploratory data analysis, unsupervised learning, and visualization for geometric data; (A.3) supervised prediction where target or features are

geometric; (A.4) hypothesis testing, including association testing, involving geometric data types. (B) model structure inference where the model is of geometric nature, i.e., model inference produces a geometric object such as a tree on data which is not necessarily of a geometric type.

(iii) Meta-modelling tasks such as composite modelling, pipelining, hyper-parameter tuning and ensembling.

It was argued that the most natural way for building a comprehensive modelling interface was through the formalism of higher-order and composite types, such as in the object oriented programming paradigm. Widely used state-of-art modelling toolboxes such as mlr [1] and sklearn [2] already formalize non-geometric aspects of this. A possible approach may include abstraction and encapsulation at different levels, as first- and higher-order objects:

(i) Geometric data types, possibly with intrinsic/extrinsic geometric methods. This abstraction coincides with J.S. Marron's idea of "object oriented data analysis".

(ii) Data containers for abstract data types, including geometric ones. This is provided by packages such as xpandas [4].

(iii) Modelling strategies, including transformers and predictors. As in mlr [1] and sklearn [2], this could follow the fit/predict/parameter interface design, with an added "inference" interface for models where model structure inference has a geometric output. Object and interface typing may be natural in the geometric setting.

(iv) Meta-modelling as first-order modelling object. Reduction and model type mutation may occur here, e.g., through transformation of a geometric to a primitive data type.

(v) For probabilistic modelling, abstraction of a first-order types "distribution-of-[geometric-type]" probablity distribution interface, such as for example in skpro [3].

(vi) Metrics, losses and utility functions involving geometric objects or geometry related predictions - such measures would be of first-order or parametric types, and will likely have to refer to intrinsic/extrinsic geometry of the data or inference objects.

Object orientation on all abstraction levels would allow quick specification of a modelling workflow, benefitting both scientific clarity and easy access by end users.

A number of interesting scientific questions around the workflow interface design and a potential higher-order modelling type language specific to geometric objects remain open, though one would hope that answers emerge in a collaborative effort which mirrors the integrative nature of this undertaking.

References

[1] Bischl, Bernd and Lang, Michel and Kotthoff, Lars and Schiffner, Julia and Richter, Jakob and Studerus, Erich and Casalicchio, Giuseppe and Jones, Zachary M, *mlr: Machine learning in R* (2011). 120–140.

[2] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, Gaël Varoquaux. *API design for machine learning software: experiences from the scikit-learn project*, ECML PKDD Workshop: Languages for Data Mining and Machine Learning (2003).

[3] Frithjof Gressmann, Franz Király, Bilal Mateen, Harald oberhauser. *Probabilistic supervised learning.* arXiv pre-print (2018).

[4] Vitaly Davydov, Franz Király. *the python/xpandas package* (2017).

Machine Learning with Topological Signatures

Roland Kwitt

(joint work with C. Hofer, S. Huber, U. Bauer, J. Reininghaus and M. Niethammer)

Over the past decade, developments from the field of algebraic topology have evolved into computationally practical methods to analyze data from a topological perspective. Arguably, the most prevalent method used in practice is *persistent homology* [6, 10] which offers a concise summary representation of topological features in data in the form of barcodes / persistence diagrams. Persistent homology not only presents a versatile approach to analyze a wide variety of data objects, but it also opens up novel pathways to address learning problems based on topological information. Methods from this field have found a broad range of applications in different areas of science, including biology, computer vision, or medicine and are now more succinctly summarized as *topological data analysis (TDA)* [4].

Despite the advantages of TDA for capturing topological invariants of data and its potential benefits for learning purposes, TDA is still somewhat disconnected from developments in machine learning. With respect to persistent homology, this can be largely attributed to the unusual structure of the resulting topological summaries (as multi-sets) and the associated, computationally expensive, metrics in that space (e.g., p-Wasserstein). In fact, barcodes or persistence diagrams cannot be used directly as input to conventional learning techniques, e.g., SVMs, without potentially sacrificing desirable theoretical properties such as stability. Recently, however, several works (e.g., [9, 8, 5, 2]) have shown advances towards bridging the gap between machine learning and TDA, predominantly in the context of *kernel-based learning techniques* [11]. This works, as kernel-methods allow to work with *non-standard* (i.e., non-Euclidean) input data, upon the definition of a suitable kernel function that 1) captures some notion of similarity between input objects and 2) satisfies certain required conditions. However, this typically comes at the cost of computational complexity, as kernel-methods do not scale well with sample size [1]. Furthermore, kernels are either constructed explicitly by mapping data into an inner-product space, or a predefined kernel function implicitly induces the

mapping. In both cases, however, the mapping is fixed *a-priori* which immediately raises the question if this is an appropriate strategy for a particular learning task.

The immanent success of deep neural networks in vision or natural language processing (e.g., [3]) has, in fact, shown that it is highly beneficial to learn task-specific *representations* of data, instead of hand-crafting a suitable representation. While this already works remarkably well for many types of input, handling data with strong geometric structure, such as graphs or manifold-valued objects poses considerable algorithmic and theoretical challenges. The aforementioned topological summaries fall exactly into this category because of their unusual structure as multi-sets together with the associated metric(s). So far, this has largely prevented principled approaches to use the output of a TDA pipeline as input to neural networks.

Nevertheless, our initial work [7] on designing a neural network module that can directly handle topological summaries has shown promising results on various (supervised) learning tasks already (e.g., classification of graphs, or 2D object shape). The idea essentially is to construct a mapping of persistence diagrams in such a way that points in the diagram are projected onto a collection of (parametrized) *structure elements* and the projections are finally summed up. On the one hand, this facilitates to *learn* task-specific representations of these diagrams via via deep neural networks, that 1) preserve certain theoretical properties (e.g., stability to some extent) and 2) allow us to handle diagrams for homology groups of different dimension jointly. On the other hand, it also presents new, interesting questions from a theoretical point of view. Further developments along these lines bear great potential to improve predictive performance on various kinds of learning problems, as TDA can offer information complementary to existing approaches. Hence, developing principled ways of bridging the gap between learning with neural networks and TDA, both in terms of a well-founded theory and practical applicability, presents a promising research direction in this field.

References

[1] Chapelle, O. (2007). Training a support vector machine in the primal. *Neural Comput.*, 19(5):1155–78.

[2] Adams, H., Emerson, T., Kirby, M., Neville, R., Peterson, C., Shipman, P., Chepushtanova, S., Hanson, E., Motta, F., and Ziegelmeier, L. (2017). Persistence images: A stable vector representation of persistent homology. *JMLR*, 18(8):1–35.

[3] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[4] Carlsson, G. (2009). Topology and data. *Bull. Amer. Math. Soc.*, 46:255–308.

[5] Carriére, M., Cuturi, M., and Outot, S. (2017). Sliced Wasserstein kernel for persistence diagrams. In *ICML*.

[6] Edelsbrunner, H., Letcher, D., and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete Comput. Geom.*, 28(4):511–533.

[7] Hofer, C., Kwitt, R., Niethammer, M., and Uhl, A. (2017b). Deep learning with topological signatures. In *NIPS*.

[8] Kusano, G., Fukumizu, K., and Hiraoka, Y. (2016). Persistence weighted Gaussian kernel for topological data analysis. In *ICML*.

[9] Reininghaus, R., Bauer, U., Huber, S., and Kwitt, R. (2015). A stable multi-scale kernel for topological machine learning. In *CVPR*.
[10] Zomorodian, A. and Carlsson, G. (2004). Computing persistent homology. In *SCG*.
[11] B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001.

Statistics on quotient spaces

Nina Miolane
(joint work with Xavier Pennec, Susan Holmes)

Statistics on quotient spaces arise when one wants to analyze data that have some invariance properties. For example, analyzing *shape* data involve analyzing the attributes of an object that are invariant with respect to rotations and translations, or more generally with respect to a Lie group of transformations. One looks at *the equivalence class* of the object in order to analyze its shape. In this talk, we show that statistics on quotient spaces are asymptotically biased. We take the running example of shape spaces and more particularly of the template shape estimation.

Known biases on shape spaces Shapes can first refer to shapes of landmarks detected on objects. Procrustean analyses study shapes of landmarks by projecting the objects in the shape space through "alignment" or "registration". In this literature, "shape" refers to a quotient by rotations, translations and scalings, while "form" refers to a quotient by rotations and translations only. Le showed that the mean "shape" has no asymptotic bias for shapes of landmarks in 2D, but an asymptotic bias appears when the noise on the objects is non-isotropic as proven by Kent and Mardia in 2D. In contrast, Lele showed that the mean "form" has an asymptotic bias even with isotropic noise in 2D. A bias has also been observed by Du, Dryden and Huang: ordinary Procrustes analysis without taking into account noise on the landmarks may compromise inference. Kume et al. also observe, study and correct the difference between the Maximum Likelihood estimate of the mean shape versus the estimate of the Procrustean analysis.

Shapes can also refer to shapes of curves. Curve data are projected in their shape space by alignment, in the spirit of a Procrustean analysis. Unbiasedness was shown for shapes of signals by Kurtek under the assumption of no noise on the objects. Allassonnière et al provide experiments showing a bias when there is noise. The bias is proven by Bigot and Charlier for curves estimated from a finite number of points in the presence of error.

Statistics on quotient spaces are biased We were missing an *abstract geometric* understanding of the bias. When does it arise? Which variables control its magnitude? Is it restricted to the mean shape or does it appear for other statistical analyses? How important is it in practice: do we even need to correct it? If so, how can we correct it? This talk addresses these questions with the geometry of the quotient space Q.

The data X_i's are generated in the finite-dimensional Riemannian manifold M by the generative model:

$$X_i = \mathrm{Exp}(g_i \cdot Y, \epsilon_i), i = 1...n \tag{1}$$

where: (i) the parameter Y is the template shape in the shape space Q, (ii) $g_i \in G$ is an element of the Lie group G acting *isometrically* on Y, (iii) ϵ_i is the noise and follows a Gaussian of variance σ^2, (iv) Exp is the Riemannian exponential on M.

The template shape Y is estimated with the Fréchet mean $\hat{Y}$ of the data projected in the quotient space Q:

$$\hat{Y} = \underset{Y \in M}{\mathrm{argmin}} \sum_{i=1}^{n} \min_{g \in G} d_M^2(Y, g \cdot X_i). \tag{2}$$

This is the estimator obtained in Procrustean analyses, or with the "max-max" algorithm used in signals / curves / (medical) images analyses.

Theorem 1. *[**Asymptotic bias on the template shape estimation** [3]]*

In the regime of an infinite number of data $n \to +\infty$, the asymptotic bias of the template's shape estimator $\hat{Y}$, with respect to the parameter Y, has the following Taylor expansion around the noise level $\sigma = 0$:

$$Bias(\hat{Y}, Y) \equiv Log_Y \hat{Y} = -\frac{\sigma^2}{2} H(Y) + \mathcal{O}(\sigma^4) + \epsilon(\sigma) \tag{3}$$

where (i) Log_Y is the Riemannian logarithm on Q at Y, hence a tangent vector at Y (ii) H is the mean curvature vector of the template shape's orbit which represents the external curvature of the orbit in M, and (iii) ϵ is a function of σ that decreases exponentially for $\sigma \to 0$.

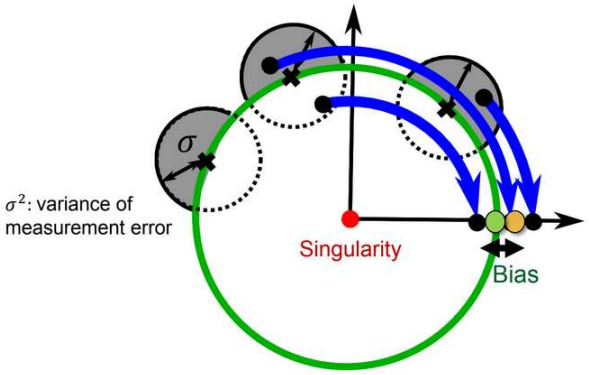


FIGURE 1. The external curvature of the template shape orbit, at the scale of σ creates the bias. The presence of the singularity of the quotient space creates the bias and has a repulsive effect of the template shape estimate.

Formulated in the Procrustean terminology, the result of Theorem 1 is: the Generalized Procrustes Analysis estimator of mean "form" is asymptotically biased. We don't consider the scalings as we assume an isometric Lie group action. This result also provides a geometric interpretation for the bias on signals and curves.

The variables controlling the bias are: (i) the distance in shape space from the template Y to a singular shape (the external curvature of orbits generally increases when Y is closer to a singularity) and (ii) the noise's scale σ. This helps determining when the bias is important and needs correction.

This bias goes beyond the template shape estimation. The next theorem shows that any Gaussian noise on the objects in M induces a non-centered skewed noise on the shapes in Q. A statistical learning that relies on a centered noise model in Q is biased. This decreases for example the performance of K-mean algorithms on shapes: clusters are less separated because of each centroid's bias.

Theorem 2. *[**Noise on shapes induced by noise on objects** [3]]*
The probability distribution function f induced by the generative model 1 on the shapes of the X_i's, $i = 1...n$, in the asymptotic regime on an infinite number of data $n \to +\infty$, has the following Taylor expansion around the noise level $\sigma = 0$:

$$f(Z) = \frac{1}{(\sqrt{2\pi}\sigma)^q} \exp\left(-\frac{d_M^2(Y,Z)}{2\sigma^2}\right)\left(F_0(Z) + \sigma^2 F_2(Z) + \mathcal{O}(\sigma^4) + \epsilon(\sigma)\right)$$

where (i) Z denotes a point in the shape space Q, (ii) F_0 and F_2 are functions of Z involving the derivatives of the Riemannian tensor at Z and the derivatives of the graph G describing the orbit O_Z at Z, and (iii) ϵ is a function of σ that decreases exponentially for $\sigma \to 0$.

The exponential in the expression of f belongs to a Gaussian distribution centered at Z and of isotropic variance $\sigma^2\mathbb{I}$. However the whole distribution f differs from the Gaussian because of the Z-dependent term in the right parenthesis. This induces a skew of the distribution away from the singularity.

We then propose an extension of the bootstrap, an iterative bootstrap on manifolds, that quantifies the bias and corrects it if needed [3]. Our results are exemplified on simulated and real data [3] and for example on the brain template shape estimation [4]. This analysis applies to finite dimensional manifolds quotiented by an isometric Lie group action. For insights on infinite dimensional Hilbert spaces, and possibly non isometric actions, we refer to the work of [1, 2].

References

[1] L. Devilliers, S. Allasonnire, A. Trouv and X. Pennec., *Template estimation in computational anatomy: Frchet means in top and quotient spaces are not consistent.* SIAM Journal of Imaging Science, 10(3):1139-1169. (2017).

[2] L. Devilliers, S. Allasonnire, A. Trouv and X. Pennec., *Inconsistency of Template Estimation by Minimizing of the Variance/Pre-Variance in the Quotient Space.* Entropy, 19(6):28. (2017).

[3] N. Miolane, S. Holmes, X. Pennec., *Template shape computation: correcting an asymptotic bias.* SIAM Journal of Imaging Science, 10(2):808-844, (2017).
[4] N. Miolane, S. Holmes, X. Pennec., *Topologically constrained template estimation control its consistency.* SIAM Journal of Geometry and Algebra (in revision). (2018).

Small versus Large Scale Features: Comparing the Appropriate Data Analysis Methods

Katharine Turner

Persistent homology captures geometrical and topological features at all different length scales. We can use persistent homology as a preprocessing step where the original data is replaced with a topological summary computed via persistent homology. Heuristically each persistent homology class corresponds to some geometric or topological feature in the data. In this talk I will compare some examples, discussing which topological summary is appropriate and what statistical methods are applicable

When comparing the persistent homology of two different samples we may be interested in using the persistent homology classes as proxies for individual "large scale" features. In this case it is well-motivated to use a bottleneck or Wasserstein distance between the persistence diagrams as these distances match up the persistent homology classes and compare the differences within each pair. As an example we can consider the persistent homology transform applied to morphology data sets such as a collection of calcanei (heel bones) of various primates. Here we have a persistence diagram for each vector in the sphere where we filter by the height function in that direction. Biological shape features will create persistent homology classes. Using the 1-Wasserstein distance we can integrate over the sphere of directions and add up the differences over a pair of bones as to when the biological shape features begin and end.

In contrast, there are also applications where we care about the distributions of the number of persistent homology classes of "short" lifetimes (such as in the analysis of point patterns). Here the features heuristically correspond to different types of local configurations. By analysing the distributions of the number of persistent homology classes with particular birth and death values we are indirectly analysing these distributions of local features. The persistent homology rank function is useful in these types of applications. For example, it can distinguish the phase type of 2D particle systems (fluid, hexatic vs crystalised) and is highly correlated to volume packing fraction in experimental sphere packing.

Smeariness in Higher Dimension – The Beast is Real!

Benjamin Eltzner
(joint work with Stephan F. Huckemann)

The central limit theorem (CLT) is among the foundations of statistics. The use of quantiles of an asymptotic distribution crucially relies on the fact, that the distribution of the difference between sample mean $\widehat{\mu}_n$ and population mean μ converges to a Gaussian distribution with a rate of $1/\sqrt{n}$. On a manifold $\mathcal{M}$ with dimension p, such as a circle, the usual definition of the mean of a distribution or sample does not work. Instead, the mean is defined as the solution to a minimization problem, using some metric d

$$\mu := \operatorname*{argmin}_{\lambda\in\mathcal{M}} \mathbb{E}[d(\lambda, X)^2] \qquad \widehat{\mu}_n := \operatorname*{argmin}_{\lambda\in\mathcal{M}} \frac{1}{n}\sum_{j=1}^{n} d(\lambda, X_j)^2 \,,$$

where, for simplicity, we assume uniqueness (a.s.). On the circle it was found by [2] that there are probability distributions, where a CLT holds with an asymptotic rate $n^{-\tau}$ where $\tau < 1/2$. The mean of such a distribution is called "smeary".

Definition 1 (Smeariness)**.** *Let μ be the population mean, $\widehat{\mu}_n$ the sample mean. A probability measure $\mathbb{P}$ is called smeary, if*

$$\exists \tau < 1/2 : \; n^{\tau} \log_{\mu}(\widehat{\mu}_n) = \mathcal{O}_P(1)$$

where log *denotes the inverse of the differential geometric exponential map.*

We explore necessary conditions for smeariness in higher dimension, prove a CLT and provide an example along with simulations.

1. Necessary Conditions for Smeariness

We refer to two theorems of [1] to point out necessary conditions for smeariness to occur. For some $q \in \mathcal{M}$ define the *cut locus*

$$C(q) := \overline{\{p \in \mathcal{M}, \text{ more than one shortest geodesic connect } q \text{ and } p\}}$$

and let $B_\varepsilon(q)$ be a geodesic ball of radius ε around q. Then we can formulate the theorem

Theorem 1 (Corollary 2.3 from [1])**.**
If supp$(\mathbb{P}) \subseteq \mathcal{M}\setminus \{q \in \mathcal{M} : \exists x \in B_\varepsilon(\mu) : \; q \in C(x)\}$, then the CLT holds for μ.

Conversely, this means that a non-empty $C(\mu)$ and a nonzero probability density at $C(\mu)$ are necessary for smeariness. This theorem is not restricted to a specific class of manifolds but holds generally and thus determines an important necessary condition for smeariness to occur.

The next question that arises is, what a probability measure at the cut locus has to satisfy to cause smeariness. Approaching this question, it is helpful to consider the following theorem

Theorem 2 (Theorem 3.3 from [1])**.**
Let $U \subset \mathbb{R}^p$ be an open neighborhood of 0. If the following conditions hold

(1) $\mathbb{E}[\operatorname{grad}_y d(\exp_\mu(y), X)^2] < \infty$ *and* $\mathbb{E}[\operatorname{Hess}_y d(\exp_\mu(y), X)^2] < \infty$ *for* $y \in U$*;*
(2) $\mathbb{P}(C(B_\varepsilon(\mu))) = \mathcal{O}(\varepsilon^{p-c})$ *for* $\varepsilon \to 0$*,* $0 \leq c < p$*;*
(3) for the Fréchet function $F(y) := \mathbb{E}[d(\exp_\mu(y), X)^2]$*,* $\operatorname{Hess} F(0)$ *is positive definite;*

then the standard central limit theorem holds if $p > 2 + c$*.*

In other words: For dimension $p > 2$ even probability densities which diverge not to quickly at the cut locus allow for a normal, non-smeary CLT.

It is clear that smeariness can only occur, if an assumption of the theorem is violated. Assumptions (1) and (2) are compelling regularity assumptions and a probability measure violating either assumption could be considered rather pathological. However, assumption (3) is a very refined technical assumption, which does not follow in a simple way from more natural assumption. Therefore, we focus on probability measures violating this assumption.

2. Smeariness on Spheres of Arbitrary dimension

First, we present an asymptotic result which is also valid for the case that condition (3) from theorem 2 does not hold. We suppress some technical assumptions for brevity.

Theorem 3 (Theorem 11 in [4], generalization of Theorem 5.23 in [3])**.** *Assume that the Fréchet function admits a power series expansion of the following form, where* $T_j > 0$ *and* $R \in SO(m)$

$$F(x) = F(0) + \sum_{j=1}^{m} T_j |(Rx)_j|^r + o(\|x\|^r) \qquad \text{where } 2 \leq r \in \mathbb{R}$$

Then, any random measurable selection of sample means $\widehat{\mu}_n$ *satisfies*

$$n^{1/2}(R^T \log_\mu(\widehat{\mu}_n))^{r-1} = W\mathcal{G} + o_P(1)$$

with a symmetric positive definite matrix W *and a multivariate normal vector* $\mathcal{G}$*. The expression* $(R^T \log_\mu(\widehat{\mu}_n))^{r-1}$ *denotes taking the product of the power of the absolute value multiplied with the original sign in each component separately.*

As an example, assume a point mass of magnitude $1 - \alpha \in (0, 1)$ at the north pole and a uniform distribution with total mass α on the southern hemisphere, illustrated in Figure 1. Then there is a critical value α_{crit} of the uniform density for every p, such that the Hessian of the Fréchet function vanishes and the first non-vanishing term is of order $r = 4$ at the mean. Thus, the mean is smeary with asymptotic rate $\tau = 1/6$.

Although all measures with smeary means known so far are very carefully constructed, the concept is more general than it appears at first sight. At finite sample size, samples from probability measures close to a smeary measure can be affected by slow convergence rates, which will render hypothesis tests unreliable. As an illustration, we perform simulations of the sample variance from the measures described above with $\alpha = \alpha_{\text{crit}} + \beta$, on $\mathbb{S}^p$. For every sample size, we draw 1000

FIGURE 1. Illustration of a probability measure on the sphere with smeary mean.

samples, determine the spherical mean for each sample and then determine the sum of squared distances of these means from the north pole. For $\beta \leq 0$ we have a unique minimum, where for the smeary case $\beta = 0$ we expect a slow decay of the empirical variance denoted by V with rate approaching $n^{-\frac{1}{3}}$, and for $\beta < 0$ we expect the rate to approach n^{-1}.

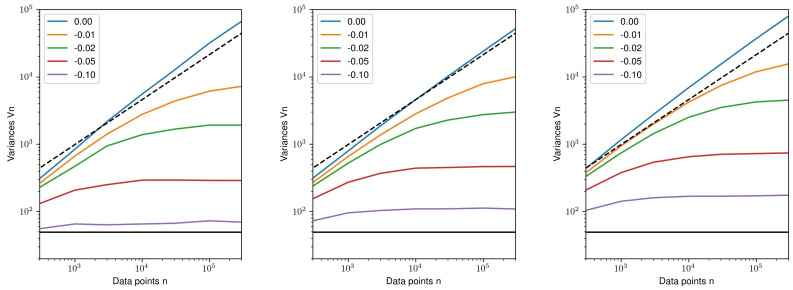


FIGURE 2. *Simulated variances V times n for different values of β for dimensions $p = 2$, 10 and 100 from left to right. Black lines $V \propto n^{-1}$ (solid) and $V \propto n^{-\frac{1}{3}}$ (dashed) for reference.*

However, Figure 2 clearly shows that the asymptotic rates follow the smeary case until fairly large sample sizes, before settling into the standard CLT behavior. This effect becomes more pronounced with increasing dimension, leading to a high dimension low sample size problem.

REFERENCES

[1] R. Bhattacharya and L. Lin, *Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces.*, The Proceedings of the American of Mathematical Society **145** (2016)

[2] T. Hotz and S. Huckemann, *Intrinsic means on the circle: Uniqueness, locus and asymptotics*, Annals of the Institute of Statistical Mathematics **67** (1) (2015), 177–193.

[3] A. van der Vaart, *Asymptotic statistics*, Cambridge Univ. Press, (2000)

[4] B. Eltzner and S. Huckemann, *A Smeary Central Limit Theorem for Manifolds with Application to High Dimensional Spheres*, arXiv:1801.06581

Probabilistic Inference on Manifolds

Stefan Sommer
(joint work with Sarang Joshi)

Statistical analysis of manifold valued data is often performed by generalizing least squares criterions and constructing data representations that mimic similar Euclidean constructions. This is for example the case for several generalizations of the Euclidean principal component analysis (PCA) procedure. PCA can be formulated as minimizing residual errors after approximating with low-dimensional linear subspaces. Procedures such as principal nested spheres (PNS/CPNS, [5]), horizontal component analysis (HCA, [8]) torus PCA (TPCA, [3]) geodesic PCA (GPCA, [4]) and barycentric subspace analysis (BSA, [7]) generalize this formulation to the nonlinear manifold setting.

In Euclidean space, fitting low-dimensional subspaces to data can equivalently be viewed as fitting Gaussian normal distributions by maximum likelihood. In essence, the log-density of the Gaussian distribution is a function of the negative square norm, and maximizing this is equivalent to minimizing squared distances. Inspired by this fact, probablistic PGA [15] and the later generalizations [9, 12] defined versions of the probabilistic PCA [14] procedure on manifolds by fitting parametric families of distributions to data.

Based on these ideas, we argue for a general probabilistic approach to statistical analysis of manifold valued data: Consider independently distributed data $y_1, \dots, y_N$ on the manifold M. Let μ_θ be a family of probability distributions $\mu_\theta \in \mathrm{Prob}(M)$ parametrized by a parameter θ. Assume now M is equipped with a fixed measure μ_0 and that μ_θ has a density. We can then let $p_\theta : M \to \mathbb{R}$ be a density such that $p_\theta \mu_0 = \mu_\theta$. From p_θ, we get a likelihood $\mathcal{L}(\theta; y_1, \dots, y_N) = \prod_{i=1}^N p_\theta(y_i)$, and we can search for a maximum likelihood estimate

$$\hat{\theta}_{\mathrm{ML}} = \mathrm{argmax}_\theta \mathcal{L}(\theta; y_1, \dots, y_N)$$

or, if we have a prior p on θ, a maximum a posteriori estimate

$$\hat{\theta}_{\mathrm{MAP}} = \mathrm{argmax}_\theta \mathcal{L}(\theta; y_1, \dots, y_N) p(\theta) \ .$$

This construction is of course natural from a probabilistic viewpoint, however, such formulations have not yet been widely explored in the manifold statistics literature. Probabilistic formulations essentially transfer the complexities of least squares constructions - projections to subspaces, existence of minimizers, recurrent geodesics, construction of linear-like subspaces - to constructions of parametric families of probability distributions. Such distributions can be defined in forms that are natural both from geometric and probabilistic viewpoints. In particular, distributions arising from stochastic processes can exploit that the infinitesimal definition of integral equations and SDEs are often naturally compatible with the differential structure of manifolds.

One example of such constructions are the anisotropic normal distributions [10, 13, 11] constructed as Brownian flows in the frame bundle of the manifold where the frames encode covariance structure. A similar example of using Brownian

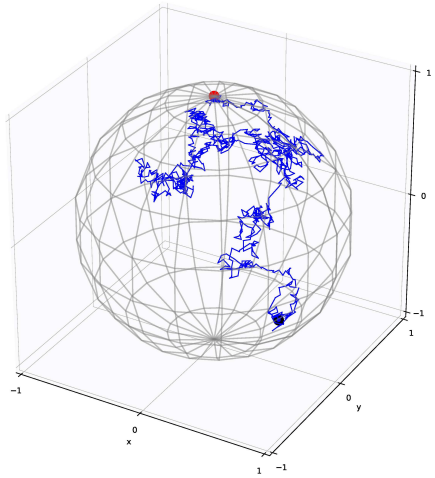


FIGURE 1. A sample from a $\mathbb{S}^2$ valued Brownian bridge from the north pole (red) to the target (black) simulated from a guided bridge scheme similar to the process (1).

motion to define a probability distribution on non-smooth spaces can be found in [6]. Both constructions use the parameter θ to encode a mean $x \in M$, and in the frame bundle construction additionally the covariance Σ. Fitting these parameters to data gives a maximum likelihood interpretation of the mean, or mean and covariance. See also [1] for a similar example of using stochastic processes to construct probability distributions in shape analysis. We are currently exploring similar constructions on Lie groups and orbit spaces.

The likelihood function can for stochastic processes be approximated by Monte Carlo sampling of bridge processes. One approach is to generalize the guided bridge simulation approach of Delyon and Hu [2] to manifolds. We explored well-posedness and existence of the guided SDE

$$dy_t = b(t, y_t)dt + \frac{\mathrm{Log}_{y_t}(v)}{T-t}dt + \sigma(t, y_t)dW_t \tag{1}$$

that uses the Riemannian Log-map to ensure the target $v \in M$ is hit a.s. under reasonable assumptions on the drift and coefficient terms b and σ. Even though Log is not continuous at the cut locus of v, the process can be shown to exist and the likelihood of v can be sought approximated from sampling y_t.

One important question arising from these considerations is properties and naturality of the probabilistic estimators, e.g. the ML mean. The Frechét mean and its corresponding manifold central limit theorem are influenced by curvature of M. In the non-smooth category, the Frechét mean exhibit stickiness or smearyness effects. It remains as an open question if the ML mean does or does not carry similar properties.

References

[1] Alexis Arnaudon, Darryl D. Holm, and Stefan Sommer. A Geometric Framework for Stochastic Shape Analysis. *submitted, arXiv:1703.09971 [cs, math]*, March 2017.

[2] Bernard Delyon and Ying Hu. Simulation of conditioned diffusion and application to parameter estimation. *Stochastic Processes and their Applications*, 116(11):1660–1675, November 2006.

[3] Benjamin Eltzner, Stephan Huckemann, and Kanti V. Mardia. Torus Principal Component Analysis with an Application to RNA Structures. *arXiv:1511.04993 [q-bio, stat]*, November 2015. arXiv: 1511.04993.

[4] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–100, January 2010.

[5] Sungkyu Jung, Ian L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, September 2012.

[6] Tom Nye. Construction of Distributions on Tree-Space via Diffusion Processes. Mathematisches Forschungsinstitut Oberwolfach, 2014.

[7] Xavier Pennec. Barycentric Subspace Analysis on Manifolds. *arXiv:1607.02833 [math, stat]*, July 2016. arXiv: 1607.02833.

[8] Stefan Sommer. Horizontal Dimensionality Reduction and Iterated Frame Bundle Development. In *Geometric Science of Information*, LNCS, pages 76–83. Springer, 2013.

[9] Stefan Sommer. Diffusion Processes and PCA on Manifolds. Mathematisches Forschungsinstitut Oberwolfach https://www.mfo.de/document/1440a/OWR_2014_44.pdf, 2014.

[10] Stefan Sommer. Anisotropic Distributions on Manifolds: Template Estimation and Most Probable Paths. In *Information Processing in Medical Imaging*, volume 9123 of *Lecture Notes in Computer Science*, pages 193–204. Springer, 2015.

[11] Stefan Sommer. Anisotropically Weighted and Nonholonomically Constrained Evolutions on Manifolds. *Entropy*, 18(12):425, November 2016.

[12] Stefan Sommer. An Infinitesimal Probabilistic Model for Principal Component Analysis of Manifold Valued Data. *arXiv:1801.10341 [cs, math, stat]*, January 2018. arXiv: 1801.10341.

[13] Stefan Sommer and Anne Marie Svane. Modelling anisotropic covariance using stochastic development and sub-Riemannian frame bundle geometry. *Journal of Geometric Mechanics*, 9(3):391–410, June 2017.

[14] Michael E. Tipping and Christopher M. Bishop. Probabilistic Principal Component Analysis. *Journal of the Royal Statistical Society. Series B*, 61(3):611–622, January 1999.

[15] Miaomiao Zhang and P.T. Fletcher. Probabilistic Principal Geodesic Analysis. In *NIPS*, pages 1178–1186, 2013.

Curvature effects in empirical means, PCA and flags of subspaces

Xavier Pennec

Because we have in practice a limited number of samples, a problem in geometric statistics is to determine the properties of the empirical Fréchet mean of n IID samples in a Riemannian manifold. In sufficiently concentrated conditions, the empirical Fréchet mean exists and is unique for each sample, so that we can define its expected moments for a fixed number of samples. Using a Taylor expansion of the Riemannian metric, we can compute the Taylor expansion of the moments of a (sufficiently concentrated) distribution. This is used in turn to practically compute the first and second order moments of empirical means of an IID n-sample.

The expected empirical mean (or more precisely its expected log at the mean of the underlying distribution) turns out to have an unexpected non vanishing term (a bias) of order 4 in the distribution extension and in 1/n with respect to the number of samples. This bias term is a double contraction of the covariant derivative of the Riemannian curvature with the covariance matrix, and vanishes for symmetric spaces:

$$\mathbf{E}\left[\log_{\bar{x}}(\bar{x}_n)^a\right] = \frac{1}{24n}(2\nabla_b R^a_{dce} + \nabla^a R_{cebd})(\mathfrak{M}_2^n)^{bc}(\mathfrak{M}_2^n)^{de} + O(\epsilon^5).$$

Likewise, the covariance of the empirical mean has a correction term in 1/n contracting twice the Riemannian curvature with the covariance:

$$\mathbf{E}\left[\log_{\bar{x}}(\bar{x}_n)^a \log_{\bar{x}}(\bar{x}_n)^b\right] = \frac{1}{n}\mathfrak{M}_2^{ab} + \frac{1}{3n}\left(R^a_{ced}\mathfrak{M}_2^{be} + R^b_{ced}\mathfrak{M}_2^{ae}\right)\mathfrak{M}_2^{cd} + O(\epsilon^3).$$

This term can be interpreted as an extended Ricci curvature: in positively curved spaces, the convergence with the number of samples is slower than in Euclidean spaces while it is accelerated in negatively curved spaces. We conjecture that these effects might be the prelude to the stickiness of the mean in limit cases where the curvature becomes singular.

The second part of the talk focuses on flags (sequences of properly nested) of affine spans for generalizing PCA to manifolds. Barycentric subspaces and affine spans are defined as the (completion of the) locus of weighted means to a number of reference points. They can be naturally nested by defining an ordering of the reference points, which allows the construction of forward or backward nested sequence of subspaces. However, forward or backward methods optimize one subspace at a time and cannot optimize the unexplained variance simultaneously for all the subspaces of the flag. In order to obtain a global criterion, PCA in Euclidean spaces is rephrased as an optimization on the flags of linear subspaces and we propose an extension of the unexplained variance criterion that generalizes nicely to flags of affine spans in Riemannian manifolds. This results into a particularly appealing generalization of PCA on manifolds, that we call Barycentric Subspaces Analysis (BSA). More details are available in [1]

References

[1] X. Pennec, *Barycentric Subspace Analysis on Manifolds*, To appear in Annals of Statistics, Institute of Mathematical Statistics. `https://arxiv.org/abs/1607.02833v2`, Oct 2017.

Tropical Sufficient Statistics for Persistent Homology

Anthea Monod

(joint work with Sara Kališnik, Juan Ángel Patiño-Galindo, and Lorin Crawford)

We show that an embedding in Euclidean space based on tropical geometry generates stable sufficient statistics for barcodes. Conventionally, barcodes are multiscale summaries of topological characteristics that capture the "shape" of data; however, in practice, they have complex structures which make them difficult to

use in statistical settings. The sufficiency result presented in this work allows for classical probability distributions to be assumed on the tropicalized representations of barcodes. This makes a variety of parametric statistical inference methods amenable to barcodes, all while maintaining their initial interpretations. More specifically, we show that exponential family distributions may be constructed.

We conceptually demonstrate sufficiency and illustrate its utility in persistent homology dimensions 0 and 1 with concrete parametric applications to HIV and avian influenza data.

References

[1] A. Monod, S. Kališnik, J.Á. Patiño-Galindo, L. Crawford *Tropical Sufficient Statistics for Persistent Homology*, arXiv:1709.02647 (2017).

Covariance Tensors on Riemannian Manifolds

Washington Mio

(joint work with Haibin Hang and Facundo Mémoli)

The mean and covariance tensor are widely used summaries of data in Euclidean space that allow for simple visualization and inference with techniques such as principal component analysis. The mean generalizes to data on metric spaces as minimizers of the Fréchet function; however, a principled formulation of covariance tensors still is lacking. Here, we discuss an approach to covariance tensors for random variables taking values on a Riemannian manifold.

To motivate our formulation, we begin with a reinterpretation of the classical covariance tensor associated with a random variable $y \in \mathbb{R}^d$ (with finite second moment) distributed according to a probability measure α. Instead of considering covariation of y only with respect to the mean, we approach covariance as a tensor field $\Sigma_\alpha \colon \mathbb{R}^d \to \mathbb{R}^d \otimes \mathbb{R}^d$ given by

$$\Sigma_\alpha(x) = \int_{\mathbb{R}^d} (y-x) \otimes (y-x)\, d\alpha(y)\,. \tag{1}$$

Σ_α encodes covariation with respect to any reference point $x \in \mathbb{R}^d$ and clearly depends only on the underlying distribution α. The term $(y-x)$ on the integrand uses the vector space structure on $\mathbb{R}^d$, so (1) does not directly extend to distributions on manifolds. To circumvent the problem, we rewrite covariance as follows. Consider the kernel function $u(x,y) = \|x-y\|^2/2$, which we may interpret as the potential energy of x relative to y. For a random $y \in \mathbb{R}^d$, the gradient field of $u(\cdot, y)$ is given by $\nabla_x u(x,y) = x - y$. Thus, we may write

$$\Sigma_\alpha(x) = \int_{\mathbb{R}^d} \nabla_x u(x,y) \otimes \nabla_x u(x,y)\, d\alpha(y)\,, \tag{2}$$

an expression that only invokes local linearization and more easily generalizes to the manifold setting.

Let (M, g) be a Riemannian manifold, $y \in M$ a random variable distributed according to a Borel probability measure α, and $u\colon M \times M \to \mathbb{R}^+$ a smooth, symmetric kernel function. (We assume that there is $A > 0$ such that $\|\nabla_x u(x, y)\|_x \leq A$, $\forall x, y \in M$, but this assumption may be relaxed.) For $k \geq 1$, we define the k-tensor field $\Sigma^k_{\alpha,u}$ at $x \in M$, as the expected value of the random variable $\otimes_k \nabla_x u(x, y) \in \otimes_k T_x M$. More formally, $\Sigma^k_{\alpha,u}$ is the section of the k-fold tensor product of the tangent bundle of M given by

$$\Sigma^k_{\alpha,u}(x) = \int_M \otimes_k \nabla_x u(x, y)\, d\alpha(y)\,. \tag{3}$$

The Fréchet function of α with respect to the kernel u is defined as

$$V_{\alpha,u}(x) = \int_{\mathbb{R}^d} u(x, y)\, d\alpha(y)\,. \tag{4}$$

Note that the 1-tensor $\Sigma^1_{\alpha,u}$ is the gradient field of $V_{\alpha,u}$, that is, $\nabla V_{\alpha,u} = \Sigma^1_{\alpha,u}$.

To state a stability result for covariance tensors, we introduce some notation. For each $x \in M$, the Riemannian structure on M induces an inner product on $\otimes_k T_x M$ given on pure tensors by $\left\langle \otimes^k_{i=1} v_i, \otimes^k_{i=1} w_i \right\rangle_x = \Pi^k_{i=1} \langle v_i, w_i \rangle_x$. We write $\| \cdot \|_x$ for the associated norm, omitting k from the notation. We denote the geodesic distance on (M, g) by d_g and write $\mathcal{P}_1(M, d_g)$ for the 1-Wasserstein space associated with (M, d_g) and w_1 for the 1-Wasserstein distance on $\mathcal{P}_1(M, d_g)$.

Theorem 1. *Let (M, g) be a complete Riemannian manifold and $\alpha, \beta \in \mathcal{P}_1(M, d_g)$. Suppose that $u\colon M \times M \to \mathbb{R}^+$ is a smooth, symmetric function that satisfies (i) $\|\nabla_x u(x, y)\|_x \leq A$, $\forall x, y \in M$ and (ii) $\|\nabla_x u(x, y_1) - \nabla_x(x, y_2)\|_x \leq L d_g(y_1, y_2)$, $\forall x, y_1, y_2 \in M$, where $A > 0$ and $L > 0$. Then, for any $k \geq 1$,*

$$\sup_{x \in M} \|\Sigma^k_{\alpha,u}(x) - \Sigma^k_{\beta,u}(x)\|_x \leq k A^{k-1} L\, w_1(\alpha, \beta)\,.$$

Remark. A consistency result for covariance fields follows as a corollary of this stability result via well-known facts about convergence of empirical measures (cf. [2]).

The covariance fields derived from potential energies associated with diffusion distances on a Riemannian manifold lead to scale spaces of covariance tensors that provide rich, informative multi-scale data summaries. Here, we only discuss the Euclidean case (cf. [1]), starting with the definition of diffusion distance. Let $K\colon \mathbb{R}^d \times \mathbb{R}^d \times (0, \infty) \to \mathbb{R}^+$ be the heat kernel that is given by

$$K(x, y, t) = \frac{1}{(4\pi t)^{d/2}} \exp\left(-\frac{\|x - y\|^2}{4t}\right). \tag{5}$$

For each $t > 0$, consider the embedding $\kappa_t\colon \mathbb{R}^d \to \mathbb{L}_2(\mathbb{R}^d)$ defined by $x \mapsto K(x, \cdot, t)$, which maps x to the isotropic Gaussian centered at x with variance $\sigma_t^2 = 2t$. The diffusion distance d_t is the metric on $\mathbb{R}^d$ induced by this embedding, up to a multiplicative factor that we introduce to simplify a few expressions. More explicitly, for any $x_1, x_2 \in \mathbb{R}^d$,

$$d_t(x_1, x_2) = \frac{1}{\sqrt{2}} \|\kappa_t(x_1) - \kappa_t(x_2)\|_2\,. \tag{6}$$

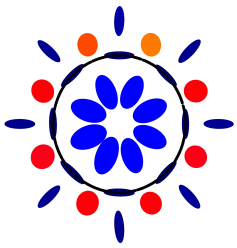


FIGURE 1. Covariance field for 1000 equally spaced points on a circle (illustration courtesy of Diego H. Díaz Martínez).

A calculation shows that $\mathrm{diam}(\mathbb{R}^d, d_t) = 1/(8\pi t)^{d/4}$. For each $t > 0$, let $u_t \colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^+$ be the kernel $u_t(x,y) = d^2_{t/2}(x,y)/2$. The k-covariance tensor and the Fréchet function of a probability measure α on $\mathbb{R}^d$ with respect to u_t will be denoted $\Sigma^k_{\alpha,t}$ and $V_{\alpha,t}$, respectively. This yields a one-parameter family of covariance tensor fields (and Fréchet functions), indexed by $t > 0$, a multi-scale summary of α. (Related 2-tensor fields have been proposed in [2].) Fig. 1 depicts the 2-tensor field at a fixed scale for a dataset comprising 1000 equally spaced points on a circle. The symmetric tensors are plotted as ellipses obtained from their eigen-decompositions. Let α_t be the solution of the heat equation $\partial_t v = \Delta v$ with initial condition α, which mollifies α to a smooth density function. Then, one can show that

$$V_{\alpha,t} = \frac{1}{(4\pi t)^{d/2}} - \alpha_t = \mathrm{diam}^2(\mathbb{R}^d, d_{t/2}) - \alpha_t \,. \tag{7}$$

If $y_1, \ldots, y_n \in \mathbb{R}^d$ are data points and $\alpha = \sum_{i=1}^n \delta_i/n$ is the associated empirical measure, then α_t is the corresponding Gaussian kernel density estimator. Thus, (7) gives an interpretation of such density estimators as Fréchet functions (cf. [1]), integrating density estimators into a hierarchy of "tensorized" moments of α.

REMARKS:

(1) The construction of covariance tensors does not directly apply to the kernel $u(x,y) = d^2_g(x,y)/2$ because it is not necessarily smooth. Nonetheless, it is possible to define covariance tensors if α is absolutely continuous with respect to the Riemannian measure since the singularities of u only occur at $y \in C_x$, the cut locus of x, which has measure zero.
(2) Persistent homology using Fréchet functions or scalar reductions of tensor fields as filtering functions may be used for extracting information about geometric organization of data in a computable manner.
(3) One may define discrete forms of covariance tensors for distributions on the vertex set of a weighted network.
(4) If $\Sigma_t = \Sigma^2_{\alpha,t}$ is everywhere non-singular, then the tensor field Σ_t^{-1} defines a new Riemannian structure on M that may be viewed as the shape of

(M, g, α) at scale $t > 0$. The condition is satisfied, for example, if α is given by a positive density function.

References

[1] D.H. Díaz Martínez, C.H. Lee, P.T. Kim, W. Mio, *Probing the Geometry of Data with Diffusion Fréchet Functions*, Applied and Computational Harmonic Analysis (2018), accepted for publication.

[2] D.H. Díaz Martínez, F. Mémoli, W. Mio, *The Shape of Data and Probability Measures*, arXiv:1509.04632v2.

Optimal matching between curves in a manifold

Marc Arnaudon

(joint work with Alice Le Brigant, Marc Arnaudon and Frédéric Barbaresco)

This talk is concerned with the computation of an optimal matching between two manifold-valued curves. Curves are seen as elements of an infinite-dimensional manifold and compared using a Riemannian metric that is invariant under the action of the reparameterization group. This group induces a quotient structure classically interpreted as the "shape space". We introduce a simple algorithm allowing to compute geodesics of the quotient shape space using a canonical decomposition of a path in the associated principal bundle. We consider the particular case of elastic metrics and show simulations for open curves in the plane, the hyperbolic plane and the sphere.

1. Introduction

A popular way to compare shapes of curves is through a Riemannian framework. The set of curves is seen as an infinite-dimensional manifold on which acts the group of reparameterizations, and is equipped with a Riemannian metric G that is invariant with respect to the action of that group. Here we consider the set of open oriented curves in a Riemannian manifold $(M, \langle \cdot, \cdot \rangle)$ with velocity that never vanishes, i.e. smooth immersions,

$$\mathcal{M} = \mathrm{Imm}([0,1], M) = \{c \in C^\infty([0,1], M) : c'(t) \neq 0 \ \forall t \in [0,1]\}.$$

It is an open submanifold of the Fréchet manifold $C^\infty([0,1], M)$ and its tangent space at a point c is the set of infinitesimal vector fields along the curve c in M,

$$T_c\mathcal{M} = \{w \in C^\infty([0,1], TM) : w(t) \in T_{c(t)}M \ \forall t \in [0,1]\}.$$

A curve c can be reparametrized by right composition $c \circ \varphi$ with an increasing diffeomorphism $\varphi : [0,1] \to [0,1]$, the set of which is denoted by $\mathrm{Diff}^+([0,1])$. We consider the quotient space $\mathcal{S} = \mathcal{M}/\mathrm{Diff}^+([0,1], M)$, interpreted as the space of "shapes" or "unparameterized curves". If we restrict ourselves to elements of $\mathcal{M}$ on which the diffeomorphism group acts freely, then we obtain a principal bundle $\pi : \mathcal{M} \to \mathcal{S}$, the fibers of which are the sets of all the curves that are identical modulo reparameterization, i.e. that project on the same "shape". We denote by $\bar{c} := \pi(c) \in \mathcal{S}$ the shape of a curve $c \in \mathcal{M}$. Any tangent vector $w \in T_c\mathcal{M}$ can

then be decomposed as the sum of a vertical part $w^{ver} \in \mathrm{Ver}_c$, that has an action of reparameterizing the curve without changing its shape, and a horizontal part $w^{hor} \in \mathrm{Hor}_c = (\mathrm{Ver}_c)^{\perp_G}$, G-orthogonal to the fiber,

$$T_c\mathcal{M} \ni w = w^{ver} + w^{hor} \in \mathrm{Ver}_c \oplus \mathrm{Hor}_c,$$
$$\mathrm{Ver}_c = \ker T_c\pi = \{mv := mc'/|c'| : \ m \in C^\infty([0,1],\mathbb{R}), m(0) = m(1) = 0\},$$
$$\mathrm{Hor}_c = \{h \in T_c\mathcal{M} : \ G_c(h, mv) = 0, \ \forall m \in C^\infty([0,1],\mathbb{R}), m(0) = m(1) = 0\}.$$

If we equip $\mathcal{M}$ with a Riemannian metric $G_c : T_c\mathcal{M} \times T_c\mathcal{M} \to \mathbb{R}$, $c \in \mathcal{M}$, that is constant along the fibers, i.e. such that

$$G_{c\circ\varphi}(w \circ \varphi, z \circ \varphi) = G_c(w, z), \quad \forall \varphi \in \mathrm{Diff}^+([0,1]), \tag{1}$$

then there exists a Riemannian metric $\bar{G}$ on the shape space $\mathcal{S}$ such that π is a Riemannian submersion from $(\mathcal{M}, G)$ to $(\mathcal{S}, \bar{G})$, i.e.

$$G_c(w^{hor}, z^{hor}) = \bar{G}_{\pi(c)}\left(T_c\pi(w), T_c\pi(z)\right), \quad \forall w, z \in T_c\mathcal{M}.$$

This expression defines $\bar{G}$ in the sense that it does not depend on the choice of the representatives c, w and z ([4], §29.21). If a geodesic for G has a horizontal initial speed, then its speed vector stays horizontal at all times - we say it is a horizontal geodesic - and projects on a geodesic of the shape space for $\bar{G}$ ([4], §26.12). The distance between two shapes for $\bar{G}$ is given by

$$\bar{d}\left(\overline{c_0}, \overline{c_1}\right) = \inf\left\{ d\left(c_0, c_1 \circ \varphi\right) \mid \varphi \in \mathrm{Diff}^+([0,1]) \right\}.$$

Solving the boundary value problem in the shape space can therefore be achieved either through the construction of horizontal geodesics e.g. by minimizing the horizontal path energy [1],[7], or by incorporating the optimal reparameterization of one of the boundary curves as a parameter in the optimization problem [2],[6],[8]. Here we introduce a simple algorithm that computes the horizontal geodesic linking an initial curve with fixed parameterization c_0 to the closest reparameterization $c_1 \circ \varphi$ of the target curve c_1. The optimal reparameterization φ yields what we will call an *optimal matching* between the curves c_0 and c_1.

2. The optimal matching algorithm

We want to compute the geodesic path $s \mapsto \bar{c}(s)$ between the shapes of two curves c_0 and c_1, that is the projection $\bar{c} = \pi(c_h)$ of the horizontal geodesic $s \mapsto c_h(s)$ - if it exists - linking c_0 to the fiber of c_1 in $\mathcal{M}$. This horizontal path verifies $c_h(0) = c_0$, $c_h(1) \in \pi^{-1}(\overline{c_1})$ and $\partial c_h/\partial s(s) \in \mathrm{Hor}_{c_h(s)}$ for all $s \in [0,1]$. Its end point gives the optimal reparameterization $c_1 \circ \varphi$ of the target curve c_1 with respect to the initial curve c_0, i.e. such that

$$\bar{d}(\overline{c_0}, \overline{c_1}) = d(c_0, c_1 \circ \varphi) = d(c_0, c_h(1)).$$

In all that follows we identify a path of curves $[0,1] \ni s \mapsto c(s) \in \mathcal{M}$ with the function of two variables $[0,1] \times [0,1] \ni (s,t) \mapsto c(s,t) \in M$ and denote by $c_s := \partial c/\partial s$ and $c_t := \partial c/\partial t$ its partial derivatives with respect to s and t. We decompose any path of curves $s \mapsto c(s)$ in $\mathcal{M}$ into a horizontal path

reparameterized by a path of diffeomorphisms, i.e. $c(s) = c^{hor}(s) \circ \varphi(s)$ where $c^{hor}_s(s) \in \mathrm{Hor}_{c^{hor}(s)}$ and $\varphi(s) \in \mathrm{Diff}^+([0,1])$ for all $s \in [0,1]$. That is,

$$c(s,t) = c^{hor}(s, \varphi(s,t)) \quad \forall s,t \in [0,1]. \tag{2}$$

The horizontal and vertical parts of the speed vector of c can be expressed in terms of this decomposition. Indeed, by taking the derivative of (2) with respect to s and t we obtain

$$c_s(s) = c^{hor}_s(s) \circ \varphi(s) + \varphi_s(s) \cdot c^{hor}_t(s) \circ \varphi(s), \tag{3a}$$

$$c_t(s) = \varphi_t(s) \cdot c^{hor}_t(s) \circ \varphi(s), \tag{3b}$$

and so if $v^{hor}(s,t) := c^{hor}_t(s,t)/|c^{hor}_t(s,t)|$ denotes the normalized speed vector of c^{hor}, (3b) gives since $\varphi_t > 0$, $v(s) = v^{hor}(s) \circ \varphi(s)$. We can see that the first term on the right-hand side of Equation (3a) is horizontal. Indeed, for any $m : [0,1] \to C^\infty([0,1], \mathbb{R})$ such that $m(s,0) = m(s,1) = 0$ for all s, since G is reparameterization invariant we have

$$\begin{aligned} G\left(c^{hor}_s(s) \circ \varphi(s),\, m(s) \cdot v(s)\right) &= G\left(c^{hor}_s(s) \circ \varphi(s),\, m(s) \cdot v^{hor}(s) \circ \varphi(s)\right) \\ &= G\left(c^{hor}_s(s),\, m(s) \circ \varphi(s)^{-1} \cdot v^{hor}(s)\right) \\ &= G\left(c^{hor}_s(s),\, \tilde{m}(s) \cdot v^{hor}(s)\right), \end{aligned}$$

with $\tilde{m}(s) = m(s) \circ \varphi(s)^{-1}$. Since $\tilde{m}(s,0) = \tilde{m}(s,1) = 0$ for all s, the vector $\tilde{m}(s) \cdot v^{hor}(s)$ is vertical and its scalar product with the horizontal vector $c^{hor}_s(s)$ vanishes. On the other hand, the second term on the right hand-side of Equation (3a) is vertical, since it can be written

$$\varphi_s(s) \cdot c^{hor}_t \circ \varphi(s) = m(s) \cdot v(s),$$

with $m(s) = |c_t(s)|\varphi_s(s)/\varphi_t(s)$ verifying $m(s,0) = m(s,1) = 0$ for all s. Finally, the vertical and horizontal parts of the speed vector $c_s(s)$ are given by

$$c_s(s)^{ver} = m(s) \cdot v(s) = |c_t(s)|\varphi_s(s)/\varphi_t(s) \cdot v(s), \tag{4a}$$

$$c_s(s)^{hor} = c_s(s) - m(s) \cdot v(s) = c^{hor}_s(s) \circ \varphi(s). \tag{4b}$$

We call c^{hor} the *horizontal part* of the path c with respect to G.

Proposition 1. *The horizontal part of a path of curves c is at most the same length as c*

$$L_G(c^{hor}) \leq L_G(c).$$

Now we will see how the horizontal part of a path of curves can be computed.

Proposition 2 (Horizontal part of a path)**.** *Let $s \mapsto c(s)$ be a path in $\mathcal{M}$. Then its horizontal part is given by $c^{hor}(s,t) = c(s, \varphi(s)^{-1}(t))$, where the path of diffeomorphisms $s \mapsto \varphi(s)$ is solution of the PDE*

$$\varphi_s(s,t) = m(s,t)/|c_t(s,t)| \cdot \varphi_t(s,t), \tag{5}$$

with initial condition $\varphi(0,\cdot) = Id$, and where $m(s) : [0,1] \to \mathbb{R}$, $t \mapsto m(s,t) := |c^{ver}_s(s,t)|$ is the vertical component of $c_s(s)$.

If we take the horizontal part of the geodesic linking two curves c_0 and c_1, we will obtain a horizontal path linking c_0 to the fiber of c_1 which will no longer be a geodesic path. However this path reduces the distance between c_0 and the fiber of c_1, and gives a "better" representative $\tilde{c}_1 = c_1 \circ \varphi(1)$ of the target curve. By computing the geodesic between c_0 and this new representative $\tilde{c}_1$, we are guaranteed to reduce once more the distance to the fiber. The algorithm that we propose simply iterates these two steps.

Data: $c_0, c_1 \in \mathcal{M}$
Result: $\tilde{c}_1$
Set $\tilde{c}_1 \leftarrow c_1$ and Gap $\leftarrow 2 \times$ Threshold;
while *Gap* $>$ *Threshold* **do**
  construct the geodesic $s \mapsto c(s)$ between c_0 and $\tilde{c}_1$;
  compute the horizontal part $s \mapsto c^{hor}(s)$ of c;
  set Gap $\leftarrow \mathrm{dist}_{L^2}\big(c^{hor}(1), \tilde{c}_1\big)$ and $\tilde{c}_1 \leftarrow c^{hor}(1)$;
end

Algorithm 1: Optimal matching.

3. Example : elastic metrics

In this section we consider the particular case of the two-parameter family of elastic metrics, introduced for plane curves by Mio et al. in [5]. We denote by ∇ the Levi-Civita connection of the Riemannian manifold M, and by $\nabla_t w := \nabla_{c_t} w$, $\nabla_t^2 w := \nabla_{c_t}\nabla_{c_t} w$ the first and second order covariant derivatives of a vector field w along a curve c of parameter t. For manifold-valued curves, elastic metrics can be defined for any $c \in T_c\mathcal{M}$ and $w, z \in T_c\mathcal{M}$ by

$$G_c^{a,b}(w,z) = \langle w(0), z(0)\rangle + \int_0^1 \big(a^2\langle \nabla_\ell w^N, \nabla_\ell z^N\rangle + b^2\langle \nabla_\ell w^T, \nabla_\ell z^T\rangle\big)\,\mathrm{d}\ell, \tag{6}$$

where $\mathrm{d}\ell = |c'(t)|\mathrm{d}t$ and $\nabla_\ell = \frac{1}{|c'(t)|}\nabla_t$ respectively denote integration and covariant derivation according to arc length. For th choice of coefficients $a = 1$ and $b = 1/2$, the geodesic equations are easily numerically solved [3] if we adopt the so-called square root velocity representation [6], in which each curve is represented by the pair formed by its starting point and speed vector renormalized by the square root of its norm. Let us characterize the horizontal subspace for $G^{a,b}$, and give the decomposition of a tangent vector.

Proposition 3 (Horizontal part of a vector for an elastic metric)**.** *Let $c \in \mathcal{M}$ be a smooth immersion. A tangent vector $h \in T_c\mathcal{M}$ is horizontal for the elastic metric* (6) *if and only if it verifies the ordinary differential equation*

$$\big((a/b)^2 - 1\big)\,\langle \nabla_t h, \nabla_t v\rangle - \langle \nabla_t^2 h, v\rangle + |c'|^{-1}\langle \nabla_t c', v\rangle\langle \nabla_t h, v\rangle = 0. \tag{7}$$

The vertical and horizontal parts of a tangent vector $w \in T_c\mathcal{M}$ are given by

$$w^{ver} = mv, \quad w^{hor} = w - mv,$$

where the real function $m \in C^\infty([0,1],\mathbb{R})$ verifies $m(0) = m(1) = 0$ and

$$\begin{aligned} &m'' - \langle \nabla_t c'/|c'|, v\rangle m' - (a/b)^2 |\nabla_t v|^2 m \\ &\qquad = \langle \nabla_t \nabla_t w, v\rangle - \left((a/b)^2 - 1\right) \langle \nabla_t w, \nabla_t v\rangle - \langle \nabla_t c'/|c'|, v\rangle \langle \nabla_t w, v\rangle. \end{aligned} \tag{8}$$

This allows us to characterize the horizontal part of a path of curves for $G^{a,b}$.

Proposition 4 (Horizontal part of a path for an elastic metric)**.** *Let $s \mapsto c(s)$ be a path in $\mathcal{M}$. Then its horizontal part is given by $c^{hor}(s,t) = c(s, \varphi(s)^{-1}(t))$, where the path of diffeomorphisms $s \mapsto \varphi(s)$ is solution of the PDE*

$$\varphi_s(s,t) = m(s,t)/|c_t(s,t)| \cdot \varphi_t(s,t), \tag{9}$$

with initial condition $\varphi(0,\cdot) = Id$, and where $m(s) : [0,1] \to \mathbb{R}$, $t \mapsto m(s,t)$ is solution for all s of the ODE

$$\begin{aligned} &m_{tt} - \langle \nabla_t c_t/|c_t|, v\rangle m_t - (a/b)^2 |\nabla_t v|^2 m \\ &\qquad = \langle \nabla_t \nabla_t c_s, v\rangle - \left((a/b)^2 - 1\right) \langle \nabla_t c_s, \nabla_t v\rangle - \langle \nabla_t c_t/|c_t|, v\rangle \langle \nabla_t c_s, v\rangle. \end{aligned} \tag{10}$$

We numerically solve the PDE of the Proposition using the following Algorithm.

Data: path of curves $s \mapsto c(s)$
Result: path of diffeomorphisms $s \mapsto \varphi(s)$
for $k = 1$ **To** n **do**
  estimate the derivative $\varphi_t(\frac{k}{n}, \cdot)$;
  solve ODE (10) using a finite difference method to obtain $m(\frac{k}{n}, \cdot)$;
  set $\varphi_s(\frac{k}{n}, t) \leftarrow m(\frac{k}{n}, t)/|c_t(\frac{k}{n}, t)| \cdot \varphi_t(\frac{k}{n}, t)$ for all t;
  propagate $\varphi(\frac{k+1}{n}, t) \leftarrow \varphi(\frac{k}{n}, t) + \frac{1}{n}\varphi_s(\frac{k}{n}, t)$ for all t;
end

Algorithm 2: Decomposition of a path of curves.

References

[1] M. Bauer, P. Harms and P. W. Michor, Almost local metrics on shape space of hypersurfaces in n-space, *SIAM J. Imaging Sci.*, 5(1) (2012), 244–310.

[2] M. Bauer, M. Bruveris, P. Harms and J. Møller-Andersen, A numerical framework for sobolev metrics on the space of curves, *SIAM J. Imaging Sci.*, 10 (2017), 47–73.

[3] A. Le Brigant, Computing distances and geodesics between manifold-valued curves in the SRV framework, *J. Geom. Mech.*, 9, 2 (2017), 131 – 156.

[4] P. W. Michor, Topics in Differential geometry, in volume 93 of *Graduate Studies in Mathematics*, American Mathematical Society, Providence, RI (2008).

[5] W. Mio, A. Srivastava and S. H. Joshi, On shape of plane elastic curves, *International Journal of Computer Vision*, 73 (2007), 307 – 324.

[6] A. Srivastava, E. Klassen, S. H. Joshi, and I. H. Jermyn, Shape analysis of elastic curves in Euclidean spaces, *IEEE PAMI*, 33, 7 (2011), 1415 – 1428.

[7] A. B. Tumpach and S. C. Preston, Quotient elastic metrics on the manifold of arc-length parameterized plane curves, *J. Geom. Mech.*, 9, 2 (2017), 227 – 256.

[8] Z. Zhang, E. Klassen and A. Srivastava, Phase-amplitude separation and modeling of spherical trajectories (2016), arXiv:1603.07066.

Procrustes Metrics on Covariance Operators and Optimal Coupling of Gaussian Processes

Victor M. Panaretos
(joint work with Valentina Masarotto and Yoav Zemel)

Covariance operators are a key object of study in *functional data analysis*: non-parametric statistics for stochastic processes, where sample paths are viewed as realisations of random elements of some infinite-dimensional separable Hilbert space $\mathcal{H}$. The spectral decomposition of a covariance operator provides the canonical means to quantify the random variation of a process X taking values in $\mathcal{H}$, and to regularise associated inference problems which are typically ill-posed.

In modern applications, it may happen that covariance operators may themselves subject to random variation, usually in situations where several different "populations" of functional data are considered, and there is strong reason to suspect that each population may present different structural characteristics. Each of the K populations will then represent the law of a random element X_k of $\mathcal{H}$, with mean function $\mu_k \in \mathcal{H}$ and covariance operator $\Sigma_k : \mathcal{H} \times \mathcal{H} \to \mathcal{H}$. And, for the purposes of inference, one will observe observe N_k realisations from each population: $\{X_k^i : i = 1, \ldots, N_k; k = 1, \ldots, K\}$. Early contributions in this area were motivated through financial and biophysical applications and led to a surge of methods and theory on *second-order variation* of functional populations. Many of these approaches, though, are intrinsically *linear*: they embed covariance operators in the space of Hilbert-Schmidt operators, and statistical inference is carried out with respect to the corresponding metric. However, covariance operators are fundamentally constrained to obey nonlinear constraints, as they are characterised as the "squares" of Hilbert-Schmidt class operators.

In the multivariate (finite dimensional) literature this problem has been long known, and well-studied, primarily due to its natural connections with important applications such diffusion tensor imaging and shape theory. Consequently, inference for populations of covariance operators has been investigated under a wide variety of possible geometries for the space of covariance matrices. Many of these metrics, however, do not easily generalise to infinite dimensional spaces, since they involve quantities such as determinants, logarithms and inverses.

Pigoli et al. [2] were the first to make important progress in the direction of considering inference for second-order variation in appropriate nonlinear spaces, motivated by the problem of cross-linguistic variation of phonetics in Romance languages. They focussed on the generalisation of the so-called Procrustes reflection-size-and-shape metric (henceforth *Procrustes metric*) and derived some of its basic properties, with a view towards initiating a programme of non-Euclidean analysis of covariance operators. In doing so, they (implicitly or explicitly) generated many further interesting research directions on the geometrical nature of this metric, its statistical interpretation, and the properties of Fréchet means with respect to this metric.

We report on recent work [1] addressing some of these questions, and furthering our understanding of the Procrustes metric and the induced statistical models and procedures, thus placing this new research direction in non-Euclidean statistics on a firm footing. The starting point is a relatively straightforward but quite consequential observation: that the Procrustes metric between two covariance operators on $\mathcal{H}$ coincides with the Wasserstein metric between two centred Gaussian processes on $\mathcal{H}$ endowed with those covariances, respectively. This connection allows us to exploit the wealth of geometrical and analytical properties of optimal transportation, and contribute in two ways. On the one hand, by reviewing and collecting some important aspects of Wasserstein spaces, re-interpreted in the Procrustean context, we elucidate key geometrical (the structure of the tangent bundle and of geodesics), topological (equivalence with the nuclear topology), and computational (descent algorithms with convergence guarantees) aspects of the space of covariances endowed with the Procrustes metric. On the other hand, we establish new results: we show existence, uniqueness, and (uniform over compacta) stability of empirical Fréchet means of covariances with respect to the Procrustes metric, and construct a tangent space principal component analysis via the notion of Gaussian optimal (multi)coupling. We also determine generative statistical models compatible with the Procrustes metric and linking with the problem of warping/registration in functional data analysis. We conclude by formulating a conjecture on the regularity of the Fréchet mean that could have important consequences on statistical inference: given $\Sigma_1, ..., \Sigma_k$ injective covariance operators on $\mathcal{H}$, we conjecture that their Fréchet mean with respect to the Procrustes metric is also injective.

References

[1] Masarotto, V., Panaretos, V.M., & Zemel, Y. (2018). Procrustes Metrics on Covariance Operators and Optimal Transportation of Gaussian Processes. `arXiv:1801.01990`

[2] Pigoli, D., Aston, J.D., Dryden, I.L., Secchi, P. (2014). Distances and Inference for Covariance Operators. *Biometrika, 101 (2): 409–422.*

Curvature concepts in probability

Theo Sturm

Various curvature concepts have been extended from Riemannian geometry to more general spaces – metric spaces or metric measures spaces – and play important roles in probability theory. We briefly discuss the three most important of them.

1. Upper Bounds for the Sectional Curvature

Let us recall the definition of upper curvature bounds in the sense of Alexandrov. For simplicity, here and in the sequel we restrict ourselves to curvature bound 0.

Definition 1. *A geodesic space (X, d) has globally nonpositive curvature iff triangles are more thin than in Euclidean space ("global NPC-space", "Hadamard space").*

Example. For simply connected Riemannian manifolds this is equivalent to nonpositive sectional curvature.

A quite intuitive, characterizing property of these spaces is the Pythagorean inequality $a^2 + b^2 \le c^2$. Of particular importance is the following quadruple characterization which easily is seen to be stable under convergence and immediately passes over to spaces of functions with values in such spaces.

Theorem 1 (Sturm 2003, Berg–Nikolaev 2008)**.** *(X, d) has globally nonpositive curvature iff*

$$d^2(x^1, x^3) + d^2(x^2, x^4) \le \sum_{i=1}^{4} d^2(x^i, x^{i+1}) \qquad (\forall x^1, x^2, x^3, x^4).$$

Example. The L^2-space of maps $f : X \to Y$ from some measure space (X, m) into a NPC space (Y, d) is NPC, too. Here $d^2(f, g) = \int_X d^2\Big(f(x), g(x)\Big)\, dm(x)$.

Theorem 2 (Cartan, Fréchet, Karcher,..., Sturm)**.**

- $\forall \mu \in \mathcal{P}_1(X) : \exists!$ *minimizer of* $z \mapsto \int [d^2(z, x) - d^2(y, x)]\, dm(x)$*, independent of* y*, and denoted by* $b(\mu)$
- $\forall \mu, \nu \in \mathcal{P}_1(X) : \quad d\big(b(\mu), b(\nu)\big) \le W_1(\mu, \nu)$

This (and straightforward generalizations) allows to define conditional expectations, martingales, etc. Of particular importance is the Law of Large Numbers.

Theorem 3 (Sturm 2003)**.** *Assume that $(Y_i)_i$ are bounded iid with distribution $\mu \in \mathcal{P}_1$. Then $\mathbb{P}$-a.s. for $n \to \infty$*

$$\frac{1}{n} \overrightarrow{\sum_{i=1,\dots,n}} Y_i \to b(\mu)$$

Here the 'inductive mean' $s_n = \frac{1}{n} \overrightarrow{\sum}_{i=1,\dots,n} Y_i$ is defined recursively: $s_1 = Y_1$ and s_n is the point $\gamma(\frac{1}{n})$ on the geodesic from $s_{n-1} = \gamma(0)$ to $Y_n = \gamma(1)$.

The convergence is exponentially fast. The rate can be estimated as in the Euclidean case, see [Kei Funano, Osaka J Math 2010].

2. Lower Bounds for the Sectional Curvature

Next we recall the definition of lower curvature bounds in the sense of Alexandrov, again for simplicity assuming that the bound is 0.

Definition 2. *A geodesic space (X, d) has nonnegative curvature iff triangles are more fat than in Euclidean space ("CAT(0) space").*

Example. For Riemannian manifolds this is equivalent to nonnegative sectional curvature.

Again, a quite intuitive, characterizing property is the Pythagorean inequality $a^2 + b^2 \geq c^2$; and a quadruple characterization is of particular importance.

Theorem 4 (Sturm 1999, Lebedeva–Petrunin 2010)**.** *A geodesic space (X, d) has nonpositive curvature iff*

$$\sum_{i=1}^{3} d^2(x^0, x^i) \geq \frac{1}{3} \sum_{1 \leq i < j \leq 3} d^2(x^i, x^j) \qquad (\forall x^0, x^1, x^2, x^3).$$

Here we will discuss two important examples.

- The Wasserstein space $(\mathcal{P}_2(X), W_2)$ which has nonnegative curvature if and only if (X, d) has so.
- The 'Space of spaces' $\{(X, d, m) : \text{metric measure space}\}/\sim$

A *metric measure space* is a triple (X, d, m) consisting of a space X, a complete separable metric d on X and a Borel probability measure on it. Two metric measure spaces are *isomorphic* if there exists a measure preserving isometry between their supports.

The L^2*-distortion distance* between two metric measure spaces (X_0, d_0, m_0) and (X_1, d_1, m_1) is defined as

$$\begin{aligned} &\Delta\Big((X_0, d_0, m_0), (X_1, d_1, m_1)\Big) \\ &\quad = \inf_m \left(\int_{X_0 \times X_1} \int_{X_0 \times X_1} \Big| d_0(x_0, y_0) - d_1(x_1, y_1) \Big|^2 dm(x_0, x_1) dm(y_0, y_1) \right)^{1/2} \end{aligned}$$

where the infimum is taken over all *couplings* of m_0 and m_1.

Theorem 5. *The metric space $(\mathbb{X}, \Delta)$ of isomorphism classes of metric measure spaces is a geodesic space with nonnegative curvature.*

The metric space $(\mathbb{X}, \Delta)$ is not complete. Its completion $\overline{\mathbb{X}}$

- is the space of equivalence classes of *pseudo metric measure spaces* (X, d, m) with X Polish, m Borel, d symmetric, measurable, triangle inequality; without restriction: $X = [0, 1]$, $m = \lambda$;
- is a convex, closed subset of $\mathbb{Y}$ (consisting of triples as above without triangle inequality), isomorphic to

  $$L^2_s([0, 1]^2, \lambda^2)/\mathrm{Inv}([0, 1], \lambda)$$

  with $\mathrm{Inv}([0, 1], \lambda)$ = set of measure preserving maps $\psi : [0, 1] \to [0, 1]$ acting on $L^2_s(\ldots)$ via $\psi^* g(s, t) = g(\psi(s), \psi(t))$.

A dense subset of $(\mathbb{X}, \Delta)$ is given by the set of metric measure spaces consisting of finitely many points, equipped with the uniform measure and a distance function. These spaces are of independent interest; each of them is completely characterized by its distance matrix.

Consider the Hilbert space $\mathsf{M}^{(n)}$ of real-valued symmetric $(n \times n)$-matrices vanishing on the diagonal, equipped with (re-normalized) l_2-norm. The permutation group S_n defines an equivalence relation by

$$f \sim f' \quad \Longleftrightarrow \quad \exists \sigma \in S_n : \, f_{ij} = f'_{\sigma_i \sigma_j} \quad (\forall i, j).$$

Theorem 6. *The quotient space* $\mathbb{M}^n = \mathsf{M}^{(n)} / \sim$ *equipped with the metric*

$$d_{\mathbb{M}^n}(f, f') = \inf\{\|f - \sigma^* f'\|_{\mathsf{M}^{(n)}} : \, \sigma \in S_n\}$$

is a complete geodesic space of nonnegative curvature. The tangent space at f *is given by* $\mathsf{T}_f \mathbb{M}^n = \mathbb{R}^{\frac{n(n-1)}{2}} / \mathsf{Sym}(f)$ *where* $\mathsf{Sym}(f) = \left\{\sigma \in S_n : \, \sigma^* f = f\right\}$ *is the* symmetry group *of* f.

3. Lower Bounds for the Ricci Curvature

Finally, let us briefly mention the powerful concept of synthetic lower Ricci bounds for *metric measure spaces*, formulated as semiconvexity of the Boltzmann entropy $\mathrm{Ent}(\nu|m) = \begin{cases} \int_X \rho \log \rho \, dm & , \text{if } \nu = \rho \cdot m \\ +\infty & , \text{if } \nu \not\ll m \end{cases}$ on the Wasserstein space.

Definition 3 (Sturm 2006, Lott–Villani 2009)**.** *A triple* (X, d, m) *has Ricci curvature* $\geq K$ *iff* $\forall \mu_0, \mu_1 \in \mathcal{P}_2(X) : \, \exists$ W_2*-geodesic* $(\mu_t)_t$ *s.t.* $\forall t \in [0, 1]$*:*

$$Ent(\mu_t|m) \leq (1-t) Ent(\mu_0|m) + t \, Ent(\mu_1|m) - \frac{K}{2} \, t(1-t) \, W_2^2(\mu_0, \mu_1).$$

The success and importance of this synthetic definition arises from the facts that

- it is equivalent to $\mathrm{Ric} \geq K \cdot g$ for Riemannian manifolds
- it is stable under convergence
- it implies in general context most of the geometric and functional inequalities which are known as consequences of lower Ricci bounds in the Riemannian case (e.g. estimates for diameter, eigenvalues, heat kernels etc.).

If the underlying metric measure space is infinitesimally Hilbertian then the heat flow is linear and the following assertions are equivalent

- (X, d, m) has Ricci curvature $\geq K$
- $W_2(P_t \mu, P_t \nu) \leq e^{-Kt} \, W_2(\mu, \nu)$ for all $t > 0$ and all μ, ν.

Dimension Reduction of Tree Data

Huiling Le

The BHV space of phylogenetic trees is a stratified space. In particular, the space $\boldsymbol{T}_{m+2}$ of trees with $m+2$ leaves has $(2m+1)!!$ m-dimensional strata, together with their bounding strata, selected from among the $\binom{M}{m}$ positive orthants in $\mathbb{R}^M$ where $M = 2^{m+2} - m - 4$. The dimensionality and structure of the space, together with the fact that tree data are usually fairly widely spread in the space, make it difficult to directly apply common Euclidean statistical techniques. Methods for constructing a principal geodesic in tree space have recently been developed in [3]. The paper [4] proposes using the locus of weighted Fréchet means to generalise to tree spaces the idea of the kth principal component in Euclidean spaces, while [5] employs tropical geometry to tackle a similar problem.

As for analysing data on manifolds, another possible way to retain some non-Euclidean structure of the tree space to a certain extent, while simplifying the structure of the data, is to use the log map to map data to the tangent cone at their Fréchet mean.

The tangent cone at a point $\boldsymbol{x} \in \boldsymbol{T}_{m+2}$ has a topology and stratification imitating that of T_{m+2} itself in the neighbourhood of $\boldsymbol{x}$. In particular, if $\boldsymbol{x}$ lies in a top-dimensional stratum, the tangent cone at $\boldsymbol{x}$ is the usual tangent space. If $\boldsymbol{x}$ lies in a stratum of co-dimension one, the tangent cone at $\boldsymbol{x}$ is an open book with three pages.

The log map, at $\boldsymbol{x}$, maps any $\boldsymbol{y}$ in the tree space to the initial segment of the geodesic from $\boldsymbol{x}$ to $\boldsymbol{y}$ rescaled to have length equal to the distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ (cf. [1] and [2]). In particular, the log map, at $\boldsymbol{x} \in \sigma$, restricted to the strata that σ bounds, is the 'identity' map. Hence, the image of points in these strata, under the log map, is not distorted.

After projecting tree data to the tangent cone at their Fréchet mean using the log map, we can then further analyse the projected data there by adapting the Euclidean methods appropriately. We use the following simple example to illustrate this idea. Suppose that the Fréchet mean of a set of data in $\boldsymbol{T}_{m+2}$ lies in a co-dimension one stratum σ. One may consider fitting a *principal spider* to the projected data as follows. Assume that the projected data are $\boldsymbol{x}_{0,1}, \cdots, \boldsymbol{x}_{0,k_0} \in \mathbb{R}^{m-1}$, $\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,k_i} \in \tau_i$, where $1 \leqslant i \leqslant 3$, $k_0 \geqslant 0$, $k_i > 0$, $\mathbb{R}^{m-1}$ is the tangent space to σ and τ_i is the ith top-dimension stratum that σ bounds. Then, the principal spider for the data could be defined as the spider formed by

$$\bigcup_{i=1}^{3} \ell_i(\hat{a}, \hat{b}_i),$$

where $\ell(a,b)$ is the intersection of line $a+tb$, in $\mathbb{R}^m$, with $\mathbb{R}^{m-1} \times \mathbb{R}_+$ and

$$(\hat{a}, \hat{b}_1, \hat{b}_2, \hat{b}_3)$$

$$= \underset{a\in\mathbb{R}^{m-1},b_i\in\mathbb{R}^{m-1}\times\mathbb{R}^i_+}{\arg\inf} \left\{ \sum_{j=1}^{k_0} d(x_{0j},a)^2 + \sum_{i=1}^{3}\sum_{j=1}^{k_i} d(x_{ij},\ell_i(a,b_i))^2 \right\}.$$

This procedure can be generalised to higher than two dimensions, for example, 2D principal open books for projected data when their Fréchet mean lies in a higher co-dimension stratum.

However, the above methodology is not the only way of tackling the problems and it raises further issues on how to generalise Euclidean statistical methodology to deal with data on a simple, but general, Euclidean cone, while taking into account features of biological data.

References

[1] D. Barden, H. Le and M. Owen, *Limiting behaviour of Fréchet means in the space of phylogenetic trees*, Annals of the Institute of Statistical Mathematics **70** (2018), 99–129.

[2] D. Barden and H. Le, *The logarithm map, its limits and Fréchet means in orthant spaces*, arxiv.org/pdf/1703.07081.pdf (2017).

[3] T.M.W. Nye, *Principal component analysis in the space of phylogenetic trees*, Ann. Statist. **39** (2011), 2716–2739.

[4] T.M.W. Nye, X. Tang, G. Weyenberg and R. Yoshida, *Principal component analysis and the locus of the Fréchet mean in the space of phylogenetic trees*, Biometrika **104** (2017), 901–922.

[5] R. Yoshida, L. Zhang and X. Zhang, *Tropical principal component analysis and its application to phylogenetics*, arxiv.org/pdf/1710.02682.pdf (2017).

Stratified spaces, fly wings, and multiparameter persistent homology

Ezra Miller

Definition 1. *A* topologically stratified space *is a Hausdorff topological space X that is a disjoint union*

$$X = M_1 \cup \cdots \cup M_\ell$$

*of manifolds (*strata*) M_i such that*

(1) $M_1 \cup \cdots \cup M_k$ is closed in X for all $k \le \ell$; and

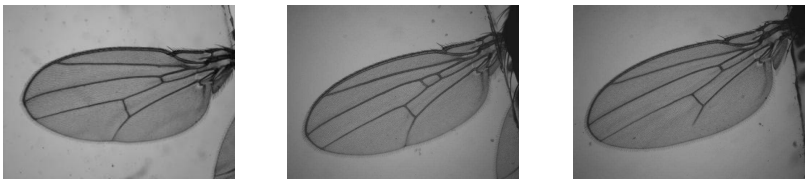
(2) for any points x, y in a stratum M_i there is a homeomorphism $X \xrightarrow{\varphi} X$ with

- *φ stratum-preserving (so $\varphi(M_k) = M_k$ for all k) and*
- *$\varphi(x) = y$.*

This notion of stratified space is more restrictive than could a priori be given—one might omit the homeomorphism condition, for example—but this definition is equivalent to the local structure of the space X being locally trivial along any fixed stratum. That is, the homeomorphism condition implies that at any point $x \in M_i$ the local structure of X looks the same as it does at $y \in M_i$.

Examples of topologically stratified spaces include all Whitney stratified spaces [GM88], in particular all real semi-algebraic varieties (and hence all real and complex algebraic varieties) [Shi97, I.2.10]. Thus polyhedral cell complexes are stratified spaces. Any planar graph embedded in $\mathbb{R}^2$ is also topologically stratified.

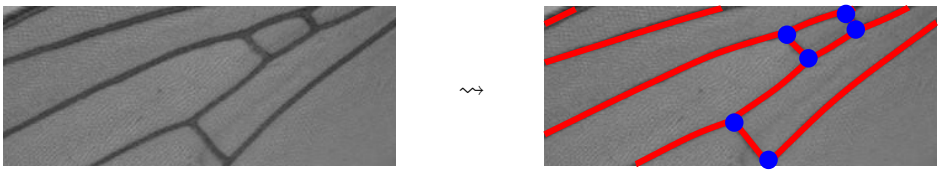
The wings of a fruit fly *Drosophila melanogaster* (images taken from [Mil15])



are such planar embedded graphs. They are naturally stratified, with strata of dimension 0 being vertices of the graph of veins and the strata of dimension 1 being the arcs that constitute the veins themselves. (In the presented dataset, the arcs are encoded as quadratic splines, which are, in particular, algebraic.) The talk presented an approach from geometric statistics to summarize these wing vein graphs in a way that respects the stratification, so as to learn from the stratification, which carries biological meaning.

The motivation for such analysis is that the wings have varying topology, so landmark-based methods do not apply. Note, for example, that the normal wing depicted on the left differs from the middle wing (which has an extra cross-vein) as well as from the wing depicted on the right (one of whose longitudinal veins fails to reach the wing boundary). The biological hypothesis to be tested posits that selecting for continuous variation of a specific sort—for the sake of argument, say selecting for longer wings—results on average in the relevant continuous change (longer wings) but also higher rates of topological variation "in a similar direction". Making this precise requires a summary that incorporates topological as well as geometric information.

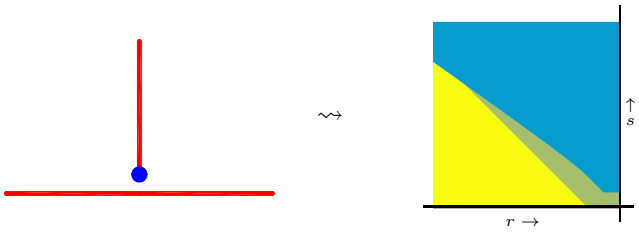
The approach that was discussed applies multiparameter persistent homology. That method was introduced around a decade ago [CZ09] but mostly developed since then in the context of discretely varying parameters. The idea for stratified fly wings is to use two real parameters. One records the radius of balls centered at the vertices (strata of dimension 0), and the other records the width of a thickening of the edges (strata of dimension 1):



(image taken from [Mil15]). The topological space X_r^s for a given radius r and thickness s is obtained from the union of the s-thickened edges by removing the r-expanded vertices. The biparameter persistent homology $\{H_i(X_r^s) \mid r, s \in \mathbb{R}_{\geq 0}\}$ summarizes the stratified fly wing.

To give an idea for what the summary looks like and how it reflects the stratification, a simple toy example was presented [Mil17, Example 1.3]. The zeroth persistent homology for the toy-model "fly wing" in the left-hand image is depicted

in the right-hand image, where each pair of parameters $(r, s) \in \mathbb{R}^2$ is colored according to the dimension of its associated vector space $H_0(X_r^s)$, namely 3, 2, or 1 proceeding up (increasing edge thickness) and to the right (decreasing disk radius):



(images produced by Ashleigh Thomas). The relations that specify the transition from vector spaces of dimension 3 to those of dimension 2 or 1 lie along a real algebraic curve, as do those specifying the transition from dimension 2 to dimension 1.

The point, in the end, is that the embedded planar wing-vein graph is summarized as an integer-valued function on the plane, regardless of the topology of the graph. These summaries lend themselves to ordinary linear statistical methods.

References

[CZ09] Gunnar Carlsson and Afra Zomorodian, *The theory of multidimensional persistence*, Discrete and Computational Geometry **42** (2009), 71–93.

[GM88] M. Goresky and R. MacPherson, *Stratified Morse theory*, Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)], 14, Springer-Verlag, Berlin, 1988.

[Mil15] Ezra Miller, *Fruit flies and moduli: interactions between biology and mathematics*, Notices of the American Math. Society **62** (2015), no. 10, 1178–1184. doi:10.1090/noti1290 arXiv:q-bio.QM/1508.05381

[Mil17] Ezra Miller, *Data structures for real multiparameter persistence modules*, 107 pages. arXiv:math.AT/1709.08155v1

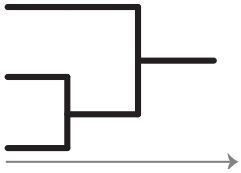
[Shi97] Masahiro Shiota, *Geometry of Subanalytic and Semialgebraic Sets*, Progress in Mathematics, vol. 150, Springer, New York, 1997. doi:10.1007/978-1-4612-2008-4

Stable signatures for dynamic metric spaces via persistent homology.

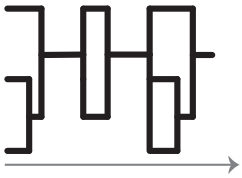
Facundo Mémoli

(joint work with Woojin Kim)

Given data as a *static* finite metric space (X, d_X), hierarchical clustering method finds a hierarchical family of partitions that captures some multi-scale features present in the dataset. These hierarchical families of partitions are called *dendrograms* (see figure on the left) and from a graph theoretic perspective, they are planar, hence their visualization is straightforward.



We now turn our attention to a problem of clustering of *dynamic* data. We model dynamic datasets as time varying finite metric spaces and study a simple generalization of the notion of dendrogram which we call *formigram* (see figure on the right)– a combination of the words *formicarium*[1]and *diagram*. Whereas dendrograms are useful for modeling situations when data points *aggregate* along a certain scale parameter, formigrams are better suited for representing phenomena when data points may also separate or *disband* and then regroup at different parameter values. One motivation for considering this scenario comes from the study and characterization of *flocking/swarming/herding* behavior of animals, convoys, moving clusters, or mobile groups (a list of numerous references is in the full paper [13]).



In contrast to dendrograms, formigrams are not always planar, so more simplification is desirable in order to easily visualize the information they contain. We do this by associating zigzag persistent homology barcodes/diagrams [3] to formigrams. We prove that the resulting signatures turn out to be (1) stable to perturbations of the input dynamic metric space and (2) still informative. The so called Single Linkage Hierarchical Clustering method [10] produces dendrograms from finite metric spaces in a stable manner: namely, if the input static datasets are close in the Gromov-Hausdorff sense, then the output dendrograms will also be close [4]. This result is further generalized for higher dimensional homological features [5]. In this paper we study to what extent one can export similar results to the case of dynamic datasets.

Overview of our results

In what follows, we omit some definitions due to a limit of length of this paper, which can be found in the full version [13]. Throughout this paper X and Y are non-empty finite sets. We denote the set of real numbers and the set of non-negative real numbers by $\mathbf{R}$ and $\mathbf{R}_+$, respectively. By a *dynamic metric spaces (DMSs) on a set* X, we mean a pair $\gamma_X = (X, d_X(\cdot))$ where $d_X(\cdot) : \mathbf{R} \times X \times X \to \mathbf{R}_+$ satisfying the following conditions: (1) for each $t \in \mathbf{R}$, the map $d_X(t) : X \times X \to \mathbf{R}_+$ is a pseudo-metric on X, (2) for any fixed $x, x' \in X$, the map $t \mapsto d_X(t)(x, x')$ is continuous, (3) there exists $t_0 \in \mathbf{R}$ such that $d_X(t_0)$ is a metric on X (in order not to have redundant points in X).

Recall that by definition a *correspondence* $R \subset X \times Y$ is mapped onto X and Y via the canonical projections to the first and second coordinates, respectively. We metrize the collection of all DMSs as follows. The structure of this metric is a hybrid between the Gromov-Hausdorff distance and the interleaving distance [2, 6] for Reeb graphs [8].

Definition 1 (Interleaving distance between DMSs)**.** *Let γ_X, γ_Y be DMSs on X and Y respectively, and $\varepsilon \geq 0$. We say that γ_X and γ_Y are* ε-interleaved *if there*

[1]A formicarium is an enclosure for keeping ants under semi-natural conditions [12].

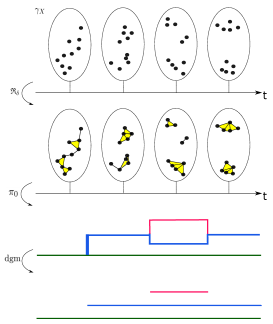


FIGURE 1. This illustrates a process through which a DMS γ_X (the dynamic point cloud of the first row) is converted into a barcode summarizing its clustering information (the last row): For a fixed $\delta \geq 0$ applying the *Rips functor* R_δ to γ_X yields a zigzag simplicial filtration (the second row). Then we apply the *connected component functor* π_0 to the zigzag simplicial filtration, obtaining a *formigram* (the third row). Via some algebraic process, one finally obtains the barcode (the last row). See [13] for details.

exists a correspondence $R \subset X \times Y$ *such that* $\forall (x,y),(x',y') \in R$, $\forall t \in \mathbf{R}$,

$$\min_{s\in[t]^\varepsilon} d_Y(s)(y,y') \leq d_X(t)(x,x') \quad \textit{and} \quad \min_{s\in[t]^\varepsilon} d_X(s)(x,x') \leq d_Y(t)(y,y').$$

The interleaving distance $d_{\mathrm{I}}^{\mathrm{dyn}}(\gamma_X,\gamma_Y)$ *between* γ_X *and* γ_Y *is defined by the infimum* $\varepsilon \geq 0$ *for which* γ_X *and* γ_Y *are* ε*-interleaved. If* γ_X *and* γ_Y *are not* ε*-interleaved for any* $\varepsilon \geq 0$*, declare* $d_{\mathrm{I}}^{\mathrm{dyn}}(\gamma_X,\gamma_Y) = +\infty$.

Given a DMS γ_X (satisfying a mild *tameness* condition [13, Definition 2.4]), for each non-negative integer k and connectivity parameter $\delta \geq 0$, we associate it with the zigzag persistent homology $\mathrm{H}_k(\mathfrak{R}_\delta(\gamma_X))$, where $\mathfrak{R}_\delta(\gamma_X)$ is the Rips zigzag filtration derived from γ_X (see Figure 1 and [13, Section D] for details).

The following stability result tells us that the assignment $\gamma_X \mapsto \mathrm{dgm}(\mathrm{H}_k(\mathfrak{R}_\delta(\gamma_X)))$ of zigzag persistence diagrams to DMSs when $k = 0$ is stable in terms of $d_{\mathrm{I}}^{\mathrm{dyn}}$ and the usual *bottleneck distance* between barcodes/persistence diagrams [7]:

Theorem 1 (Stability theorem)**.** *For any two tame DMSs* γ_X *and* γ_Y*, and any* $\delta \geq 0$*:*

$$d_{\mathrm{B}}\big(\mathrm{dgm}(\mathrm{H}_0(\mathfrak{R}_\delta(\gamma_X))), \mathrm{dgm}(\mathrm{H}_0(\mathfrak{R}_\delta(\gamma_Y)))\big) \leq 2\, d_{\mathrm{I}}^{\mathrm{dyn}}(\gamma_X,\gamma_Y).$$

We remark that the lower bound can be computed in polynomial time [3, 9, 11].

In the way to prove Theorem 1, we introduce (a) the notion of *formigrams*, both as a summary (akin to dendrograms) of the dynamic clustering behavior of a DMS and as an object whose algebraic interpretation (via its zigzag persistence barcode) is parsimonious (see the last two rows in Figure 1); (b) a notion of distance $d_{\mathrm{I}}^{\mathrm{F}}$ between formigrams which mediates between $d_{\mathrm{I}}^{\mathrm{dyn}}$ and the bottleneck distance between barcodes; and motivated by practical applications (c) a smoothing operation on formigrams. In particular, in order to prove Theorem 1 we make use of recent stability results for zigzag persistence due to Botnan and Lesnick [1].

Theorem 1 above together with the available results for static finite metric spaces suggests that such stability might extend beyond 0-dimensional homology. Interestingly, there is a family of counter-examples indicating that stability, as expressed by Theorem 1, is a phenomenon which seems to be essentially tied to

clustering (i.e. H_0) information. We refer the reader to [13, Theorem 1.3, Figure 2] for details.

Acknowledgement

This work was partially supported by NSF grants IIS-1422400 and CCF-1526513.

References

[1] Magnus Bakke Botnan and Michael Lesnick. Algebraic stability of persistence modules. *arXiv preprint arXiv:1604.00655*, 2016.

[2] Peter Bubenik and Jonathan A Scott. Categorification of persistent homology. *Discrete & Computational Geometry*, 51(3):600–627, 2014.

[3] Gunnar Carlsson and Vin De Silva. Zigzag persistence. *Foundations of computational mathematics*, 10(4):367–405, 2010.

[4] Gunnar Carlsson and Facundo Mémoli. Characterization, stability and convergence of hierarchical clustering methods. *Journal of Machine Learning Research*, 11:1425–1470, 2010.

[5] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot. Gromov-Hausdorff stable signatures for shapes using persistence. In *Proc. of SGP*, 2009.

[6] Frédéric Chazal, David Cohen-Steiner, Marc Glisse, Leonidas J. Guibas, and Steve Oudot. Proximity of persistence modules and their diagrams. In *Proc. 25th ACM Sympos. on Comput. Geom.*, pages 237–246, 2009.

[7] David Cohen-Steiner, Herbert Edelsbrunner, and John Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007.

[8] Vin De Silva, Elizabeth Munch, and Amit Patel. Categorified Reeb graphs. *Discrete & Computational Geometry*, 55(4):854–906, 2016.

[9] Herbert Edelsbrunner and John Harer. *Computational Topology - an Introduction*. American Mathematical Society, 2010.

[10] N. Jardine and R. Sibson. *Mathematical taxonomy*. John Wiley & Sons Ltd., London, 1971. Wiley Series in Probability and Mathematical Statistics.

[11] Nikola Milosavljević, Dmitriy Morozov, and Primoz Skraba. Zigzag persistent homology in matrix multiplication time. In *Proceedings of the Twenty-seventh Annual Symposium on Computational Geometry*, SoCG '11, pages 216–225, New York, NY, USA, 2011. ACM. URL: `http://doi.acm.org/10.1145/1998196.1998229`, `doi:10.1145/1998196.1998229`.

[12] Wikipedia. Formicarium — Wikipedia - the free encyclopedia. `https://en.wikipedia.org/wiki/Formicarium`, 2017. [Online; accessed 03-June-2017].

[13] Woojin Kim and Facundo Memoli. Stable Signatures for Dynamic Metric Spaces via Zigzag Persistent Homology. arXiv preprint arXiv:1712.04064, 2017.

[14] Zane Smith, Woojin Kim and Facundo Memoli. Computational examples about flocking, formigrams, and zigzag barcodes. `https://research.math.osu.edu/networks/formigrams`, 2017.

Scaling-rotation statistics for symmetric positive-definite matrices

Sungkyu Jung

(joint work with Armin Schwartzman, David Groisser and Brian Rooks)

We discussed a geometric structure on $\mathrm{Sym}^+(p)$, the set of $p \times p$ symmetric positive-definite (SPD) matrices, $p \geq 2$. Eigen-decomposition determines both a stratification of $\mathrm{Sym}^+(p)$, defined by eigenvalue multiplicities, and fibers of the eigen-composition map $F : SO(p) \times \mathrm{Diag}^+(p) \to \mathrm{Sym}^+(p)$, $F((U, D)) = UDU^{-1}$ [1]. This leads to the notion of scaling-rotation distance [2], a measure of the minimal amount of scaling and rotation needed to transform an SPD matrix, X, into

another, Y, by a smooth curve in $\mathrm{Sym}^+(p)$. A systematic characterization and analysis of minimal smooth scaling- rotation (MSSR) curves, images in $\mathrm{Sym}^+(p)$ of minimal-length geodesics connecting two fibers in $SO(p)\times\mathrm{Diag}^+(p)$, were given. The length of such a geodesic connecting the fibers over X and Y is what we define to be the scaling-rotation distance from X to Y.

This scaling-rotation geometric framework coincides with identifying $\mathrm{Sym}^+(p)$ with the quotient space $SO(p) \times \mathrm{Diag}^+(p)\backslash \sim$, where the equivalence relation $\sim$ is given by F; $(U_1, D_1) \sim (U_2, D_2)$ if and only if $F((U_1, D_1) = F((U_2, D_2))$. A lift of the MSSR curve between X and Y is in fact the minimal-length path between a lift of X and that of Y among all continuous paths between them, which turns out to be a geodesic. Allowing for the path in $SO(p) \times \mathrm{Diag}^+(p)$ discontinuity within fibers results in a minimal-length piecewise-smooth scaling-rotation curve in $\mathrm{Sym}^+(p)$. The length of such a curve gives a notion of scaling-rotation metric ρ, and $(\mathrm{Sym}^+(p), \rho)$ is a metric space.

In an application area of diffusion tensor imaging, a tensor is defined as a 3×3 SPD matrix M, and often visualized by the corresponding ellipsoid, whose surface coordinates $x \in \Re^3$ satisfy $x^T M^{-1} x = 1$. The scaling-rotation geometric framework provides a means of smooth interpolation between two SPD matrices, or tensors, by an MSSR curve between X and Y. When the multiset of eigenvalues of X coincides with the multiset of eigenvalues of Y, and if the eigenvalues are distinct, and the difference between eigenvalues are sufficiently large, then the scaling-rotation interpolation is of a pure rotation of constant angular velocity. This prevents "swelling" of tensor (ellipsoid) when interpolating two "skinny" tensors. As a comparison, suppose that the interpolation is given by the shortest geodesic between X and Y, where the geodesic is defined under the affine-invariant Riemannian inner product on $\mathrm{Sym}^+(p)$. Such an interpolation is of the form $f_{AI}(t) = X^{1/2} \exp(t \log(X^{-1/2} Y X^{-1/2})) X^{1/2}$, where exp and log are matrix exponential and its inverse. If set of the eigenvector matrices of X is disjoint from the set of eigenvector matrices of Y, then the angular velocity of the eigenvector matrix of f_{AI} is not constant. Data examples, omitted from this abstract, confirm this. Will the advantage of scaling-rotation framework remain true when the smoothness requirement is relaxed to the piecewise-smoothness? The answer is yes, if the minimal piecewise-smooth curve is indeed smooth. A formal algebraic analysis on the conditions on X, Y, for which both MSSR curves is shortest among all piecewise-smooth curves, is an open problem.

References

[1] David Groisser, Sungkyu Jung, and Armin Schwartzman. Geometric foundations for scaling-rotation statistics on symmetric positive-definite matrices: minimal smooth scaling-rotation curves in low dimensions. *Electron. J. Stat.* 11:1092-1159, 2017.

[2] Sungkyu Jung, Armin Schwartzman, and David Groisser. Scaling-rotation distance and interpolation of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* 36 1180-1201.

S-reps and Their Statistics

Stephen Pizer

S-reps are a rich geometric representation of anatomic objects that are suited for statistics of shape analysis. They are skeletal models that are quasi-medial and are stable so that there is correspondence of the primitives, called spoke vectors, across objects in an anatomic population. An example of an s-rep for a hippocampus is shown in Fig. 1 – the spokes are continuous but are shown densely sampled.

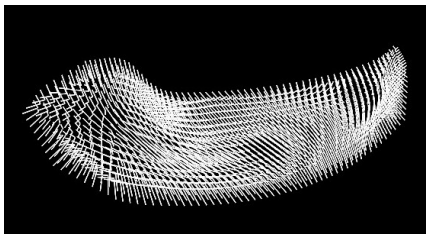


Figure 1. An s-rep for a hippocampus.

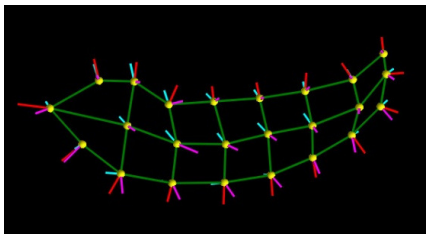


Figure 2. An s-rep as represented in the computer.

The s-rep captures the important shape property of object boundary direction U as it varies along the boundary, the relevant shape property of object width r (actually half-width) as it varies along the object, as well as positional information along the object. Thereby it provides improved statistical performance, as compared to other object representations, as shown in a variety of empirical studies as to its application to classification and provision of a prior for segmentation from 3D images.

S-reps can be produced for any amount of essential branching and any topology. However, we have focused on objects in 3D with no essential branching and with either spherical topology and a slabular geometry (the three major axes have notably different lengths) or a generalized cylinder topology (with curvilinear center curve dilated into curved cylinder with an ϵ-radius). Very many objects take one of these two forms and have been successfully represented using s-reps, for example, Slabular: the hippocampi, lateral ventricles, putamen, cerebral cortex (even though the cortex is heavily folded), bladder, prostate, heart, lung, muscles; Generalized cylinder: various arteries, the rectum.

As illustrated in Fig. 3, the unbranching s-rep skeleton in 2D is a folded curve with circular topology such that the two sides of the curve are pasted together. In 3D in its slabular form the skeleton can be understood as formed by two sheets of plastic wrap pasted together and connected along a fold curve. In its generalized cylinder form the skeleton is formed by a curved cylinder with ϵ-radius

"Spoke" vectors, going from each point on the skeleton to the object boundary form the s-rep. The s-rep is fit to an object boundary given as data in a way such that

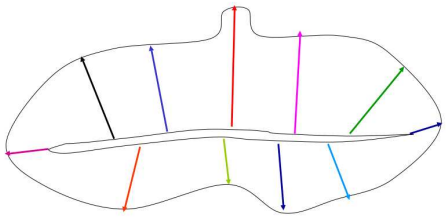


FIGURE 3. An s-rep in 2D. In its mathematical form the spoke vectors are continuous along the skeleton. The two sides of the skeleton follow the same positional locus.

(1) the spokes fill the object interior
(2) spokes do not cross each other
(3) spokes from the fold go to crest points on the boundary, where the crest exists
(4) the spokes from points on the skeleton that are the same in ambient space are close to equal in length
(5) spokes intersect the boundary nearly orthogonally
(6) the swing of spokes follows a radial shape operator [1] that, in analogy to the well-known shape operator describing the swing of normal on the boundary, describes the swing of the spokes on the skeleton.

The approximate nature of the fit of the spoke ends to the boundary and of conditions 4 and 5 makes it possible for the branching topology to be given as a precondition and the fit to be stable and rather tight to the boundaries, and this suits the s-rep for statistics, unlike the medial form of skeletal models in which its bushy skeleton, highly sensitive to boundary noise, makes statistical analysis extremely hard to achieve.

For computer representation sampled spokes of the s-rep are used, and a mathematically careful means of spoke interpolation [2] using the aforementioned radial shape operator yields the spokes at any desired density that is used in fitting to boundary data at all the interpolated spoke ends. The fitting to an input boundary requires the user only to provide the number of spokes along the long axis of the object and that number across the 2nd widest axis of the object. Given that, the regular spacing of the spokes is determined in a way that produces correspondence across a training sample of s-reps used in statistics.

Each spoke in a computer-represented (discrete) s-rep consists of a length r, a spoke direction $\mathbf{U}$, and a skeletal point $\underline{\mathbf{p}}$. The (length, direction) form of the representation yields more direct characterization of the desired object features, produces more well-behaved geodesics on the abstract manifold on which an s-rep lives, and empirically has been shown to produce better statistical analysis than a Euclidean representation of the spokes.

The lengths of the n spokes of an s-rep live abstractly on $\mathbb{R}^n$ (for the logs of the n spoke lengths), and the directions $\mathbf{U}$ of those spokes live on $(S^2)^n$. The tuple of n spoke positions on the skeleton are understood, according to [3], after centering each skeleton's n $\underline{\mathbf{p}}$ values on its center of mass, as a spatial scale, computed as the Euclidean norm γ of the centered points, and a point on S3n-4. This representation of a tuple of spatial points has been found empirically to yield better statistical performance than the Euclidean representation. After taking the log of the spatial scale, the spatial scale lives on $\mathbb{R}^1$. Thus an s-rep is understood to live on the Cartesian product of a polysphere $(S^2)^n \times S^{3n-4}$ and $\mathbb{R}^n$.

Probability estimation analysis of s-reps are accomplished by the methods described in the abstract by Marron in this Proceedings. Here the method for classification of s-reps will be sketched. The s-reps in the two training classes are pooled, and a polar system for principal nested spheres (PNS) is computed from that pool. Then each training s-rep is transformed into Euclideanized coordinates by compiling the PNS scores for each dimension reduction. The tuples of these Euclideanized coordinates is therefore analyzed by the Euclidean method, Distance-Weighted Discrimination (DWD) [5] to produce a separation direction in Euclidean space. The Euclideanized training cases are then projected onto this direction to form a histogram for each class. These histograms are then used to compute the class probabilities for a new s-rep after it has been Euclideanized using the polar system derived in training. This approach is also suited to other representations living on the Cartesian product of a polysphere and a Euclidean space.

Classification into control and diseased classes of a number of brain structures using this method has yielded superior results over other object representations and their associated statistical analysis techniques [4]. Likewise, high quality segmentations in 3D of a number of anatomic structures from a number of 3D medical image types have been produced by a variant on posterior optimization in which the prior (anatomic shape statistics) is computed based on s-reps [2].

Future work will include analyzing the polysphere statistics using PPCA [6], doing multiscale analysis of s-reps according to the ideas of Mio (see this Proceedings), producing an s-rep variant that can handle 3D objects with a cusp, such as the caudate nucleus [4], evaluating a variety of classifications of brain structures [4], further development of the s-rep for generalized cylinders, and extending the s-reps to objects with other topologies or with essential branches.

References

[1] J. Damon. Determining the Geometry of Boundaries of Objects from Medial Data. *Int J Comput Vision* 63: 45–64, (2005).
[2] J. Vicory.
[3] D. G. Kendall. The Diffusion of Shape. *Advances in Applied Probability*, 9, 3:428–430 (1977)
[4] J. Hong.

[5] J. S. Marron, M. J. Todd and J. Ahn. Distance-Weighted Discrimination. *Journal of the American Statistical Association*, 102, 480: 1267–1271, (2007).
[6] B. Eltzner, S. Jung, S. F. Huckemann. Dimension Reduction on Polyspheres with Application to Skeletal Representations. *Geometric Science of Information 2015 proceedings*, 22–29, (2015).

On the Geometry of Latent Variable Models

Søren Hauberg

Latent variable models (LVMs) describe the distribution of data $\mathbf{y} \in \mathcal{Y} = \mathbb{R}^D$ through a low-dimensional random variable $\mathbf{x} \in \mathcal{X} = \mathbb{R}^d, (d \ll D)$ and a (generally nonlinear) stochastic mapping $f : \mathcal{X} \to \mathcal{Y}$. Here we discuss the random Riemannian geometry induced by this stochastic mapping. The presented results was first stated in [1, 2].

To make the discussion explicit, we consider a Gaussian Process (GP) LVM [3] where f has component-wise conditionally independent Gaussian process entries,

$$f_i(\mathbf{x}) \sim \mathcal{GP}(m_i(\mathbf{x}), k(\mathbf{x}, \mathbf{x}')), \qquad \forall i = 1, \dots, D. \tag{1}$$

Here m_i and k are the mean and covariance functions of the i^{th} GP. Note that we assume the same covariance function across all dimensions as this simplify future calculations. The key presented results holds regardless of this simplification.

Assuming k is sufficiently smooth covariance then the image of a sample from f is a smooth d-dimensional immersed manifold. Note that this manifold is only locally diffeomorphic to d-dimensional Euclidean space, and it may globally self-intersect. It is then natural to consider the pull-back metric $\mathbf{M} = \mathbf{J}^\top \mathbf{J}$ over $\mathcal{X}$, where $\mathbf{J} \in \mathbb{R}^{D \times d}$ is the Jacobian of f. This defines a Riemannian metric over $\mathcal{X}$. Since f is stochastic, $\mathbf{M}$ is a stochastic object as well.

Since Gaussian variables are closed under differentiation, then $\mathbf{J}$ follows a GP,

$$\mathbf{J} \sim \prod_{j=1}^{D} \mathcal{N}(\mu(j,:), \Sigma) = \prod_{j=1}^{D} \mathcal{N}(\partial \mathbf{K}_{\mathbf{x},*}^\top \tilde{\mathbf{K}}_{\mathbf{x},\mathbf{x}}^{-1} \mathbf{Y}_{:,j}, \partial^2 \mathbf{K}_{*,*} - \partial \mathbf{K}_{*,\mathbf{x}}^\top \mathbf{K}_{\mathbf{x},\mathbf{x}}^{-1} \partial \mathbf{K}_{*,\mathbf{x}}), \tag{2}$$

where we use standard notation for GPs [4]. It then follows that $\mathbf{M}$ at a given point is governed by a non-central Wishart distribution [5]

$$\mathbf{M} \sim \mathcal{W}_d(D, \Sigma, \mathbb{E}[\mathbf{J}]^\top \mathbb{E}[\mathbf{J}]). \tag{3}$$

The entire metric by definition follows a generalized Wishart process [6].

Since the metric is a stochastic variable, we cannot apply standard Riemannian geometry to understand the space $\mathcal{X}$ (e.g. curvature is stochastic, geodesics are solutions to a stochastic differential equation, etc.). We can, however, inspect the leading moments of the metric

$$\mathbb{E}[\mathbf{M}] = \mathbb{E}[\mathbf{J}^\top \mathbf{J}] = \mathbb{E}[\mathbf{J}]^\top \mathbb{E}[\mathbf{J}] + D\,\Sigma \qquad = \mathcal{O}(D) \tag{4}$$

$$\text{var}[M_{ij}] = D(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}) + \mu_j^\top \Sigma \mu_j + \mu_i^\top \Sigma \mu_i \qquad = \mathcal{O}(D) \tag{5}$$

which we see both grow linearly with the dimension of $\mathcal{Y}$. This motivate the question as to how the pull-back metric behaves in high dimensions, $D \to \infty$. To ensure that the inner product of $\mathcal{Y}$ converges to the usual L^2 inner product in the limit $D \to \infty$ we let

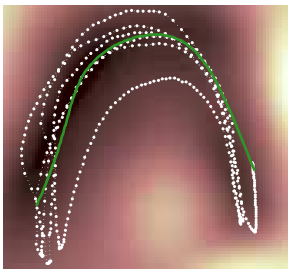
$$\langle \mathbf{a}, \mathbf{b} \rangle_{\mathcal{Y}} = \frac{1}{D} \sum_{i=1}^{D} a_i b_i \xrightarrow[D \to \infty]{} \int a_t b_t \mathrm{d}t. \tag{6}$$

Then the natural pull-back becomes $\tilde{\mathbf{M}} = \frac{1}{D} \mathbf{J}^\top \mathbf{J}$, which has moments

$$\mathbb{E}[\tilde{\mathbf{M}}] = \mathbb{E}\left[\frac{1}{D} \mathbf{J}^\top \mathbf{J}\right] = \frac{1}{D} \mathbb{E}[\mathbf{J}]^\top \mathbb{E}[\mathbf{J}] + \Sigma \qquad = \mathcal{O}(1) \tag{7}$$

$$\mathrm{var}[\tilde{M}_{ij}] = \frac{1}{D}(\Sigma_{ij}^2 + \Sigma_{ii}\Sigma_{jj}) + \frac{1}{D^2} \mu_j^\top \Sigma \mu_j + \frac{1}{D^2} \mu_i^\top \Sigma \mu_i \quad = \mathcal{O}\left(\frac{1}{D}\right) \tag{8}$$

In the limit $D \to \infty$ we, thus, see that the variance vanishes and the metric becomes fully deterministic even if the underlying manifold is a stochastic object.



Implications and Extensions. This simple-to-prove result is rather surprising: even if we only have stochastic information about the underlying data manifold, its metric is deterministic. Furthermore, from Eq. 7 we see that this deterministic metric correspond to the (usual) pull-back metric of the mean f plus an additional term capturing the uncertainty of the manifold. This imply that the metric is large in regions of low data density (where the manifold is uncertain), and consequently, that geodesics will tend to avoid such regions. One such example is shown in the figure. Here human motion capture data $\mathbf{y}$ is used to estimate a two-dimensional manifold [1]. In the figure white points correspond to low-dimensional representations of the data, the green curve is an example geodesic computed under the expected metric, and the background color is proportional to the volume measure induced by the expected metric. We see that the metric is "larger" in regions of low data density and that geodesics consequently follow the structure of the data. The latter is a useful property when analyzing real data as distance-based data distribution will adapt well to the data [2].

From a practical point of view, geodesics can be computed in $\mathcal{X}$ by numerically solving the usual system of ordinary differential equations under the expected metric. The solution will be a curve in $\mathcal{X}$, which correspond to a GP in $\mathcal{Y}$. As such, geodesics remain stochastic objects, but they can be determined by solving a set of deterministic equations.

The presented derivations rely on the dimensions of $f(\mathcal{X})$ being conditionally independent, which is a common assumption. It can be eased upon: if the dimensions are (imperfectly) correlated, then the variance will still decrease, albeit at a slower rate than D^{-1}. Consequently, as a general rule of thumb, *the stochastic*

pull-back metric of an uncertain manifold immersed in a high-dimensional space is well approximated by the (deterministic) expected metric.

Acknowledgments. SH was supported by a research grant (15334) from VILLUM FONDEN. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 757360).

References

[1] A. Tosi, S. Hauberg, A. Vellido, and N. Lawrence, *Metrics for Probabilistic Geometries*, In The Conference on Uncertainty in Artificial Intelligence (2014).
[2] G. Arvanitidis, LK. Hansen, and S. Hauberg, *Latent Space Oddity: on the Curvature of Deep Generative Models*, In the International Conference on Learning Representations (2018).
[3] N. Lawrence, *Probabilistic non-linear principal component analysis with Gaussian process latent variable models*, Journal of machine learning research 6. Nov (2005): 1783-1816.
[4] CE. Rasmussen, and CKI. Williams, *Gaussian Processes for Machine Learning*, University Press Group Limited (2006).
[5] RJ. Muirhead, *Aspects of Multivariate Statistical Theory*, John Wiley & Sons (2005).
[6] AG. Wilson, and Z. Ghahramani, *Generalised Wishart Processes*, In The Conference on Uncertainty in Artificial Intelligence (2011).

Focus Group Discussions

Do Tran, John Kent, Ruriko Yoshida, Sarang Yoshi, Stefan Anell

There were five discussion groups on topics spanning the whole range of the workshop.

1. Dirty CTLs – Reporter: Do Tran

The topic of this group is mostly of a theoretical nature, namely the non-standard asymptotic theory of smeary and sticky means. The discussion in this group mostly centered on some concrete questions related to the talk by Xavier Pennec. In this talk it was pointed out, that the variance of the sample Fréchet mean on a manifold acquires a curvature dependent term

$$\mathbf{E}\left[\log_{\bar{x}}(\bar{x}_n)^a \log_{\bar{x}}(\bar{x}_n)^b\right] = \frac{1}{n}\mathfrak{M}_2^{ab} + \frac{1}{3n}\left(R^a_{ced}\mathfrak{M}_2^{be} + R^b_{ced}\mathfrak{M}_2^{ae}\right)\mathfrak{M}_2^{cd} + O(\epsilon^3).$$

Since these terms of order $1/n$ can balance out for negative curvature, it was concluded, that the additional term might be an indicator of stickiness. To make such a conjecture more precise, it was proposed to define a sequence of probability measures on a sequence of manifolds $\mathcal{M}_m$, such that the curvature at the population mean diverges to negative infinity and investigate, under which conditions the limits $n \to \infty$ and $m \to \infty$ commute and what the consequences are.

Furthermore, the sample mean acquires a bias term

$$\mathbf{E}\left[\log_{\bar{x}}(\bar{x}_n)^a\right] = \frac{1}{24n}(2\nabla_b R^a_{dce} + \nabla^a R_{cebd})(\mathfrak{M}_2^n)^{bc}(\mathfrak{M}_2^n)^{de} + O(\epsilon^5).$$

However, the interpretation of this bias term is not immediately clear. To achieve a clearer interpretation of the bias term, simulations in simple toy models and a reformulation in terms of the sectional curvature tensor were proposed.

Aside, the question was raised, whether asymptotic statistics should be replaced by considerations towards non-asymptotic confidence sets as presented by Thomas Hotz. This approach was found appealing from a conceptual point of view and deserving of more general elaboration.

2. PCA on manifolds: tensor fields, gradient flows, and scale space – Reporter: John Kent

The topic of this group has a clear methodological focus, bridging the gap between theory and application. The discussion brought up the notions of principal flow, which is an integrating vector field of principal components obtained by a local tangent covariance field:

$$\Sigma_h(x) = (1/\sum_i w_i) \sum_I w_i \log_x(Y_i - I) \log_x(Y_i - I)^T$$

for $w_i = w(d^2(Y_i, x)/h)$ and h a scale function.

Let $\lambda(x)e(x)$ denote the field of top eigenvectors scaled by corresponding eigenvalue, representing locally the direction of maximal variation and $\gamma(t)$ the principal flow. An integral curve starting at some point can be numerically determined by a greedy algorithm or as a solution to variational problem

$$\max \int_0^1 \langle \gamma'(t), \lambda(\gamma(t))e(\gamma(t)) \rangle dt$$

maximizing accumulated variation along the flow.

A number of questions were raised:

(1) On the principal flow
   (a) How far does one follow a principal curve/flow? Until it "turns on itself"? Could there be a parameter describing the "dying out" of the curve/flow?
   (b) Is additional regularization needed in order to avoid self-intersection or winding?
   (c) Can one have curves from different local Fréchet means using the notion of persistence diagram for the Fréchet potential to discover means at different scales?
   (d) Is it possible to use a single notion of scale for both the starting point(s) and the local covariance?

(2) On generalization of the principal flow to higher dimensions
   (a) A higher dimensional generalization is not clear: Should a second flow be defined by parallel transport of first flow? Should it be locally defined by the second eigenvector field?
   (b) Is there a canonical definition of a principal submanifold instead of a flow?

(3) And some wider questions

(a) Nestedness should play a key role: what is the backwards approach when using covariance tensor fields? "Commutative PCA": when does going backward give the same results as going forward?
(b) Should PCA or classification be done in case of considerable curvature?
(c) Can one do multi scale analysis on manifolds similar to linear spaces?

3. Tropical geometry on tree spaces – Reporter: Ruriko Yoshida

The topic of this group is theoretical in nature but it is close to the methodology on tree spaces as well as persistent homology. This group first reviewed some background on tropical geometry. To clarify nomenclature, it was pointed out that tropical algebra is concerned with tropical operations, while tropical geometry encompasses the study of solution sets of systems of tropical polynomial equations, i. e. tropical algebraic varieties. Furthermore, it was discussed how tropical lines and linear spaces are constructed. After clarifying these basic notions, the connection between tropical Grassmannians and tree spaces was reviewed and finally a number of open questions were determined.

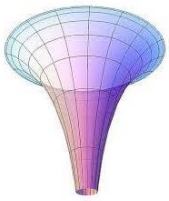
(1) How sensitive is tropical PCA to outliers? How sensitive is it to perturbation?
(2) When following a tropical line segment on tree space, how do tree topologies change along the line segment?
(3) Are there any relations between deep learning and tropical geometry?

Implementation of a toy model in R was started, which should serve to compute a tropical line segment in order to investigate these questions by numerical experiments.

4. Stochastic processes on manifolds – Reporter: Sarang Joshi

The topic of this group is relevant both theoretically and methodologically, as stochastic process theory is the theoretical foundation for the development of numerous methods. The discussion determined two major applications for stochastic process theory on manifolds which are of particular interest to the discussants.

(1) Use transition densities of SDE's on manifolds to build parametric families of distributions for parametric inference.
(2) Statistics for time-evolving manifold-valued data: Longitudinal Geometric Statistical Analysis.



Diving into the technical problems of stochastic process theory on manifold, it was pointed out that, because of curvature, Brownian motion can blow up in finite time, even if the manifold is geodesically complete. To clarify the meaning of this statement, blow-up time ξ was defined as

$$T_n = \inf\{t > 0, d(x_0, x_t) > n\} \qquad \xi = \lim_{n\to\infty} T_n$$

and a blow-up is said to occur when $\xi < \infty$. An example given by Marc Arnaudon

is the manifold defined by the line element $ds^2 = dr \otimes dr + e^{(1+r)^4} d\theta \otimes d\theta$. It was further noted that adding drift $dX_t = -\mathrm{Log}_{X_t}(y) + dB_t$ cannot stabilize a stochastic process on this manifold.

To improve on the notion of geodesic completeness, a manifold is called stochastically complete, if and only if $\xi = \infty$. To achieve this, curvature need not be absolutely bounded only imposing curvature growth bounds. One such bound introduce by Marc Arnaudon is expressed as

Theorem 1. *If $Ric(x) \geq -c(1 + d^2(x_0, x))$ then $\mathcal{M}$ is stochastically complete.*

As an example, also Kendall Shape space is stochastically complete as it is compact.

Open questions that were brought up are

(1) Is the LDDMM landmark manifold stochastically complete?
(2) What about the continuity at the cut locus and using the Log map for bridge construction? This problem is possibly solved, since the cut locus always has co-dimension ≥ 1 and therefore it should be possible to use mollification.

5. Statistics with persistent homology – Reporter: Stefan Anell

The topic of this group is mostly methodological, touching both theoretical and applied questions. The discussion in this group revolved around two main subjects. The first topic are problems arising for large data sets. Especially from the computational side large data sets can become difficult to manage and it may be worthwhile to develop new techniques. In this context, it would be interesting to understand the behavior of persistent homology features under subsampling.

The second topic that was discussed was a higher dimensional generalization of persistent homology, when more than one scale can be independently varied. In this case, features may have higher dimensional persistence regions, which might be difficult to handle. Instead, it could be beneficial to consider scale dependent summary features of the homology. In this context, some parts of the later talk by Ezra Miller were foreshadowed.

Reporter: Benjamin Eltzner

Participants

Prof. Dr. Stéphanie Allassonnière
CMAP UMR 7641
École Polytechnique
Route de Saclay
91128 Palaiseau Cedex
FRANCE

Stefan Anell
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
37077 Göttingen
GERMANY

Prof. Dr. Marc Arnaudon
Institut de Mathématiques
Université de Bordeaux I
CNRS: UMR 5251
351 Cours de la Libération
33405 Talence Cedex
FRANCE

Prof. Dr. Martin Bauer
Department of Mathematics
Florida State University
208 Love Building
1017 Academic Way
Tallahassee, FL 32306-4510
UNITED STATES

Prof. Dr. Ming Yen Cheng
Department of Mathematics
Hong Kong Baptist University
Kowloon
HONG KONG

Prof. Dr. Herbert Edelsbrunner
Institute of Science and Technology
Austria
(IST Austria)
Am Campus 1
3400 Klosterneuburg
AUSTRIA

Dr. Benjamin Eltzner
Institut für Mathematische Stochastik
Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Aasa Feragen
Department of Computer Science
University of Copenhagen
Sigurdsgade 41, 2.01E
2100 København
DENMARK

Prof. Dr. Eduardo Garcia-Portugués
Departamento de Matematicas
Universidad Carlos III de Madrid
Avenida de la Universidad, 30
28911 Leganes, Madrid
SPAIN

Prof. Dr. Ellen Gasparovic
Department of Mathematics
Union College
Bailey Hall 202
Schenectady, NY 12308
UNITED STATES

Prof. Dr. Aleksei Glazyrin
School of Mathematical & Statistical Sciences
The University of Texas - Rio Grande Valley
LHSB 2.520
One West University Blvd.
Brownsville, TX 78520
UNITED STATES

Matthias Glock
Institut für Mathematik
Technische Universität Ilmenau
Postfach 100565
98684 Ilmenau
GERMANY

Prof. Dr. Heather A. Harrington
Mathematical Institute
Oxford University
Andrew Wiles Building
Woodstock Road
Oxford OX2 6GG
UNITED KINGDOM

Prof. Dr. Soren Hauberg
Section for Cognitive Systems
DTU Compute
Technical University of Denmark
Richard Petersens Plads, Building 321
2800 Kgs. Lyngby
DENMARK

Prof. Dr. Thomas Hotz
Institut für Mathematik
Technische Universität Ilmenau
Postfach 100565
98684 Ilmenau
GERMANY

Prof. Dr. Stephan Huckemann
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstrasse 7
37077 Göttingen
GERMANY

Prof. Dr. Sarang Joshi
Scientific Computing and Imaging Institute
University of Utah
72 S. Central Campus Drive
Salt Lake City UT 84112
UNITED STATES

Prof. Dr. Sungkyu Jung
Department of Statistics
University of Pittsburgh
Pittsburgh, PA 15260
UNITED STATES

Prof. Dr. John T. Kent
Department of Statistics
University of Leeds
Leeds LS2 9JT
UNITED KINGDOM

Dr. Franz J. Király
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Line Kühnel
Department of Computer Science
University of Copenhagen
Universitetsparken 1
2100 København Ø
DENMARK

Dr. Alfred Kume
School of Mathematics, Statistics and Actuarial Science
University of Kent
Canterbury CT2 7NZ
UNITED KINGDOM

Prof. Dr. Roland Kwitt
Department of Computer Sciences
Universität Salzburg
Jakob-Haringer-Straße 2
5020 Salzburg
AUSTRIA

Prof. Dr. Huiling Le
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Anton Jussi Olavi Mallasto
Department of Computer Science
University of Copenhagen
Sigurdsgade 41, 0.02
2200 København N
DENMARK

Prof. Dr. James Stephen Marron
Department of Statistics and
Operations Research
University of North Carolina
Chapel Hill, NC 27599-3260
UNITED STATES

Dr. Facundo Mémoli
Department of Mathematics
The Ohio State University
100 Mathematics Building
231 West 18th Avenue
Columbus, OH 43210-1174
UNITED STATES

Prof. Dr. Ezra Miller
Department of Mathematics
Duke University
P.O.Box 90320
Durham, NC 27708-0320
UNITED STATES

Prof. Dr. Washington Mio
Department of Mathematics
Florida State University
Tallahassee, FL 32306-4510
UNITED STATES

Dr. Nina Miolane
INRIA Sophia Antipolis
Equipe Asclepios
Bâtiment Fermat
2004 Route des Lucioles
06410 Sophia-Antipolis
FRANCE

Prof. Dr. Anthea Monod
Department of Systems Biology
Columbia University
1130 St. Nicholas Ave., 8th Floor
New York NY 10032
UNITED STATES

Dr. Tom Nye
School of Mathematics and Statistics
Newcastle University
Newcastle upon Tyne NE1 7RU
UNITED KINGDOM

Dr. Megan Owen
Department of Mathematics &
Computer Science
Lehman College
The City University of New York
250 Bedford Park Blvd W.
Bronx, NY 10468-1589
UNITED STATES

Prof. Dr. Victor Panaretos
Institut de Mathématiques
Station 8
École Polytechnique Fédérale de
Lausanne
1015 Lausanne
SWITZERLAND

Dr. Xavier Pennec
INRIA Sophia Antipolis
2004 Route des Lucioles
06902 Sophia-Antipolis Cedex
FRANCE

Prof. Dr. Stephen M. Pizer
Department of Computer Science
University of North Carolina at Chapel
Hill
Chapel Hill, NC 27599-3175
UNITED STATES

Dr. Yvo Pokern
Department of Statistical Science
University College London
Gower Street
London WC1E 6BT
UNITED KINGDOM

Robin Richter
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstraße 7
37077 Göttingen
GERMANY

Dr. Stefan Sommer
The Image Section
Department of Computer Science
Universitetsparken 1
2100 København
DENMARK

Prof. Dr. Karl-Theodor Sturm
Institut für Angewandte Mathematik
Universität Bonn
Endenicher Allee 60
53115 Bonn
GERMANY

Dr. Fabian Telschow
University of California, San Diego
Division of Biostatistics
9500 Gilman Drive
La Jolla, CA 92093-0631
UNITED STATES

Do Tran
Department of Mathematics
Duke University
Durham, NC 27708-0320
UNITED STATES

Dr. Katharine Turner
Mathematical Sciences Institute
Australian National University
Union Lane
Acton ACT 2601
AUSTRALIA

Prof. Dr. Eva B. Vedel Jensen
Matematisk Institut
Aarhus Universitet
Ny Munkegade 118
8000 Aarhus C
DENMARK

Johannes Wieditz
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstraße 7
37077 Göttingen
GERMANY

Dr. Andrew Wood
School of Mathematical Sciences
The University of Nottingham
University Park
Nottingham NG7 2RD
UNITED KINGDOM

Jie Xu
Department of Mathematics
Boston University
MCS 145
111 Cummington Mall
Boston, MA 02215
UNITED STATES

Prof. Dr. Ruriko Yoshida
Department of Operations Research
Naval Postgraduate School
1411 Cunningham Road
Monterey CA 93943
UNITED STATES