NUS National University of Singapore | IMS Institute for Mathematical Sciences

SCIENTIFIC REPORTS

Interactions of Statistics and Geometry (ISAG) II

14 Oct 2024–25 Oct 2024

Organizing Committee

Stephan Huckemann

Universität Göttingen

Ezra Miller

Duke University

Zhigang Yao

National University of Singapore

CONTENTS PAGE

TESTING FOR UNIQUENESS OF DESCRIPTORS

BENJAMIN ELTZNER

Classification AMS 2020: 62F12; 62F05; 62F40; 62H11; 62J02

Keywords: Non-unique descriptors; M-estimators; hypothesis test; model selection

1. Introduction

The Central Limit Theorem (CLT), which states that for data in a real vector space the mean of a random sample converges to the population mean if the population distribution has finite variance, is a staple of statistics. It is the essential underpinning for the wide applicability of confidence sets and hypothesis tests like the t-test. Therefore, the field of asymptotic statistics has developed which aims at generalizing the CLT and related results to more general descriptors than the mean on a vector space. One crucial assumption in nearly all the generalized CLTs is that the population descriptor is unique.

However, already in simple systems, like the intrinsic mean on a circle, the assumption of a unique population descriptor is not generic. This problem has long been known and it has recently gathered some attention [3], [5], [1]. In this report, we give an abridged overview of the paper [3]. We present the hypothesis test developed in this paper to determine with confidence whether the population descriptor is unique or whether non-uniqueness cannot be ruled out. Then we show two examples from our research, where the test can be applied to check the viability of the model.

2. Hypothesis Test

When designing a hypothesis test to distinguish unique from non-unique population descriptors, it is important to consider which properties the test should ideally have. We posit the following three desirable properties

(1) The test should work for an *arbitrary number of minima*.
(2) The test should require *no knowledge of minima positions*.
(3) The "worst error" would be assuming a unique population descriptor, when in reality there are two or more; this should be the error of first kind.

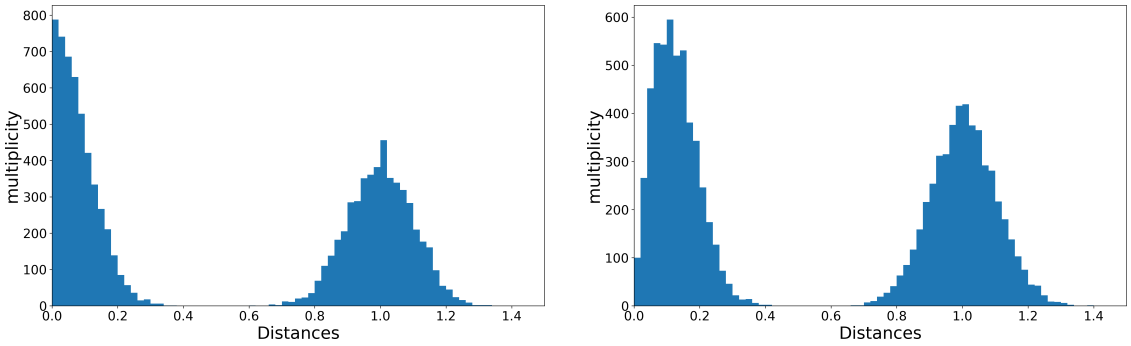
Especially the final point has important implications, namely the null hypothesis of our test is going to be that the investigated descriptor is non-unique, so if the test rejects, one can go ahead with standard methods. For the population descriptor set E we write:

$$H_0 : |E| \geq 2 \qquad\qquad H_1 : |E| = 1\,.$$

Remark 2.1. *Note that H_1 is not a single point, but all of the descriptor space. The parameter space of H_0 is much larger than that of H_1, but H_0 typically contains only a null set of probability distributions. Thus, in terms of the set of probability distributions, H_0 is typically much "smaller" than H_1, as usual for hypothesis tests.*

We use the framework of M-estimators, which are minimizers of a loss function, to formulate our test. From a given data sample, we draw many (e.g. $B = 10\,000$) bootstrap samples and calculate a *bootstrap estimator* for every bootstrap sample: $\mu^{*1}_{n,n}, \ldots, \mu^{*B}_{n,n}$. Bootstrap descriptors μ^{*j} form clusters around local minima of the sample loss function due to CLT. One of the clusters contains the sample descriptor $\widehat{\mu}_n$. *The test rejects if less than* $\frac{\alpha B}{2}$ *bootstrap descriptors* μ^{*j} *are in another cluster than* $\widehat{\mu}_n$. The factor $1/2$ is explained by asymptotic theory in [3].

The hypothesis test requires a reliable clustering method which can at least distinguish the cluster containing $\widehat{\mu}_n$ from all other clusters. Since a distinction of other clusters is not needed, we propose the significant simplification to consider only the distances $d_j := d(\widehat{\mu}_n, \mu^{*j})$ of bootstrap descriptors from the sample descriptor. As illustrated in Figure 1, the cluster containing $\widehat{\mu}_n$ will be the first mode in the distribution of d_j but in dimension larger than 1 it starts with a rising slope. All other modes will correspond to different clusters, some of them possibly lumped together by only taking the distance.



(A) Typical distance histogram in 1d (B) Distance histogram in 2d or higher

FIGURE 1. In dimensions higher than 1, the cluster containing $\widehat{\mu}_n$ starts with a rising slope due to the spherical volume element.

The problem now boils down to finding the onset of the second mode and counting all points from there. We use the multiscale method by [2] to identify rising and falling slopes. Since the first falling slope marks the end of the cluster containing $\widehat{\mu}_n$, we identify the onset of the next rising slope after the first falling slope as d_+. Then we simple count all $d_j > d_+$, which yields the following test

Hypothesis Test 2.2.
Null hypothesis and alternative:

$$H_0 : |E| \geq 2, \quad H_1 : |E| = 1.$$

Test statistic:

$T := \frac{2}{B}|\{d_j : d_j > d_+\}|$.
Recall: d_+ *is the onset of the rising slope after the first falling slope.*

Rejection regions and p-values:

Reject if $T < \alpha$, *p-value* $p = \min\{1, T\}$.

In [3], it is shown that the test has exactly size α if the population has two global descriptors and is conservative if it has more. It is also shown that the power of the test asymptotically goes to 1 under the alternative for $n \to \infty$ and $B \to \infty$.

3. Applications

As first application, we consider the growth of actin fibers in blood platelets after spreading on a substrate [6]. We use the Filament Sensor [4] to detect fibers in fluorescence microscopy images of the platelets taken in 10-second intervals, see Figure 2, and consider total fiber length over time. To these time series we then fit an exponential saturation model detailed in [6], [3].

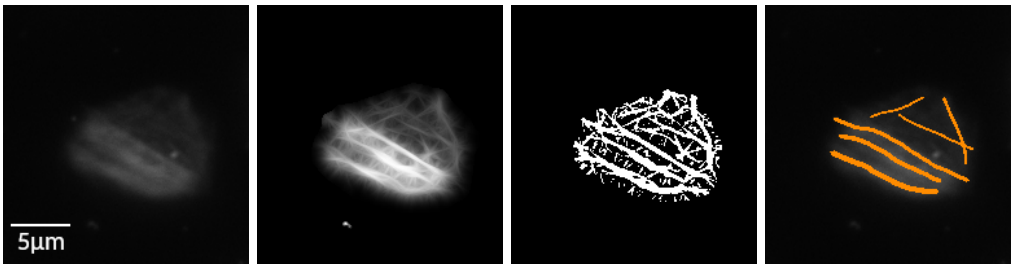
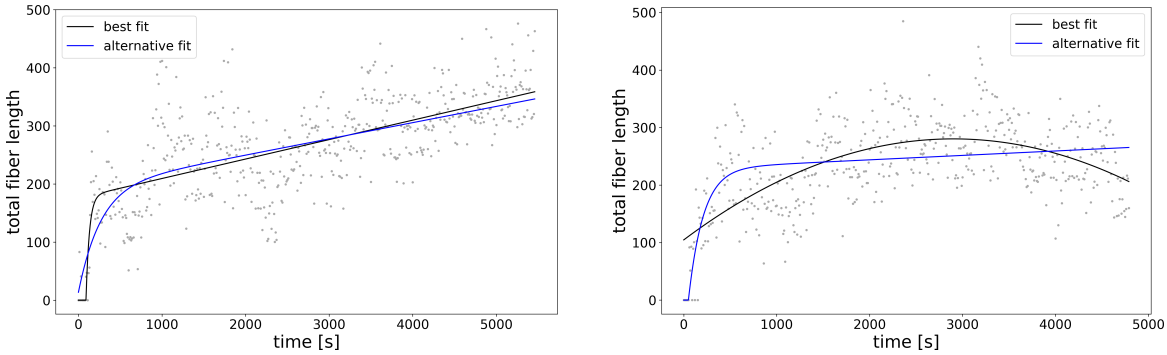


Figure 2. An overview of the image filtering, binarization and filament detection of the Filament Sensor on a microscopy image of a platelet.

For most platelets, our test rejects the null hypothesis of non-uniqueness. In Figure 3, we show two examples of platelets where alternate solutions exist and in one case the null is not rejected. However, in both cases, one of the growth curves has a too early onset of growth, since the time coordinate is chosen such that time $t = 0$ marks the time when the platelet sets down on the substrate and starts spreading. In [6] we constrained the parameter space of the model to avoid an onset of growth before $t = 0$, which leads to unique fits for all platelets.

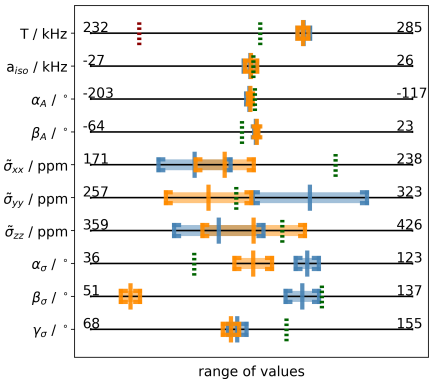


(A) Test does not reject, $p = 0.227$

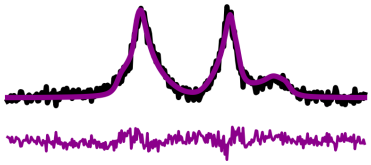
(B) Test rejects, $p = 0.044$,
but not with multiple testing correction!

Figure 3. Time series and fits for two different platelets.

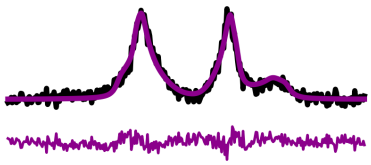
As a second application, we consider parameter estimation from ENDOR spectra [7]. Two local minima of the loss function emerged in the parameter fitting, leading to different parameters but very similar fitted spectra, see Figure 4. From structural chemistry one can determine that only the parameters shown in blue here correspond to a viable conformation. In this case, our hypothesis test also clearly rejects the null hypothesis of non-uniqueness, so the blue parameters are also statistically preferred.



(A) Estimated parameter bounds



(B) Fit for blue parameters



(C) Fit for orange parameters

FIGURE 4. Estimated parameters with confidence bands for two local minima of the loss function and the corresponding fitted spectra.

Conclusion. We have developed an unprecedented hypothesis test for uniqueness of descriptors with statistical size and power guarantees, which is highly model agnostic in the sense that only weak assumptions are made on the model and the population. In two applications, we have shown that non-uniqueness can especially arise if the parameters space of a model is too broad and can give rise to fitting results which contradict basic known facts about the system under investigation. We propose using our test in complex models, where it can indicate that the model is not suitably restrained, in the simplest case by using a too broad parameter space.

REFERENCES

[1] M. Blanchard and A. Q. Jaffe. Fréchet mean set estimation in the Hausdorff metric, via relaxation. *Bernoulli*, 31(1):432 – 456, 2025. doi: 10.3150/24-BEJ1734.

[2] L. Dümbgen and G. Walther. Multiscale inference about a density. *The Annals of Statistics*, 36(4):1758–1785, 2008. doi: 10.1214/07-AOS521.

[3] B. Eltzner. M-variance asymptotics and uniqueness of descriptors. *under review, arXiv:2011.14762*, 2020. doi: 10.48550/arXiv.2011.14762.

[4] B. Eltzner, C. Wollnik, C. Gottschlich, S. F. Huckemann, and F. Rehfeldt. The Filament Sensor for Near Real-Time Detection of Cytoskeletal Fiber Structures. *PLOS ONE*, 10 (5):1–28, 2015. doi: 10.1371/journal.pone.0126346.

[5] S. N. Evans and A. Q. Jaffe. Limit theorems for Fréchet mean sets. *Bernoulli*, 30(1): 419 – 447, 2024. doi: 10.3150/23-BEJ1603.

[6] A. K. Paknikar, B. Eltzner, and S. Köster. Direct characterization of cytoskeletal reorganization during blood platelet spreading. *Progress in Biophysics and Molecular Biology*, 144:166 – 176, 2019. doi: 10.1016/j.pbiomolbio.2018.05.001.

[7] H. Wiechers, A. Kehl, M. Hiller, B. Eltzner, S. F. Huckemann, A. Meyer, I. Tkach, M. Bennati, and Y. Pokern. Bayesian optimization to estimate hyperfine couplings from ^{19}F ENDOR spectra. *Journal of Magnetic Resonance*, 353:107491, 2023. doi: 10.1016/j.jmr.2023.107491.

UNIVERSITY OF GÖTTINGEN
Email address: beltzne@uni-goettingen.de

HUBER MEANS ON RIEMANNIAN MANIFOLDS

SUNGKYU JUNG

The work presented here was jointly investigated with Jongmin Lee, Pusan National University, who was another participant of the workshop.

Classification AMS 2020: 62R20, 62R30.

Keywords: Central limit theorem, Covariance estimation, Hypothesis testing, Riemannian center of mass, Robust statistics, Statistics on manifolds

Determining the mean of a dataset has been a core problem in statistics for a long time, forming the foundation for various statistical inferences and computations. Traditional approaches to mean estimation typically employ the L_2 loss function, as exemplified by the Fréchet mean, which extends to data residing in spaces beyond Euclidean vector spaces. However, these methods are often highly sensitive to outliers, particularly when applied to manifold-valued data, which have become increasingly prevalent in modern scientific research.

In this work, we introduce the Huber means on Riemannian manifolds. The Huber mean, defined as the minimizers of the expected Huber loss, offers a robust alternative to the Frèchet mean by combining elements of L_2 and L_1 losses. This dual nature makes the Huber mean highly resistant to outliers while maintaining efficiency under heavy-tailed distributions. The Huber mean serves as a natural generalization of Huber's M-estimator [2] to the manifold setting, and can be viewed as a robust extension of the Fréchet mean.

The *Huber loss* function, introduced by [2], combines elements of both L_2 and L_1 losses. For a cutoff constant $c > 0$, the Huber loss function is defined for $x \geq 0$ as follows:

$$\rho_c(x) = \begin{cases} x^2 & \text{if } x \leq c, \\ 2c(x - \frac{c}{2}) & \text{if } x > c. \end{cases}$$

When $c \simeq 0$, the Huber loss closely resembles $2c$ times the L_1 loss, since $c \simeq 0$ implies $\rho_c(x) = 2cx - c^2 \simeq 2cx$. As $c \to \infty$, the Huber loss converges pointwise to the L_2 loss. We extend the definitions of (pseudo) Huber losses with $c = 0$ and $c = \infty$ by setting $\rho_0(x) = \tilde{\rho}_0(x) := L_1(x) = x$ and $\rho_\infty(x) = \tilde{\rho}_\infty(x) := L_2(x) = x^2$ for a comprehensive study.

The population (or sample) Huber mean is defined as any minimizer of the expected Huber loss for P_X (or P_n, respectively):

Definition 0.1. *Given a prespecified constant $c \in [0, \infty]$, the population Huber mean set with respect to P_X is*

$$E^{(c)} := \textit{argmin}_{m \in M} F^c(m), \ F^c(m) := \int \rho_c\{d(X, m)\} dP.$$

For given n deterministic observations $(x_1, x_2, ..., x_n) \in M^n$, the sample Huber mean set is

$$E_n^{(c)} := \textit{argmin}_{m \in M} F_n^c(m), \ F_n^c(m) := \frac{1}{n} \sum_{i=1}^{n} \rho_c\{d(x_i, m)\}.$$

To ensure the existence of Huber means, we impose an integrability condition:
(A1) For some $m \in M$, $F^c(m) = \int \rho_c\{d(X,m)\}dP < \infty$.

Theorem 0.2 (Existence of the population Huber means). *For a given $c \in [0,\infty]$, assume that P_X satisfies Assumption (A1). Then, the population Huber mean exists, i.e., $E^{(c)} \neq \phi$.*

The geometric median, minimizing the sum of absolute deviations, may not be unique in Euclidean spaces, and similarly, Huber means are not necessarily unique. The Fréchet mean can also lack uniqueness on manifolds with nonzero curvature. This raises the key question: under what conditions is the Huber mean unique?

Ideally, uniqueness would follow from the convexity of F^c on M, but no non-constant continuous function can be convex on a compact, boundaryless manifold [6]. To address this, we ensure the Huber mean lies in a strongly convex subset of M and establish the strict convexity of F^c there.

(A2) For the prespecified $c \in [0,\infty]$, there exists $p_0 \in M$ such that $\mathsf{supp}(P_X) \subseteq B_{r_0}(p_0)$, where $\mathsf{supp}(P_X)$ denotes the support of P_X, and

$$r_0 = \begin{cases} \frac{1}{2}\min\{\frac{\pi}{2\sqrt{\Delta}}, r_{\text{inj}}(M)\} & \text{if } 0 \le c < \frac{\pi}{\sqrt{\Delta}}, \\ \frac{1}{2}\min\{\frac{\pi}{\sqrt{\Delta}}, r_{\text{inj}}(M)\} & \text{if } \frac{\pi}{\sqrt{\Delta}} \le c \le \infty, \end{cases} \tag{0.1}$$

where $\Delta < \infty$ denotes for the supremum of the sectional curvatures of M, and $r_{\text{inj}}(M)$ its injectivity radius.

Theorem 0.3 (Uniqueness of population Huber means). *For a prespecified constant $c \in [0,\infty]$, suppose that P_X satisfies Assumptions (A1) and (A2). If P_X does not degenerate to any single geodesic, the the population Huber mean with respect to P_X is unique.*

The sample Huber mean set $E_n^{(c)}$ for $c \in [0,\infty]$ is *strongly consistent* with $E^{(c)}$, as stated next (see [4] for the choice of terminology). Given n random observations $X_1, X_2, ..., X_n \overset{i.i.d.}{\sim} P_X$, the sample Huber mean set $E_n^{(c)}$ is a random closed set.

Theorem 0.4 (Strong consistency). *For a given $c \in [0,\infty]$, if P_X satisfies Assumption (A1), then with probability 1,*

$$\lim_{n\to\infty} \sup_{m \in E_n^{(c)}} d(m, E^{(c)}) = 0,$$

where $d(m, E^{(c)}) := \inf_{p\in E^{(c)}} d(m,p)$.

We next establish a central limit theorem for Huber means. Throughout, we assume that for a prespecified $c \in (0,\infty]$, the population Huber mean m_0^c for c with respect to P_X is unique, and so is the sample Huber mean m_n^c with probability 1 for every sample size n. In a local coordinate chart $(\phi_{m_0^c}, U)$ centered at m_0^c (i.e., $\phi_{m_0^c}(m_0^c) = \mathbf{0} \in \mathbb{R}^k$), let $\Sigma_c(\mathbf{x}) := \mathsf{Var}[\mathsf{grad}\rho_c\{d(X, \phi_{m_0^c}^{-1}(\mathbf{x})\}]$ and $H_c(\mathbf{x}) = E[\mathbf{H}\rho_c\{d(X, \phi_{m_0^c}^{-1}(\mathbf{x}))\}]$, where grad and $\mathbf{H}$ refer to the Euclidean gradient and the Euclidean Hessian, respectively. We write $\Sigma_c := \Sigma_c(\mathbf{0})$ and $H_c := H_c(\mathbf{0})$, and assume a set of regularity conditions, typically appeared in related works such as [1, 3, 5]. It is generally challenging to verify the asymptotic normality of an estimator on manifolds, due to their nonlinearity. To overcome the difficulty, a "linearization" of manifolds by utilizing local coordinate charts is used to state a central limit theorem for m_n^c.

Theorem 0.5 (Central limit theorem). *For a given $c \in (0, \infty]$, suppose that P_X satisfies Assumptions (A1), (A3), and (A4). Then,*
(a) $m_n^c \to m_0^c$ almost surely as $n \to \infty$, and
(b) $\sqrt{n}\phi_{m_0^c}(m_n^c) \to N_k(\mathbf{0}, H_c^{-1}\Sigma_c H_c^{-1})$ in distribution as $n \to \infty$.

We next evaluate the breakdown point of the sample Huber mean, demonstrating its high robustness. The breakdown point of the sample Huber mean at $\mathbf{X}$ is given by $\epsilon^*(m_n^c, \mathbf{X}) = \min_{1 \le k \le n}\{\frac{k}{n} : \sup_{\mathbf{Y}_k} d(m_n^c(\mathbf{X}), m_n^c(\mathbf{Y}_k)) = \infty\}$, where the supremum is taken over all possible $\mathbf{Y}_k$. The higher the breakdown point is, the more resistant the Huber mean is to outliers.

Theorem 0.6 (Breakdown point). *Let $\mathbf{X} = (x_1, x_2, ..., x_n)$ be a collection of observations on M. If M is unbounded and all isometric transformations on M are transitive, then for any $c \in [0, \infty)$, $\epsilon^*(m_n^c, \mathbf{X}) = [\frac{n+1}{2}]/n$, where $[\cdot]$ denotes the floor function.*

The Huber mean possesses a breakdown point of 0.5, which is the highest possible breakdown point among all isometric-equivariant estimators.

REFERENCES

[1] Bhattacharya, R. and Patrangenaru, V. (2003), "Large sample theory of intrinsic and extrinsic sample means on manifolds I," *Annals of Statistics*, 31, 1–29.

[2] Huber, P. J. (1964), "Robust Estimation of a Location Parameter," *Annals of Mathematical Statistics*, 35, 73–101.

[3] Huckemann, S. F. (2011a), "Inference on 3D Procrustes means: Tree bole growth, rank deficient diffusion tensors and perturbation models," *Scandinavian Journal of Statistics*, 38, 424–446.

[4] — (2011b), "Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth," *Annals of Statistics*, 39, 1098–1124.

[5] Jung, S., Rooks, B., Groisser, D., and Schwartzman, A. (2024), "Averaging symmetric positive-definite matrices on the space of eigen-decompositions," *to appear in Bernoulli (arXiv preprint arXiv:2306.12025)*.

[6] Yau, S.-T. (1974), "Non-existence of continuous convex functions on certain Riemannian manifolds," *Mathematische Annalen*, 207, 269–270.

SEOUL NATIONAL UNIVERSITY
Email address: sungkyu@snu.ac.kr

THE INFLUENCE OF THE CUT LOCUS ON THE CLT FOR FRÉCHET MEANS

HUILING LE

Classification AMS 2020: Primary 60F05; Secondary 62R30

Keywords: Central limit theorem; Cut locus; Fréchet mean; Intrinsic; Parallel transport; Riemannian manifold.

This talk presents the result of the recent paper [3] (joint work with Thomas Hotz and Andrew Wood): a general result on the CLT for sample Fréchet means on compact Riemannian manifolds when the support of a distribution meets the cut locus of its Fréchet mean. We (i) clarify when non-standard behaviour of the Fréchet mean in compact Riemannian manifolds occurs; and (ii) to characterise the non-standard behaviour when it does occur. In particular, whether or not a non-standard term arises in the CLT depends on whether the co-dimension of the cut locus of the Fréchet mean is one or greater than one: in the former case a non-standard term appears but not in the latter case.

This result generalises the result of Bhattacharya and Lin [1], where the authors considered the case where the cut locus of the Fréchet mean is at least 2, and the result of Hotz and Huckemann [2], where the authors considered the intrinsic Fréchet mean on the circle.

REFERENCES

[1] R.N. Bhattacharya and L. Lin. Omnibus CLTs for Fréchet means and nonparametric inference on non-Euclidean spaces. *Proc. Am. Math. Soc.*, 145, 413–428, 2017.

[2] T. Hotz and S. Huckemann. Intrinsic means on the circle: uniqueness, locus and asymptotics. *Ann. Inst. Math. Sci.*, 67, 177–193, 2015.

[3] T. Hotz, H. Le and A.T.A. Wood Central limit theorem for intrinsic Fréchet means in smooth compact Riemannian manifolds. *Probability Theory and Related Fields*, 189, 1219–1246, 2024.

UNIVERSITY OF NOTTINGHAM, UK
Email address: huiling.le@nottingham.ac.uk

PROBING THE SHAPE OF METRIC AND NETWORKED DATA THROUGH OBSERVABLES

WASHINGTON MIO AND ECE KARACAM

Classification AMS 2020: 62R20, 62R30

Keywords: statistics on metric spaces, principal observable analysis, metric space covariance

We discuss a method for analyzing datasets in compact metric spaces through a technique referred to as principal observable analysis, which may be viewed as a metric counterpart to principal component analysis in Euclidean space. The underlying metric space is denoted (X, d_X) and data is modeled as a Borel probability measure μ on X, which may represent a (theoretical) population model or empirical data. Thus, the objects of interest are *metric measure spaces* given by triples (X, d_X, μ). This framework also includes networked data by viewing the set of nodes of a network as a metric space equipped with the shortest path distance.

PCA in d-dimensional Euclidean space $\mathbb{R}^d$ (or more generally, in Hilbert spaces) is based on maximizing the variance of orthogonal projections of μ to subspaces of $\mathbb{R}^d$. It is well known that these projections can be constructed iteratively from projections to 1-dimensional subspaces. As such, one possible approach to metric versions of PCA is to replace 1-D linear projections with mappings $f\colon X \to \mathbb{R}$ for which we have some control on the metric distortions across all scales, as is the case for orthogonal projections. This motivates us to define an *observable* as a 1-Lipschitz mapping $f\colon X \to \mathbb{R}$; that is, a map that satisfies

$$|f(x) - f(y)| \leq d_X(x, y), \tag{0.1}$$

for any $x, y \in X$. The underlying philosophy is that, in the aggregate, such observables retain substantial information about the shape of (X, d_X, μ).

We say that an observable f is μ-centered if $\int_X f(x)\, d\mu(x) = 0$ and denote the space of all μ-centered observables by $O^c_\mu(X)$. The *observable covariance operator* $\Sigma_\mu\colon O^c_\mu(X) \times O^c_\mu(X) \to \mathbb{R}$ is defined by

$$\Sigma_\mu(f, g) := \int_X f(x)g(x)d\mu(x). \tag{0.2}$$

$\Sigma_\mu(f, g)$ measures the correlation between the μ-centered observables f and g.

A *first principal observable* (PO1) for (X, d, μ), denoted ϕ_1, is a μ-centered observable f that maximizes the variance $\sigma^2(f) = \Sigma_\mu(f, f)$. In other words,

$$\phi_1 := \operatorname*{arg\,max}_{f \in O^c_\mu(X)} \Sigma_\mu(f, f)\,. \tag{0.3}$$

Inductively, assuming that $\phi_1, \ldots, \phi_{n-1}$ have been constructed, define an nth *principal observable* ϕ_n as a μ-centered metric observable that maximizes the variance among those

μ-orthogonal to the subspace spanned by $\phi_1, \ldots, \phi_{n-1}$. More precisely,

$$\phi_n := \arg\max_{f \in O^c_\mu(X)} \Sigma_\mu(f, f)\,, \tag{0.4}$$

subject to the constraints $\int_X f(x)\phi_i(x)d\mu(x) = 0$, for $1 \leq i \leq n-1$. The existence of observables follows from general compactness arguments based on the Arzelà-Ascoli Theorem.

REMARKS.

(1) Principal observable analysis provides a dimension reduction and vectorization method for metric data. For a fixed integer $n > 0$, the first n principal observables define a map $\phi\colon X \to \mathbb{R}^n$ given by $x \mapsto (\phi_1(x), \ldots, \phi_n(x))$. Note, however, that the natural metric in $\mathbb{R}^n$ to analyze the reduced data is not the Euclidean metric; rather, it is the ℓ_∞-metric for the map ϕ to be 1-Lipschitz, our guiding principle for the construction of observables.
(2) Unlike PCA, the variance of the reduced data is not additive over the various principal components. In a Hilbert space this holds because principal components are not just uncorrelated but also orthogonal with respect to the underlying inner product, a structure that is not present in metric spaces.
(3) Principal observables also yield basis functions for the representation of functions $g\colon X \to \mathbb{R}$. Unlike PCA, the number of basis functions is not limited by the dimension of X. For example, for the unit interval $I = [0, 1]$, there are infinitely many non-trivial principal observables for the uniform distribution μ on I.

We address the stability of the observable covariance operator in a general setting where both the underlying compact metric space and the probability distribution can vary. Note that for two different distributions μ and ν, even if defined on the same space X, their spaces of centered distributions are distinct so that their observable covariance operators Σ_μ and Σ_ν are not defined on the same domain. This leads us to analyze the stability of covariance in a Gromov-Hausdorff framework. To this end, for $p \geq 1$, we equip $O^c_\mu(X)$ with the metric induced by the L_p-norm denoted $\|\cdot\|_{p,\mu}$ that turns $O^c_\mu(X)$ into a bounded, but generally non-complete, metric space. We then equip $O^c_\mu(X) \times O^c_\mu(X)$ with the product metric

$$d_{p,\mu}((f_1, g_1), (f_2, g_2)) := \max\{\|f_1 - f_2\|_{p,\mu}, \|g_1 - g_2\|_{p,\mu}\}. \tag{0.5}$$

We state our stability result for covariance in terms of the *(functional) Gromov-Hausdorff distance* (cf. [1]), denoted d_{GH} and calculated with respect to the metric $d_{p,\mu}$.

Theorem 0.1 (Stability Theorem)**.** *Let $\mathcal{X} = (X, d_X, \mu)$ and $\mathcal{Y} = (Y, d_Y, \nu)$ be metric measure spaces such that μ and ν have finite p-moments, $p \geq 1$. Then,*

$$d_{GH}(\Sigma_\mu, \Sigma_\nu) \leq 2\max\{1, D_X + D_Y\} d_{GW,p}(\mu, \nu),$$

where $d_{GW,p}$ denotes the Gromov-Wasserstein p-distance, and D_X and D_Y are the diameters of (X, d_X) and (Y, d_Y), respectively.

Here, we use Sturm's version of the Gromov-Wasserstein p-distance [3], which is based on both metric and probabilistic couplings, as opposed to the original definition due to Mémoli which is based on expected distortions of probabilistic couplings [2].

Consistency of the observable covariance operator follows as a corollary of the Stability Theorem. If $(x_i)_{i=1}^{\infty}$ are independent samples from μ, and μ_n is the empirical measure $\mu_n = \sum_{i=1}^{n} \delta_{x_i}/n$, then $\lim_{n\to\infty} d_{GH}(\Sigma_\mu, \Sigma_{\mu_n}) = 0$ almost surely. It is also possible to get estimates for the convergence rate using results from [4].

REFERENCES

[1] J. Curry, W. Mio, T. Needham, O.B. Okutan, and F. Russold. Stability and Approximations for Decorated Reeb Spaces. *Symposium on Computational Geometry* (SoCG), Ahtnes, Greece, 2024.

[2] F. Mémoli. Gromov-Wasserstein Distances and the Metric Approach to Object Matching. *Found. Comput. Math.* 11(4):417-487, 2011.

[3] K.-T. Sturm. On the geometry of metric measure spaces. *Acta Math.* 196 (1) 65 - 131, 2006.

[4] J. Weed and F. Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *Bernouilli* 25(4A), 2620-2648, 2019.

DEPARTMENT OF MATHEMATICS, FLORIDA STATE UNIVERSITY
Email address: wmio@fsu.edu

ADVANCES IN GEOMETRIC STATISTICS WITH FLAG SPACES

XAVIER PENNEC
(JOINT WORK TOM SZWAGIER, DIMBIHERY RABENORO)

Classification AMS 2020: 62R30, 60D05, 62H25, 14M15, 58C99, 62H11

Keywords: Flags, subspaces, manifolds, PCA, CLT

Generalizing PCA to manifolds: Barycentric Subspaces Analysis. Statistically reduction of the dimension is a key issue in numerous problems. When data belong to a manifold, we first need to define families of parametric subspace in manifolds which could play the role of principal subspaces. Geodesic shooting along the main eigenvectors of the covariance matrix at the Fréchet mean point generates a Geodesic Subspace (GS) in tangent PCA. The point and tangent vectors defining that GS that can also be optimized to best fit the data such as proposed in Principal Geodesic Analysis (PGA) and Geodesic PCA (GPCA). To restore the full symmetry between the parameters, we proposed in [1] to use the *Exponential Barycentric subspace* (EBS) defined as the locus of weighted means of $k+1$ reference points (with positive or negative weights). The EBS is locally a stratified spaces of maximal dimension k comprised of critical points of the weighted variance satisfying the barycentric equation $\sum_i \lambda_i log_x(x_i) = 0$. Its metric completion is called the *affine span* Aff$(x_0, \dots x_k)$. Such spaces generalise the notion of Geodesic Subspaces which can be shown to be the limit when reference points coalesce towards a local 1-jet.

Barycentric subspaces and affine spans can naturally be nested by defining an ordering of the reference points. This allows the construction of forward or backward nested sequence of subspaces. However, these methods optimized for one subspace at a time and cannot optimize the explained (or unexplained) variance simultaneously for all the subspaces of the flag. In order to obtain a global criterion, PCA in Euclidean spaces was rephrased in [1] as an optimization on the flags of linear subspaces of the accumulated unexplained variance criterion. This generalizes nicely to flags of affine spans in Riemannian manifolds and gives a particularly appealing generalization of PCA on manifolds, called Barycentric Subspaces Analysis (BSA).

The curse of isotropy: From PCA to Principal Subspaces analysis. Considering PCA as an optimization of flags spaces gives an interesting geometric point of view, even in Euclidean spaces. Indeed, one usually consider the succession of unidimensional eigenmodes for the interpretation of the data in PCA, but more general flags with higher dimensional subspaces naturally arise with the geometric point of view. They correspond to covariance matrix with repeated eigenvalues, in which case eigenspaces are stable but eigenvectors may be freely rotated within each eigenspace. This raises an an important issue about the interpretation of PCA modes, called the curse of isotropy [2]: principal components associated with equal eigenvalues show large intersample variability and are arbitrary combinations of potentially more interpretable components.

Most users overlook the problem because empirical eigenvalues are almost surely distinct in practice due to sampling errors with a finite number of samples. In [2], we propose to identify datasets that are likely to suffer from the curse of isotropy by introducing a generative Gaussian model with repeated eigenvalues and comparing it to traditional models via the principle of parsimony. This yields an explicit criterion to detect the curse of isotropy in practice. We notably argue that in a dataset with 1000 samples, all the eigenvalue pairs with a relative eigengap lower than 21% should be assumed equal. This demonstrates that the curse of isotropy cannot be overlooked. In this context, we propose to transition from fuzzy principal components to more interpretable principal subspaces. The final methodology, coined *principal subspace analysis* is extremely simple and shows promising results on a variety of datasets from different fields.

A geometric formulation of CLT for flags. Estimating principal subspaces rather than eigenvectors raises the question of the uncertainty of the estimated flag: with a statistical point of view, one thus looks for confidence regions for principal subspaces. The previous works of Anderson and Tyler were limited to confidence regions on individual eigenvectors or on one single eigenspace at a time. In [3], we develop an asymptotic method to infer the collection of all principal subspaces together, i.e. the full flag of eigenspaces of this covariance matrix. Our approach is based on the Riemannian homogeneous geometry of the flag manifold. However, even if flags generalize Grassmmann and Steifel manifolds, they are generally not symmetric, and the Riemannian logarithm is not known in closed form. To get around this problem, we develop and approach based on the embedding of the flag manifold in the product of Grassmannians, where we can show a central limit theorem and a χ^2 distribution of the Mahalanobis distance.

REFERENCES

[1] X. Pennec, *Barycentric subspace analysis in manifolds* , **Annals of Statistics 46(6A)** (2018), p.2711-2746.

[2] T. Szwagier, X. Pennec, *The curse of isotropy: from principal components to principal subspaces*, **arXiv:2307.15348**, (2023).

[3] D. Rabenoro, X. Pennec, *A geometric framework for asymptotic inference of principal subspaces in PCA*, **arXiv:2209.02025**, (2022).

UNIVERSITÉ CÔTE D'AZUR AND INRIA, EPIONE TEAM, 2004 RTE DES LUCIOLES, BP93, F-06092 SOPHIA-ANTIPOLIS CEDEX, FRANCE

Email address: xavier.pennec@inria.fr

OBJECT CORRESPONDENCE FOR SHAPE STATISTICS

STEPHEN M. PIZER, JS MARRON, MOHSEN TAHERI AND JARED VICORY

Abstract

We describe a representation targeted for anatomic objects that is designed to enable strong locational correspondence within object populations and thus to provide powerful object statistics. The method generates fitted coordinate frames on the boundary and *in the interior of objects* and produces alignment-free geometric features from them. It accomplishes this by understanding an object as the diffeomorphic deformation of an ellipsoid and using a skeletal representation (which has swept curvilinear cross-sections) fitted throughout the deformation to produce a model of the target object, where the object is provided initially in the form of a boundary mesh. We call this object representation the *evolutionary s-rep*. Via classification performance on hippocampi shape between individuals with a disorder vs. others, we compare our method to two state-of-the-art methods for producing object representations that are intended to capture geometric correspondence across a population of objects and to yield geometric features useful for statistics, and we show notably improved classification by this new representation on clinical data as to infants hippocampal shape's ability to diagnose autism. The geometric features that are derived from each of the representations, especially via fitted coordinate frames, is discussed. Finally, we briefly discuss an s-rep variation in which object cross-sections are ellipses, and we show how, unlike all known object representations, it guarantees object means that avoid self-intersections.

1. OVERVIEW

If one wants to do statistics, such as classification or computations of means, on object shape, such as the ones shown in Fig. 1, it is important that the features used reflect as much as possible agreement on the localization of geometric properties held in common across the population. This characteristic is called *correspondence*. We compare two categories of object representations used for producing such object correspondences, given smooth-boundaried, protrusion-free object samples in the training and test sets, all specified by a relatively dense boundary mesh. Of the 3D object representations that have been claimed to be promising for statistics, the two most promising categories are the ones we study:

(1) Diffeomorphisms over 3D space derived after alignment via an LDDMM algorithm from the object mesh vertices [Durrleman 2014] or from the binary images describing the objects [Zhang 2019]; these are represented respectively by momentum images or initial velocity images, i.e., 3D arrays of vectors, with the 3D space covered by the array containing the objects (see Fig. 5).

(2) The skeletal representation called evolutionary s-reps described in this paper and with more history and details in [Pizer 2022]; in 3D these are represented

by a 2D skeletal grid in which at each grid-point a collection of directions and lengths from the skeleton to the boundary are provided (Fig. 2). From these, fitted-frame-based curvatures and inter-grid-point lengths are derived that prevent the need for pre-alignment. Though the methods can apply to objects of a wide variety of topologies and geometries, here we restrict ourselves to ones of spherical topology for which a swept sequence of possibly cross-sectional curved surfaces do not intersect within the object [Damon 2021] and which have no protrusions or indentations (see Fig. 1).

FIGURE 1. Target objects used in this study: A hippocampus (left) and a smoothed mandible (middle), both of which have no protrusions. Also shown on the right is the mandible, which has protrusions.

Our method of achieving statistical correspondence over a population of objects is achieved by using a representation seen to richly reflect object geometric properties throughout the *whole closure of the interior of the object,* to fit this to members of the object population, to provide corresponding local 3D coordinate systems within these regions so as to avoid preliminary alignment of objects, and to build the statistical approach of interest based on features derived from the coordinate systems. The statistical strength of our method and the common ones of boundary point distributions and of LDDMM-derived 3D arrays of vectors are each measured by classification performance on infants' hippocampi discriminated into autistic and control classes. Past experiments have shown that the simple pre-aligned boundary point distribution model (PDM) provides inferior statistical effectiveness than the two derived representations listed earlier, which are designed to capture a richer set of geometric properties than boundary locations alone. The pre-alignment required by both the LDDMM and PDM approaches propagates its error into the determination of correspondence. Also, they both ignore geometry related to the object interior, in particular (see Fig. 2, left), a) the curvature information of that 2D object and b) the bulge seen in its middle. Both the curvature and width features can be understood by fitting the boundary with a skeletal axis equipped with spoke vectors from the axis to the boundary (Fig. 2, 2nd and 4th from left, in 2D and 3D respectively). Moreover, we will show how the skeleton and its spokes can be used to avoid pre-alignment and to capture interior geometric features.

FIGURE 2. From left to right: 1) 2D object with curvature and a bulge, showing as dashed its skeletal (interior) axis. 2) "Spoke" vectors from the skeleton to the boundary capturing object width. Together the axis and the spokes form an s-rep [Pizer 2022]. 3) A 3D object, a hippocampus. 4) The "s-rep" for the hippocampus. 5) The s-rep for an ellipsoid, the base object for deformation into the target object. Both 3D s-reps show the object's skeleton as the top side of a folded surface with spherical topology. The s-reps' spines (skeletons of the skeletons) are shown in bold, and their fold curves are shown in yellow. Curves called "veins" proceed from the spine to both halves of the fold.

The s-reps approach is built on the intuition that object width and curvature of the object interior as they vary across the object are especially indicative of object shape and the way in which it corresponds across a population. In our recent work the s-rep computation, for any training or test object, can be understood as a sequence of diffeomorphic stages ending at the target object's s-rep and starting with an s-rep for the closure of the interior of the most basic form of this shape representation, namely an ellipsoid common over the population. The s-rep of the ellipsoid is its Blum medial skeleton, which is analytically known.

To reflect important geometric properties, the mapping from the ellipsoid s-rep to the target object s-rep must have spoke vectors that remain straight and the so-called radial distances (fractions of the distance along spokes) map onto the same radial distances in the target object's s-rep. The vertices (local maxima of Gaussian curvature) of the ellipsoid, which correspond to the vertices of the skeleton, must map onto ones of the target object, and the crests of the ellipsoid, which correspond to the fold curves of the skeleton, must map onto crests of the target object. This has led us to our new "evolutionary" method producing an ellipsoid-based diffeomorphism, which maintains these properties all the way through the stages of the diffeomorphism.

The means of novelly deriving fitted coordinate frames in the closure of a target object's interior depends on the s-rep because they are computed by differential geometry [E. Cartan, 1907] on the onions skins formed as the level sets of radial distance (Fig, 3). These frames throughout the object allow alignment-independent geometric features to be derived there to represent curvatures and vectors specifying inter-point relationships, both in the coordinates of the local frame. Section 2 shows how these frames and the geometric features are computed, so as to be in correspondence across the population.

Section 3 overviews the evolutionary method of fitting an s-rep to each of the objects in a sample set. Section 4 gives the results of our methods' comparison experiment on three object modeling approaches after overviewing the classification method and data we used for that comparison, the geometric features derived from the various methods,

and how we measured classification performance. Section 5 describes how non-self-intersecting object means can be produced from a form of s-rep whose cross-sections are planar ellipses. Section 6 summarizes our method and discusses the lessons on producing correspondence that were learned in this work and available generalizations to a wider class of anatomic objects.

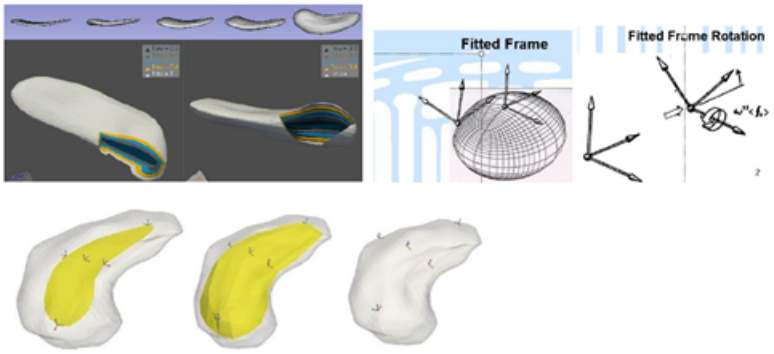


FIGURE 3. Top left: The onion skins for a hippocampus. Top middle: fitted frames on two points of an ellipsoid boundary. Top right: Rotation of a fitted frame in its own local coordinate system. Bottom: Fitted frames. Left: on the skeleton. Middle: on the onion skin halfway between the skeleton and the boundary. Right: on the boundary.

2. CORRESPONDENCE AND GEOMETRIC FEATURES VIA S-REPS

Our s-rep based intra-object coordinates (Fig. 4) lead to a correspondence that is insensitive to uniform widening 1) across the s-rep spokes, 2) along the long axis of the object, and 3) across the skeleton from one fold side to the other. The first of these coordinates, denoted $\tau_2 \in [0,1]$, being 0.0 at skeleton and 1.0 at boundary, is Damon's [Damon 2008a] *radial distance*, the fraction of the distance along the spoke from the skeleton to the boundary. The second coordinate, denoted θ, is cyclic along the spine of the s-rep, passing along the top side of the spine and back along the bottom side of the spine. The third coordinate, denoted $\tau_1 \in [0,1]$, captures distance from the spine along the veins as a fraction of their length, together with a flag indicating which side of the spine the vein is. All of these coordinates being fractions, they already are not sensitive to uniform scaling respectively in the three aforementioned directions.

Every position in the closure of an object interior has a unique value of (θ, τ_1, τ_2), and importantly, if the s-rep fitting is adequately reflective of the object geometry, the tuple provides the inter-object correspondences.

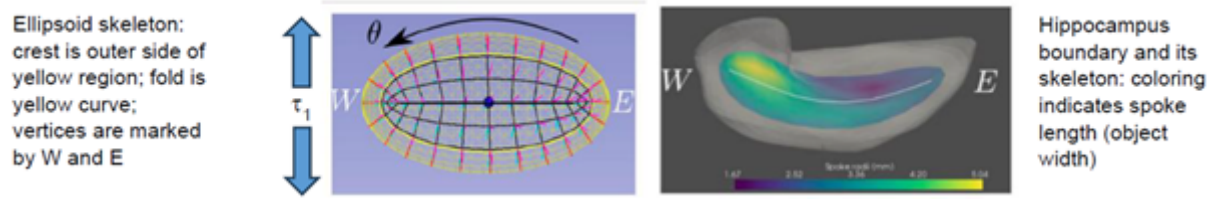


FIGURE 4. Object coordinates for ellipsoid and hippocampus.

In the studies for this paper the fitted frames are calculated along each spoke for τ_2 (radial distance) values of 0.0 (the skeleton), ± 0.25, ± 0.5, ± 0.75, and ± 1.0 (the boundary position of the spoke), where the negative values refer to spokes on the bottom side. The normal to the onion skin there and the pure θ direction in the tangent plane to the onion skin are computed using small shifts of the relevant spokes, with the shifted spokes computed using the spoke interpolation described in [Liu 2021].

The geometric features that we will use to characterize any object are all derived from its s-rep. There are three types of features: local frame curvatures of the onion skin in the $\Delta\theta$ and $\Delta\tau_1$ directions at any onion skin point, local vectors describing positional shifts between onion skin points, and, at each skeleton point (with the topside and bottomside skeleton points taken as being different) the spoke vector and the frame curvature between the two ends of each spoke (τ_2=0 and 1.0). Each feature at an onion skin point uses the coordinate system provided by the frame at that point. The spoke-related features use the coordinate system provided by the frame at the skeletal end of the spoke. The onion skin points used are along the respective spokes.

The hippocampi used in this experiment are spatially sampled with 61 interior skeletal points, of which 6 are along the spine ends' extensions, and 24 skeletal fold points. The features there, each Euclideanized and mean centered to allow Euclidean statistics to be applied [Jung, PNS], form a tuple of size 8,076.

To produce the competing representations of LDDMM (Large-Deformation Diffeomorphic Metric Mapping) features on a boundary mesh, the publicly available program called *Deformetrica* [Durrleman 2014] used the same boundary point arrays input to our method to deform the mean of the input object points to the corresponding target object points. The alternative works on binary images derived from the mesh. Both yield an energy-minimizing diffeomorphism over all of 3-space that carries the locations in the source object to their corresponding locations in the target object. Its momentum result, or a corresponding representation of initial velocities, is represented over an object-containing 3D Cartesian grid (see Fig. 5), with each vector Euclideanized and mean centered.

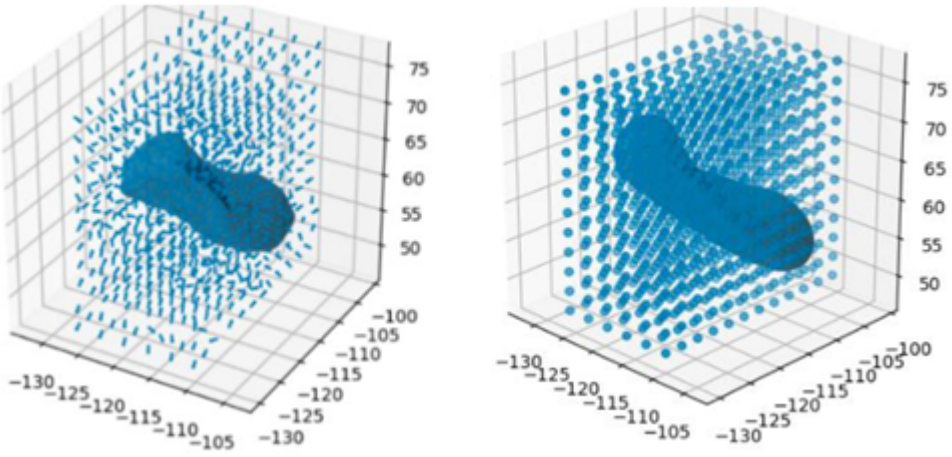


FIGURE 5. Left: Mean hippocampus in the grid of vectors. Right: Target hippocampus in the grid of vectors.

3. Fitting an S-rep to an Object Boundary via a Novel Evolutionary Method

The Pizer/Vicory evolutionary method for fitting an s-rep to a target object operates on the principle that the input boundary mesh representing each object in the population can be smoothly mapped to a common ellipsoid and that the s-rep analytically derived from that ellipsoid can be diffeomorphically *mapped back* to the target object. For the method to apply to any object of spherical topology, producing no singularities, we use the conformalized mean curvature flow method (CMCF) [Kazhdan 2012] for a forward flow. Very early in the CMCF mapping stages, whatever protrusions and indentations that were on the target object disappear. Soon thereafter, the curvatures of the object straighten out. See Fig. 6, where the mandible resolves into what we call a "bent hotdog", followed by its straightening and shortening into an ellipsoid.

Figure 6. Various stages of CMCF applied to a mandible. Far left: the mandible. Left of center: the bent hotdog. Right of center: the largely straightened hotdog. Far right: the resulting ellipsoid.

The CMCF is applied in stages from the target object. It yields a sphere in the limit, but it nears an ellipsoid on the way, which can be diffeomorphically mapped to the mean of the training objects' ellipsoids in an approach respecting the s-rep properties [Damon 2021]. At each stage of CMCF the boundary mesh computed is tested for nearness to an ellipsoid. When it is near enough, a final stage fits the flow-applied points to that proper ellipsoid.

The evolutionary method (Fig. 7) reverses the forward flow in stages and in doing so, transforms the ellipsoid's s-rep to that of the target object, all the while respecting geometric correspondences of the boundaries, spokes, vertices, crests, and s-rep folds.

We thus have an s-rep for each target in the population, from which the geometric features can be computed. Also from these stage-based transformations, a new diffeomorphism composing the inter-stage diffeomorphisms can be computed. Fig. 8 shows examples of the s-reps from the evolutionary method, both for simulated objects formed as a bent ellipsoid and for two of the hippocampus samples.

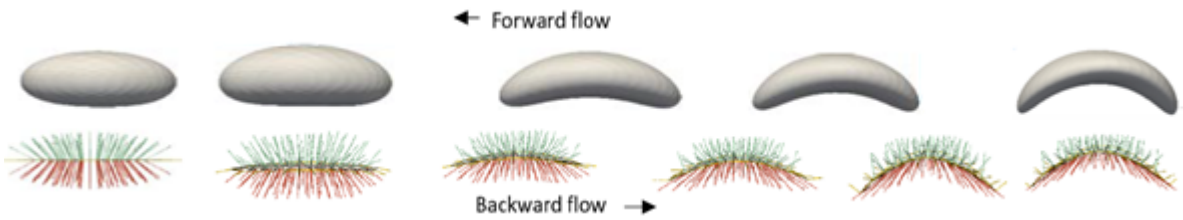


Figure 7. The s-rep fitting method on a bent ellipsoid. Right to left is forward (CMCF) flow. Left to right is the backward (s-rep fitting) flow, from the ellipsoid to the target bent ellipsoid.

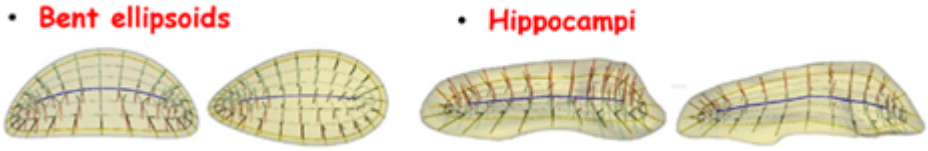


FIGURE 8. S-rep fitting results from the Pizer/Vicory evolutionary method.

4. TESTING VIA CLASSIFICATION PERFORMANCE

4.1. **Materials.** The objects with which we compare the classification performance of our three methods are hippocampi of 6-month olds (see Fig. 2) in two classes: those that developed autism symptoms in later years and those that did not. There are 34 cases in which autism developed and 143 in which it did not. For all the representations, we have chosen to have a number of starting features from each sample: 8,076 and 1296, respectively, to be much larger than the number of training samples: 177. By PCA each instance in these sets are reduced in dimension to 1 less than the number of data samples. The challenging classification problem is to predict from a hippocampus whether the young patient will later develop autism. The best of our previous methods on single hippocampi for doing this classification have yielded areas under the ROC (AUCs) of 0.6.

Two forms of the LDDMM method were evaluated. The first uses objects represented as boundary meshes and yields a momentum grid. Using a program due to Zhang [2019], the second uses objects represented as binary images and yields a grid of initial velocity vectors. However, to avoid a statistically invalid evaluation where the full data is used in forming the geometric features, for the LDDMM methods the mean was computed over a random quarter of both classes of the data and the evaluation was produced by the random holdout method on the remaining ¾ of the data. This process was repeated 6 times, and the resulting AUC and AUPR measures (see Table 1) were averaged over the repetitions. We checked and verified that the results were not materially affected by this form of averaging by repeating the experiment with the standard approach in the literature, in which the mean is computed from the whole data set and the random holdouts are applied to the whole data set.

4.2. **Methods of Comparison.** Each representation's data set is analyzed using a classification method to yield an Area Under an ROC Curve (AUC) as well as an Area Under the Precision, Recall Curve (AUPR). The AUC is a commonly used measure; the AUPR was chosen due to its lower sensitivity to the imbalance between the number of cases of each class. The method of random holdouts was applied using the classification method entitled Distance Weighted Discrimination (DWD) [Marron 2007]. For each holdout, the non-held-out cases are used to produce a separation direction in feature space, upon which all of the training cases are projected, forming two histograms, one for each class. Using those histograms, the public software for the method called SMOTE [Chawla 2002] was used to form an AUC and an AUPR for the collection of holdouts [Liu 2021]. The AUCs over 1000 holdouts produce the overall AUC and AUPR for that object representation method.

4.3. **Results.** Table 1 gives the AUC and AUPR values for each of the four object representations. While it is not possible to attach statistical significance levels to these values, due to the high correlation of the various random holdouts, the results indicate that the superior method for producing a representation aimed for statistical analysis is our evolutionary s-rep fitting method proposed in this paper. Of course, this result is on only one anatomic structure with respect to only one mental disorder, but it nevertheless suggests the superiority of representing the object interior using fitted frames, as well as the superiority of fitting s-reps to object boundary meshes by deformation of their interior closure from an ellipses in a way recognizing maintenance of s-rep relevant properties throughout the deformation.

Representation	**AUC**	**AUPR**
Evolutionary s-reps	**0.73**	**0.38**
Mesh diffeomorphism momenta	0.53	0.23
Binary diffeomorphism initial velocities	0.58	0.27

TABLE 1. AUCs and AUPRs for the Pizer/Vicory s-reps and diffeomorphism momenta and initial velocities, respectively. For both measures, a larger value shows better classification performance.

4.4. **Means Guaranteed Not to Self-Intersect.** Like all object representations that we know of, including our evolutionary s-reps, means of objects, each of which has no self-intersections can yield a mean that has self-intersections. A variant of s-reps due to Taheri [2024] in which the cross-sections consist of planar ellipsoids has yielded an approach through which the mean can be guaranteed to have no self-intersections.

5. CONCLUSIONS

This paper described a novel means of generating fitted frames in the closure of an object's interior and then generating alignment-independent geometric curvature and positional spacing primitives from those frames. This appears intellectually to be a notable advance, since in a number of human anatomic objects and disease categories [Pizer 2022, Liu 2023, Taheri 2023] it has shown its strength for generating statistically useful shape features, not only locally but across inter-object and intra-object locations. The method is based on s-reps fitted to mesh-represented object boundaries in the population in a way designed to produce interior correspondence across the instances, at least according to geometric features. This novel form of s-rep fitting operates by evolution through each of a succession of stages by warping an ellipsoid into the object such that the s-rep geometry relative to the warping object is maintained. The results of our evaluation confirm this behavior via its superior performance in classification based on shape, albeit in only one population of anatomic objects. In particular, it suggests that an object representation that highly recognizes the shape properties of the object interior, not just globally within the object but locally as well, can produce better statistics than one that is based on the limited properties of the object boundary alone. The method given here need not be restricted to objects with spherical topology. For

example, extension to objects with toroidal topology might be developed. Also, in the object interior location's correspondences produced by the s-rep-based coordinates other than geometric features, such as biological or image intensity ones can be provided; they can be used with the geometric features to produce even better correspondences. The details of the s-rep fitting method and of the comparison experiment are provided in the companion paper [Vicory 2024]. The software underlying the method [Vicory 2018, 2023] will very shortly be made available at the Slicer/SALT website: https://salt.slicer.org.

Acknowledgements. We appreciate help on this project and/or in writing this paper from James Fishbaugh, Mohsen Taheri, Md Asadullah Turja, and Ankur Sharma. Funding: This work was done with the partial support of NIH grant R01EB021391.

A longer version of this paper can be found at http://arxiv.org/abs/2407.14357

REFERENCES

[1] Cartan, E. La structure des groupes de transformations continus et la théorie du trièdre mobile, Bull. Soc. Math. France, t. 34: 250-284, or. Oeuvres complètes, Partie III, Vol. 1, 145-178, 1910.
[2] Cates, J., Fletcher, P. T., Styner, M., Shenton, M., & Whitaker, R. Proceedings 20. Shape modeling and analysis with entropy-based particle systems. In Information Processing in Medical Imaging: 20th International Conference, IPMI 2007, 333–345, 2007.
[3] Chawla, NV et al. SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research 16, 321-357, 2002.
[4] Cootes, T F, & Taylor, CJ. Statistical models of appearance for medical image analysis and computer vision. In Medical Imaging 2001: Image Processing, 4322, 236–248, SPIE.2001.
[5] Cury, C, JA Glaunès, R Toro, M Chupin, G Schumann, V Frouin, J-P Poline, O Colliot, and the Imagen Consortium. Statistical Shape Analysis of Large Datasets Based on Diffeomorphic Iterative Centroids. *Front. Neurosci.*, 12 November 2018, Sec. Brain Imaging Methods, 12, 2018, https://doi.org/10.3389/fnins.2018.00803
[6] Damon, JN. Geometry and medial structure. Ch. 3 in [Siddiqi 2008], 2008a.
[7] Damon, J. "Swept regions and surfaces: Modeling and volumetric properties." Theoretical Computer Science 392.1-3, 66–91, 2008b.
[8] Damon, JN. Thoughts on Ellipsoidal Models. Personal Communication, 2001.
[9] Dryden IL, KV Mardia. *Statistical Shape Analysis*, Wiley, 1998. Also available in a second edition, 2016.
[10] Durrleman, S., Prastawa, M., Charon, N., Korenberg, J. R., Joshi, S., & Gerig, G. Morphometry of anatomic shape complexes with dense deformations and sparse parameters. Neuroimage 101, 35–49, doi: 10.1016/j.neuroimage.2014.06.043. 2014.
[11] Hong, JP, J Vicory, J Schulz, M Styner, JS Marron, SM.Pizer. Non-Eudlidean classification of medically imaged objects via s-reps. *Medical Image Analysis*, no. 31, 37–45, 2016.
[12] Kazhdan, M, J Solomon, M Ben-Chen. Can mean-curvature flow be modified to be non-singular?. *Comput. Graphics Forum*, 31, 1745–1754. doi: 10.1111/j.1467-8659. 2012.03179.x. 2012.
[13] Liu, Z, JP Hong, J Vicory, JN Damon, SM Pizer. Fitting unbranching skeletal structures to objects. *Medical Image Analysis*, 2021.
[14] Liu, Z, J Damon, JS Marron, S Pizer. Geometric and Statistical Models for Analysis of Two-Object Complexes. *Int. J. Comp. Vis*, 1126-1123, 2023.
[15] Marron, JS, MJ Todd, J Ahn. Distance weighted discrimination. *Journal of the American Statistical Association*, 102 (480), 1267–1271, 2007.
[16] Pizer, SM and JS Marron. Objects statistics on curved manifolds. *Statistical Shape and Deformation Analysis*, no. (G. Zheng, S. Li, and others, eds.), 2017.

[17] Pizer, SM, JS Marrron, JN Damon, J Vicory, A Krishna, Z Liu, M Taheri. Skeletons, Object Shape, Statistics. *Frontiers in Computer Science*, 18 October 2022, Sec. Computer Vision https://doi.org/10.3389/fcomp.2022.842637, 2022.
[18] Styner, M., et al. (2006, 01). Statistical shape analysis of brain structures using SPHARMPDM. *The insight journal*,1071, 242-250, 2026.
[19] Taheri, M, SM Pizer, J Schulz et al. Fitting the discrete swept skeletal representation to slabular objects. *Under review for journal publication*, PREPRINT, Research Square [https://doi.org/10.21203/rs.3.rs-2927062/v1]. 2023.
[20] Taheri, M, SM Pizer, J Schulz. The mean shape under the relative curvature criterion. http://arxiv.org/abs/2404.01043, 2024.
[21] Tu, L; M Styner; J Vicory, S Elhabian; B Paniagua; JC Prieto; Dan Yang; R Whitaker; SM Pizer. Skeletal Shape Correspondence through Entropy. *IEEE Trans. Med. Img.*, 37, 1–11, 2016.
[22] Vicory J, Pascal L, Hernandez P, Fishbaugh J, Prieto J, Mostapha M, Huang C, Shah H, Hong J, Liu Z, Michoud L, Fillion-Robin JC, Gerig G, Zhu H, Pizer SM, Styner M, Paniagua B. Title of the paper. SlicerSALT: Shape AnaLysis Toolbox. Shape Med Imaging., 2018 Sep;11167:65-72. doi: 10.1007/978-3-030-04747-4_6. Epub 2018 Nov 23. PMID: 31032495; PMCID: PMC6482453. 2018.
[23] Vicory J, Han Y, Prieto JC, Allemang D, Leclercq M, Bowley C, Scheirich H, Fillion-Robin JC, Pizer SM, Fishbaugh J, Gerig G, Styner M, Paniagua B. SlicerSALT: From Medical Images to Quantitative Insights of Anatomy. Shape Med Imaging. 2023 Oct;14350:201-210. doi: 10.1007/978-3-031-46914-5_16. Epub 2023 Oct 31. PMID: 38250732; PMCID: PMC10798161. 2023.
[24] Vicory, J, N. Tapp-Hughes, J Zhang, Z Liu, SM Pizer. Fitting, Extraction, and Evaluation of Geometric Features. Companion paper to this paper, 2024.
[25] Zhang, M. and Fletcher, P.T., 2019. Fast diffeomorphic image registration via fourier-approximated lie algebras. *International Journal of Computer Vision*, 127, 61–73.

UNIV. OF NORTH CAROLINA AT CHAPEL HILL, UNIV. OF STAVANGER, AND KITWARE, INC.
Email address: pizer@cs.unc.edu

TOWARD A SCALING-ROTATION GEOMETRY OF SYMMETRIC POSITIVE DEFINITE MATRICES

ARMIN SCHWARTZMAN

Classification AMS 2020: 62R30, 62E20

Keywords: Fréchet mean; Riemmanian manifold; scaling-rotation distance

Symmetric positive definite matrices are familiar in statistics as covariance matrices. They also appear as data objects, particularly in brain imaging, such as in Diffusion Tensor Imaging [1, 3] and Tensor Based Morphometry [9, 11]. Many geometric frameworks have been developed for analysis of such data objects, including the log-Euclidean framework [2], affine-invariant framework [5], log- Cholesky framework [10], and Procrustes framework [4].

While these geometric frameworks account for the positive-definiteness constraint of these data objects, it is not clear which, if any, is most "natural" for describing deformations of symmetric positive definite matrices. We begin with the premise that a natural representation of such matrices is by their eigenvalue-eigenvector decomposition, because changes in that coordinate system can be directly interpreted as scaling and rotation in space.

Let $\mathrm{Sym}^+(p)$ be the space of $p \times p$ symmetric positive definite matrices, and let $\overline{\mathrm{Sym}^+}(p)$ be its closure, the space of $p \times p$ symmetric positive *semi*-definite matrices. Under the parametrization given by the eigenvalue-eigenvector decomposition, $\overline{\mathrm{Sym}^+}(p)$ takes the form of a Whitney-stratified manifold with 2^p strata, where strata are distinguished by the multiplicity of the eigenvalues. Half of these strata lie on the boundary of $\overline{\mathrm{Sym}^+}(p)$, corresponding to matrices of rank less than p, while the other half lie on the interior of $\mathrm{Sym}^+(p)$, corresponding to matrices of full rank equal to p. The dimension of the interior strata can be computed explicitly.

This geometry can be used to define a *scaling-rotation* (SR) distance on $\mathrm{Sym}^+(p)$, introduced by [8], measuring scaling of eigenvalues and rotation of eigenvectors. When restricted to the top stratum, corresponding to the set of symmetric positive definite matrices whose eigenvalues are all distinct, this distance is shown to be a Riemannian metric and turns this set into a complete Riemannian manifold. The geodesics according to this geometry are *scaling-rotation* curves that smoothly transform one symmetric positive definite matrix to another by scaling its eigenvalues and rotating its eigenvector. The shortest geodesic is called the *minimal* scaling-rotation curve.

When extended to the full space $\mathrm{Sym}^+(p)$, the SR distance is no longer a metric but only a *semi*-metric, because the triangle inequality does not hold. This is because paths traversing through lower strata may be shorter than those remaining in higher strata. An implication of this result is that Fréchet means cannot be directly computed, as they require properly defined metrics.

For a set of random points on $\mathrm{Sym}^+(p)$, we formally define the SR mean set to be the set of Fréchet means in $\mathrm{Sym}^+(p)$ with respect to the SR distance. Since computing such

means requires a difficult optimization, we also define an extension of the Fréchet mean set called *partial SR* (PSR) mean set, lying in the space of eigen-decompositions as a proxy for the SR mean set. The PSR mean set corresponds to a set of elements of $\mathrm{Sym}^+(p)$ whose eigendecompositions minimize the average squared scaling-rotation distance to the original points. It is easier to compute and its projection to $\mathrm{Sym}^+(p)$ often coincides with SR mean set. In the eigenvalue-eigenvector parametrization, the set of partial scaling-rotation means may have up to $2^{p-1}p!$ elements. However, if the support of the distribution of random points is small enough (smaller than the regular convexity radius of the quotient space induced by the sign changes and permutation of the eigenvalues), then the corresponding eigenvalue-eigenvector composition on $\mathrm{Sym}^+(p)$ is unique. Following the techniques in [6, 7], we also establish strong consistency of the sample PSR means as estimators of the population PSR mean set, and a central limit theorem.

In an application to multivariate tensor-based morphometry, we demonstrate that a two-group test using the proposed PSR means can have greater power than the two-group test using the usual affine-invariant geometric framework for symmetric positive-definite matrices.

REFERENCES

[1] Alexander, D.C. Multiple-fiber reconstruction algorithms for diffusion MRI. *Ann. N.Y. Acad. Sci.*, 1064, 113–133, 2005.

[2] Vincent Arsigny, Pierre Fillard, Xavier Pennec, and Nicholas Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29, 328-347, 2006.

[3] Batchelor, P.G., Moakher, M., Atkinson, D., Calamante, F. and Connelly, A. A rigorous framework for diffusion tensor calculus. *Magn. Reson. Med.*, 53, 221–225, 2005.

[4] I. L. Dryden, Alexey Koloydenko, and Diwei Zhou. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. *Annals of Applied Statistics*, 3, 1102-1123, 2009.

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

[6] Huckemann, S. Inference on 3D Procrustes means: Tree bole growth, rank deficient diffusion tensors and perturbation models. *Scand. J. Stat.*, 38, 424–446, 2011a.

[7] Huckemann, S. Intrinsic inference on the mean geodesic of planar shapes and tree discrimination by leaf growth. *Ann. Statist.*, 39, 1098–1124, 2011b.

[8] Jung, S., Schwartzman, A. and Groisser, D. Scaling-rotation distance and interpolation of symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.*, 36, 1180–1201, 2015.

[9] Lepore, F., Brun, C., Chou, Y.-Y., Chiang, M.-C., Dutton, R., Hayashi, K., Luders, E., Lopez, O., Aizenstein, H., Toga, A.W., Becker, J. and Thompson, P. Generalized tensor-based morphometry of HIV/AIDS using multivariate statistics on deformation tensors. *IEEE Trans. Med. Imag.*, 27, 129–141, 2008.

[10] Zhouchen Lin. Riemannian geometry of symmetric positive definite matrices via Cholesky decomposition. *SIAM Journal on Matrix Analysis and Applications*, 40, 1353-1370, 2019.

[11] Paquette, N., Shi, J., Wang, Y., Lao, Y., Ceschin, R., Nelson, M., Panigrahy, A. and Lepore, N. Ventricular shape and relative position abnormalities in preterm neonates. *NeuroImage Clin.*, 15, 483–493, 2017.

HALICIOĞLU DATA SCIENCE INSTITUTE, UNIVERSITY OF CALIFORNIA, SAN DIEGO, LA JOLLA, CA 92093, UNITED STATES

Email address: armins@ucsd.edu

KUNITA FLOWS, SHAPE STOCHASTICS, AND PHYLOGENETIC INFERENCE

STEFAN SOMMER

classification AMS 2020: 60H10, 60J70, 62P10

Keywords: Kunita flows, shape stochastics, phylogenetic inference

The Brownian motion model of trait evolution is widely used in biology to model the evolution of phenotypic traits. While such traits are often specific and low-dimensional, modelling the evolution of the entire morphology requires high-dimensional, correlated shape processes. If the number of morphological features used to represent the shapes is increased to represent full continuous shapes, the models become infinite-dimensional.

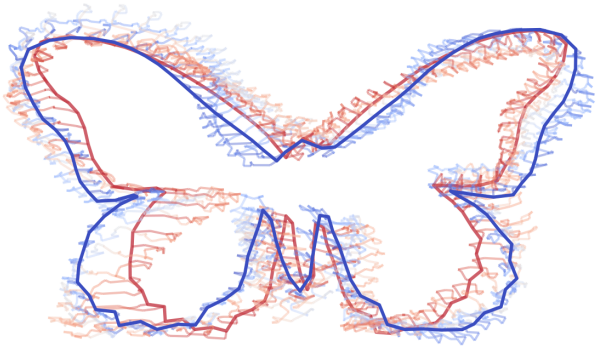


FIGURE 1. Outline of two butterflies connected with a bridge, here a Kunita flow applied to the landmarks of one butterfly and conditioned on hitting the landmarks of the other butterfly at a fixed time. The correlation between nearby points can be observed.

We presented an axiomatic approach to shape stochastics, that asks for shape stochastic processes that

- apply to multiple representations of shapes, e.g. landmarks, curves, surfaces and images,
- are independent of specific discretization,
- preserve shape structure,
- model correlation between points,
- is invariant to acting similarity groups, e.g. rigid body transformations.

A consequence of these properties is that such shape processes often cannot be modelled linearly, thus leading to non-linear and state dependent processes. This is the case because close points on the shape must be correlated to preserve the shape structure, and this correlation between points will change if initially far away points are

brought closer during the evolution of the process. Figure 1 illustrates the correlation between nearby points on the shape. The axiomatic approach to shape stochastics is further detailed in [5].

One class of shape processes that satisfies these properties are Kunita flows [3, 4]. We outlined the use of Kunita flows to model shape evolution, and how they can be used to condition shape evolution on phenotypic traits. This applies also to branching processes that models the evolution governed by a phylogenetic tree.

With shape observations, we can condition on the observed shapes at the leaves of the tree, and then use the resulting likelihood to infer parameters of the process, e.g. the correlation structure of the Kunita flow. The conditioning on the full shape observation is explored in [2]. For finite dimensional observations, e.g. finite numbers of landmarks on the shape, we perform the statistical inference using the backwards filtering, forwards guiding approach of van der Meulen and Schauer [6]. We outlined the application of this algorithm for inference of the parameters of the Kunita flow from landmark data, and its implementation in the software package Hyperiax [1]

References

[1] Hyperiax. `https://github.com/ComputationalEvolutionaryMorphometry/hyperiax/`, 2024.

[2] Elizabeth Louise Baker, Gefan Yang, Michael L. Severinsen, Christy Anna Hipsley, and Stefan Sommer. Conditioning non-linear and infinite-dimensional diffusion processes. arXiv:2402.01434, February 2024.

[3] Hiroshi Kunita. *Lectures on Stochastic Flows and Applications.* Springer, Berlin Heidelberg, 1st edition edition, January 1986.

[4] Hiroshi Kunita. *Stochastic Flows and Stochastic Differential Equations*. Cambridge University Press, April 1997.

[5] Stefan Sommer, Gefan Yang, and Elizabeth Louise Baker. Stochastics of shapes and Kunita flows. *submitted*, 2025.

[6] Frank van der Meulen and Moritz Schauer. Automatic Backward Filtering Forward Guiding for Markov processes and graphical models. arXiv:2010.03509, October 2022.

Department of Computer Science, University of Copenhagen, Copenhagen, Denmark
Email address: `sommer@di.ku.dk`

MODULI SPACES OF METRICS AND LOCALLY SYMMETRIC SPACES

WILDERICH TUSCHMANN

Classification AMS 2020: 53C20

Keywords: Moduli Spaces of Metrics, Locally Symmetric Spaces

In my talk, I provided a gentle introduction to the study of moduli spaces of metrics on Riemannian manifolds and more singular spaces and, in the flat and Ricci flat case, related upon their connection to certain locally symmetric spaces and orbifolds of potential interest in applications. In particular, I presented in detail the following results:

Theorem 0.1. *In each dimension $4k+3$, $k \geq 1$, there exist infinite sequences of closed smooth simply connected manifolds of pairwise distinct homotopy type for which the moduli space of Riemannian metrics with nonnegative sectional curvature has infinitely many path components.*

For further details, see [1]. Moreover, in this regard we also have (compare [2]):

Theorem 0.2. *In every dimension $4k+1$, $k \geq 2$, there are infinite sequences of closed manifolds with pairwise nonisomorphic integral cohomology for which the moduli space of metrics of nonnegative sectional curvature has infinitely many path components.*

The following theorem from [3] is so far the only result about the higher homotopy (and cohomology) groups of the moduli spaces of Ricci nonnegative metrics on closed manifolds:

Theorem 0.3. *Let M be a simply connected closed smooth manifold which admits a metric with nonnegative Ricci curvature and T be a torus of dimension $k \geq 4$, $k \neq 8, 9, 10$. Then the moduli space of nonnegatively Ricci curved metrics on $M \times T$ has non-trivial higher rational cohomology groups and non-trivial higher rational homotopy groups.*

In particular, this also allows to infer:

Corollary 0.4. *In every dimension $n \geq 4, n \neq 5$ there exist closed smooth manifolds M^n for which the third rational cohomology group and the third rational homotopy group of the moduli space of metrics with nonnegative sectional curvature on M^n are non-trivial.*

In his Ph.D. thesis [4], the present author's former student David Degen used these facts as a starting point to prove, among others, the following results:

Theorem 0.5. *The moduli space of Ricci flat metrics on the $K3$ manifold K is simply connected and its second Betti number is positive.*

Theorem 0.6. *The moduli space of Ricci flat metrics, including orbifold metrics, on the $K3$ manifold K is simply connected and its fourth Betti number is at least 1.*

Crossing the $K3$ manifold with flat tori and simply connected round spheres, respectively, this also yields:

Corollary 0.7. *In every dimension $n \geq 4$ there exist closed smooth manifolds M^n for which the second rational homotopy group of the moduli space of Ricci flat metrics on M^n is non-trivial.*

Corollary 0.8. *In every dimension $n \geq 4, n \neq 5$ there exist simply connected closed smooth manifolds M^n for which the second rational homotopy group of the moduli space of nonnegatively Ricci curved metrics on M^n is non-trivial.*

To conclude (see [5]), let us also embark upon the moduli spaces of flat metrics on Bieberbach manifolds, giving rise to certain distinguished subsets of the spaces of positive definite symmetric matrices, which, hence, do also bear, but hitherto unexplored, close relations to non-Euclidean statistics:

Theorem 0.9. *If Γ is a Bieberbach group on Euclidean n-space $\mathbb{R}^n$, its moduli space of flat metrics on $\Gamma\backslash\mathbb{R}^n$ is homeomorphic to a locally symmetric space or orbifold of noncompact type.*

References

[1] Anand Dessai, Stephan Klaus, and Wilderich Tuschmann. Nonconnected moduli spaces of nonnegative sectional curvature metrics on simply connected manifolds. *Bull. Lond. Math. Soc.*, 50(1), 96–107, 2018.

[2] Anand Dessai. Moduli space of nonnegatively curved metrics on manifolds of dimension $4k + 1$. *Algebr. Geom. Topol.* 22 325–347, 2022.

[3] Wilderich Tuschmann and Michael Wiemeler. On the topology of moduli spaces of non-negatively curved Riemannian metrics. *Math. Ann.* 384, No. 3-4, 1629–1651, 2022.

[4] David Degen. On the Global Topology of Moduli Spaces of Riemannian Metrics with Holonomy $Sp(n)$. Doctoral thesis, *Karlsruher Institut für Technologie (KIT)*, 183 pages, 2023. https://publikationen.bibliothek.kit.edu/1000155796.

[5] Wilderich Tuschmann. Moduli spaces of flat metrics on Bieberbach manifolds. In preparation.

MY ADDRESS

Email address: tuschmann@kit.edu

TOPOLOGICAL DEEP LEARNING: THE PAST, PRESENT, AND FUTURE

GUO-WEI WEI

Classification AMS 2020: 55N99; 68W01; 57M99; 55T05; 52-02

Keywords: Topological deep learning, Artificial intelligence, Persistent homology, Persistent Laplacians, Algebraic topology, Geometric topology, Differential geometry

Artificial intelligence (AI) represents one of the most transformative advancements in human history. Deep learning (DL), a subfield of machine learning (ML), has revolutionized AI by enabling machines to learn complex patterns through multi-layered neural networks. This has positioned DL as a crucial area of study within computer science, statistics, and mathematics, all of which are fundamental to AI, ML, and DL. While calculus, linear algebra, probability, and optimization form the core mathematical foundations of ML and DL, advanced mathematical areas such as geometry, topology, algebra, and combinatorics also play significant roles and hold the future of AI, ML, and DL. Concepts like the Wasserstein metric and geometric measure theory are integral to numerous ML and DL algorithms. Notably, the intersection of DL and algebraic topology has led to the emergence of topological deep learning (TDL), a field that offers innovative methods for analyzing complex datasets, particularly high-dimensional, nonlinear structures that are challenging for traditional physical and statistical approaches. The term "Topological Deep Learning" was first introduced in 2017 to describe the integration of topological features into deep neural network input pipelines [1] and has since evolved to encompass a broader range of methodologies that apply topological concepts to deep learning [2].

TDL combines the expressive power of deep neural networks with the mathematical rigor of topology. Traditional ML methods often struggle to capture the geometric and structural properties of data, a limitation that TDL addresses by leveraging topological features and representations. TDL methods are broadly categorized as observational or interventional [3]. Observational methods focus on understanding deep learning models through their topological properties, while interventional methods incorporate topological structures to enable architectures to effectively handle data with inherent topological properties.

A key component of TDL is persistent homology [4, 5], a technique from algebraic topology [6] that connects abstract topology and geometry. Persistent homology performs multiscale analysis by quantifying topological invariants, such as loops, voids, and higher-dimensional structures, within datasets. Integrating persistent homology into deep learning frameworks allows for the extraction of meaningful topological signatures, improving tasks such as classification, regression, clustering, and anomaly detection. TDL also explores the direct incorporation of topological structures into neural networks, activation functions, and loss functions. Examples such as topological transformers [7] and simplicial neural networks [3] highlight the versatility of this

approach[8]. It is worth noting that AlphaFold 2, recognized with the 2024 Nobel Prize in Chemistry, employed transformers to predict protein structures.

TDL offers several advantages, including increased interpretability compared to traditional ”black-box” deep learning models. By extracting intuitive topological signatures, TDL provides insights into data structures and model predictions [1]. Furthermore, TDL demonstrates robustness to noise and excels at handling high-dimensional data by focusing on inherent data topologies, revealing patterns often obscured by noise or irrelevant features. TDL models also exhibit strong generalization capabilities [9, 10] due to their emphasis on fundamental geometric and topological properties. Applications of TDL span various fields, including biology, chemistry, neuroscience, social networks, and computer vision [2, 11], with examples including advancements in disease diagnosis, drug design, chip design, and graph analysis. Its efficacy is demonstrated by its success in the D3R Grand Challenges for computer-aided drug design [9, 12], its discovery of SARS-CoV-2 evolution mechanisms [13, 14], and its accurate predictions of emerging dominant SARS-CoV-2 variants [15, 10].

Despite its potential, TDL is a relatively nascent field with significant opportunities and challenges. Integrating domain-specific knowledge into TDL models can further enhance performance and interpretability. Future research may focus on effectively incorporating prior knowledge while advancing topological theories beyond homology. Exploring new topological formulations such as Laplacians and Dirac operators, and expanding topological domains to cell complexes, path complexes, hypergraphs, knots, links, and tangles, will further strengthen the field [16]. TDL will also be benefited from other mathematical fields, such as differential geometry [17] and geometric topology [18]. This comprehensive approach will ensure TDL's continued relevance and effectiveness in addressing complex real-world problems and driving innovation in data-driven research.

In conclusion, TDL represents a trending paradigm that merges with the computational power of deep learning with the mathematical richness of topology, including algebraic topology, differential topology, and geometric topology. Its robustness, interpretability, and versatility make it a valuable tool for analyzing complex datasets across diverse domains. Just as partial differential equations (PDEs) significantly shaped applied mathematics for decades, TDL has the potential to inspire future generations of mathematicians and computer science researchers, paving the way for new solutions in AI, mathematics, and beyond.

REFERENCES

[1] Zixuan Cang and Guo-Wei Wei. Topologynet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions. *PLoS computational biology*, 13(7):e1005690, 2017.

[2] Theodore Papamarkou, Tolga Birdal, Michael M Bronstein, Gunnar E Carlsson, Justin Curry, Yue Gao, Mustafa Hajij, Roland Kwitt, Pietro Lio, Paolo Di Lorenzo, et al. Position: Topological deep learning is the new frontier for relational learning. In *Forty-first International Conference on Machine Learning*, 2024.

[3] Felix Hensel, Michael Moor, and Bastian Rieck. A survey of topological machine learning methods. *Frontiers in Artificial Intelligence*, 4:681108, 2021.

[4] G. Carlsson. Topology and data. *Am. Math. Soc*, 46(2):255–308, 2009.

[5] Herbert Edelsbrunner, John Harer, et al. Persistent homology-a survey. *Contemporary mathematics*, 453:257–282, 2008.
[6] Tomasz Kaczynski, Konstantin Mischaikow, and Marian Mrozek. *Computational Homology*, volume 157 of *Applied Mathematical Sciences*. Springer-Verlag, New York, 2004.
[7] Dong Chen, Jian Liu, and Guo-Wei Wei. Multiscale topology-enabled structure-to-sequence transformer for protein–ligand interaction predictions. *Nature Machine Intelligence*, 6(7):799–810, 2024.
[8] Chi Seng Pun, Si Xian Lee, and Kelin Xia. Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review*, 55(7):5169–5213, 2022.
[9] Duc Duy Nguyen, Zixuan Cang, Kedi Wu, Menglun Wang, Yin Cao, and Guo-Wei Wei. Mathematical deep learning for pose and binding affinity prediction and ranking in d3r grand challenges. *Journal of computer-aided molecular design*, 33(1):71–82, 2019.
[10] Jiahui Chen, Yuchi Qiu, Rui Wang, and Guo-Wei Wei. Persistent Laplacian projected Omicron BA.4 and BA.5 to become new dominating variants. *Computers in Biology and Medicine*, 151:106262, 2022.
[11] Jacob Townsend, Cassie Putman Micucci, John H Hymel, Vasileios Maroulas, and Konstantinos D Vogiatzis. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nature communications*, 11(1):3230, 2020.
[12] Duc Duy Nguyen, Kaifu Gao, Menglun Wang, and Guo-Wei Wei. MathDL: mathematical deep learning for D3R Grand Challenge 4. *Journal of computer-aided molecular design*, 34:131–147, 2020.
[13] Jiahui Chen, Rui Wang, Menglun Wang, and Guo-Wei Wei. Mutations strengthened SARS-CoV-2 infectivity. *Journal of Molecular Biology*, 432:5212–5226, 2020.
[14] Rui Wang, Jiahui Chen, and Guo-Wei Wei. Mechanisms of SARS-CoV-2 evolution revealing vaccine-resistant mutations in Europe and America. *The journal of physical chemistry letters*, 12(49):11850–11857, 2021.
[15] Jiahui Chen and Guo-Wei Wei. Omicron BA. 2 (B. 1.1. 529.2): high potential for becoming the next dominant variant. *The journal of physical chemistry letters*, 13(17):3840–3849, 2022.
[16] Xiaoqi Wei and Guo-Wei Wei. Persistent topological laplacians—a survey. *Mathematics*, 13(2):208, 2025.
[17] Zhe Su, Yiying Tong, and Guo-Wei Wei. Persistent de rham-hodge laplacians in eulerian representation for manifold topological learning. *AIMS Mathematics*, 9(10):27438–27470, 2024.
[18] Li Shen, Jian Liu, and Guo-Wei Wei. Evolutionary khovanov homology. *AIMS Mathematics*, 9(9):26139–26165, 2024.

DEPARTMENT OF MATHEMATICS, MICHIGAN STATE UNIVERSITY
Email address: weig@msu.edu

GLMY THEORY AND TOPOLOGICAL STATISTICS

JIE WU

Classification AMS 2020: 18G85, 55N31, 62R40, 05C20, 05C65,

Keywords: digraphs, hypergraphs, super-hypergraphs, high-order interaction networks, GLMY theory, path homology, embedded homology, topological data analysis

We give an introduction to the new research field GLMY theory that may become the topological component of future topological statistics, presented in a way through the views of the topological approaches beyond simplicial complex to data and complex networks.

Topological dada analysis (TDA) is a research area born in 2009 by a landmark paper of Gunnar Carlsson [7], after the poineering works [1, 11, 23]. TDA was popularlized since then, with leading to the birth of the new area of topological deep learning (TDL) from the first exploration [6] in 2017. The commonly used methodolgy in TDA is the persistent homology (PH) of the simplicial complex modeling on point cloud data using Vietoris-Rips complex construction together with the persistence from the Euclidean distance giving rise to multiscaled topological feature of data. Such an approach achieves great success in data analytics, and also becomes a popular tool in topological machine learning (TML) and topological deep learning. Mathematically, this is a simplicial complex approach, where the classical Čech nerve theorem guarantees the theory to effectively detect the geometric shape of data.

The most challenge in network science is to uderstand the mechanism of high-order interactions in complex systems, which has been intensively studied with achieving fruitful results [2, 4, 10, 15, 16]. Simplicial complex becomes an important tool for networks beyond pairewise interactions, which can overcome some of the problems encountered by other lower dimensional representations. However, simplicial complexes are still quite limited by the requirement on the existence of all subfaces. Hypergraph becomes a choice of many researchers as a more general modeling on unconstrained description of high-order interactions. From the topological view, a hypergraph can be considered as "*a simplicial complex with some missing faces*", which gives rise a mathematical question whether simplicial homology theory can be extended to hypergraphs. Developed from the path homology theory of digraphs [14] introduced by Shing-Tung Yau et al in 2012—the foundational work of GLMY theory, this question received an affirmative answer in [5] with introducing the notion of embedded homology on hypergraphs as a natural extension of simplicial homology. The embedded homology of hypergraph performs effetively on protein-ligand binding affinity prodiction in drug design [19, 20].

The notion of hypergraph has been extended as super-hypergraph introduced in [13], where a super-hypergraph could be viewed as a multiset version of hypergraph. In this work, we developed the embedded homology theory of super-hypergraphs that can unifies various aspects of topological approaches for data science, by being applicable

both to point cloud data and to graph data, including networks beyond pairwise interactions.

IdopNetwork [9] is a statistical modeling introduced by Rongling Wu et al in 2019, which gives an informative dynamic ominidirectional and personalized network reconstruction. The intergration of IdopNetwork and the path homology of digraphs (GLMY homology) achieved various successful applications in biology and medicine [8, 12, 18, 22].

Hypergraphes are still limited that could not decsribe some phenomena of interactions between different hyperedges in biology and social sciences. In [21], we introduced the notion of interaction complex (IntComplex) as a mathematical modeling for describing the phenomena of interactions between hyper-edges using binary trees, and applied the embedded homology of super-hypergraphs for establising a topological theoy on interaction complexes, which gives an updated developemnt of GLMY theory. Then, in the work [17], we developed a generalized statistical mechanics model to reconstruct bidirectional, signed, and weighted hypernetworks that characterize how constituent agents are influenced by their own strategies, the strategies of co-existing agents, and strategies of interactions between other agents, as well as how directed pairwise interactions are influenced by individual agents, and then integrated this new stastical mechanics modeling with GLMY theory on IntComplex to dissect the topological architecture of hypernetworks in terms of nodes, links and hyperlinks, which provides a generic tool for unveiling hidden patterns in complex systems across a wide spectrum of physical and biological scenarios.

REFERENCES

[1] S. Framed Barannikov Morse complexes and its invariants. *Adv. Soviet Math.*, 22, 93–115, 1994.

[2] F. Battiston, E. Amico, A. Barrat, et al. The physics of higher-order interactions in complex systems. *Nat. Phys.*, 17, 1093–1098, 2021.

[3] Federico Battiston, Giulia Cencetti, Iacopo Iacopini, Vito Latora, Maxime Lucas, Alice Patania, Jean-Gabriel Young, and Giovanni Petri Networks beyond pairwise interactions: Structure and dynamics *Physics Reports*, 874, 1-92, 2020.

[4] S. Boccaletti, P. De Lellis, C.I. del Genio, K. Alfaro-Bittner, R. Criado, S. Jalan, and M. Romance The structure and dynamics of networks with higher order interactions *Physics Reports*, 1018, 1–64, 2023.

[5] Stephane Bressan, Jingyan Li, Shiquan Ren, and Jie Wu the embedded homology of hypergraphs and applications *Asian J. Math.* 23 (3), 479-500, 2019.

[6] Z. Cang, and G.-W. Wei TopologyNet: Topology based deep convolutional and multi-task neural networks for biomolecular property predictions *PLoS Comput. Biol.*, 13(7), e1005690, 2017.

[7] G. Carlsson topology and data *AMS Bulletin*, 46(2), 255–308, 2009.

[8] Jincan Che, Huiying Gong, Shen Zhang, Xiang Liu, Yu Wang, Claudia Gragnoli, Christopher Griffin, Jie Wu, Shing-Tung Yau, and Rongling Wu IdopNetwork as a genomic predictor of drug response *Drug Discovery Today*, 30 (1), 104252, 2025.

[9] Chen, C., Jiang, L., Fu, G. et al. An omnidirectional visualization model of personalized gene regulatory networks *npj Systems Biology and Applications*, 5, Article number 38, 2019.

[10] J. Domingo, G. Diss, and B. Lehner Pairwise and higher-order genetic interactions during the evolution of a tRNA *Nature*, 558, 117–121, 2018.

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

[12] Huiying Gong, Hongxing Wang, Yu Wang, Shen Zhang, Xiang Liu, Jincan Che, Shuang Wu, Jie Wu, Xiaomei Sun, Shougong Zhang, Shing-Tung Yau, and Rongling Wu, Topological change of soil

microbiota networks for forest resilience under global warming *Physics of Life Reviews*, 50, 228-251, 2024.
[13] Jelena Grbić, Jie Wu, Kelin Xia, and Guo-Wei Wei aspects of topological approaches for data science *Foundations of Data Science*, 4(2), 165-216, 2022.
[14] Alexander Grigor'yan, Yong Lin, Yuri Muranov, and Shing-Tung Yau, homologies of path complexes and digraphs, arXiv:1207.2834, 2012
[15] Grilli, J., Barab´as, G., Michalska-Smith, M. et al. Higher-order interactions stabilize dynamics in competitive network models *Nature*, 548, 210–213, 2017.
[16] Jonathan M. Levine, Jordi Bascompte, Peter B. Adler, and Stefano Allesina Beyond pairwise mechanisms of species coexistence in complex communities *Nature*, 546, 56–64, 2017.
[17] Li Feng, Huiying Gong, Shen Zhang, Xiang Liu, Yu Wang, Chengwen Xue, Christopher H. Griffin, Jie Wu, Shing-Tung Yau, and Rongling Wu Hypernetwork modeling and topology of high-order interactions for complex systems *PNAS*, 121(40), e2412220121, 2024.
[18] Li Feng, Dengcheng Yang, Sinan Wu, Chengwen Xue, Mengmeng Sang, Xiang Liu, Jie Wu, Claudia Gragnoli, Christopher Griffin, Chen Wang, Shing-Tung Yau, and Rongling Wu Network modeling and topology of aging *Physics Reports*, 1101, 1-65, 2025.
[19] Xiang Liu, Huitao Feng, Jie Wu, and Kelin Xia Persistent spectral hypergraph based machine learning (PSH-ML) for protein-ligand binding affinity prediction *Briefings in Bioinformatics*, 22(5), bbab127, 2021.
[20] Xiang Liu, Xiangjun Wang, Jie Wu, and Kelin Xia hypergraph based persistent cohomology (HPC) for machine learning in drug design *Briefings in Bioinformatics*, 22(5), bbaa411, 2021.
[21] Xiang Liu, Ran Liu, Jingyan Li, Rongling Wu, and Jie Wu IntComplex for high-order interactions arXiv:2412.02806, 2024.
[22] Shuang Wu, Xiang Liu, Ang Dong, Claudia Gragnoli, Christopher Griffin, Jie Wu, Shing-Tung Yau, and Rongling Wu the metabolomic physics of complex diseases *PNAS*, 120 (42), e2308496120, 2023.
[23] A. Zomorodian and G. Carlsson Computing persistent homology *Discrete Comput. Geom.*, 33, 249–274, 2005.

BEIJING INSTITUTE OF MATHEMATICAL SCIENCES AND APPLICATIONS, NO. 544, HEFANGKOU VILLAGE, HUAIROU TOWN, HUAIROU DISTRICT, BEIJING 101408, CHINA
Email address: wujie@bimsa.cn