Snapshot of Modern Mathematics: Biological Shape Analysis with Geometric Statistics and Learning

Saiteja Utpala*   Nina Miolane†

January 17, 2022

Abstract

The advances in biomedical imaging techniques have enabled us to access the *3D shapes* of a variety of structures: organs, cells, proteins. Since biological shapes are related to physiological functions, shape data may hold the key to unlock outstanding mysteries in biomedicine. This snapshot introduces the framework of Geometric Statistics and Learning and its applications to biomedicine. We show how this mathematics allows performing computations on shape data, which then enables running statistical learning algorithms on biological shapes.

1 Introduction

Vast quantities of biological imaging data are currently been generated by high-throughput imaging systems. In this context, statistical learning is poised to play a major role in making sense of the wealth of incoming information. Foundational mathematical research defining the appropriate learning tools to study biological features, such as the biological shapes shown in Figure 1, is thus critical and timely.

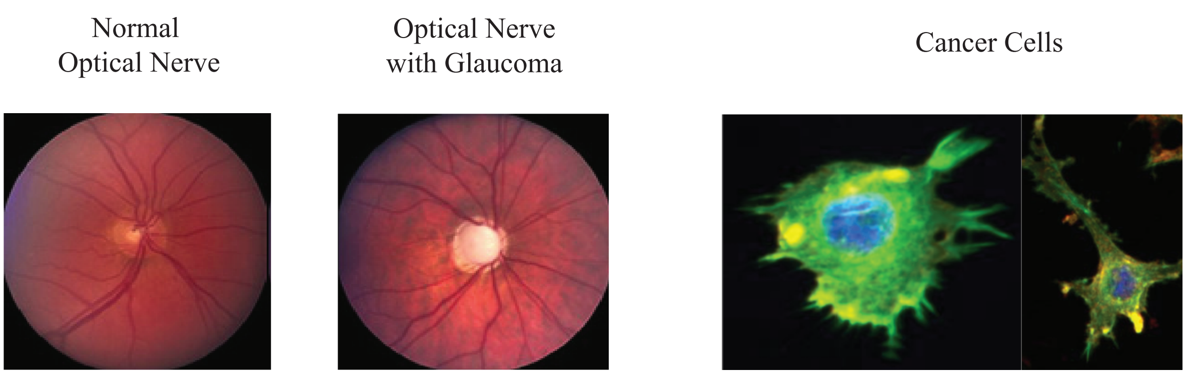

Figure 1: Left: Optical nerve heads in monkeys showing a normal anatomy, or a pathological shape corresponding to Glaucoma [PE15]. Right: Cancer cells observed by fluorescence microscopy. Cancers cells typically exhibit a irregular boundary, in contrast to healthy cells. Courtesy of Ashok Prasad, Associate Professor of Chemical and Biological Engineering at Colorado State University.

From a statistical perspective, *shapes* are fascinating data objects, as they can be modelled as points in a high-dimensional curved space, called a manifold. This makes computations on shape data possible through "Riemannian geometry", a mathematical theory that conveniently describes shape spaces, but that also contributed to significant advances throughout the history of sciences — notably as the foundational mathematics of "General Relativity". We introduce Riemannian geometry and the associated field of "Geometric Statistics and Learning" to showcase the potential of this mathematics for advances in biomedical shape analysis and computational medicine. The illustrations are generated

*Saiteja Utpala is a Staff Research Associate of Computer Engineering at the University of California, Santa Barbara.
†Nina Miolane is an assistant professor of Electrical and Computer Engineering at University of California, Santa Barbara. Nina Miolane is partially supported by the NSF SCALE MoDL Grant 2134241.

with the software Geomstats, a comprehensive Python toolbox for computations and learning for data leveraging (Riemannian) Geometry [MGL+20][1].

2 Riemannian Geometry of Shape Spaces

Riemannian geometry can be leveraged to elegantly describe shape spaces as manifolds. This leads to the question: what is a manifold?

2.1 What is a Manifold?

A differential manifold M of dimension m is a generalization of a vector space of dimension m: M is a space that looks like a m-dimensional vector space in the neighborhood of every point and is locally similar enough to a vector space to allow one to do calculus. Globally, a manifold looks curved, as opposed to a vector space that is flat — see Figure 2 (left). We also say that the manifold is nonlinear, in contrast to the vector space which is linear.

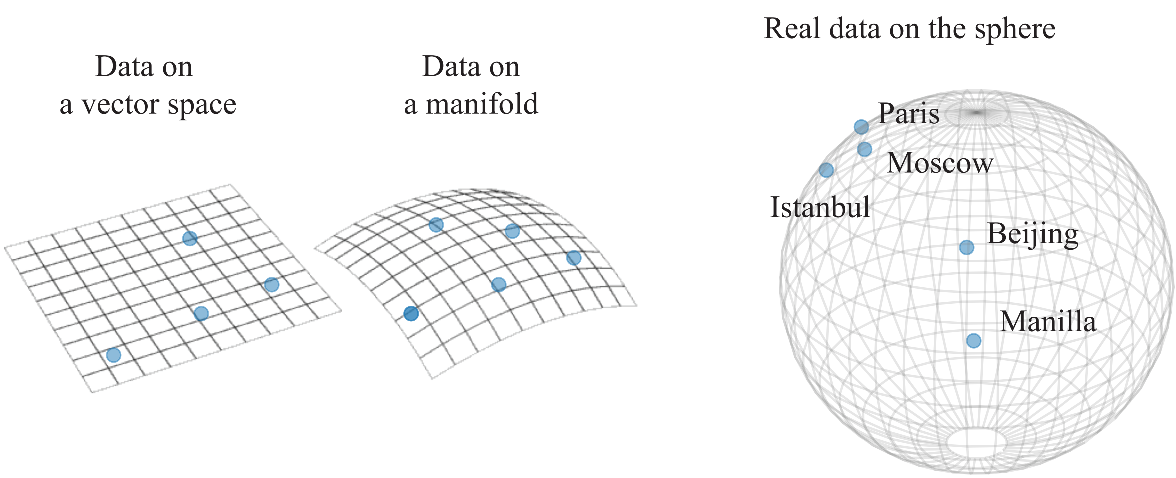

Figure 2: Left: Comparison of a vector space with a manifold, that is a curved generalization of a vector space. Right: A classical example of manifold: a sphere, such as the surface of the earth. Visualization generated with the software Geomstats [MGL+20].

The most classical textbook example of manifolds is the surface of the earth, i.e. the sphere. To a person walking on the earth, the earth looks flat: locally, the earth looks like a 2-dimensional vector space. However, globally, the earth is curved — see Figure 2 (right).

2.2 Example: The Space of Triangles Shapes

Let us show how manifolds can describe shape spaces by considering simple shapes: shapes of triangles in 2D. The triangle *object* can be represented as the 2D coordinates of three points. The triangle *shape* is everything that remains, once we have filtered out the triangle's position, orientation and scale. In other words, triangles that only vary by their position in space, orientation or overall size are considered to have the same shape. Mathematically, the shape of a triangle is therefore an "equivalence class" under translation, rotation and scaling transformations. The shape is an element of a "quotient space" which is then called: the shape space.

The mathematician David Kendall formalized these ideas in the 1980's. Interestingly, Kendall began studying shapes while motivated by statistical questions in Archeology such as the study of the shape of Stonehenge monument [Ken89]. David Kendall showed that the space of triangles is a 2-dimensional manifold, that resembles the sphere of radius 0.5 — see Figure 3. We observe that the equator of this sphere corresponds to flat triangles, and the first meridian to isocele triangles.

[1] Geomstats: https://github.com/geomstats/geomstats

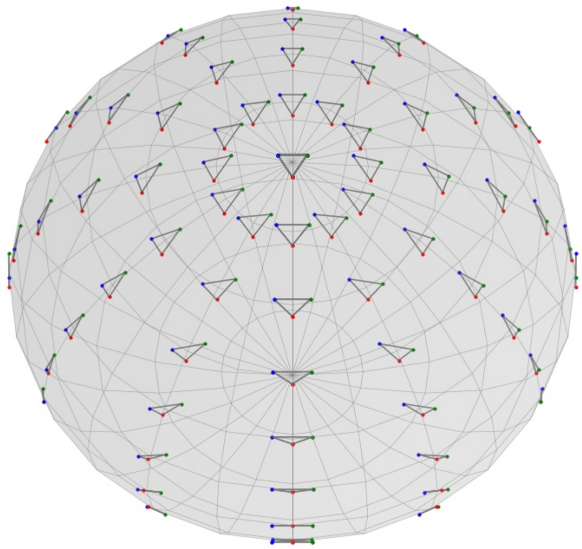

Figure 3: Visualization of the shape space for 2D triangles generated with the software Geomstats [MGL+20]. Courtesy of Elodie Maignant, Geomstats contributor and co-winner of the ICLR 2021 international challenge of Computational Geometry and Topology [MCL+21]. The triangles are superimposed on the manifold to provide intuition on this shape space.

2.3 Computing with Shapes Through Computations on Manifolds

We have established that shape data, such as triangle shapes, belong to a nonlinear manifold. We show here how we can compute with shape data and for instance, compute meaningful distances between two (triangle) shapes. Since shape spaces are manifolds, performing computations on manifolds allows performing computations on shape spaces.

Operations on manifolds can be defined by extending standard computations on vector spaces. The elementary operations on a vector space are: inner-product, addition, subtraction and multiplication by a scalar. The inner-product between vectors leads to the notions of angles between vectors, norms of vectors and distances between points. We can add a vector to a point, subtract two points to get a vector, or multiply a vector by a scalar value, see Figure 4 (left). For points on a manifold, like the sphere, the same operations are not permitted. Indeed, adding a vector to a point will not give a point that belongs to the manifold. We need to generalize them, as illustrated in Figure 4 (right) and detailed below.

On a manifold M like the sphere, we can define an inner product on the tangent spaces of each point of the manifold, while requiring that the inner-product changes smoothly when changing tangent spaces. This smooth set of inner products is called Riemannian metric, and a manifold M equipped with a Riemannian metric is called a Riemannian manifold. A Riemannian metric allows us to determine angle between tangent vectors (of the same tangent space) and their norms. The Riemannian metric also induces a distance along a curve $\mathcal{C}$ of the manifold M. Then, the curve between two points of minimal length is called a *geodesic*, which is the generalization of straight line on a Euclidean space. On the sphere, the geodesics are the great circles.

The addition of a tangent vector ν to a point p on a Riemannian manifold is called the Riemannian exponential of the vector at the point. Intuitively, this is the point that is reached after a time step 1 on geodesic starting from p with initial tangent vector ν at p. The subtraction of the point p_2 to point p_1 is called Riemannian logarithm of p_2 with respect p_1. Intuitively, this corresponds to the tangent vector at p_1 that is needed to reach p_2 in a unit time step with a geodesic. These operations are shown in Figure 4.

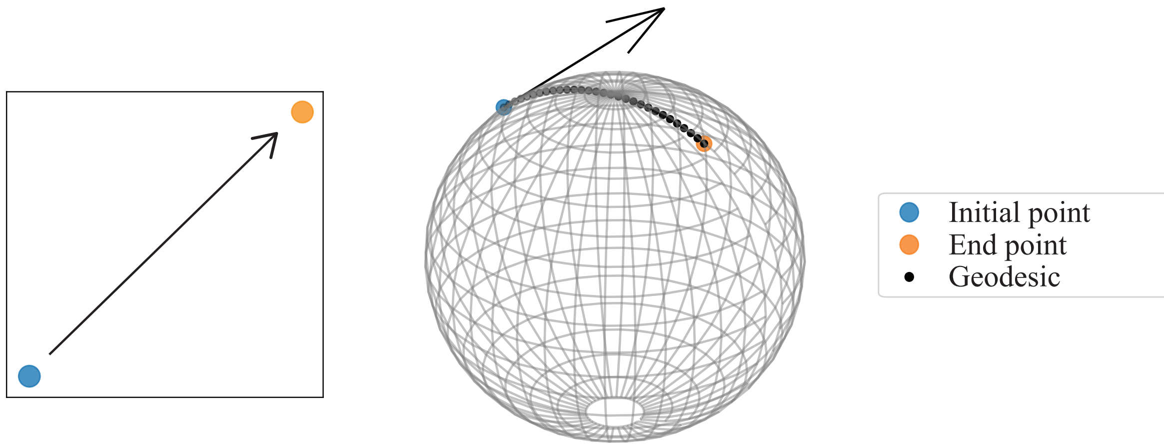

Figure 4: Left: operations on a vector space: Addition of the black vector to the blue point gives the orange point. Subtraction of the orange point to the blue point gives the black vector. Right: Generalization of these operations on a manifold. Visualizations generated with the software Geomstats [MGL+20].

3 Geometric Statistics and Learning on (Biological) Shape Spaces

Now that we have defined the basic operations on a manifold, we turn to performing statistical computations on them. Statistics is a branch of mathematics that is concerned with the collection and analysis of data, and forms the foundations of machine learning and deep learning algorithms. Yet, the mathematical theory of (traditional) statistics is defined for data that are numbers or vectors, i.e. that belong to a vector space. What happens if our data are shapes, and belong to a shape space that is a manifold, like the sphere in Figure 3? We need a theory of statistics that is, by construction, compatible with a manifold and any structure we could equip this manifold with. This theory is called "Geometric Statistics" and lies at intersection of two major fields of mathematics: Geometry and Statistics. We can get intuition about statistics on manifolds, by generalizing traditional statistics on vector spaces.

3.1 Example: The Mean

In traditional statistics, we know how to compute the mean $\bar{x}$ from a set of vectors x_i for $i = 1, ..., n$, like n velocity vectors $x_i \in \mathbb{R}^3$ of blood cells in a vessel:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{1}$$

We observe that the very definition of mean, formulated as weighted sum of the data elements, is itself linear. If we apply it on a (nonlinear) manifold, we obtain an element $\bar{x}$ that does not belong to the manifold, see Figure 5 (left). Thus, the definition of the mean needs to be generalized, as follows:

$$\bar{x} = \underset{x}{\operatorname{argmin}} \sum_{i=1}^{n} \operatorname{dist}(x_i, x)^2. \tag{2}$$

This is the definition of the Fréchet mean. The Fréchet mean of data on a manifold now provides an element of the manifold, see Figure 5 (right).

3.2 Beyond the Mean

Beyond the extension of the mean's definition, many operations and statistical learning methods can be extended to manifolds. As a matter of fact, Riemannian geometry and Geometric Statistics provide

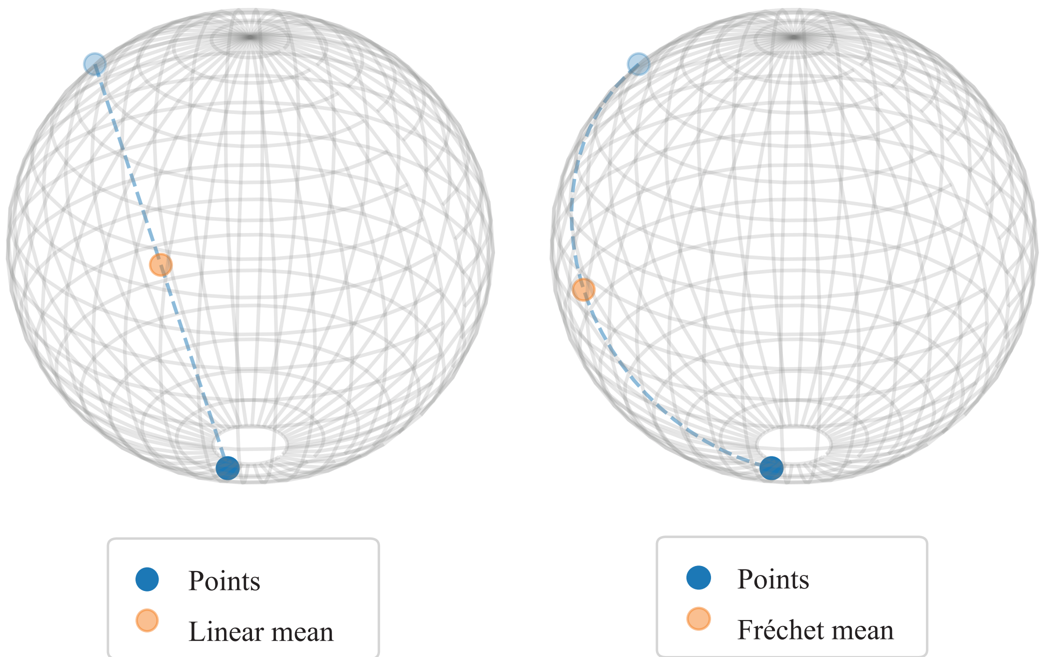

Figure 5: Left: Traditional mean does not provide an element of the manifold. Right: Fréchet mean belongs to the manifold. Visualizations generated with the software Geomstats [MGL+20].

the theoretical and computational building blocks supporting the recent trend of Geometric (Deep) Learning [BBCV21, HIZ20]. There is a wide interest in the machine learning and deep learning community to extend traditional learning algorithms to data that belong to manifolds, such as the shape spaces presented here.

4 Discussion and Conclusion

While machine learning and deep learning have been remarkably successful at solving a massive set of problems on data types including images and texts, it is only recently that they have started to be generalized to geometric data — such as the shapes presented in this snapshot, but also: point sets, graphs and simplicial complexes. As novel methods are being published at an increasing rate in this field, there is a need for mathematics and statistics that can ground new algorithms within a principled framework — in order to study their theoretical properties. In this context, Geometric Statistics may hold the key to analyze this literature from a mathematical perspective.

Just as benchmark datasets such as MNIST [Den12] supported the growth of deep learning and comparison of methods, we suggest that a suite of geometric benchmark datasets should be provided — covering the range of possible geometric characteristics such as positive or negative curvature manifolds. This would allow comparing new geometric (deep) learning methods, not only in terms of their statistical properties, but also in terms of the geometric regime that optimizes the performances. Evaluating properties such as the uncertainty of prediction algorithms will be even more critical in the context of biomedicine, where the highest standards of reliable learning are being held.

References

[BBCV21] Michael M. Bronstein, Joan Bruna, Taco Cohen, and Petar Velickovic. Geometric deep learning: Grids, groups, graphs, geodesics, and gauges. *CoRR*, abs/2104.13478, 2021.

[Den12] Li Deng. The mnist database of handwritten digit images for machine learning research. *IEEE Signal Processing Magazine*, 29(6):141–142, 2012.

[HIZ20] Mustafa Hajij, Kyle Istvan, and Ghada Zamzmi. Cell complex neural networks. *CoRR*, abs/2010.00743, 2020.

[Ken89] David G. Kendall. A survey of the statistical theory of shape. *Statistical Science*, 4(2), May 1989.

[MCL+21] Nina Miolane, Matteo Caorsi, Umberto Lupo, Marius Guerard, Nicolas Guigui, Johan Mathe, Yann Cabanes, Wojciech Reise, Thomas Davies, António Leitão, Somesh Mohapatra, Saiteja Utpala, Shailja Shailja, Gabriele Corso, Guoxi Liu, Federico Iuricich, Andrei Manolache, Mihaela Nistor, Matei Bejan, Armand Mihai Nicolicioiu, Bogdan-Alexandru Luchian, Mihai-Sorin Stupariu, Florent Michel, Khanh Dao Duc, Bilal Abdulrahman, Maxim Beketov, Elodie Maignant, Zhiyuan Liu, Marek Cerný, Martin Bauw, Santiago Velasco-Forero, Jesús Angulo, and Yanan Long. ICLR 2021 challenge for computational geometry & topology: Design and results. *CoRR*, abs/2108.09810, 2021.

[MGL+20] Nina Miolane, Nicolas Guigui, Alice Le Brigant, Johan Mathe, Benjamin Hou, Yann Thanwerdas, Stefan Heyder, Olivier Peltre, Niklas Koep, Hadi Zaatiti, Hatem Hajri, Yann Cabanes, Thomas Gerald, Paul Chauchat, Christian Shewmake, Daniel Brooks, Bernhard Kainz, Claire Donnat, Susan Holmes, and Xavier Pennec. Geomstats: A python package for riemannian geometry in machine learning. *Journal of Machine Learning Research*, 21:1–9, 2020.

[PE15] Victor Patrangenaru and Leif Ellingson. *Nonparametric Statistics on Manifolds and Their Applications to Object Data Analysis.* CRC Press, September 2015.