Mathematisches Forschungsinstitut Oberwolfach

Mini-Workshop: Asymptotic Statistics on Stratified Spaces

Organised by
Aasa Feragen, Copenhagen
Stephan Huckemann, Göttingen
J.S. Marron, Chapel Hill
Ezra Miller, Durham

28 September – 4 October 2014

ABSTRACT. Statistical analysis of non-Euclidean data such as data on manifolds is an active and established topic of research, for instance, in the statistical analysis of shape. However, many types of data naturally reside in metric spaces which are *not* smooth manifolds as a whole, rather they are unions of manifold strata of varying dimensions. These spaces form a key general family of geometric spaces for data analysis. Statistics in stratified spaces has recently found great interest in applications and mathematical theory building. While the fundamental theory is still in its beginnings, as a centerpiece the derivation and investigation of statistics and their asymptotics has materialized. Only a few basic results are known, but it is clear that the geometric constraints imposed by stratified spaces lead to unexpected asymptotic behavior of standard statistical properties, such as "stickiness" of means, see [4]. It is the scope of the proposed workshop to better understand fundamental relations between asymptotic behavior of statistical descriptors and global as well as local geometric and topological structures. This investigation calls for an intense collaboration of the fields involved: statistics & stochastics; geometry & topology; combinatorics, algorithms & numerics. This workshop sought to bring together world-leading scientists and high-potential early career researchers working in this field to collaborate on a focused set of fundamental questions.

Introduction by the Organisers

The mini-workshop *Asymptotic Statistics on Stratified Spaces* organized by Aasa Feragen, Stephan Huckemann, J.S. Marron and Ezra Miller, had 17 participants, 10 of whom were junior participants.

Many statistical problems of high current interest deal with the analysis of data sampled from spaces that are naturally non-Euclidean. For example, phylogenetic trees in evolutionary biology as well as anatomical tree-like structures such as the lung airway system or blood vessels can be viewed as sampled from stratified spaces with discontinuous non-positive curvatures (cf. [3, 11]). Stratified spaces appear naturally when data have varying topological structure but continuous interpolation between different topological structures is meaningful, as is the case with anatomical tree- or graph shapes or phylogenetic trees. Stratified spaces also appear naturally as quotients under group actions in shape analysis, for instance when data distances are to be invariant under symmetries such as translation, rotation, scaling (cf. [6]).

When data reside in stratified spaces, fundamental statistical concepts, even simple ones such as "mean" or "principal component", have no canonical generalization. The non-differentiability (due to the stratification) combined with non-positivity of curvature at lower-dimensional strata results in particular statistical phenomena such as "stickiness of intrinsic means", cf. [4] (intrinsic mean may stick to lower dimensional strata under arbitrary small perturbations). Smooth curvature may also exhibit unexpected limiting behaviors; already in directional statistics on the circle, mass distributed around the antipodal of an intrinsic mean can cause "smeariness" of intrinsic means, in some sense the effect opposite to stickiness, cf. [5] (suitable mass near the antipodal may reduce asymptotic rates arbitrarily).

Not only does geometry have an impact on defining statistical properties; it may also be non-trivial to compute distances and thus statistical descriptors in stratified spaces. Optimization near points of negative curvature faces combinatorical challenges of computation at lower-dimensional strata adjacent to vast numbers of higher-dimensional manifold strata, cf. [9, 10]. While in the context of shape analysis curvature positive with respect to the top space ensures that intrinsic means of distributions not restricted to singular strata are likewise not singular, cf. [7], means may no longer be unique, but can be non-trivially set valued. In particular, positive curvature may lead to non-uniqueness of shortest paths and ambiguity which can lead to NP completeness, even for computing distances, cf. [2]. Understanding and solving computational problems often requires an interplay of geometry, stochastics, combinatorics and optimization, cf. [1, 8–10].

To address these problems, the workshop brought together young and world-leading specialists in the fields of statistics, geometry, topology, combinatorics, and numerics currently grappling with related issues from their specific scientific viewpoint. Due to the interdisciplinary nature of our research, the workshop started out with five introductory lectures on current key problems:

- Object oriented data analysis
- Dimension reduction on manifolds
- Geometric structure and statistical problems in tree spaces
- Diffusion in tree space
- Central limit theorem on T_4

The introductory lectures were held on the first day and the first half of the second day, with extremely engaged participation from the audience in the form of extended discussions of the presented problems. The introductory lectures were followed by response lectures from the remaining participants, who presented alternative viewpoints and partial solutions to the proposed problems.

The size and informal structure of the workshop resulted in a week of lively discussion and collaboration, leading to new insights and results, some of which were presented in the talks. This included the asymptotic confidence intervals on the three-spider presented by Thomas Hotz and Huiling Le, as well as very insightful discussions on topics such as diffusion on non-linear spaces, the embeddability of nonlinear distance functions in linear spaces, and a realization that kernel PCA can be viewed as slicing with varieties in higher dimensions thus potentially allowing for backward principal nested subspace analysis. We have no doubt that many of the new ideas generated at this workshop will spur new collaborations and papers in the near future.

References

[1] M. Bacak. Computing medians and means in hadamard spaces. *preprint, http://arxiv.org/abs/1210.2145*, 2013.

[2] A. Feragen. Complexity of computing distances between geometric trees. In *SSPR/SPR*, pages 89–97, 2012.

[3] A. Feragen, P. Lo, M. de Bruijne, M. Nielsen, and F. Lauze. Towards a theory of statistical tree-shape analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, page in press, 2013.

[4] T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, S. Skwerer, *Sticky central limit theorems on open books*, The Annals of Applied Probability, 23(6), 2238–2258, 2013.

[5] Thomas Hotz and Stephan Huckemann. Intrinsic means on the circle: Uniqueness, locus and asymptotics. *arXiv.org*, page 1108.2141, 2011.

[6] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic principal component analysis for Riemannian manifolds modulo Lie group actions (with discussion). *Statistica Sinica*, 20(1):1–100, 2010.

[7] Stephan Huckemann. On the meaning of mean shape: Manifold stability, locus and the two sample test. *Annals of the Institute of Mathematical Statistics*, 64(6):1227–1259, 2012.

[8] E. Miller, M. Owen, and J.S. Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. *arXiv:math.MG/1211.7046*, 2012

[9] T.M.W. Nye. Principal components analysis in the space of phylogenetic trees. *The Annals of Statistics*, 39(5):2716–2739, 2011.

[10] Megan Owen and J Scott Provan. A fast algorithm for computing geodesic distances in tree space. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(1):2–13, 2011.

[11] Sean Skwerer, Elizabeth Bullitt, Stephan Huckemann, Ezra Miller, Ipek Oguz, Megan Owen, Vic Patrangenaru, Scott Provan, and J.S. Marron. Tree-oriented analysis of brain artery structure, 2013. preprint.

Mini-Workshop: Asymptotic Statistics on Stratified Spaces

Table of Contents

Abstracts

Introduction to Object Oriented Data Analysis

J. S. Marron

Object Oriented Data Analysis is the statistical analysis of populations of complex objects. In the special case of Functional Data Analysis, these data objects are curves, where standard Euclidean approaches, such as principal components analysis, have been very successful. Challenges in diverse applications, such as image analysis and genetics have motivated the statistical analysis of populations of more complex data objects which naturally lie in non-Euclidean spaces, such as manifolds and manifold stratified spaces. These contexts for Object Oriented Data Analysis create several potentially large new interfaces between mathematics and statistics.

1. Discussion

The terminology Object Oriented Data Analysis (OODA) was coined in the statistics literature by Wang and Marron (2007) [8]. That paper was a direct consequence of a meeting on machine learning and statistics, at Oberwolfach in November of 2004.

Good discussion of the current state of the art can be found in Marron and Alonso (2014) [3]. The general concept can be understood through consideration of the *atom* of a statistical analysis, often called the *experimental unit.* In elementary statistics courses, the focus is on numbers as experimental units, and the goal is analysis of the population structure of a data set of numbers. In multivariate analysis, vectors are the atoms, i.e. the data objects. A currently active research area in statistics is *functional data analysis*, where the data objects are functions, and the goal is to understand the variation in a set of curves. See Ramsay and Silverman (2002, 2005) [4,5] for good introduction to this area. The OODA concept is to simply extend this progression to more general types of data objects. For example, the data objects could be a data set of images, with each perhaps represented as a matrix of pixel values representing gray levels or colors.

In many medical imaging applications, interest tends to focus on something of interest in each data image, such as an organ, which can appear in different parts of the images that comprise the data set. In such cases, a more reasonable choice of data object is some sort of shape representation. The statistical analysis of shapes as data objects presents special challenges, because the natural data space in non-Euclidean, often most usefully understood as a curved manifold. One example of this is the landmark based approach to shape analysis, see Dryden and Mardia (1998) [2] for a particularly lucid explanation. The idea is to represent landmark locations as point is Euclidean space, then to quotient out the group actions of translation, rotation and often scaling. This results in data objects which become equivalence classes, which are usefully viewed as lying in a curved data space, often a high dimensional sphere. Another approach to shapes in images,

which give generally better representations in contexts where there are no clearly definable landmarks, which correspond well across data cases, is the medial and skeletal representations of Siddiqi and Pizer (2008) [6]. Natural data spaces in such cases are different, usually being high dimensional products of S^2. Analogs of Principal Component Analysis (PCA) in non-Euclidean spaces were an active topic of discussion at this meeting.

Another type of data object comes from the task of analyzing populations of tree structured objects. Early approaches to this challenge used purely combinatorial methods, see Wang and Marron (2007) [8] and Aydin et al (2009) [1]. More recent approaches, see Skwerer et al (2013) [7] have used the phylogenetic tree space ideas of Billera, Holmes and Vogtmann (2001) , which results in statistical analysis of a set of data objects lying on manifold stratified spaces. New ideas in this direction were an active subject of discussion at this workshop.

2. A Deep Open Problem

During the course of the current workshop, both from the talk of Washington Mio, and through informal discussion, it became apparent that there is a *methodological hole* in the usual tool set used in OODA. This comes in terms of various types of *data summarization*. The first idea about that terminology in most people's minds, are the population center point, often some version of the mean vector μ, and some notion of spread such as the covariance matrix, Σ. Such summaries are very *high level* in the sense that there are usually other aspects of interest of the distribution, such as clusters or other types of non-Gaussian behavior. PCA, at the level of full loadings and scores, can also be viewed as a data summary, but because it still contains a large amount of data information, it can be viewed as a *low level* summary of the data.

Aspects of Mio's talk fit into this framework in a couple of ways. The local covariance function and local mean function are in some sense even lower level than PCA. However, the critical points of these functions determine interesting summaries at a higher level. This type of summary, plus any related distributional quantities, such as cluster centers as well as an indication of local variation in each cluster, is available, but so far seems to have only been considered in an ad hoc way.

The open problem proposed here is to develop a systematic and principled approach to finding such data summaries, which fill the hole between the often uninformatively high level summaries of μ and Σ, and the too detailed low level summary of PCA.

References

[1] B. Aydın, G. Pataki, H. Wang, E. Bullitt, and J. S. Marron. A principal component analysis for trees. *The Annals of Applied Statistics*, 3(4):1597–1615, 2009.
[2] I. L. Dryden and K. V. Mardia. *Statistical analysis of shape*. Wiley, 1998.
[3] J Steve Marron and Andrés M Alonso. Overview of object oriented data analysis. *Biometrical Journal*, 2014.

[4] James O Ramsay and Bernard W Silverman. *Applied functional data analysis: methods and case studies*, volume 77. Springer, 2002.

[5] JO Ramsay and BW Silverman. Functional data analysis. 2005.

[6] K. Siddiqi and S. Pizer. *Medial representations: mathematics, algorithms and applications*, volume 37. Springer, 2008.

[7] Sean Skwerer, Elizabeth Bullitt, Stephan Huckemann, Ezra Miller, Ipek Oguz, Megan Owen, Vic Patrangenaru, Scott Provan, and JS Marron. Tree-oriented analysis of brain artery structure. *Journal of Mathematical Imaging and Vision*, pages 1–18, 2013.

[8] Haonan Wang, JS Marron, et al. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

Dimension reduction for directions and 2D shapes

SUNGKYU JUNG

(joint work with Stephan Huckemann, J. S. Marron, and Thomas Hotz)

Principal Component Analysis (PCA) is an effective method of analyzing main modes of variation in a dataset, which also gives a basis for dimension reduction. There has been a growing need for PCA-like analysis of data that naturally lie on a curved manifold. Examples of such data situations include directions in a unit hypersphere S^d and landmark-based shapes in Kendall's shape space $\Sigma_m^k = S^d/\mathrm{SO}(m)$, $d = mk - m - 1$. Several extensions of PCA for manifold data [1–3,5] have been proposed, and they provide submanifolds of any dimension $1 \leq d_0 \leq d$, or a series of geodesics that "spans" the submanifolds. These geodesics take the role of basis in vector space, and provide projection scores of data analogous to principal component scores in PCA. Different methods have different fitting strategies. In particular, Principal Nested Spheres (PNS [3]), applied to S^d, allows the submanifolds to be non-geodesic, but only requires the submanifolds to be self-similar. When the object to be fitted in PNS is resticted to be only geodesic submanifolds, we call the sub-method and its result of analysis by *Geodesic PNS*.

There is a close relationship between Geodesic PNS and Horizontal Component Analysis (HCA, [5]), when the data $\mathcal{X} = \{x_1, \ldots, x_n\}$ are in S^d. To see this, denote the k-dimensional geodesic submanifolds of S^d by $A_k \cong S^k$, $k \leq d$. Geodesic PNS fits the nested structure $S^d \supset A_{d-1} \supset \cdots \supset A_1$ to the data. The PNS fitting is a *backward* successive fashion. That is, A_{d-1} is first fitted to the data, by minimizing the sum of squared residuals $\sum_{i=1}^n \rho^2(x_i, A_{d-1})$, where $\rho(x, A) = \inf_{y \in A} \rho(x, y)$, $\rho(x, y) = \arccos(|\langle x, y \rangle|)$, $x, y \in S^d \subset \mathbb{R}^{d+1}$. Then each data point x_i is projected onto A_{d-1}, leading to a projected data set on A_{d-1}. By the trivial identification of A_{d-1} with S^{d-1}, smaller-dimensional submanifolds are fitted successively. HCA, on the other hand, is a *forward* generalization of PCA, where the first geodesic γ_1, a 1-dimensional submanifold, is fitted by minizing residual variance. The family of second geodesics $\Gamma_2 = \{\gamma_2^{(t)}, t \in \mathbb{R}\}$ is formed and fitted by parallel transport of a geodesic $\gamma_2^{(0)}$, which is orthogonal to γ_1 at the intersection $\gamma_1(0)$. On S^d, there exists A_2 such that $A_2 = s(\Gamma_2) := \{\gamma(s) \in S^d : \gamma \in \Gamma_2, s \in \mathbb{R}\}$. This relation can be extended to the kth family of geodesics Γ_k $(k = 1, \ldots, d)$ where $s(\Gamma_k) \subset S^d$ is a geodesic submanifold of dimension k. Therefore, for S^d,

the underlying structure of Geodesic PNS equals that of HCA. Moreover, given the sequence Γ_k and $A_k = s(\Gamma_k)$, $k = 1, \ldots, d$, both methods provide the same projection score for any point in S^d, modulo linear and circular translation.

Despite the resemblance, we emphasize that the estimation strategy of Geodesic PNS differs from that of HCA. For a given dataset $\mathcal{X}$, we expect that $\widehat{A}_k \neq s(\widehat{\Gamma}_k)$ for all $k \leq d$, $d > 2$. The difference is in whether the dimension of "principal" submanifold is reduced (backward PCA and PNS) or increased (forward PCA and HCA). In the exceptional case of S^2, the unit sphere in $\mathbb{R}^3$, we have $\widehat{A}_k = s(\widehat{\Gamma}_k)$ for $k = 1, 2$, and HCA becomes the same as Geodesic PNS in all aspects. In fact, many methods coincide when they are applied to S^2; See [4].

PNS is developed using two basic principles: i) self-similarity of (possibly non-geodesic) submanifolds, and ii) the backward successive fitting strategy. For the shape data, these principles are applied to develop a novel non-Euclidean intrinsic version of PCA directly taking into account all of the data. Suppose data are observed in Σ_2^d, the Kendall's shape space for d-landmark planar shapes. Every principal component (PC) represents the data in a lower dimensional space, that is similar in curvature to the original data space. For a sequence of PCs of varying dimensions we follow a nested backward fitting scheme as proposed in [3]. In Σ_2^d, lower-dimensional shape spaces Σ_2^k thus give PCs in the following sequence of embedded spaces

$$(1) \qquad\qquad \Sigma_2^d \supset \Sigma_2^{d-1} \supset \cdots \supset \Sigma_2^3.$$

Given a dataset $\mathcal{X} = \{x_1, \ldots, x_n\} \in \Sigma_2^d$, the fitting of nested submanifolds mostly depends on the means of *projections* onto Σ_2^{d-1}. Different projections give different shapes with $d-1$ landmarks. Should the projection be data-dependent? What would be the optimal projection? To answer the question of finding the data-dependent optimal projection, we introduce a flexible framework of dimension reduction, called reduction of object features (ROOF).

The ROOF is a general concept for dimension reduction of *object* data. While it has a potential to describe various dimension reduction techniques for other types of object data, such as trees as data objects [6], let us focus on the shape analysis. In shape analysis, we take the object as the outline that is given by k landmarks, modulo translation, scaling and rotation. An object space $(\mathcal{S}, \rho)$ with an appropriate metric ρ makes it possible compare a d-landmark shape in Σ_2^d and a k-landmark shape in Σ_2^k, $d \neq k$. The shape spaces $T_k = \Sigma_2^k$ ($k = 3, \ldots, d$) are understood as self-similar feature spaces of dimension k. The object space and the feature spaces are linked by mappings $\Phi_k : T_k \to \mathcal{S}$, which are data-dependent. The goal of ROOF, given a dataset $\mathcal{X}$, is then to find a set of projections $y_1, \ldots, y_n \in T_k$ and a mapping Φ_k such that the new features $\mathcal{Y}_k = \{y_i\}$ contain as much information as possible compared to the original features $\mathcal{X}$; It is equivalent to solve the following least-squares problem.

$$(2) \qquad\qquad \min_{y_i \in T_k, \Phi_k} \sum_{i=1}^n \rho^2(\Phi_d(x_i), \Phi_k(y_i)).$$

Successively solving (2), *i.e.*, first by setting $k = d - 1$, then replacing (x_i, d, k) by $(y_i, d - 1, d - 2)$, and so forth, gives an estimate of the sequence of the lower-dimension PCs (1). In many real and simulated examples, we found that the lower-dimensional approximations $\mathcal{Y}_k \subset \Sigma_2^k$ efficiently reduce the dimension of the feature space, while preserving large amount of variation contained in the original data $\mathcal{X}$. For example when the PNS is applied to $\mathcal{Y}_k$ and $\mathcal{X}$ for $k < d$, the major modes of variation in $\mathcal{Y}_k$ are comparable in its interpretation to those in $\mathcal{X}$. The results of ROOF differs from those of other PCA extensions, such as PNS [3], HCA [5] and Geodesic PCA [2]: The PCA extensions provides k-dimensional approximation of $\mathcal{X}$ in Σ_2^d as d-landmark shapes; The ROOF for shape data provides reduction of number of landmarks, with k-landmark shapes $(\mathcal{Y}_k, k < d)$ as approximations of the original shapes.

References

[1] P Thomas Fletcher, Conglin Lu, Stephen M Pizer, and Sarang Joshi. Principal geodesic analysis for the study of nonlinear statistics of shape. *Medical Imaging, IEEE Transactions on*, 23(8):995–1005, 2004.

[2] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic PCA for Riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 20(1):1–58, 2010.

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

[4] Sungkyu Jung, Mark Foskey, and J. S. Marron. Principal Arc Analysis on direct product manifolds,. *Ann. App. Statist.*, 5(1):578–603, 2011.

[5] Stefan Sommer. Horizontal dimensionality reduction and iterated frame bundle development. In *Geometric Science of Information*, pages 76–83. Springer, 2013.

[6] Haonan Wang and JS Marron. Object oriented data analysis: Sets of trees. *The Annals of Statistics*, 35(5):1849–1873, 2007.

On tree-space PCA

Aasa Feragen

(joint work with Sean Cleary, Megan Owen, and Daniel Vargas)

BHV tree-space [1][1], or the space of trees on a fixed leaf label set, has proven useful for statistical analysis of many different types of data including phylogenetic trees, airway trees, blood vessel trees and dendrograms [2, 5, 8, 9]. In this talk we discuss some interesting properties of first principal components in BHV tree-space.

The 2-dimensional open book on 3 leaves, or $\mathcal{T}_3 \times \mathbb{R}$, see Fig. 1, is a subspace of any BHV tree-space $\mathcal{T}_n$, $n \geq 4$. It is also a subspace of the BHV space of 3 leaves when the lengths of the root and pendant edges are included in the tree-space. This makes the open book an excellent setting for producing simple examples of tree-space phenomena.

---

[1]BHV is short for for Billera, Holmes, and Vogtmann, the authors of the paper [1] that first defined the space of trees on a fixed leaf label set.
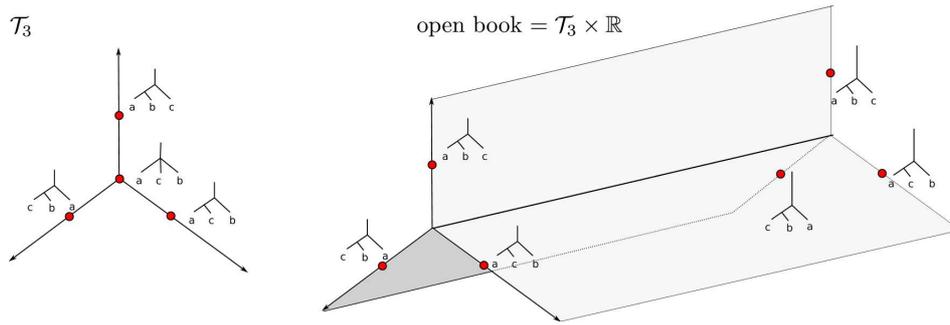
FIGURE 1. **Left:** BHV tree-space on 3 leaves $\{a, b, c\}$. **Right:** The open book with 3 sheets.

1. DEFINITION OF FIRST PRINCIPAL COMPONENT

Given a finite dataset $X = \{x_i | i = 1, \ldots, N\}$ in the BHV tree-space on n leaves $\mathcal{T}_n$, the Fréchet mean of X is defined as the point minimizing the sum of squared distances to the data points [6]:

$$\hat{x} = \operatorname{argmin}_{x \in \mathcal{T}_n} \sum_{i=1}^{N} d(x_i, x)^2.$$

The first principal component of X, which we denote $PC1(X)$, is defined [2] as the geodesic segment $\gamma_{a_0 b_0}$ minimizing the squared projection error $\varphi(X, \gamma_{ab})$:

$$PC1(X) = \gamma_{a_0 b_0}, \text{ where } a_0, b_0 = \operatorname{argmin}_{a,b \in \mathcal{T}_n} \varphi(X, \gamma_{ab}),$$

where $\varphi(X, \gamma_{ab})$ is the squared error of projection onto the geodesic segment γ_{ab} connecting the points $a, b \in \mathcal{T}_n$.

$$\varphi(X, \gamma_{ab}) = \sum_{i=1}^{N} d(x_i, \operatorname{pr}_{\gamma_{ab}}(x_i))^2.$$

This definition is analogous to the definition of first principal component on manifolds due to Huckemann et al. [3] and Sommer et al. [4], except for the restriction to geodesic segments, which is due to the problem of parametrizing geodesic rays in tree-space [2, 5].

2. THE FIRST PRINCIPAL COMPONENT IS STICKY

Just like the Fréchet mean [7], the principal component $PC1(X)$ can be "sticky", defined as follows: *The first principal component $PC1(X)$ for a finite sample $X \subset \mathcal{T}_n$ sticks to a subset $S \subset \mathcal{T}_n$ if $PC1(X) \subset S$ and sufficiently small permutations $\tilde{X}$ of X lead to $PC1(\tilde{X}) \subset S$.*

Stickiness of $PC1(X)$ is shown by the example of Fig. 2 on the 2-dimensional open book with 3 leaves. It is clear that $PC1(X) \subset S$ where S is the spine of the open book, and the same holds for small perturbations of X.

Stickiness of $PC1(X)$ indicates that, just as for sticky means, first principal components of a topologically diverse dataset will tend to be contained in the

higher codimension strata of tree-space. It is not clear how parallel transport along such principal components might be defined, which has consequences for extending techniques from manifold statistics [4] to the open problem of defining second principal components.
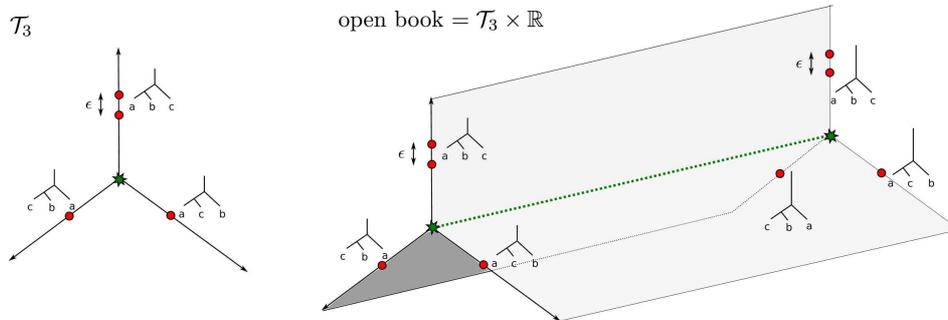

FIGURE 2. **Left:** The Fréchet mean (star-shaped point) of the dataset in $\mathcal{T}_3$ is sticky [7]. **Right:** The first principal component (the dotted line geodesic connecting two star-shaped endpoints) of the dataset in the open book sticks to the spine of the book.

3. THE FIRST PRINCIPAL COMPONENT DOES NOT ALWAYS CONTAIN THE FRÉCHET MEAN

In Euclidean space, the Fréchet mean always lies on the first principal component. On curved manifolds, this is known not to be the case [3]. In Fig. 3 we give an example on the open book, where the Fréchet mean does *not* lie on the first principal component, showing that this is also not the case in BHV tree-spaces.

This has consequences for the definition and interpretation of the fraction of variance captured by a principal component, which is frequently used to measure the success of dimensionality reduction via PCA in Euclidean space [10].

4. THE FIRST PRINCIPAL COMPONENT DOES NOT VARY CONTINUOUSLY WITH THE DATA

In Euclidean space, the first principal component of a sample X varies continuously with the sample, in the sense that small perturbations of X lead to small perturbations of principal components. This, however, is not always the case in BHV tree-space. We illustrate this with the example in Fig. 4, where we let $X = \{x_1, x_2, x_3\}$, with the tree x_3 having interior branches of length a. For $a = 0$, $PC1(X)$ is clearly given by the option shown in (b), which leads to a projection error of 0. For varying $a > 0$, the figures (b)-(e) illustrate all possible options for $PC1(X)$. The only way that $PC1(X)$ can vary continuously away from (b) as a moves from 0 is through option (c), which is never optimal.
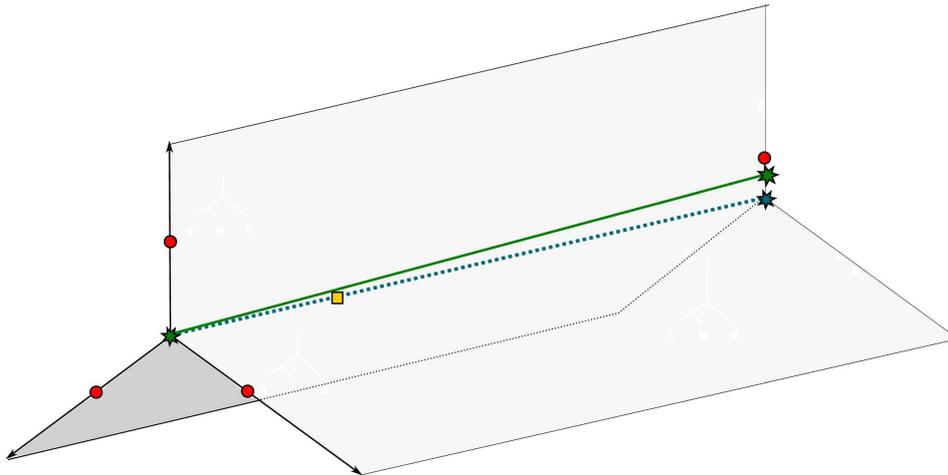
FIGURE 3. Let X consist of the circular points in the figure. The Fréchet mean $\mu(X)$ is the quadratic point, which sits on the spine. However, $PC1(X)$ is *not* the dotted line segment along the spine, as the solid line segment connecting two star-shaped points can easily be shown to give a lower least squares error.
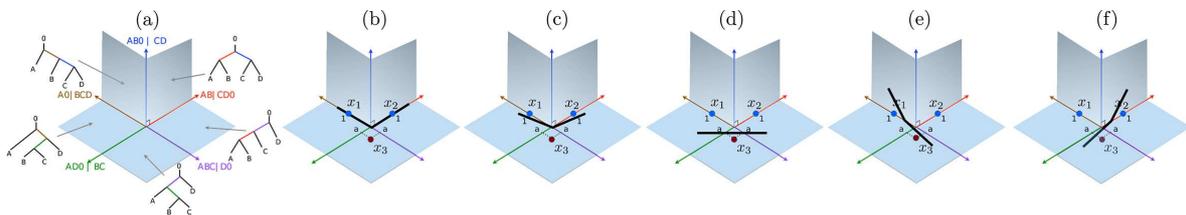
FIGURE 4. For $a = 0$ the geodesic segment shown in (b) is optimal. As the lengths a increase, (b) remains a better option than (c) until one of the solutions (d)-(f) takes over.

5. SUMMARY

In this talk, we have reviewed some interesting properties of first principal components in tree-space. In some cases, the observed behavior may, while mathematically interesting, seem practically counterproductive at a first glance. However, we do not believe that they are. In datasets with moderate topological variation, we do not expect stickiness to lead to uninformative statistics, as already observed with Fréchet means of airway trees appearing in medical imaging [2]. Moreover, several of the observed properties such as discontinuous dependency on the dataset, or failure to contain the Fréchet mean, also happen on nonlinear manifolds, where both means and principal components are still of great utility.

REFERENCES

[1] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.

[2] A. Feragen, M. Owen, J. Petersen, M.M.W. Wille, L.H. Thomsen, A. Dirksen, and M. de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, Vol. 7917, pp 74-85, 2013.

[3] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–58, 2010.

[4] S. Sommer. Horizontal Dimensionality Reduction and Iterated Frame Bundle Development. *Geometric Science of Information (GSI)*, Lecture Notes in Computer Science, Vol. 8085, pp 76-83, 2013.

[5] T. M. W. Nye. Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, 39(5):2716–2739, 2011.

[6] H. Karcher. Riemannian center of mass and mollifier smoothing. *Communications on Pure and Applied Mathematics*, 30(5):509–541, 1977.

[7] 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 theorem on open books. *Annals of Applied Probability*, vol. 23 no. 6 (2013), pp. 2238-2258.

[8] S. Skwerer, E. Bullitt, S. Huckemann, E. Miller, I. Oguz, M. Owen, V. Patrangenaru, J.S. Provan, and J.S. Marron. Tree-Oriented Analysis of Brain Artery Structure. *Journal of Mathematical Imaging and Vision*, vol. 50 no. 1-2 (2014), pp. 126–143.

[9] K. Kobayashi and M. Orita. Permutation test for dendrograms and its application to the analysis of mental lexicons. Preprint, arXiv1403.2845 (2014).

[10] T.M.W. Nye. An Algorithm for Constructing Principal Geodesics in Phylogenetic Treespace. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pages 304-315, 2014.

Construction of distributions on tree-space via diffusion processes
TOM M. W. NYE

The set of rooted edge-weighted trees on a fixed set of leaves $\{1, 2, \ldots, N\}$ forms a geodesic metric space known as BHV tree-space [1], denoted $\mathcal{T}_N$. Data consisting of sets of points in tree-space arise in evolutionary biology and from various medical imaging techniques. Construction of non-trivial probability distributions on BHV tree-space for which the density function can be evaluated is a challenging problem. For example, the volume of the unit radius ball around a point $x_0 \in \mathcal{T}_N$ varies with x_0 and is difficult to compute. It follows that the set of uniform distributions on each ball of radius r around any x_0 does not form a tractable family of probability distributions. Nonetheless, in order to develop probability models on tree-space, it would be highly desirable to construct distributions representing a cloud of dispersed density around a central point. One approach is to construct distributions by fixing a 'source' point $x_0 \in \mathcal{T}_N$ and running Brownian motion from x_0 for a duration t_0 of time. We denote the resulting distribution $B(x_0, t_0)$.

For $N = 3$ and $N = 4$ analytic solutions for the density function of $B(x_0, t_0)$ have been obtained previously [3]. For $N = 3$, tree-space is particularly simple: it consists of three copies of the positive real line $[0, \infty)$ glued together at the origin. This space is often called the 3-*spider*. Brownian motion the 3-spider is

the same as a regular Brownian motion on each arm, but at the origin the diffusing particle moves onto each arm with equal probability. The heat equation can be easily solved on $\mathcal{T}_3$ by folding the two arms which do not contain x_0 together, and applying the reflection principle for Brownian motion on the real line. The solution consists of a linear combination of Gaussians on each arm of the spider. The graph below shows the solution for increasing t_0.
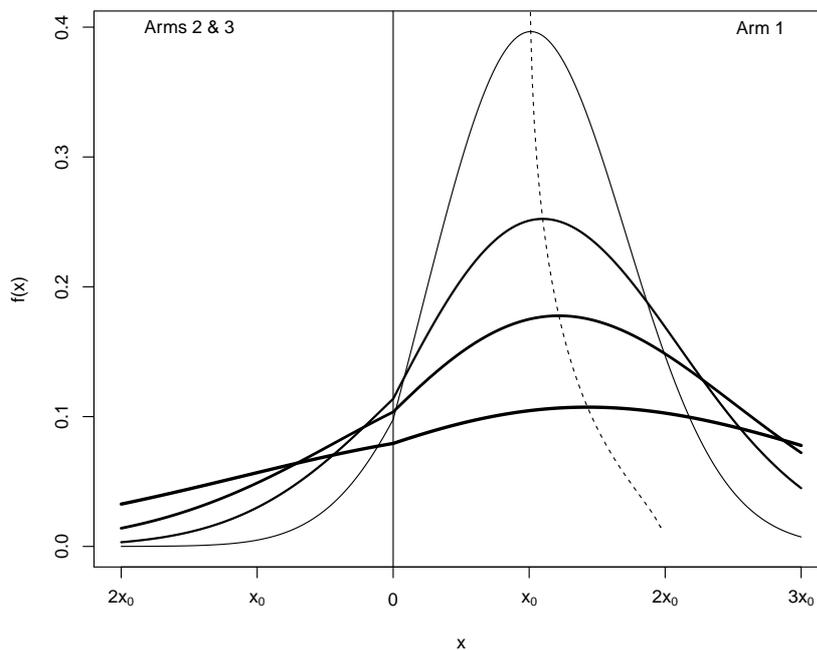
It is interesting to note from the graph that there are three possible summary statistics for the location of the distribution: the source x_0, the mode, and the Fréchet mean. The dotted line shows the behaviour of the mode with increasing t_0. It approaches $2x_0$ on the arm containing the source as $t_0 \to \infty$. On the other hand, the Fréchet mean moves to the origin at some finite value of t_0, and then sticks there as t_0 increases further. In some situations, 'stickiness' of the Fréchet mean [2] might be undesirable, so modelling data as being drawn from $B(x_0, t_0)$ and using the source x_0 or the mode as a summary of location offers an interesting alternative.

For $N > 4$ obtaining an analytic solution to the heat equation on $\mathcal{T}_N$ becomes very difficult. Instead the author has developed an alternative approach based on simulation. An appropriate random walk on tree-space from x_0 can be defined, and the distribution of points after m steps of the walk is denoted $W(x_0, t_0; m)$. It can be shown that $W(x_0, t_0; m) \to B(x_0, t_0)$ in distribution as $m \to \infty$. Given a data set on $\mathcal{T}_N$, parameter inference for x_0, t_0 can then be performed by forward-simulating random walks. The details of this approach will be published elsewhere.

References

[1] L. Billera, S. Holmes, and K. Vogtmann, "Geometry of the space of phylogenetic trees," *Advances in Applied Mathematics*, vol. 27, pp. 733–767, 2001.

[2] T. Hotz, S. Huckemann, H. Le, J.S. Marron, J.C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, and S. Skwerer, "Sticky central limit theorems on open books," *The Annals of Applied Probability*, vol. 23(6), pp. 2238–2258, 2013.

[3] T. M. W. Nye and M. C. White, "Diffusion on Some Simple Stratified Spaces," *Journal of Mathematical Imaging and Vision*, vol. 50, pp. 115–125, 2014.

Fréchet Means in the Space of Phylogenetic Trees

Huiling Le

(joint work with Dennis Barden and Megan Owen)

The concept of Fréchet means of random variables on a metric space is a generalisation of the least mean-square characterisation of Euclidean means: a point is a Fréchet mean of a probability measure μ on a metric space $(\boldsymbol{M}, d)$ if it minimises the Fréchet function for μ defined by

$$x \mapsto \frac{1}{2} \int_{\boldsymbol{M}} d(x, x')^2 d\mu(x').$$

Among other applications, this has recently been used in the statistical analysis of phylogenetic trees, as motivated by [3] and [4]. The space $\boldsymbol{T}_m$ of phylogenetic trees with m leaves was first introduced in [3]. For each fixed m, the space is a topologically stratified space and also a $CAT(0)$ space. Thomas Hotz's notes, produced during the SAMSI 2010-11 Program on 'Analysis of Object Data', on the behaviour of Fréchet means on a 'spider', as well as the ensuring work of [5] and [1], demonstrates that the results on the limiting behaviour of sample Fréchet means in Riemannian manifolds can not be directly generalised to stratified spaces. In particular, when a population mean there lies in a stratum of positive co-dimension, the sample Fréchet means exhibit a range of non-standard behaviour. See also [2].

To be able to analyse this further in the case of $\boldsymbol{T}_m$, we recall that the main tool used in [5] for obtaining the limiting distribution for sample Fréchet means in *open books* is the so-called *folding map*. These folding maps connect the asymptotic behaviour on the open books with those of certain sequences of means of Euclidean random variables. Thus, they link the limiting distribution of sample Fréchet means on the open books with the limiting distributions of the sample means of the related Euclidean random variables.

Analogous to the folding maps for open books, the crucial step for studying the limiting behaviour of sample Fréchet means in $\boldsymbol{T}_m$, the space of phylogenetic trees with m leaves, is the *log map*, or equivalently the *translated log map*. By the log map, we mean a generalisation of the inverse of the exponential map on a Riemannian manifold. In particular, the log map at each tree T^* is a map from $\boldsymbol{T}_m$ to the tangent cone at T^* and is expressed in terms of the lengths and initial tangent vectors of the geodesics starting from T^*. The translated log map is a

composition of the log map with parallel transport, so that the images of the log maps at all trees in a given stratum are transported to a common Euclidean cone. However, the translated log maps depend also on the base tree T^*. Contrasting with the case for Riemannian manifolds, the log map at each tree in $\boldsymbol{T}_m$ is neither a diffeomorphism, nor second order differentiable. In particular, a new phenomenon in $\boldsymbol{T}_m$ is that a single unit vector in the tangent cone at T^* usually results in infinitely many geodesics, inherited from the structure of the space. For analysis of the properties of the log map required for our study of asymptotic behaviour of sample Fréchet mean, we use the geometric structure of geodesics in $\boldsymbol{T}_m$ obtained in [6] and [7], which enables us to express the log map on $\boldsymbol{T}_m$ in a form usable for our investigation.

Using the results on the log map, we first characterise the conditions for a tree in $\boldsymbol{T}_m$ to be the Fréchet mean of a given probability measure μ on $\boldsymbol{T}_m$. The characterisation obtained varies with the dimension of the stratum on which the Fréchet mean lies. It is expressed, in the directions orthogonal to the stratum where the Fréchet mean lie, in terms of directional derivatives while, on the subspace tangent to the stratum, it is expressed in terms of the projection of the translated log map.

By taking advantage of the special structure of tree space in the neighbourhood of any given stratum, we demonstrate the non-classical behaviour of sample Fréchet means arising from the global topological structure of the space. In particular, we show that, despite the log map being neither a diffeomorphism nor second order differentiable, when the Fréchet mean of μ lies on a top-dimensional stratum the sample Fréchet means behaves in a similar way to those in Riemannian manifolds. Under suitable conditions, we establish a central limit theorem for *iid* random variables having probability measure μ that has its Fréchet mean lying in a top-dimensional stratum, where the role of the global topological structure of the space is played through the covariance structure of the limiting distribution. When the Fréchet mean of μ lies on a stratum of co-dimension one, under certain conditions the limiting distribution of sample Fréchet means can take one of three possible forms, distinguished by the nature of its support. This support may be either the $(m-1)$-dimensional Euclidean space containing the stratum of co-dimension one where the Fréchet mean lies, or a half Euclidean space of dimension m whose boundary contains that stratum of co-dimension one, or the union of two such half spaces. In contrast, when a Fréchet mean lies in a stratum of co-dimension at least two, the support of the limiting distribution can take various different forms. Nevertheless, in all these cases, the limiting distributions are linked closely with Gaussian distributions in Euclidean spaces. In particular, these results also improve our results, obtained in [1], on the limiting behaviour of sample Fréchet means when $m = 4$ and the Fréchet mean of μ is at the cone point.

References

[1] D. Barden, H. Le and M. Owen, *Central limit theorems for Fréchet means in the space of phylogenetic trees*, Electron. J. Probab. **18** (2013), no. 25.

[2] B. Basrak, *Limit theorems for the inductive mean on metric trees*, J. Appl. Prob. **47** (2010) 1136–1149.

[3] L.J. Billera, S.P. Holmes and K. Vogtmann, *Geometry of the space of phylogenetic trees*, Advances in Applied Mathematics **27** (2001), 733–767.

[4] S. Holmes, *Statistics for phylogenetic trees*, Theoretical Population Biology **63** (2003), 17–32.

[5] T. Hotz, S. Huckemann, H. Le, J.S. Marron, J.C. 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.

[6] M. Owen, *Computing geodesic distances in tree space*, SIAM J. Discrete Math., **25** (2011), 1506–1529.

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

Algebraic Combinatorial Methods in Statistics and Machine Learning

FRANZ J. KIRÁLY

(joint work with Duncan A.J. Blythe, Martin Kreuzer, Louis Theran, Ryota Tomioka)

A common generative sparsity assumption for data is to lie in some sub-manifold or sub-variety of n-space, that is, observations corresponding to a data matrix $X \in \mathbb{R}^{N \times n}$ where the rows are vectors in $\mathcal{M} \subseteq \mathbb{R}^n$, potentially distorted by statistical noise. Often $\mathcal{M}$ is unknown, in which case the central question arising is estimating $\mathcal{M}$, given the data matrix X and possible knowledge on X. Closely related to it is denoising X. For example, if $\mathcal{M}$ is known to be a r-dimensional sub-vector space of $\mathbb{R}^n$, and the noise on X is i.i.d. Gaussian, a least-squares estimator is given by the span of the first r principal, or right singular vectors associated to X. A denoising of a row of X is given by orthogonal projection.

In the talk, two more general cases were discussed: (a) the case where $\mathcal{M}$ is a r-dimensional sub-vector space of $\mathbb{R}^n$, but some entries of X are missing, and (b) the case where $\mathcal{M}$ is a sub-manifold, or an algebraic sub-variety of $\mathbb{R}^n$. As presented, both cases are amenable to estimation using certain ideas from algebra and combinatorics, as will be briefly discussed below.

(a) Algebraic-Combinatorial Low-Rank Matrix Completion

If $\mathcal{M}$ is linear and some entries of X are missing, the problem is known as low-rank matrix completion - as X will be an incomplete matrix of rank r. Both an estimation of $\mathcal{M}$ or of one of the missing entries of X is in general hard, and similarly is a further denoising of an observed entry. Existing literature is mostly focused on optimization-based estimation of $\mathcal{M}$; in the talk, the first systematic theory and methodology for imputing single entries of X and obtaining entry-wise

error estimates was presented, see [1]. The theory is centered around objects and results of the following algebraic combinatorial kind:

Definition: Let $C \subseteq [m] \times [n]$ a set of indices. C is called dependent if the map $\Omega : \mathcal{M}^N \to \mathbb{C}^{\#C}, X \mapsto (X_{ij}, (ij) \in C)$ is not surjective. C is called circuit (of rank r) if it is minimally dependent (intuitively, all but one entries of a circuits leave no degree of freedom for the remaining one).

Theorem: Assume the rows of X are in (very) general position. Assume the entry $X_{k\ell}$ can be completed from the $X_{ij}, (ij) \in E$. Then $(k\ell) \in E$, or there is a circuit C such that $(k\ell) \in C$ and $C \in E \cup (k\ell)$.

Theorem: Assume the rows of X are in (very) general position. Let $C \subseteq [m] \times [n]$ be a circuit (of rank r). Then there is an up-to-scaling-unique polynomial θ_C, such that $\theta_C(X_{ij}, (ij) \in C) = 0$ if and only if the $X_{ij}, (ij) \in E$ can be completed to a rank r matrix X.

For theoretical purposes, the circuit polynomials generalize the determinants as certifying functions for low-rankness for the case of incomplete matrices. For practical purposes, they can be used to fill in the missing entries, by finding circuits which are supported on observed entries except one, and similarly to denoise observed ones by finding circuits with support in the observed entries. Algorithms for this and error estimation are described in [1] and the follow-up papers. I conclude with two open questions:

Open problem: Characterize circuits and circuit polynomials in rank $r = 2$.

Open problem: If $N = n$, and the observed positions E are the edges of a uniformly random $2r$-regular graph: are all entries of X completable with high probability, i.e., for the asymptotics $n = N \to \infty$?

(b) Low-Rank Approximation with Cross-Kernel Matrices
If $\mathcal{M}$ is non-linear, e.g. an algebraic manifold, kernel methods offers a variety of well-scaling algorithms to obtain information from the data matrix X. Most of these involve the so-called kernel matrix $K(X, X) = (k(x_i, x_j))_{ij} \in \mathbb{R}^{N \times N}$, where x_i is the i-th row of X, and $k : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ is a positive semi-definite kernel function, e.g., the degree d inhomogenous polynomial kernel $k(x, y) = (x^\top y + 1)^d$. For example, a possible estimate for $\mathcal{M}$ which can be readily deduced from existing literature is the vanishing locus of the kernel PCA projection residual, which is the positive semi-definite polynomial $\rho(x) = k(x, x) - \kappa(x)^\top (PK(X, X)P)^{-1}\kappa(x)$, where $\kappa(x) = (k(x, x_i))_i \in \mathbb{R}^N$, where P denotes projection on the principal eigenvalues of $K(X, X)$, and $(PK(X, X)P)^{-1}$ the pseudo-inverse of the projection.

The computational bottleneck of evaluating $\rho(x)$, similarly to most kernel based algorithms, consists in tasks related to the eigenvalue decomposition of $K(X, X)$, whose computational cost is $\Theta(N^3)$. An idea presented in the talk is to use an approximation based on the cross-kernel matrix $K(X, Z) = (k(x_i, z_j))_{ij} \in \mathbb{R}^{N \times M}$, where z_j are rows of some matrix $Z \in \mathbb{R}^{M \times n}$, see [2]. The approximation takes the

form of $K(X, X) \approx K \cdot K^\top$, where $K = K(X, Z) \cdot K(Z, Z)^{-1/2}$. Computation of K costs only $\Theta(M^2 N + M^3)$, and the singular value decomposition of K can be used to approximate the eigenvalue decomposition of $K(X, X)$. For the polynomial kernel, one can show the following exact asymptotics in M:

Theorem: Assume the rows of Z are in (very) general position (this holds for example when the entries are i.i.d. Gaussian). Then, there is some M_0 such that if $M \geq M_0$, then $K(X, X) = K \cdot K^\top$.

Note that the statement above does not depend on X, only on Z which can be prescribed by the statistician. Unfortunately, M_0 can be quite big and grow exponentially in d. Though if $\mathcal{M}$ is an algebraic variety, a better bound is available.

Theorem: Assume both the rows of X and Z are in $\mathcal{M}$, with the rows of Z in (very) general position under this constraint. If $M \geq \binom{n+d+1}{d} - \dim \mathrm{I}(\mathcal{M})_{\leq d}$, then $K(X, X) = K \cdot K^\top$.

Here $\mathrm{I}(\mathcal{M})_{\leq d}$ denotes the (degree-bounded part of the) so-called ideal of $\mathcal{M}$, the vector space of all polynomials in n variables vanishing on $\mathcal{M}$. Heuristically, the smaller dimension $\mathcal{M}$ has and the less complex it is, the closer its dimension will be to $\binom{n+d+1}{d}$, allowing to choose smaller M.

Numerical experiments (see [2]) show that the approximation, while not anymore exact, is still quite good for even much smaller for i.i.d. random Z when M is small, and for other kernels which are not polynomial. As currently no related proven statement is known to me, I would like to formulate these as open problems:

Open problem: Prove the following asymptotic conjecture inspired by results from compressed sensing, or a suitably modified variant for the polynomial kernels. Consider an algebraic manifold $\mathcal{M} \subseteq \mathbb{R}^n$ of dimension D, and the polynomial kernel k of degree d. Let $x, y \in \mathcal{M}$ arbitrary but fixed. Let $z_1, \ldots, z_M$ with i.i.d. standard normal entries. Assume $M \geq \lambda D d \log n$, with $\lambda \geq 1$. Then, with global C, c,

$$P\left(\|k(x, y) - K(x, Z) \cdot K(Z, Z)^{-1} \cdot K(Z, y)\| \geq \varepsilon\right) \leq C \varepsilon^{-c\lambda}.$$

Open problem: For the Gaussian kernel $k(x, y) = \exp\{\|x - y\|^2 / 2\sigma^2\}$, prove an approximation result of the type $K(X, X) \approx K(X, Z) \cdot K(Z, Z)^{-1} \cdot K(Z, X)$, in terms of $\mathcal{M}$, X and Z, asymptotic in M.

References

[1] Franz J. Király, Louis Theran, Ryota Tomioka, *The Algebraic Combinatorial Approach for Low-Rank Matrix Completion.* arXiv 1211.4116, 2012.

[2] Franz J. Király, Martin Kreuzer, Louis Theran, *The Algebraic Combinatorial Approach for Low-Rank Matrix Completion.* arXiv 1211.4116, 2014.

On Geometry Underlying Borel Measures on Euclidean Space

Washington Mio

(joint work with Diego Diaz Martinez and Facundo Mémoli)

We discuss the notion of multiscale covariance tensor fields associated with a Borel probability measure α on Euclidean space $\mathbb{R}^d$, as introduced in [1]. Scale dependence is controlled by a kernel function $K(x, y, \sigma) \geqslant 0$, where $x, y \in \mathbb{R}^d$ and $\sigma > 0$ is the scale parameter. The idea is that, at scale σ, the kernel delimits the horizon of an observer positioned at x by attributing weight $K(x, y, \sigma)$ to the point y. Covariation of the weighted data is measured relative to every point $x \in \mathbb{R}^d$, not just relative to the mean as is common practice, thus giving rise to a multiscale covariance field. We prove stability and consistency theorems for covariance fields and investigate geometric properties of α that are contained in these fields.

We define the multiscale tensor field associated with a Borel measure α as the mapping $\Sigma_\alpha \colon \mathbb{R}^d \times (0, \infty) \to \mathbb{R}^d \otimes \mathbb{R}^d$ given by

$$(1) \qquad \Sigma_\alpha(x, \sigma) = \int_{\mathbb{R}^d} (y - x) \otimes (y - x) K(x, y, \sigma)\, d\mu(y),$$

provided that the integral is convergent. For any (x, σ), the 2-tensor $\Sigma_\alpha(x, \sigma)$ yields a non-negative bilinear form. If α is a probability measure, Σ_α may be interpreted as a multiscale covariance field. If $y_1, \ldots, y_n$ are $\mathbb{R}^d$-valued random variables sampled from α, the empirical covariance tensor field is the covariance field associated with the empirical measure $\widehat{\alpha}_n = \frac{1}{n} \sum_{i=1}^n \delta_{y_i}$, which is given by

$$(2) \qquad \Sigma_{\widehat{\alpha}_n}(x, \sigma) = \frac{1}{n} \sum_{i=1}^n (y_i - x) \otimes (y_i - x) K(x, y_i, \sigma).$$

In this note, we focus on the isotropic Gaussian kernel

$$G(x, y, \sigma) = \frac{1}{(2\pi\sigma^2)^{d/2}} \exp\left(\frac{\|y - x\|^2}{2\sigma^2} \right)$$

and the truncation kernel

$$T(x, y, \sigma) = \chi_{(x, \sigma)}(y),$$

where $\chi_{(x, \sigma)}$ is the characteristic function of the ball of radius σ centered at $x \in \mathbb{R}^d$.

To state the stability and consistency results, we introduce some notation. We let $m_p(\alpha)$ be the pth moment of α and

$$\Psi_\sigma(c) = c \left(\frac{4c^8}{\sigma^8} + \frac{2c^6}{\sigma^6} + \frac{13c^4}{\sigma^4} + \frac{12c^2}{\sigma^2} + 18 \right)^{1/2}.$$

Theorem 1 (Stability). *Let α and β be Borel probability measures on $\mathbb{R}^d$, $\Sigma_\alpha(x, \sigma)$ and $\Sigma_\beta(x, \sigma)$ the multiscale covariance fields of α and β associated with the Gaussian kernel, and $c > 0$. If $m_p(\alpha) < c$ and $m_p(\beta) < c$, $1 \leqslant p \leqslant 10$, then*

$$\sup_{x \in \mathbb{R}^d} \left\| \Sigma_\alpha(x, \sigma) - \Sigma_\beta(x, \sigma) \right\| \leq \frac{2\Psi_\sigma(c)}{(2\pi\sigma^2)^{d/2}}\, d_W(\alpha, \beta),$$

for all $\sigma > 0$, where d_W denotes Wasserstein 2-distance.

Theorem 2 (Consistency). *Let α be a Borel probability measure on $\mathbb{R}^d$ and $y_1, \ldots, y_n$ be i.i.d. random variables sampled from the probability measure α. If α has sufficiently many finite moments, then for each $\sigma > 0$ and $\varepsilon > 0$,*

$$\mathrm{Prob}_{\alpha^{\otimes n}}\left(\sup_{x \in \mathbb{R}^d} \left\|\Sigma_{\widehat{\alpha}_n}(x, \sigma) - \Sigma_\alpha(x, \sigma)\right\| \geq \varepsilon\right) \xrightarrow{n \uparrow \infty} 0.$$

The next results illustrate the fact that many geometric properties of submanifolds of $\mathbb{R}^d$ can be recovered from Σ_α at small scales. Let $C \subset \mathbb{R}^2$ be a smooth simple closed curve. We consider the singular measure α on $\mathbb{R}^2$ supported on C induced by arc length. For $x_0 \in C$, the arc-length parametrization of C near x_0 may be written as $X(s) = s - \frac{\kappa^2 s^3}{6} + O(s^4)$ and $Y(s) = \frac{\kappa s^2}{2} + \frac{\kappa_s s^3}{6} + O(s^4)$, where $X(s)$ and $Y(s)$ are coordinates along the tangent and normal directions to C at x_0, respectively.

Proposition 3. *Let $\sigma > 0$ be small. For any $x_0 \in C$, in the coordinates specified above, the multiscale covariance matrix for the truncation kernel is given by*

$$(3) \qquad \Sigma_\alpha(x_0, \sigma) = \begin{pmatrix} \frac{2\sigma^3}{3} - \frac{\kappa^2 \sigma^5}{20} + O(\sigma^6) & \frac{\kappa_s \sigma^5}{15} + O(\sigma^6) \\ \frac{\kappa_s \sigma^5}{15} + O(\sigma^6) & \frac{\kappa^2 \sigma^5}{10} + O(\sigma^6) \end{pmatrix},$$

In particular, the curvature of C may be recovered from Σ_α.

Now we consider a compact surface $S \subset \mathbb{R}^3$. Given a non-umbilic point $x_0 \in S$, choose a Cartesian coordinate system centered at x_0 so that the x-axis and y-axis are in the directions of maximal and minimal sectional curvatures at x_0, respectively, and the z-axis is normal to the surface at x_0.

Proposition 4. *Let $\sigma > 0$ be small, $x_0 \in S$ be non-umbilic, and let α be the surface area measure on S. Then, in the coordinate system described above, the multiscale covariance matrix for the truncation kernel is given by*

$$\Sigma_\alpha(x_0, \sigma) = \begin{bmatrix} A_1 & O(\sigma^7) & O(\sigma^8) \\ O(\sigma^7) & A_2 & O(\sigma^8) \\ O(\sigma^8) & O(\sigma^8) & A_n \end{bmatrix},$$

where $A_1 = \frac{\pi \sigma^4}{4} + \frac{\pi}{192}(-3\kappa_1^2 - 6\kappa_1\kappa_2 + \kappa_2^2)\sigma^6 + O(\sigma^7)$, $A_2 = \frac{\pi \sigma^4}{4} + \frac{\pi}{192}(\kappa_1^2 - 6\kappa_1\kappa_2 - 3\kappa_2^2)\sigma^6 + O(\sigma^7)$, $A_n = \frac{3\kappa_1^2 + 2\kappa_1\kappa_2 + 3\kappa_2^2}{96}\sigma^6 + O(\sigma^7)$, and κ_1 and κ_2 are the principal curvatures of S at x_0.

Remark. It follows from the proposition that the determinant of the covariance is $\left(3\kappa_1^2 + 2\kappa_1\kappa_2 + 3\kappa_2^2\right)\frac{\pi^3}{1536}\sigma^{14} + O(\sigma^{15})$ and the trace is $\frac{\pi}{2}\sigma^4 + \frac{\pi}{48}(\kappa_1 - \kappa_2)^2\sigma^6 + O(\sigma^7)$. Thus, κ_1 and κ_2 can be recovered from the spectrum of $\Sigma_\alpha(x_0, \sigma)$ as a function of σ. Indeed, from the Taylor expansions of these two functions one can extract the values of $(\kappa_1 - \kappa_2)^2$ and $3\kappa_1^2 + 2\kappa_1\kappa_2 + 3\kappa_2^2$, which in turn determine κ_1 and κ_2.

We conclude with some simple illustrations of applications of multiscale covariance fields (for the Gaussian kernel) to learning arrangements of submanifolds of $\mathbb{R}^d$ from data. Let $p_1, \ldots, p_n$ be points representing a subspace of $\mathbb{R}^d$ that is a finite union of connected submanifolds in general position. The goal is to cluster

the points so that each component of the arrangement is represented as a cluster. For a fixed scale σ, we represent the ith data point by the covariance tensor $\Sigma_{\widehat{\alpha}_n}(p_i, \sigma))$ and cluster these tensors using a single linkage scheme based on the Euclidean metric $d(\Sigma_1, \Sigma_2) = \|\Sigma_1 - \Sigma_2\|$. Figure 1 shows two examples of data rep-
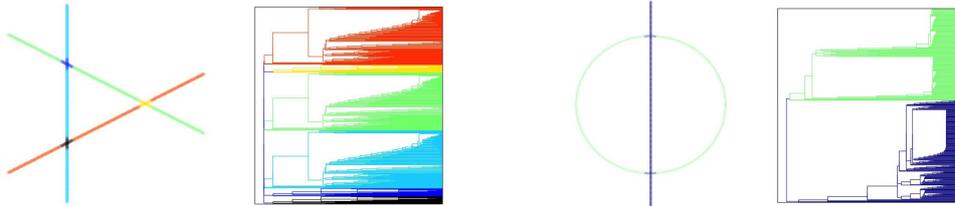

FIGURE 1. Single linkage clusterings using covariance tensors and the corresponding dendrograms.

resenting arrangements of curves in the plane and the corresponding single linkage dendrograms, which show the hierarchy of cluster fusions. The clusters displayed correspond to a fixed level in the dendrogram. Note that the method is able to separate the various components of the arrangement, as well as find neighborhoods of their intersection points. Figure 2 shows a similar example involving lines and
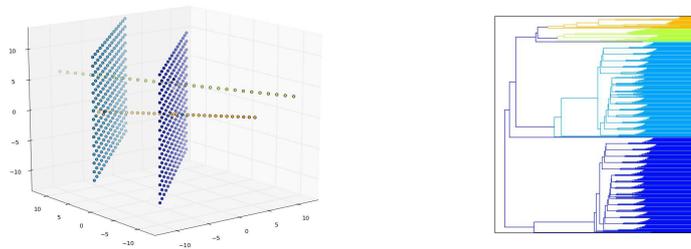

FIGURE 2. Clustering two planes and two lines.

planes in $\mathbb{R}^3$.

REFERENCES

[1] D. Diaz Martinez, F. Mémoli, W. Mio *Multiscale Covariance Fields, Local Scales, and Shape Transforms.* In Proc. Geometric Science of Information (Paris, France, 2013), LNCS 8085, 794–801, Springer, 2013.

Frölicher spaces as a setting for tree spaces and stratified spaces

PETER W. MICHOR

(joint work with Thomas Hotz and Andreas Kriegl)

Frölicher spaces were introduced under the name 'espaces lisses' (smooth spaces) in [4] and [5]; they were called Frölicher spaces in [7, section 23]. They furnish a very simple vehicle for extending the notion of smooth mappings from manifolds to singular spaces and they give a cartesian closed category.

Frölicher spaces. A *Frölicher space*, also called a *smooth space* or a space with *smooth structure*, is a triple $(X, \mathcal{C}_X, \mathcal{F}_X)$ consisting of a set X, a subset $\mathcal{C}_X$ of the set of all mappings $\mathbb{R} \to X$, and a subset $\mathcal{F}_X$ of the set of all functions $X \to \mathbb{R}$, with the following two properties:

- A function $f : X \to \mathbb{R}$ belongs to $\mathcal{F}_X$ if and only if $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$ for all $c \in \mathcal{C}_X$.
- A curve $c : \mathbb{R} \to X$ belongs to $\mathcal{C}_X$ if and only if $f \circ c \in C^\infty(\mathbb{R}, \mathbb{R})$ for all $f \in \mathcal{F}_X$.

Note that a set X together with any subset $\mathcal{F}$ of the set of functions $X \to \mathbb{R}$ generates a unique Frölicher space $(X, \mathcal{C}_X, \mathcal{F}_X)$, where we put in turn:

$$\mathcal{C}_X := \{c : \mathbb{R} \to X : f \circ c \in C^\infty(\mathbb{R}, \mathbb{R}) \text{ for all } f \in \mathcal{F}\},$$
$$\mathcal{F}_X := \{f : X \to \mathbb{R} : f \circ c \in C^\infty(\mathbb{R}, \mathbb{R}) \text{ for all } c \in \mathcal{C}_X\},$$

so that $\mathcal{F} \subseteq \mathcal{F}_X$. The set $\mathcal{F}$ will be called a *generating set of functions for the Frölicher space*.

Likewise, a set X together with any subset $\mathcal{C}$ of the set of curves $\mathbb{R} \to X$ generates a unique Frölicher space $(X, \mathcal{C}_X, \mathcal{F}_X)$, where we put in turn:

$$\mathcal{F}_X := \{f : X \to \mathbb{R} : f \circ c \in C^\infty(\mathbb{R}, \mathbb{R}) \text{ for all } c \in \mathcal{C}\},$$
$$\mathcal{C}_X := \{c : \mathbb{R} \to X : f \circ c \in C^\infty(\mathbb{R}, \mathbb{R}) \text{ for all } f \in \mathcal{F}_X\},$$

so that $\mathcal{C} \subseteq \mathcal{C}_X$. The set $\mathcal{C}$ will be called a *generating set of curves for the Frölicher space*.

Smooth mappings. A mapping $\phi : X \to Y$ between two Frölicher spaces is called *smooth* if one of the following three equivalent conditions hold:

- For each $c \in \mathcal{C}_X$ the composite $\phi \circ c$ is in $\mathcal{C}_Y$.
- For each $f \in \mathcal{F}_Y$ the composite $f \circ \phi$ is in $\mathcal{F}_X$.
- For each $c \in \mathcal{C}_X$ and for each $f \in \mathcal{F}_Y$ the composite $f \circ \phi \circ c$ is in $C^\infty(\mathbb{R}, \mathbb{R})$.

Note that $\mathcal{F}_Y$ can be replaced by any generating set, as well as $\mathcal{C}_X$. The set of all smooth mappings from X to Y will be denoted by $C^\infty(X, Y)$. Then we have $C^\infty(\mathbb{R}, X) = \mathcal{C}_X$ and $C^\infty(X, \mathbb{R}) = \mathcal{F}_X$. Obviously, Frölicher spaces and smooth mappings form a category.

Theorem. [7, 23.2] *The category of Frölicher spaces and smooth mappings has the following properties:*

- *Complete, i.e., arbitrary limits exist. The underlying set is formed as in the category of sets as a certain subset of the cartesian product, and the smooth structure is generated by the smooth functions on the factors.*
- *Cocomplete, i.e., arbitrary colimits exist. The underlying set is formed as in the category of sets as a certain quotient of the disjoint union, and the smooth functions are exactly those which induce smooth functions on the cofactors.*
- *Cartesian closedness, which means: The set $C^\infty(X,Y)$ carries a canonical smooth structure generated by all functions of the form*

$$C^\infty(X,Y) \xrightarrow{\;C^\infty(c,f)\;} C^\infty(\mathbb{R},\mathbb{R}) \xrightarrow{\;\lambda\;} \mathbb{R}$$

where $c \in C^\infty(\mathbb{R},X)$ and $f \in C^\infty(Y,\mathbb{R})$, or in a generating sets, and where $\lambda \in C^\infty(\mathbb{R},\mathbb{R})'$. With this structure the exponential law holds:

$$C^\infty(X \times Y, Z) \cong C^\infty(X, C^\infty(Y,Z)).$$

Corollary. *Canonical mappings are smooth, for Frölicher spaces X, Y, Z:*

$$\mathrm{ev} : C^\infty(X,Y) \times X \to Y, \quad \mathrm{ev}(f,x) = f(x)$$
$$\mathrm{ins} : X \to C^\infty(Y, X \times Y), \quad \mathrm{ins}(x)(y) = (x,y)$$
$$(\;\;)^\wedge : C^\infty(X, C^\infty(Y,Z)) \to C^\infty(X \times Y, Z)$$
$$(\;\;)^\vee : C^\infty(X \times Y, Z) \to C^\infty(X, C^\infty(Y,Z))$$
$$\mathrm{comp} : C^\infty(Y,Z) \times C^\infty(X,Y) \to C^\infty(X,Z)$$
$$C^\infty(\;\;,\;\;) : C^\infty(Y,Y_1) \times C^\infty(X_1,X) \to$$
$$\to C^\infty(C^\infty(X,Y), C^\infty(X_1,Y_1))$$
$$(f,g) \mapsto (h \mapsto f \circ h \circ g)$$

Natural topologies on Frölicher spaces. [3, section 1] On a Frölicher space $(X, \mathcal{C}_X, \mathcal{F}_X)$ we consider the following two topologies:

- The final topology with respect to all smooth curves in $\mathcal{C}_X$; it is denoted by $\tau_\mathcal{C}$.
- The inital topology with respect to all smooth functions in $\mathcal{F}_X$; we denote it by $\tau_\mathcal{F}$

The identity mapping $(X, \tau_\mathcal{C}) \to (X, \tau_\mathcal{F})$ is obviously continuous. A Frölicher space is called *balanced* if these two topologies coincide and are Hausdorff.

Related concepts.

- *Holomorphic Frölicher spaces.* As curves one has to take mappings from the complex unit disk $\mathbb{D}$, and complex valued functions such that each composition is holomorphic $\mathbb{D} \to \mathbb{C}$. Stein manifolds are holomophic Frölicher spaces whereas compact complex manifolds are not. See [7, 23.5].
- *Sikosrki spaces.* Here one specifies an algebra of 'smooth' functions with certain properties. One can also specify sheafs of 'smooth' functions.

- *Diffeological spaces.* Here one specifies mappings from open sets in all $\mathbb{R}^n$'s with appropriate conditions. These were introduces by Sourieau, see the recent book [6].

There are natural functors from the categories of Sikorski spaces and of diffeological spaces into the category of Frölicher spaces, which are right and left adjoints. See [9] for a comparison.

Theorem. *Tree spaces in the sense of [1] are balanced Frölicher spaces.*

This follows from the fact that a tree space $\mathcal{T}$ is always a closed subspace of $\mathbb{R}^N$, where different quadrants always meet at non-trivial angles. As generating set of functions one can take the restrictions of linear functions on $\mathbb{R}^N$. This is called the standard Frölicher structure.

The following two examples are fundamental to understanding tree spaces, see e.g. [8].

Example of a treespace: the 3-spider.
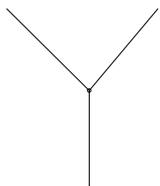
A generating set of functions consists of all linear functions on $\mathbb{R}^2$ or on $\mathbb{R}^3$. Smooth curves in $\mathcal{C}_X$ then have to stop in all derivatives when they change sheets. Functions $f \in \mathcal{F}_X$ are then smooth on each closed sheet.

The open book as part of tree space.
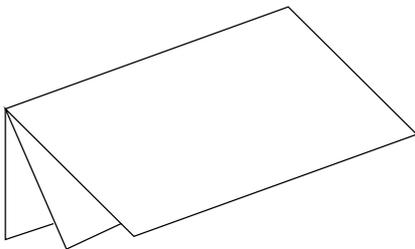
A generating set of functions consists again of all linear functions on $\mathbb{R}^n$. Smooth curves in $\mathcal{C}_X$ can meet the spine S only tangentially; more precisely, the first non-vanishing derivative of the normal component has to be of even order. Functions $f \in \mathcal{F}_X$ are smooth on each closed sheet.

Example of a non-Hausdorff orbit space: adjoint action of $SL(2,\mathbb{R})$.
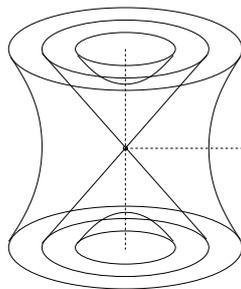
The adjoint action of $SL(2,\mathbb{R})$ on its Lie algebra $\mathfrak{sl}(2,\mathbb{R})$ has as orbits the connected components of the 'spheres' with respect to the Killing form, which is isomorphic to Minkowski space $\mathbb{R}^{1,2}$. The orbits are as follows. The double light cone decomposes in three orbits: the future light cone, the past one (these two are not closed orbits), and 0.
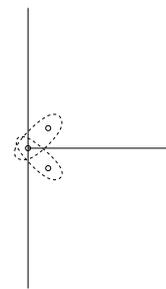
The other orbits are: The two parts of each two-sheeted hyperboloid, and the one sheeted hyperboloids. The orbit space X can be visualized as a vertical line, a horizontal half-line, and two further points (corresponding to the open light cones)

which cannot be separated in the quotient topology from the intersection point depicting the equivalence class of 0.

The structure of a Frölicher space on X is generated by the set $\mathcal{C}$ of projections to X of all smooth curves in $\mathbb{R}^{1,2}$. A smooth curve can go from the vertical half-line through one of the nonclosed orbits to the horizontal half-line, but through 0 it can only go infinitely flat (in $\mathbb{R}^{1,2}$). The functions $f \in \mathcal{F}_X$ are those such that $f \circ \pi$ is in $C^\infty(\mathbb{R}^{1,2}, \mathbb{R})$. The topology $\tau_{\mathcal{F}}$ is strictly coarser then the quotient topology: The closure of each non-closed point contains all 3 points. We get curves in $\mathcal{C}_X$ which are not in $\mathcal{C}$, namely, a curve in $\mathcal{C}_X$ can now also go smoothly with nontrivial speed through 0 from vertical to horizontal. The final topology $\tau_{\mathcal{C}}$ is finer: the two non-closed points become closed, too; so $\tau_{\mathcal{C}}$ is T_1 but still not T_2. The space X is not balanced.

The geodesic Frölicher structure on tree spaces. Then we can put the following Frölicher structure on X: Let us take as generating set $\mathcal{C}$ the union of the space $\mathcal{C}_X$ of smooth curves for the standard Frölicher structure on X with the set of all curves $\gamma : \mathbb{R} \to X$ such that $s \mapsto \gamma(\tan(s)) = \tilde{\gamma}(s)$ is a geodesic between the points $\tilde{\gamma}(-\pi/2)$ and $\tilde{\gamma}(\pi/2)$ which is parameterized proportional to arclength. That means, we put:

$$\mathcal{F}_X^{\mathrm{geo}} = \{f : X \to \mathbb{R} \,:\, f \circ \gamma \in C^\infty(\mathbb{R}, \mathbb{R}) \,\forall \gamma \in \mathcal{C}\},$$
$$\mathcal{C}_X^{\mathrm{geo}} = \{c : \mathbb{R} \to X \,:\, f \circ c \in C^\infty(\mathbb{R}, \mathbb{R}) \,\forall f \in \mathcal{F}_X^{\mathrm{geo}}\}.$$

Then $(X, \mathcal{C}_X^{\mathrm{geo}}, \mathcal{F}_X^{\mathrm{geo}})$ is a Frölicher space by the general construction.

We define $T_x^i X$ as the quotient of $\{c \in \mathcal{C}_X^{\mathrm{geo}} \,:\, c(0) = x\}$ by the equivalence relation $c_1 \sim c_2 \iff (f \circ c_1)'(0) = (f \circ c_2)'(0) \,\forall f \in \mathcal{F}_X^{\mathrm{geo}}$, and call this the inner tangent space at $x \in X$. For a tree-space $T_x^i X$ is the tangent space at x of the stratum containing x in its interior.

We may define $T_x^c X$ as the quotient of the set of all geodesics $\gamma : [0, 1] \to X$ with $\gamma(0) = x$, parameterized proportional to arclength, modulo the equivalence relation $\gamma_1 \sim \gamma_2 \iff \gamma_1 = \gamma_2$ near 0. We call $T_x^c X$ the *conical tangent space*. It contains all vectors pointing from x into higher strata which are bounded by the stratum of x.

Geodesic Frölicher structures on certain metric spaces. Let X be a geodesic metric space, i.e., between any two points there exists a unique geodesic realizing the distance (see e.g. [2]).

If we generate a Frölicher structure only by the set $\mathcal{C}$ of geodesics, even in $\mathbb{R}^n$ we do not get the usual structure. Besides C^∞-function we also get homogeneous rational functions in $\mathcal{F}_X$, and more.

Let us take as generating set $\mathcal{F}$ of functions squares of geodesic distances $x \mapsto d(y_i, x)^2$, where y_i runs through a subset of points in X. If $X = \mathbb{R}^n$ and y_i are $n + 1$ generic points, the resulting Frölicher structure is the usual one. If X is a tree-space, the resulting Frölicher structure seems to be the $(\mathcal{C}_X^{\mathrm{geo}}, \mathcal{F}_X^{\mathrm{geo}})$ structure.

References

[1] L. J. Billera, S. P. Holmes, and K. Vogtmann, *Geometry of the space of phylogenetic trees*, Adv. in Appl. Math., **27(4)** (2001), 733–767.

[2] M. R. Bridson and A. Haefliger, *Metric spaces of non-positive curvature*, volume 319 of *Grundlehren der Mathematischen Wissenschaften*, Springer-Verlag, Berlin (1999).

[3] A. Cap, *K-theory for convenient algebras*, PhD thesis, Universität Wien (1991). `http://www.mat.univie.ac.at/~michor/cap_diss.pdf`.

[4] A. Frölicher, *Categories cartésiennement fermées engendrées par des monoides*, Cahier Top. Géom. Diff., **21** (1980), 367–375.

[5] A. Frölicher, *Applications lisses entre espaces et variétés de Fréchet*, C. R. Acad. Sci. Paris **293** (1981), 125–127.

[6] P. Iglesias-Zemmour, *Diffeology*, volume 185 of *Mathematical Surveys and Monographs*, American Mathematical Society, Providence, RI (2013).

[7] A. Kriegl, P. W. Michor. *The Convenient Setting for Global Analysis*, Surveys and Monographs 53, AMS, Providence (1997).

[8] T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, S. Skwerer, *Sticky central limit theorems on open books*, The Annals of Applied Probability **23(6)** (2013), 2238–2258.

[9] A. Stacey. Comparative smootheology. *Theory Appl. Categ.*, 25:No. 4, 64–117, 2011.

Diffusion Processes and PCA on Manifolds

Stefan Sommer

Approaching dimensionality reduction on differentiable manifolds with affine connection from a probabilistic viewpoint, we develop a generalization of Principal Component Analysis (PCA) that does not rely on parametric representations of principal submanifolds. The method fits a class of diffusions processes arising as horizontal stochastic flows in the frame bundle to observed data by maximum likelihood. The probabilistic interpretation removes the reliance of previous methods on explicitly constructed submanifolds that are not totally geodesic. In addition, projections to dense geodesics are avoided thus giving a well-defined construction on tori where projections do not exist.

1. Background

Conventional PCA uses the inner product structure of Euclidean space, a fact that makes generalization of the procedure to differentiable manifolds and stratified spaces difficult. Existing non-Euclidean extensions of PCA include Principal Geodesic Analysis (PGA, [1]) that parametrizes low-dimensional principal components with geodesic sprays from a Frechét mean; Geodesic PCA (GPCA, [2]) that finds principal geodesic curves minimizing residual errors; Principal Nested Spheres (PNS, [3]) that finds low-dimensional spheres; and Horizontal Component Analysis (HCA, [4]) that uses development of curves in the frame bundle of the manifold to construct horizontal subspaces that project to the manifold. These approaches capture different properties of Euclidean PCA but they are all limited by the fact that totally geodesic submanifolds do not exist in general in

non-Euclidean spaces. This is in contrast to the Euclidean case where linear subspaces are totally geodesic. In addition, subspaces maximizing captured variance are not equivalent to subspaces minimizing residual errors. Even for distributions with local support, recurring and dense geodesics can make projections minimizing residual errors undefined [5].

2. Diffusion PCA

The principal subspaces that in Euclidean PCA maximize variance of orthogonally projected data samples can be found by eigendecomposing the sample covariance matrix C. The principal axes are given by unit eigenvectors u_j of C, and with $U = (u_1, \ldots, u_d)$, the principal components $x_n = U^T(y_n - \mu)$ are projections of the centered data to the span of the principal axes. Conventionally, u_j are ordered according to decreasing eigenvalues.

In [6], a probabilistic formulation of PCA was developed as a maximum likelihood estimate (MLE) of the matrix W in the latent variable model

$$y = Wx + \mu + \epsilon \ .$$

The latent variables x are assumed normally distributed $\mathcal{N}(0, I)$, and, in contrast to factor analysis, the noise ϵ is isotropic $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$. The marginal distribution of the observed data is then again Gaussian $\mathcal{N}(\mu, C_\sigma)$, $C_\sigma = WW^T + \sigma^2 I$. The MLE for W is up to rotation given by $W_{ML} = U(\Lambda - \sigma^2 I)^{1/2}$, $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_d)$ with the usual PCA solution recovered as $\sigma^2 \to 0$. The principal components x_n are defined as the mean of x conditional on the sample y_n and given by $E[x_n|y_n] = (W^T W + \sigma^2 I)^{-1} W_{ML}^T(y_n - \mu)$. This definition again approaches the orthogonal projections used in PCA as $\sigma^2 \to 0$.

Importantly for our purpose, the probabilistic formulation makes no reference to linear subspaces. Thus a generalization to nonlinear manifolds can be obtained without constructing submanifolds that in general cannot be totally geodesic. The main difficulty in generalizing the method instead lies in the latent variable model being additive and using normal distributions, neither which are directly transferable to manifolds.

We now define *diffusion PCA* (DPCA) that rephrases the latent variable model using stochastic paths and diffusions processes. This formulation naturally extends to differentiable manifolds with affine connection. Let W_t be a Wiener process in $\mathbb{R}^d$ and let X_t be given by the $\mathbb{R}^n$ valued stochastic differential equation (SDE) $dX_t = \sigma \circ W_t$ with source $X_0 = \delta_{(0,\ldots,0)}$. Here the $n \times d$ matrix σ is stationary so that X_t is a driftless diffusion with infinitesimal generator $\sigma \sigma^T$. Through the process of stochastic development [7], the process maps to a stochastic process U_t in the frame bundle FM of a manifold M^n with affine connection: If $H_1, \ldots, H_n$ are the horizontal vector fields on FM, U_t satisfies $dU_t = H_i(U_t)\sigma_j^i \circ dW_t^j$, and the source is a point $(p, u) \in FM$, $p \in M$. Trough the bundle projection $\pi : FM \to M$, U_t projects to a manifold valued diffusion $\pi(U_t)$ which is unique given its generator L and initial distribution [7].

(a) sample paths, elliptic hyperboloid, $-2x^2 + y^2 = -z^2$

(b) sample paths, cylinder, $x^2 + 0y^2 = -z^2$

(c) sample paths, sphere, $x^2 + y^2 = -z^2$

(d) $\mathbb{R}^2 \cong T_p M$, mean path (red/dark grey) and density from (a)

(e) $\mathbb{R}^2 \cong T_p M$, mean path (red/dark grey) and density from (b)

(f) $\mathbb{R}^2 \cong T_p M$, mean path (red/dark grey) and density from (c)
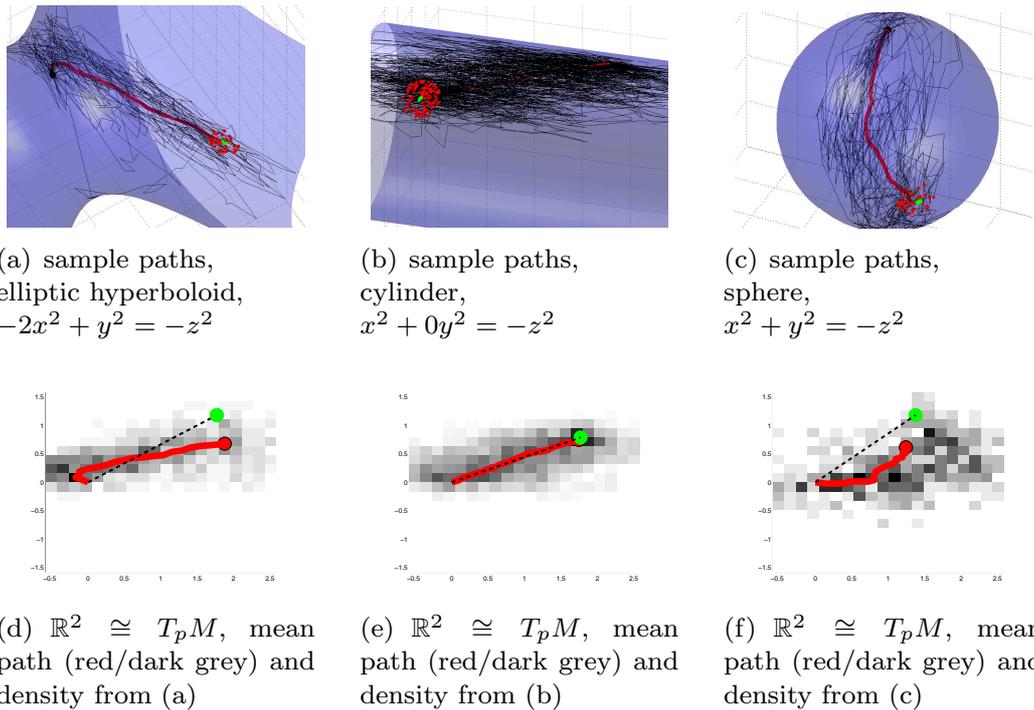
FIGURE 1. Anisotropic diffusion on quadratic hypersurfaces, source $p = (0, 0, 1)$; var. major axis: 4; minor axis: 1.2. (a),(b),(c): sample paths ending near $\mathbf{x} = \mathrm{Exp}_p x$, $x \in T_p M$ (green/light grey dot). (d),(e),(f): mean sample path anti-development $\hat{x}_i(t)$ (red/dark grey) with path densities (background) and shortest path from source $(0, 0)$ to x (dashed line). The paths all reach $\mathbf{x}$ through different anti-developments in $\mathbb{R}^2$. Negative curvature (a),(d): mean path deviates from shortest path by moving along minor diffusion axis before major axis. Zero curvature (b),(e): mean path and shortest path align. Positive curvature (c),(f): mean path aligns with major diffusion axis before moving along minor axis. This case resembles the HCA construction, see text.

Consider the map Diff. : $FM \to \mathrm{Dens}(M)$ that maps (p, u) to $\pi(U_1)$ where the FM diffusion $dU_t = H_i(U_t) \circ dW_t^i$ is started at (p, u) and $W_t \in \mathbb{R}^n$. We let $\Gamma \subset \mathrm{Dens}(M)$ be the image Diff.(FM), i.e. the set of densities resulting from point-sourced diffusions in FM stopped at time $t = 1$. In diffusion PCA, the observed data is assumed to be distributed according to $\mu \in \Gamma$ so that $y \sim \pi(U_1)$ for a diffusion $U_1 \in FM$.

Let μ_0 be a fixed measure on M (e.g. a Riemannian volume form) and for $\mu = p\mu_0 \in \Gamma$ define the log-likelihood

$$\ln \mathcal{L}(\mu) = \sum_{i=1}^{N} \ln p(y_i)$$

for a set of samples $y_1, \ldots, y_N \in M$. Now let $(p, u) \in FM$ be a maximum for $\ln \mathcal{L}(\text{Diff.}(p, u))$. Then (p, u) is an MLE of y_i generalizing the probabilistic PCA formulation to the non-Euclidean case. We denote (p, u) a diffusion PCA. Note that Diff. is not injective as multiple frames at p may lead to the same diffusion process. The MLE is hence not unique though a different formulation of Diff. can correct this (see open questions below).

3. The Principal Components

In probabilistic PCA, the mean of the latent variables conditional on the observed data $E[x_n|y_n]$ converges to the principal components as $\sigma^2 \to 0$. With non-zero curvature, single vectors cannot summarize the observations in this way because of the path dependences of the diffusion, see Figure 1. Instead, the mean sample paths reaching y_i $\hat{x}_i(t) = E[x(t)|x(1) = y_i]$ take the role of the latent variables in probabilistic PCA. Note that given the source $(p, u) \in FM$, the sample paths can be equivalently viewed as paths on M or as paths in $\mathbb{R}^n$ trough (anti-)development. Examples of mean paths are illustrated in Figure 1. In $\mathbb{R}^n$, the data can be further summarized by integrating out the time dependence from $\hat{x}_i$ giving $\tilde{x}_i = \int_0^1 \frac{d}{dt}\hat{x}_i(t)dt = \hat{x}_i(1)$.

If M is Riemannian, let $u_0 \in OM$ be an orthonormal frame at p such that the matrix of uu^T in the u_0 basis is diagonal with decreasing diagonal. Anti-developing $\hat{x}_i$ with base (p, u_0) gives $\mathbb{R}^n$ valued paths with the major variation residing in the low coordinates. The vectors $\tilde{x}_i$ here provide a Euclideanization of the data similar to those provided by PGA, GPCA, and HCA. In general, the linearization will differ from the linearizations provided by the existing methods.

In effect, the complicated geometric problems arising when defining parametric subspaces of non-linear manifolds and projecting data are removed with the probabilistic approach. In particular, recurring and dense geodesics on tori prevent a well-defined notion of projection as closest point on a geodesic as required by existing methods. With diffusion PCA, the mean path $\hat{x}_i$ is defined without projecting to a submanifold. Note in Figure 1 (f) that $\hat{x}_i$ resembles the axis-aligned curves of HCA [4] indicating a new characterization of HCA as approximating diffusion processes in positively curved spaces.

4. The Geometry of Γ

The set $\text{Dens}(M) = \{\mu \in \Omega^n(M) : \int_M \mu = 1, \mu > 0\}$ is an infinite dimensional manifold in the Fréchet topology of smooth functions [8], and it can be equipped with the Fisher-Rao metric $G_\mu^{FR}(\alpha, \beta) = \int_M \frac{\alpha}{\mu}\frac{\beta}{\mu}\mu$ [9]. In information geometry, finite dimensional submanifolds of $\text{Dens}(M)$ are called statistical manifolds. In coordinates, the matrix form of the metric is the Fisher Information Matrix. If Γ locally has the structure of a statistical manifold, the MLE in diffusion PCA can be found by a gradient flow with respect to the gradient inherited from the

Fisher-Rao metric, i.e.

$$(1) \qquad \dot{\theta}_s = -\nabla_\theta \left(\sum_{i=1}^{N} \ln p_\theta(y_i) \right)$$

where θ is a local chart for Γ and p_θ is the distribution for a given value of θ.

5. OPEN QUESTIONS

At the workshop, we identified a number of questions related to the diffusion PCA construction that are open for further research. These include finding the precise topological and geometric structure of Γ as a subset of $\mathrm{Dens}(M)$. We conjecture that Γ in the Riemannian case can be parametrized by the bundle of symmetric positive-definite covariant tensors of order 2. Conditions for the convergence of the gradient flow (1) to the MLE estimate needs to be identified. In addition, a scaled Brownian motion can be added to the diffusion PCA model similarly to the isotropic error ϵ in probabilistic PCA. The exact form of this construction and the convergence as $\sigma^2 \to 0$ remains to be explored.

REFERENCES

[1] P.T. Fletcher, Conglin Lu, S.M. Pizer, and Sarang Joshi, *Principal geodesic analysis for the study of nonlinear statistics of shape*, Medical Imaging, IEEE Transactions on (2004).

[2] Stephan Huckemann, Thomas Hotz, and Axel Munk, *Intrinsic shape analysis: Geodesic PCA for riemannian manifolds modulo isometric lie group actions*, Statistica Sinica **20** (2010), no. 1, 1–100.

[3] Sungkyu Jung, Ian L. Dryden, and J. S. Marron, *Analysis of principal nested spheres*, Biometrika **99** (2012), no. 3, 551–568 (en).

[4] Stefan Sommer, *Horizontal dimensionality reduction and iterated frame bundle development*, Geometric Science of Information (Frank Nielsen and Frederic Barbaresco, eds.), Lecture Notes in Computer Science, Springer Berlin Heidelberg, January 2013, pp. 76–83.

[5] Stephan Huckemann, Thomas Hotz, and Axel Munk, *Rejoinder to the discussion of ???intrinsic shape analysis: Geodesic principal component analysis for riemannian manifolds under isometric lie group actions???*, Statistica Sinica **20** (2010), no. 1.

[6] Michael E. Tipping and Christopher M. Bishop, *Probabilistic principal component analysis*, Journal of the Royal Statistical Society. Series B (Statistical Methodology) **61** (1999), no. 3, 611–622.

[7] Elton P. Hsu, *Stochastic analysis on manifolds*, American Mathematical Soc., 2002 (en).

[8] Richard S. Hamilton, *The inverse function theorem of nash and moser*, Bulletin of the American Mathematical Society **7** (1982), no. 1, 65–222.

[9] Klas Modin, *Information geometry and the Fisher-Rao metric on the space of probability distributions*, Math in the Cabin - Shape Analysis Workshop in Bad Gastein (Bad Gastein, Austria), July 2014, `https://hal.archives-ouvertes.fr/hal-01076953`.

Multiple Principal Components Analysis in Tree Space

MEGAN OWEN

(joint work with Sean Cleary, Aasa Feragen, Daniel Vargas)

Tree-shaped data arise in a number of applications, including images of anatomical trees in medical imaging, and phylogenetic (evolutionary) trees in evolutionary biology. One question associated with this data is how to understand the resulting distributions. Through the use of the Billera-Holmes-Vogtmann (BHV) treespace [2] as a geometric framework for statistical analysis of tree-shaped data, the mean, variance, and first principal component have been defined, and algorithms given for their computations [1, 3, 5–7]. Here we present a method for defining and computing multiple principal components.

The tree data objects we consider are leaf-labelled trees, with no interior vertices of degree 2, and with an m-dimensional vector associated with each edge, containing information about the length, shape, or other properties of the edge. The BHV treespace $\mathcal{T}_n$ contains all such trees with a fixed leaf-label set of size n. It is a right-angled polyhedral complex, in which the points in each orthant correspond to all trees with the same topology, or branching order, but different edge lengths. See Figure 1. The metric is that induced by the Euclidean metric on each orthant. The BHV treespace is a globally non-positively curved space (CAT(0) space) [2], and so has unique geodesics.
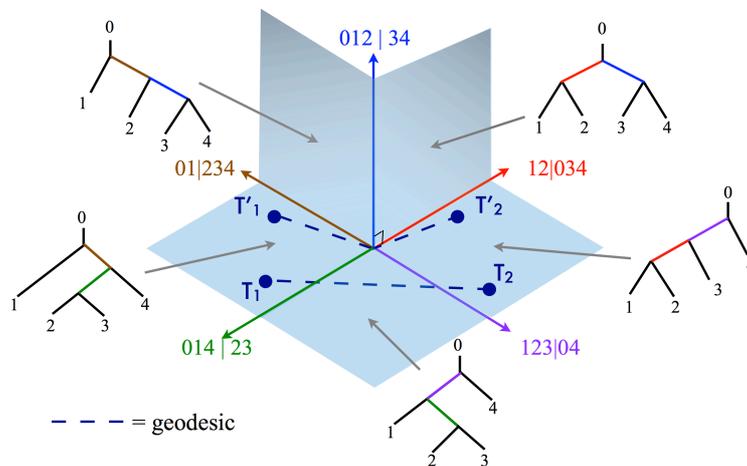
FIGURE 1. Five of the 15 right-angled orthants in the treespace $\mathcal{T}_5$, with a 1-dimensional vector associated with each edge of each tree, representing the length of the edge. The trees T_1 and T'_1, and also T_2 and T'_2, have the same topologies, but different branch lengths, resulting in different geodesics between the pairs (T_1, T_2) and (T'_1, T'_2) .

We consider a set of points $X = \{x_1, ..., x_r\}$ in $\mathcal{T}_n$. The first principal component of X in $\mathcal{T}_n$ has been previously defined as the geodesic γ minimizing the sum of

squared distances between x_i and the projection of x_i onto γ, denoted by $Pr_\gamma(x_i)$ [3, 7]. That is, define the first principal component of X, $PC1(X)$ as

$$(1) \qquad PC1(X) = argmin_\gamma \sum_{i=1}^{r} d(x_i, Pr_\gamma(x_i))^2$$

This follows a similar definition of the first principal component on manifolds [4,8]. In practice, this optimization is done over geodesic segments, parametrized by their end points, instead of infinite lines.

In Euclidean space, we can compute the second principal component by projecting all data points onto a hyperplane orthogonal to the first principal component, and then computing the first principal component in this lower dimensional space. In the BHV treespace, however, we can not extend the hyperplane into all the necessary orthants in an consistent fashion. Because of this and other effects of the stratification of the space, we choose to instead define multiple second principal components.

1. SECOND PRINCIPAL COMPONENTS

To compute the second principal components, we wish to ignore, as much as possible, the variation in the direction of the first principal component. We thus group the data points X by where they project onto the first principal component (Figure 2). For each group j, or *bin*, we compute use the algorithm GeoPhytter [7] to compute the first principal component of just these data points. This becomes the jth second principal component. This algorithm can be repeated for each second principal component to generate third principal components, and so on, if desired.
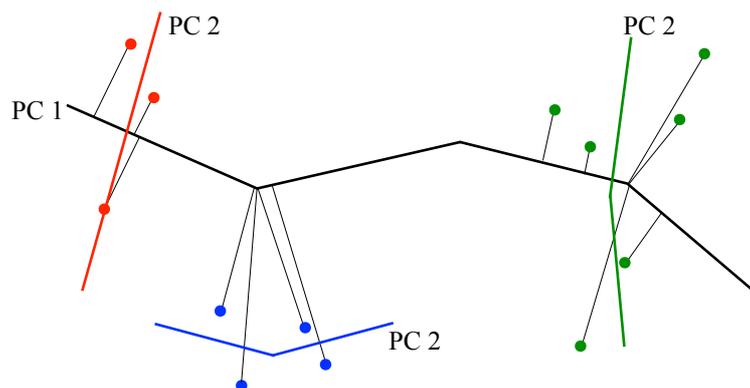
FIGURE 2. A cartoon of the binning process. The red, blue, and green circles represent the data points. They are grouped into bins according to where they project on the first principal component. The second principal components for each bin are represented by lines of the same colour as the bin's data points.

The specifics of how the points are grouped is as follows. The data points X_i are projected onto the first principal component, which is equivalent to projecting

them onto a line in the Euclidean case. We then use the 1-dimensional clustering method of smoothing the projected points using kernel density estimation (KDE). The minima of the smoothed function are chosen to delineate the bins. Any bins containing too few points are merged with neighboring bins.

REFERENCES

[1] M. Bacak. Computing medians and means in Hadamard spaces. To appear SIAM J. Optim., *preprint http://arxiv.org/abs/1210.2145*, 2014.
[2] L. J. Billera, S. P. Holmes, and K. Vogtmann. Geometry of the space of phylogenetic trees. *Advances in Applied Mathematics*, 27(4):733–767, 2001.
[3] A. Feragen, M. Owen, J. Petersen, M.M.W. Wille, L.H. Thomsen, A. Dirksen, and M. de Bruijne. Tree-space statistics and approximations for large-scale analysis of anatomical trees. In *Information Processing in Medical Imaging*, Lecture Notes in Computer Science, Vol. 7917, pp 74-85, 2013.
[4] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: geodesic PCA for Riemannian manifolds modulo isometric Lie group actions. *Statistica Sinica*, 20(1):1–58, 2010.
[5] E. Miller, M. Owen, J.S. Provan. Polyhedral computational geometry for averaging metric phylogenetic trees. *Preprint, http://arxiv.org/abs/1211.7046*, 2012.
[6] T. M. W. Nye. Principal components analysis in the space of phylogenetic trees. *Annals of Statistics*, 39(5):2716–2739, 2011.
[7] T.M.W. Nye. An Algorithm for Constructing Principal Geodesics in Phylogenetic Treespace. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 11, no. 2, pages 304-315, 2014.
[8] S. Sommer. Horizontal Dimensionality Reduction and Iterated Frame Bundle Development. *Geometric Science of Information (GSI)*, Lecture Notes in Computer Science, Vol. 8085, pp 76-83, 2013.

Asymptotic confidence sets for the Fréchet mean on the 3-spider

THOMAS HOTZ

(joint work with Huiling Le)

Using previously obtained results for the asymptotic distribution of the Fréchet mean [2], we derive corresponding asymptotic confidence regions. In order to allow for a generalisation to other stratified spaces such as the tree spaces in [1], we describe the general construction – which is commonly used in statistics – in detail.

Constructing asymptotic confidence regions as acceptance regions of asymptotic tests. We consider the following abstract framework: let $\mathcal{P}$ be a family of probability distributions on a measure space $(\Omega, \mathcal{A})$. We are interested in some parameter $\theta \in \Theta$ depending on the unknown distribution $\mathbf{P} \in \mathcal{P}$ via some parameter function $L : \mathcal{P} \to \Theta$, $L(\mathbf{P}) = \theta$, where Θ denotes the parameter set. For $n \in \mathbb{N}$, our observations are modelled by some random variable $X^{(n)} : \Omega \to \mathfrak{X}^{(n)}$ taking values in some measure space $(\mathfrak{X}^{(n)}, \mathcal{B}^{(n)})$; typically, when assuming independent and identically distributed (i.i.d.) data, n is the number of observations and $\mathfrak{X}^{(n)}$ is an n-fold product space. We aim to construct an asymptotic $(1 - \alpha)$-confidence region for θ, where $1 - \alpha \in (0, 1)$ is the fixed, prespecified confidence level.

Now assume that for each $\mathbf{Q} \in \mathcal{P}$ we are given a (measurable) test statistic $T_n^{\mathbf{Q}} : \mathfrak{X}^{(n)} \to \mathfrak{T}^{\mathbf{Q}}$, taking values in some measure space $(\mathfrak{T}^{\mathbf{Q}}, \mathcal{C}^{\mathbf{Q}})$, whose asymptotic distribution under $\mathbf{Q}$ when n tends to infinity is known, i.e. $T_n^{\mathbf{Q}} = T_n^{\mathbf{Q}}(X^{(n)})$ converges under $\mathbf{Q}$ in distribution to some random variable $T_*^{\mathbf{Q}}$. Then, picking sets $M^{\mathbf{Q}}$ with $\mathbf{Q}(T_*^{\mathbf{Q}} \in M^{\mathbf{Q}}) \geq 1 - \alpha$, we get

$$\mathbf{Q}(T_n^{\mathbf{Q}} \in M^{\mathbf{Q}}) \to \mathbf{Q}(T_*^{\mathbf{Q}} \in M^{\mathbf{Q}}) \geq 1 - \alpha.$$

One therefore may reject the hypothesis $\mathbf{P} = \mathbf{Q}$ if $T_n^{\mathbf{Q}} \notin M^{\mathbf{Q}}$, resulting in the probability of falsely rejecting the hypothesis converging to a value $\leq \alpha$, i.e. one thereby obtains an asymptotic test of significance level α.

From this, one immediately constructs an asymptotic confidence region for θ as the acceptance region of the tests, *viz.* all $\mathbf{Q}$ corresponding to hypotheses which do not get rejected given the data at hand. More precisely, setting

$$C_n = \{L(\mathbf{Q}) \,:\, T_n^{\mathbf{Q}} \in M^{\mathbf{Q}}\},$$

one gets

$$\mathbf{P}(\theta \in C_n) \geq \mathbf{P}(T_n^{\mathbf{P}} \in M^{\mathbf{P}}) \to \mathbf{P}(T_*^{\mathbf{P}} \in M^{\mathbf{P}}) \geq 1 - \alpha,$$

i.e. the probability of C_n covering the true parameter θ will in the limit be at least $1 - \alpha$, in other words C_n is indeed an asymptotic $(1 - \alpha)$-confidence region.

The Fréchet mean on the 3-spider. The *3-spider* is the metric graph obtained by glueing three real half-lines together at their end-points. More formally, it is the set

$$S = \{0\} \cup \big(\{1, 2, 3\} \times (0, \infty)\big)$$

with the metric specified by

$$d\big((j, x), 0\big) = x \quad \text{and} \quad d\big((j, x), (k, y)\big) = \begin{cases} |x - y| & \text{if } j = k, \\ x + y & \text{else.} \end{cases}$$

The point 0 is called the *singular* point, $S \setminus \{0\}$ which comprises the three open half-lines contains the *regular* points.

We consider i.i.d. random elements $X_1, \ldots, X_n$ on S which are square-integrable, i.e. for which $\mathbf{E}^{\mathbf{P}} d(X_1, 0)^2 = \int_S d(x, s)^2 \, d\mathbf{P}^{X_1}(x)$ is finite. We are interested in constructing an asymptotic confidence region for the (population) *Fréchet mean*

$$\theta = \operatorname{argmin}_{s \in S} \mathbf{E}^{\mathbf{P}} d(X_1, s)^2.$$

Since S is a CAT(0) space, θ exists and is unique; we will give a direct proof below.

Folding and stickiness of the Fréchet mean. The behaviour of Fréchet means on S is best studied using the *folding maps* $F^j : S \to \mathbb{R}$ for $j = 1, 2, 3$ that identify the "other" two half-lines with the negative real axis. More precisely, $F^j(0) = 0$, $F^j\big((j, x)\big) = x$, and $F^j\big((k, x)\big) = -x$ for $k \neq j$. This gives rise to the "folded data" $Y_i^j = F^j(X_i)$ which are real-valued, square-integrable random variables.

There are now several cases possible:

(i) $\mathbf{E}^{\mathbf{P}} Y_1^j > 0$ for some j. This is equivalent to $\theta = (j, \mathbf{E}^{\mathbf{P}} Y_1^j)$, i.e. to θ being a *regular* point on the jth open half-line $\{j\} \times (0, \infty)$; it implies $\mathbf{E}^{\mathbf{P}} Y_1^k < 0$ for $k \neq j$. In all other cases $\theta = 0$.

(ii) $\mathbf{E}^{\mathbf{P}} Y_1^j < 0$ for all $j = 1, 2, 3$. Then θ is called *sticky* since θ remains at 0 under all sufficiently small perturbations of $\mathbf{P}^{X_1}$ where "small" is defined via the Wasserstein metric, say.

(iii) $\mathbf{E}^{\mathbf{P}} Y_1^j = 0$ for some j and $\mathbf{E}^{\mathbf{P}} Y_1^k < 0$ for $k \neq j$. Then θ is called *half-sticky* since it may move to the jth open half-line under small perturbations but not to the others.

(iv) $\mathbf{E}^{\mathbf{P}} Y_1^j = \mathbf{E}^{\mathbf{P}} Y_1^k = 0$ for some $j \neq k$. Then θ is called *non-sticky* since it may move between the jth and kth half-lines under small perturbations. However, this case can only occur if the other open half-line, the lth with $j \neq l \neq k$, say, is a $\mathbf{P}^{X_1}$-null set, in particular $\mathbf{E}^{\mathbf{P}} Y_1^l < 0$ then.

Note that these observations prove existence and uniqueness of θ.

Asymptotics on the 3-spider. If we fix j then the Y_i^j, $i \in \mathbb{N}$ form a sequence of i.i.d. square-integrable random variables. They thus fulfill a central limit theorem if $\mathbf{P}^{X_1}$ is not concentrated on a single point, i.e. if it is not the Dirac measure at θ; to be more precise, under this condition the distribution of the renormalised partial sums

$$Z_n^j = \frac{\sqrt{n}}{\hat{\sigma}_n^j}(\bar{Y}_i^j - \mathbf{E}^{\mathbf{P}} Y_1^j)$$

converges weakly to the standard normal distribution $\mathcal{N}(0, 1)$ when n tends to infinity where $\bar{Y}_n^j = \frac{1}{n}\sum_{i=1}^n Y_i^j$ is the sample mean and $\hat{\sigma}_n^j = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (Y_i^j - \bar{Y}_1^j)^2}$ is the sample standard deviation of $Y_1^j, \ldots, Y_n^j$. This will be the essential ingredient to obtain the required test statistics.

In the singular case, $\mathbf{E}^{\mathbf{P}} Y_1^j < 0$ implies that $\mathbf{P}(\bar{Y}_n^j \geq 0)$ converges to zero, meaning that the probability of finding the sample Fréchet mean, i.e. the Fréchet mean with respect to the empirical distribution of $X_1, \ldots, X_n$, on the jth open half-line converges to zero. In particular if the mean is sticky the probability of the sample Fréchet mean not being identical to the population Fréchet mean converges to zero! This is a phenomenon which does not occur in Euclidean spaces; it is due to the "infinite curvature" at the singular point.

Constructing the confidence region. The situation at hand may be cast in the framework described at the beginning by choosing Ω to be the sequence space $S^{\mathbb{N}}$ with $\mathcal{A}$ the corresponding product σ-algebra when S is endowed with its Borel σ-algebra $\mathcal{B}^{(1)}$. $X^{(n)} : \Omega \to \mathfrak{X}^{(n)} = S^n$ is the projection onto the first n entries, i.e. $X^{(n)} = (X_1, \ldots, X_n)$, with $\mathcal{B}^{(n)}$ the Borel σ-algebra on S^n. The probability distributions on Ω under consideration are then

$$\mathcal{P} = \left\{ \bigotimes_{i \in \mathbb{N}} \mu : \mu \text{ probability measure on } (S, \mathcal{B}^{(1)}) \text{ with } \int_S d(x, 0)^2 \, \mathrm{d}\mu(x) < \infty \right\}.$$

For $\mathbf{Q} \in \mathcal{P}$, we let $L(\mathbf{Q}) = \operatorname{argmin}_{s \in S} \mathbf{E}^{\mathbf{Q}} d(X_1, s)^2$, i.e. the Fréchet mean with respect to μ if $\mathbf{Q} = \otimes_{i \in \mathbb{N}} \mu$.

We now have to specify a test statistic $T_n^{\mathbf{Q}}$ and a set $M^{\mathbf{Q}}$ for every $\mathbf{Q} \in \mathcal{P}$; all test statistics will take values in $\mathfrak{T}^{\mathbf{Q}} = (-\infty, \infty]$ with its Borel σ-algebra. We will treat each of the cases above separately.

(i) In the regular case $L(\mathbf{Q}) \in \{j\} \times (0, \infty)$ we are essentially in the situation of a classical one-sample test for the mean; we thus choose $T_n^{\mathbf{Q}} = Z_n^j$ with the expectation of course taken under the hypothesis, i.e. with respect to $\mathbf{Q}$, and $M^{\mathbf{Q}} = (-q_{1-\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}})$ where $q_{1-\frac{\alpha}{2}}$ denotes the $(1 - \frac{\alpha}{2})$-quantile of the standard normal distribution $\mathcal{N}(0, 1)$. $\mathbf{Q}$ will then be rejected if $\bar{Y}_n^j$ deviates too much from $\mathbf{E}^{\mathbf{Q}} Y_1^j$.

(ii) In the sticky case, we choose $T_n^{\mathbf{Q}} = 1$ if $\bar{Y}_n^j \geq 0$ for some $j = 1, 2, 3$ and 0 otherwise; $M_n^{\mathbf{Q}} = \{0\}$ will then ensure that $\mathbf{Q}$ is rejected if the sample Fréchet mean is unequal to 0.

(iii) For half-sticky $L(\mathbf{Q})$ with $\mathbf{E}^{\mathbf{Q}} Y_1^j = 0$ we set $T_n^{\mathbf{Q}} = \infty$ if $\bar{Y}_n^k \geq 0$ for some $k \neq j$ and $T_n^{\mathbf{Q}} = Z_n^j$ else; choosing $M^{\mathbf{Q}} = (-q_{1-\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}})$ then results in rejecting $\mathbf{Q}$ if the sample Fréchet mean is on another (kth) half-line or if $\bar{Y}_n^j$ deviates too much from 0.

(iv) In the non-sticky case with $\mathbf{E}^{\mathbf{Q}} Y_1^j = \mathbf{E}^{\mathbf{Q}} Y_1^k = 0$ for $j \neq k$ and $j \neq l \neq k$, we set $T_n^{\mathbf{Q}} = \infty$ if $Y_i^l > 0$ for some i and again $T_n^{\mathbf{Q}} = Z_n^j$ else, so that once more choosing $M^{\mathbf{Q}} = (-q_{1-\frac{\alpha}{2}}, q_{1-\frac{\alpha}{2}})$ ensures rejection of $\mathbf{Q}$ if the sample Fréchet mean deviates too much from 0 or if we observe a point on the lth open half-line.

With this choice, the general framework laid out in the beginning applies and we obtain the desired asymptotic $(1 - \alpha)$-confidence region for the population Fréchet mean.

Discussion. As proposed, we constructed asymptotic confidence regions for the (population) Fréchet mean. In general, depending on the observations, these may differ qualitatively by being either (a) a proper subset, namely a "subinterval", of one of the open half-lines, or (b) comprising only the singular point, or (c) comprising an entire connected neighbourhood of the singular point, or a neighbourhood of the singular point excluding (d) one or (e) even two half-lines. From these different scenarios one may also infer some properties of the Fréchet mean, namely one may reject the hypothesis of it being at the origin in case (a) or it being regular in case (b). For the other cases, one may resort to observing that if $\bar{Y}_n^j$ is negative and far enough from the origin for all $j = 1, 2, 3$, one may reject the hypothesis of a half-sticky Fréchet mean, and similar for non-stickiness.

Algorithmically, the computation of these confidence regions essentially requires only the computation of the "classical" confidence regions for $\mathbf{E}^{\mathbf{P}} Y_1^j$ based on the sample mean and sample standard deviation of $Y_1^j, \ldots, Y_n^j$ for all $j = 1, 2, 3$; this can of course be done efficiently. Note also that the latter confidence regions are (in a certain asymptotic sense) optimal, so that we expect the confidence regions above to be optimal (in a similar sense) as well; observe, for instance, that having a

sticky Fréchet mean implies that the probability of the confidence region containing any other point than 0 converges to zero, similarly for half-sticky Fréchet means.

References

[1] L. J. Billera, S. P. Holmes, and K. Vogtmann, *Geometry of the space of phylogenetic trees*, Adv. in Appl. Math. **27(4)** (2001), 733–767.

[2] T. Hotz, S. Huckemann, H. Le, J. S. Marron, J. C. Mattingly, E. Miller, J. Nolen, M. Owen, V. Patrangenaru, S. Skwerer, *Sticky central limit theorems on open books*, The Annals of Applied Probability **23(6)** (2013), 2238–2258.

Stickiness and Smeariness
Stephan F. Huckemann
(joint work with Benjamin Eltzner)

Let Q be a Whitney-stratified space with connected top manifold stratum Q^* and singular part $Q^0 = Q \setminus Q^*$. Further, consider a distance $\rho : Q \times Q \to [0, \infty)$. For random variables $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$ on Q we have then the Frechét population and sample means

$$E_\rho(X) = \operatorname*{argmin}_{q \in Q} \mathbb{E}\left[\rho(q, X)^2\right], \quad \hat{E}_\rho^{(n)} = \operatorname*{argmin}_{q \in Q} \sum_{j=1}^n \rho(q, X_j)^2 \, .$$

We say that (Q, ρ) is *manifold stable* if for all random variables X on Q

$$\mathbb{P}\{X \in Q^*\} > 0 \Rightarrow E_\rho(X) \subset Q^* \, .$$

Obviously, BHV-spaces with the intrinsic distance are not manifold stable and one can show [2] that

(1) spaces arising as the quotient of a Riemannian manifold modulo a proper isometric Lie group action, with the canonical metric are manifold stable,

(2) Kendall's shape spaces Σ_m^k with the full Procrustes metric are not manifold stable for $m \geq 3$.

Let (Q, ρ) be manifold stable with $d = \dim(Q^*)$. A random variable X on Q with $E_\rho(X) = \{\mu\}$ and $\mathbb{P}\{X \in Q^*\} > 0$ is *k-th order smeary* ($k \geq 0$) if for all $X_1, \ldots, X_n \overset{i.i.d.}{\sim} X$, every measurable section $\hat{\mu}_n \in E_\rho^{(n)}$ and local chart $\phi : U \to \mathbb{R}^d$ near $\mu = \phi^{-1}(0)$, $\mu \in U \subset Q^*$ we have that

$$n^{\frac{1}{2(k+1)}} \phi(\hat{\mu}_n) \text{ has a non-trivial limiting distribution} \, .$$

While every X on a Euclidean space is 0-th order smeary, on S^1 for every order $k \in \mathbb{N}$ a k-th order smeary X can be found, cf. [1].

Now, consider illustrating examples on Q, the circle $S^1 = [-\pi, \pi)$ with the obvious topology joined with $T = [0, \infty)$ attached orthogonally at the center point 0 of S^1 (a circle with a vertical stick at the north pole) with the canonical intrinsic distance ρ. For $0 < a < \pi/2$ consider a random variable X_d with measure comprising the uniform distribution on S^1 along $[-\pi, -\pi + a] \cup [\pi - a, \pi]$ and a point mass at height $d \geq 0$ on T of mass $1 - a/\pi$. We then have that $E_\rho(X_0) = [-a, a]$

while $E_\rho(X_d) = \{0\}$ is fully sticky for $0 < d < \frac{\pi - a/2}{\pi - a}$. If the point mass splits into three point masses of equal weight, one at $-d/2$ one at height $d/2$ on T and one at d then $E_\rho(X_d) = [0, a]$ is half sticky for small $d > 0$. If the uniform mass on the interval collapses to a point and the other point moves up in T, then there is no stickiness any more. The same phenomenon (removing stickiness) is obtained when the sphere is replaced by a surface of revolution with a cusp at the base point of T: there is a tangent space at the origin.

Underlying these examples and the discussion, the following questions have been posed:

(1) On a non-positive curvature manifold, is there only non-positive k-smeariness?

(2) On a non Hausdorff Lie group quotient, how do manifold stability and stickiness relate?

(3) Are there cases where strict inequality holds in Ziezold's strong consistency [3]

$$\bigcap_{n=1}^{\infty} \overline{\bigcup_{k=n}^{\infty} \hat{E}_\rho^{(k)}} \subset E_\rho \, ?$$

(4) Find descriptors for the degree of (full) stickiness, e.g by a (Wasserstein) radius of a largest ball of sticky measures.

(5) Can (all) stickiness be "removed" by changing the geometry? Are there classes of stickiness where this is not possible?

(6) Can only subsets of Q^0 be sticky?

(7) Under which conditions (on the tangent cone, say) is stickiness possible?

References

[1] T. Hotz and S. Huckemann, *Intrinsic means on the circle: uniqueness, locus and asymptotics*, Ann. Inst. Stat. Math. (2014), to appear.

[2] S. Huckemann, *On the meaning of mean shape: manifold stability, locus and the two sample test*, Ann. Inst. Stat. Math. **64** (2012), 1227–1259.

[3] H. Ziezold, *Expected figures and a strong law of large numbers for random elements in quasi-metric spaces*, Transaction of the 7th Prague Conference on Information Theory, Statistical Decision Function and Random Processes **A** (1977), 1591–602.

Topological definition of stickiness for means in arbitrary metric spaces

Ezra Miller

(joint work with Stephan F. Huckemann, Jonathan C. Mattingly, and James Nolen)

The following notion is [1, Definiton 7.10] verbatim. Its details, merits, implications, and relations to previous concepts of stickiness formed the basis for a lively discussion. Nothing else was written on the board during the discussion.

Definition 1. Let $\mathcal{M}$ be a set of measures on a metric space $\mathcal{X}$. Assume $\mathcal{M}$ has a given topology. A *mean* is a continuous assignment $\mathcal{M} \to \{\text{closed subsets of } \mathcal{X}\}$. A measure μ *sticks* to a closed subset $C \subseteq \mathcal{X}$ if every neighborhood of μ in $\mathcal{M}$ contains an open subset consisting of measures whose mean sets are contained in C.

Continuity implies that the mean set is contained in C if μ sticks to C.

References

[1] S. Huckemann, J. Mattingly, E. Miller, and J. Nolen. Sticky central limit theorems at isolated hyperbolic planar singularities. Preprint, 35 pages. arXiv:math.PR/1410.6879

Optimization in Phylogenetic Treespace

SEAN SKWERER

(joint work with J. S. Marron and Scott Provan)

In the first half of the talk BHV Treespace [2] is introduced, the structure of treespace geodesics [4] is discussed, and the subdivision of treespace into regions where the form of geodesics to a fixed point is constant, i.e. the vistal subdivision of treespace from [3], is described briefly. The second half of the talk focuses on the elements of finding improvements and demonstrating optimality in convex optimization. This half of the talk draws from the PhD dissertation of the author [5].

Optimization problems on BHV treespace are challenging because the geometry of treespace creates difficult in two essential parts of optimization (1) making progress towards optimality and (2) verifying optimality. In treespace a metrically small neighborhood can actually be quite large in a certain sense. In constructing the space, the topological identification of the shared faces of orthants may create points in the closure of many orthants. In terms of trees, the neighborhood around a tree, X, is comprised not only of trees with the same topology as X but also trees for which it is possible to obtain the topology of X by contracting some edges. However, the list of tree topologies in a neighborhood can be quite large. For example, if X has no interior edges, then X can be obtained from any tree.

There is a broad class of algorithms, known as proximal point algorithms, with variants for non-positively curved metric space introduced in [1], which have been proven to converge to an optimal point for convex problems on such spaces. The general proximal point algorithm does not overcome the complexity of neighborhoods around degenerate trees mentioned previously. However, for certain classes of optimization problems, such as the Fréchet mean problem, there are practical versions of proximal point algorithms. Rather than focus on these types of algorithms, this talk focuses on studying the differential properties, i.e. the rates of change of functions on treespace. In particular the focus is on directional derivatives i.e. rates of change along geodesics from a point to points in its neighborhood.

This analysis facilitates more precise specification optimality conditions and an efficient algorithm for verifying that a point on a lower dimensional face of an orthant $\mathcal{O}$ is the minimizer of the Fréchet function within $\mathcal{O}$.

Directional Derivatives and Convex Functions

Let $\mathcal{T}_r$ denote the phylogenetic treespace with leaf indexes $\{0, 1, ..., r\}$. A treespace geodesic is denoted by a continuous function from the unit interval to treespace, $\gamma : [0, 1] \to \mathcal{T}_r$. A function $F(X)$ is (strictly) convex [6], that is $F \circ \gamma : [0, 1] \to \mathbb{R}$ is (strictly) convex for every geodesic $\gamma(\lambda)$ in $\mathcal{T}_r$.

Definition 1. *The* directional derivative from X to Y *is*

$$(1) \qquad F'(X, Y) = \lim_{\alpha \to 0} \frac{F(\Gamma(X, Y; \alpha)) - F(X)}{\alpha}$$

When both X and Y are in the relative interior of the same maximal orthant of treespace, where the gradient at X is well defined in $\mathcal{O}(Y)$, the directional derivative can be expressed in terms of the gradient at X inside $\mathcal{O}(Y)$. However when $\mathcal{O}(X) \subset \mathcal{O}(Y)$, the gradient at X is not well defined in $\mathcal{O}(Y)$. Analysis of the directional derivative in the later situation, is the main focus here.

The following lemma is used in the proof of Lem. 3, which states that the directional derivative of a convex function on BHV Treespace is also a convex function.

Lemma 2. *[5, Lem. 2.4.10] Let Y^0 and Y^1 be a points in $\mathcal{T}_r$ such that $\mathcal{O}(X) \subseteq \mathcal{O}(Y^0)$ and $\mathcal{O}(X) \subseteq \mathcal{O}(Y^1)$. Let $Y^t = \Gamma(Y^0, Y^1; t)$ be the point which is proportion t along the geodesic from Y^0 to Y^1. The point which is α proportion along the geodesic from X to Y^t is t proportion along the geodesic between the point $\Gamma_{X,Y^0}(\alpha)$ and the point $\Gamma_{X,Y^1}(\alpha)$, that is $\Gamma(X, Y^t; \alpha) = \Gamma(\Gamma(X, Y^0; \alpha), \Gamma(X, Y^1; \alpha); t)$.*

Lemma 3. *[5, Lem. 2.4.11] The directional derivative $F'(X, Y)$ is a convex function of Y over the set of Y such that $\mathcal{O}(X) \subseteq \mathcal{O}(Y)$ and X and Y share a vistal facet.*

Proof. Let Y^0 and Y^1 be a points in $\mathcal{T}_r$ such that $\mathcal{O}(X) \subseteq \mathcal{O}(Y^0)$ and $\mathcal{O}(X) \subseteq \mathcal{O}(Y^1)$. Let Y^t be the point which is proportion t along the geodesic from Y^0 to Y^1. Let $\Gamma_{XY^t}(\alpha) : [0, 1] \to \mathcal{T}_r$ be a function which parameterizes the geodesic from X to Y^t. Using Lem. 2 and the strict convexity of F together yields

$$(2) \qquad F(\Gamma_{XY^t}(\alpha)) < F(\Gamma_{XY^0}(\alpha))(1 - t) + F(\Gamma_{XY^1}(\alpha))t$$

The directional derivative from X in the direction of $\Gamma_{XY^t}(\alpha)$ is

$$(3) \qquad F'(X, Y^t) = \lim_{\alpha \to 0} \frac{F(\Gamma_{XY^t}(\alpha)) - F(X)}{\alpha}$$

Substituting for $F(\Gamma_{XY^t}(\alpha))$ using the inequality on line (2) yields,

$$(4) \qquad F'(X, Y^t) \leq \lim_{\alpha \to 0} \frac{F(\Gamma_{XY^0}(\alpha))(1 - t) + F(\Gamma_{XY^1}(\alpha))t - F(X)}{\alpha}$$

Note that strict inequality may not hold even though the Fréchet function is convex because in the limit the value may approach an infimum. Simplifying by separating the fraction and limit reveals that the directional derivative is convex in Y,

$$(5) \quad F'(X, Y^t) \leq (1-t) \lim_{\alpha \to 0} \frac{F(\Gamma_{XY^0}(\alpha)) - F(X)}{\alpha} + t \lim_{\alpha \to 0} \frac{F(\Gamma_{XY^1}(\alpha)) - F(X)}{\alpha}$$

$$(6) \qquad\qquad = (1-t) F'(X, Y^0) + t F'(X, Y^1)$$

$$\square$$

Fréchet Function Minimiziation

For a given data set of n phylogenetic trees in $\mathcal{T}_r$, $T^1, T^2, \ldots, T^n$, the (sample) *Fréchet function* is the sum of squares of geodesic distances from the data trees to a variable tree X. A geodesic $\gamma : [0,1] \to \mathcal{T}_r$ is the shortest path between its endpoints. The geodesic from X to T^i is characterized by a geodesic support, $(\mathcal{A}^i, \mathcal{B}^i) = \left((A_1^i, B_1^i), \ldots, (A_{k^i}^i, B_{k^i}^i)\right)$ [4]. Given the geodesic supports $(\mathcal{A}^1, \mathcal{B}^1), \ldots, (\mathcal{A}^n, \mathcal{B}^n)$ the Fréchet function is

$$(7) \quad F(X) = \sum_{i=1}^{n} d(X, T^i)^2 = \sum_{i=1}^{n} \left(\sum_{l=1}^{k^i} (\|x_{A_l^i}\| + \|B_l^i\|)^2 + \sum_{e \in C^i} (|e|_T - |e|_{T^i})^2 \right)$$

Theorem 4. *[5, Thm. 2.4.17] (Decomposition Theorem for Fréchet Function Directional Derivatives) Let $X, Y \in \mathcal{T}_n$, with $\mathcal{O}(X) \subseteq \mathcal{O}(Y)$ and with X and Y in a common multi-vistal cell, V_{XY}, let Y_X be the projection of Y onto $\mathcal{O}(X)$, and let $Y_\perp$ be the projection of Y onto $\mathcal{O}^\perp(X)$ at X. Then,*

$$(8) \qquad\qquad F'(X, Y) = F'(X, Y_X) + F'(X, Y_\perp)$$

The optimality condition for a point on a lower dimensional face of treespace can be expressed in terms of directional derivatives. In that case the optimality condition is

$$(9) \qquad F'(X, Y) \geq 0 \qquad\qquad \text{for all } Y \text{ such that } \mathcal{O}(X) \subseteq \mathcal{O}(Y)$$

By using Thm. 4 to separate the directional derivative into the contribution from the component of Y in $\mathcal{O}(X)$, and the component of Y which is perpendicular to $\mathcal{O}(X)$ the optimality condition becomes

$$(10) \quad [\nabla F(X)]_e = 0 \qquad\qquad\qquad \text{for all } e : |e|_X > 0$$

$$(11) \quad F'(X, Y) \geq 0 \quad \text{for all } Y \text{ such that the component of } Y \text{ in } \mathcal{O}(X) \text{ is } 0$$

Recursively applying a similar decomposition to the directional derivative yields a nested optimality condition.

Theorem 5. *Consider trees $Y^0, \ldots, Y^k$ such that (i) $\mathcal{O}(Y^0) \subset \ldots \subset \mathcal{O}(Y^k) = \mathcal{O}(E)$, and (ii) $Y^{i+1} - Y^i \perp \mathcal{O}(Y^i)$ for $i = 0, \ldots, k-1$. Let E^i be the set of positive edges in Y^i for $i = 0, \ldots, k$. Define a set of edge length difference vectors P^i for $i = 1, \ldots, k$ with the component for edge e having value $p_e^i = |e|_{Y^i} - |e|_{Y^{i-1}}$. Denote*

the unit simplex in the orthant $\mathcal{O}(E^i \setminus E^{i-1})$ by $\Delta^i = \{P \in \mathcal{O}(E^i \setminus E^{i-1}) | \sum p_e^i = 1\}$. The minimizer of $F(X)$ in $\mathcal{O}(E)$ is Y^0 if and only if

$$(12) \qquad\qquad \nabla F(Y^0) = 0$$

$$(13) \qquad\qquad F'(Y^{i-1}, Y^i) \geq 0 \ \text{for } i = 1, \ldots, k$$

$$(14) \qquad\qquad \nabla F'(Y^{i-1}, Y^i) \perp \Delta^i \ \text{for } i = 1, \ldots, k.$$

The optimality condition in Thm. 5 can be used as the logical basis for an algorithm which finds the minimizer of the Fréchet function $F(X)$ in the closure of a fixed orthant.

References

[1] M. Bacak, textitThe proximal point algorithm in metric spaces, Israel Journal of Mathematics, **194**, (2012), 689–701
[2] L. Billera, S. Holmes and K. Vogtmann, *Geometry of the Space of Phylogenetic Trees*, Adv. in Appl. Math, **27** (2001), 733–767
[3] E. Miller, M. Owen, and S. Provan, *Polyhedral Computational Geometry for Averaging Metric Phylogenetic Trees*, http://arxiv.org/abs/1211.7046
[4] M. Owen and S. Provan, *A Fast Algorithm for Computing Geodesic Distances in Tree Space*, Computational Biology and Bioinformatics, **8** (2011), 2–13
[5] S. Skwerer, *Tree Oriented Data Analysis*, http://arxiv.org/abs/1409.5501
[6] K.T. Sturm, *Probability Measures on Metric Spaces of Nonpositive Curvature, in Heat kernels and analysis on manifolds, graphs, and metric spaces: lecture notes from a quarter program on heat kernels, random walks, and analysis on manifolds and graphs*, **338** (2003),357–390

Reporters: Aasa Feragen and Stephan Huckemann

Participants

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

Prof. Dr. Aasa Feragen
Department of Computer Science
University of Copenhagen
Universitetsparken 5
2100 Copenhagen
DENMARK

Jun.-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
Goldschmidtstr. 7
37077 Göttingen
GERMANY

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

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

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

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

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

Prof. Dr. Peter W. Michor
Fakultät für Mathematik
Universität Wien
Nordbergstr. 15
1090 Wien
AUSTRIA

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

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

Dr. Sean Skwerer
Collaborative Center f. Statistics in
Science
Yale University, Biostatistics
300 George St.
New Haven CT 06511
UNITED STATES

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

Max Sommerfeld
Institut für Mathematische Stochastik
Georg-August-Universität Göttingen
Goldschmidtstr. 7
37077 Göttingen
GERMANY