Snapshots of modern mathematics
from Oberwolfach

№ 8/2022

Biological Shape Analysis with Geometric Statistics and Learning

Saiteja Utpala ● Nina Miolane[1]

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 unlocking outstanding mysteries in biomedicine. This snapshot introduces the mathematical framework of geometric statistics and learning and its applications to biomedicine.

1 Introduction

Statistics is the branch of mathematics that is concerned with the collection and analysis of data, and thus it forms the foundations of machine learning and deep learning algorithms[2]. Vast quantities of biological imaging data are currently being 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 irregularly-shaped cancer cells shown in Figure 1, is therefore important and timely.

[1] Nina Miolane is partially supported by the NSF SCALE MoDL Grant 2134241.

[2] To read more about machine learning, particularly in the context of medical imaging, see Snapshot 15/2019 Deep Learning and Inverse Problems by S. Arridge *et al.*

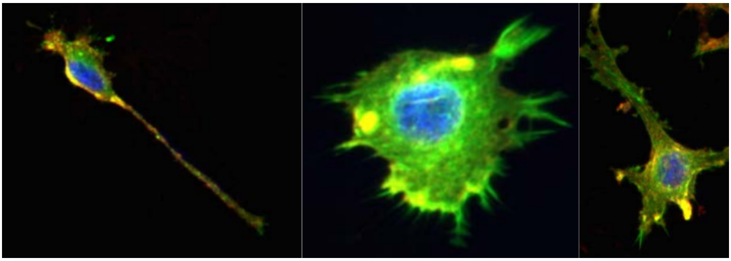

Figure 1: Cancer cells observed by fluorescence microscopy. Cancer cells typically exhibit an irregular boundary, in contrast to healthy cells. (See Chapter 3 in the book [1], available here: https://www.ncbi.nlm.nih.gov/books/NBK9553/.)

From a statistical perspective, shapes are fascinating data objects, as they can be modelled as points in some curved space, potentially of many dimensions. To do computations on shape data requires us to take account of the geometry of these spaces, and to adapt our statistical tools accordingly. In this snapshot, we introduce shape spaces and the geometry needed to study them, 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.

2 The Geometry of Shape Spaces

We will begin this section by considering a simple example of a shape space, and then we will introduce some more general ideas.

2.1 Example: The Space of Triangle Shapes

Let us consider simple shapes: triangles in 2D. A triangle 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, all similar triangles are considered to have the same shape. Mathematically, the shape of a triangle is therefore what is called an "equivalence class" under translation, rotation and scaling transformations. That is, we can group together all similar triangles and choose one representative element to work with.

The mathematician David Kendall formalized these ideas in the 1980s. Interestingly, Kendall began studying shapes motivated by statistical questions related to archeology such as the study of the shape of the Stonehenge monument

[6]. Kendall showed that the space of triangles carries in a very natural way the 2-dimensional structure of a sphere, shown in Figure 2. We observe that the equator of this sphere corresponds to flat triangles (a triangle with all three points aligned, effectively a line segment), and the first meridian to isoceles triangles. Part of the intuition behind the appearance of the sphere here is that all scaling has been removed, in the same way that if we consider vectors in the plane and consider the length to be normalised to 1, we obtain a circle.[3]

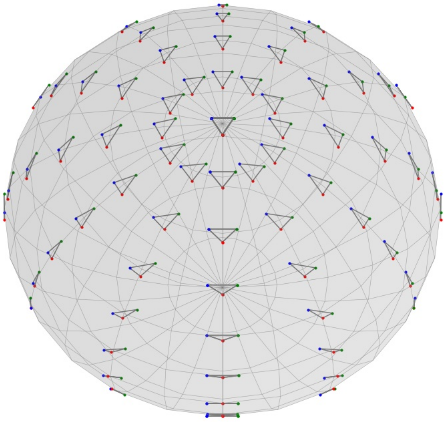

Figure 2: Visualization of the shape space for 2D triangles. Courtesy of Elodie Maignant, Geomstats contributor and co-winner of the ICLR 2021 Challenge for Computational Geometry and Topology [4]. The triangles are superimposed on the manifold to provide intuition on this shape space.

2.2 Manifolds

One of the most fundamental notions in geometry, introduced by Bernhard Riemann (1826–1866), is that of a *manifold*. The basic idea is that an n-dimensional manifold is a space which near every point looks like the Euclidean space $\mathbb{R}^n$. Thus, in two dimensions, a manifold locally looks like the plane $\mathbb{R}^2$ in the vicinity of any of its points. The most intuitive example is that of the sphere, for instance, the surface of the earth, as shown in Figure 3, on the right.

[3] A 2-dimensional vector is a mathematical quantity with a magnitude and a direction. Vectors can be pictured as arrows in the plane, where the direction is measured as the angle anti-clockwise from the horizontal, and the magnitude is the length of the arrow. For more details, see https://en.wikipedia.org/wiki/Euclidean_vector.

To a person walking on the earth, the earth looks flat: locally, the earth looks like a 2-dimensional plane. However, globally, the earth is curved.

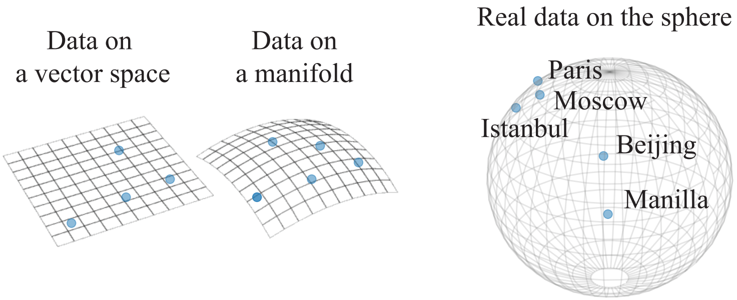

Figure 3: Left: Comparison of a vector space (a flat space) with a manifold (a curved space). Right: A classical example of manifold, namely, a sphere, such as the surface of the earth.

2.3 Computing on a Manifold

Like the triangle example that we have seen, it is generally the case that shape data belong to a manifold. If we want to be able to do computations with our shape data, that means we need to understand how to do computations on a manifold. In any Euclidean space, we know how to measure the distance between points (a function that measures distances is called a *metric*) and how to measure angles. Riemann proposed a way of generalising these ideas to manifolds, by introducing the concept that is now known as a Riemannian metric. A Riemannian metric provides a manifold with notions of lengths and angles, and starting from this we can investigate any aspect of geometry. For instance, we can ask for the curve on the manifold between two points of minimal length, called a *geodesic*, which is the generalization of a straight line in a Euclidean space. On the sphere, the geodesics are the *great circles*. Considering the embedding of the sphere in the 3D Euclidean space, these are the circles that result from the intersection of the surface of the sphere with any plane that passes through the centre.

3 Geometric Statistics and Learning on (Biological) Shape Spaces

Now that we have outlined the basic idea behind operations on a manifold, we turn to performing statistical computations on them. The mathematical theory of (traditional) statistics is defined for data points that are numbers or vectors, that is, data that belongs to a Euclidean space. What happens if our data are shapes, and belong to a shape space that is a manifold, like the sphere in Figure 2? We refer also to the left-hand side of Figure 3, which shows examples of data points on a vector space and a manifold. We need a theory of statistics that is, by construction, compatible with a manifold. This theory is called "geometric statistics" and lies at intersection of two major fields of mathematics: geometry and statistics. Let us now use the example of the mean to obtain some intuition as to what must be changed in order to do statistical computations on a curved space.

3.1 Example: The Mean

In traditional statistics, we know how to compute the mean $\bar{x}$ from a set of data points x_i for $i = 1, ..., n$:

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

We observe that the very definition of mean, formulated as a weighted sum of the data elements, is itself linear. If we apply it on a (nonlinear) manifold, we obtain an element $\overline{x}$ that does not necessarily belong to the manifold. Figure 4 (left) shows two points on the sphere, that can be thought of as 3D vectors starting from the origin of Euclidean 3D space that contains the sphere.

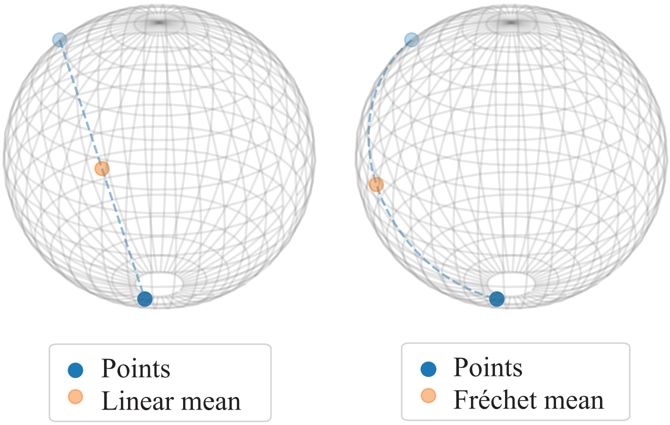

Figure 4: Left: The traditional mean does not lie on the manifold. Right: The Fréchet mean does belong to the manifold.

Thus, the definition of the mean needs to be changed so that the so-obtained mean value lies in our space. One way to achieve this is as follows:

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

where "argmin" stands for "argument minimum", and it refers to points x on the manifold that may realize the minimum of the sum. This is the definition of the *Fréchet mean*. The Fréchet mean of data on a manifold now provides, by definition, an element of the manifold, see Figure 4 (right). The Fréchet mean is an extension of the usual mean on a flat space, in the sense that it is a well-known fact that the mean of a dataset has the property of being the point that minimizes the sum of the squared distances to the data. Notably, the Fréchet might not exist or not be unique. We invite the reader to think about examples where this could happen, such as with data points located at the poles of the sphere.

3.2 Beyond the Mean

Beyond the extension of the definition of the mean, many operations and statistical learning methods can be extended to manifolds. Figure 5 illustrates, for example, the generalization of the notion of "addition" of a vector (in black) to a point (in blue) that gives another point (in orange). On the left, we see the definition of addition on a vector space, while on the right, we see it for a manifold. As a matter of fact, Riemannian geometry and geometric statistics provide the theoretical and computational building blocks supporting the recent trend of geometric (deep) learning [2]. Statistics and machine learning are two branches of data analysis that go hand-in-hand. If we can generalize statistics to manifolds, then we will be able to generalize a very wide range of machine learning and deep learning methods to manifolds as well. There is much interest in the machine learning and deep learning community for extending 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 to point sets, graphs and simplicial complexes [2, 5]. As novel methods are being published at an increasing rate in this field, there is a need for mathematics and statistics that

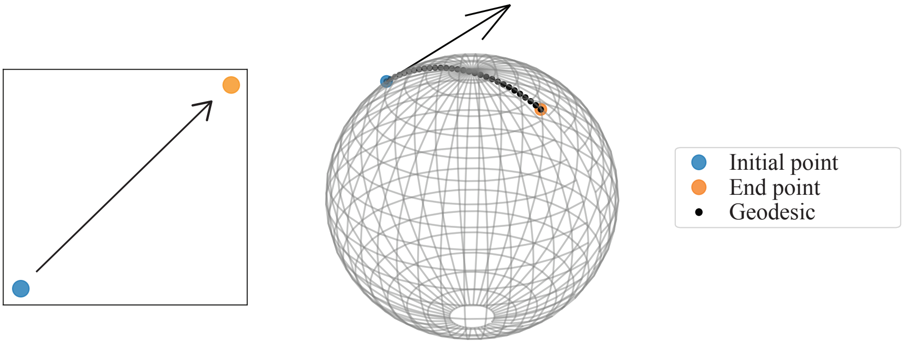

Figure 5: Left: Addition of the black vector to the blue point in the plane gives the orange point. Subtraction of the orange point from the blue point gives the black vector. Right: Generalization of these operations to the surface of the sphere. Here the vector points along the direction of the "tangent line" to the geodesic curve.

can ground new algorithms within a rigorous framework, in order to study their theoretical properties. In this context, geometric statistics may hold the key to analysing this literature from a mathematical perspective.

Just as benchmark datasets such as MNIST [3] 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 comparison of new geometric (deep) learning methods, not only in terms of their statistical properties, but also in terms of the geometric regime that optimizes their performance. Evaluating properties such as the uncertainty of prediction algorithms will be even more critical in the context of biomedicine, where the highest standards of reliability are necessary.

Image credits

Figure 1 Image kindly provided by Ashok Prasad, Colorado State University.

The other illustrations were generated with the “Geomstats” software [7].

References

[1] A. I. Baba and C. Câtoi, *Comparative Oncology*, The Publishing House of the Romanian Academy, 2007.

[2] M. M. Bronstein, J. Bruna, T. Cohen, and P. Velickovic, *Geometric deep learning: Grids, groups, graphs, geodesics, and gauges*, Computing Research Repository in arXiv (2021), https://arxiv.org/abs/2104.13478.

[3] L. Deng, *The MNIST database of handwritten digit images for machine learning research*, IEEE Signal Processing Magazine **29** (2012), no. 6, 141–142.

[4] N. Miolane et al., *ICLR 2021 challenge for computational geometry & topology: Design and results*, Computing Research Repository in arXiv (2021), https://arxiv.org/abs/2108.09810.

[5] M. Hajij, K. Istvan, and G. Zamzmi, *Cell complex neural networks*, Computing Research Repository in arXiv (2020), https://arxiv.org/abs/2010.00743.

[6] D. G. Kendall, *A survey of the statistical theory of shape*, Statistical Science **4** (1989), no. 2, 87–99.

[7] N. Miolane, N. Guigui, A. Le Brigant, J. Mathe, B. Hou, Y. Thanwerdas, S. Heyder, O. Peltre, N. Koep, H. Zaatiti, H. Hajri, Y. Cabanes, T. Gerald, P. Chauchat, C. Shewmake, D. Brooks, B. Kainz, C. Donnat, S. Holmes, and X. Pennec, *Geomstats: A Python package for Riemannian geometry in machine learning*, Journal of Machine Learning Research **21** (2020), 1–9, ISSN 15337928.

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 the University of California, Santa Barbara.*

Mathematical subjects
Geometry and Topology, Probability Theory and Statistics

Connections to other fields
Computer Science, Life Science

DOI
10.14760/SNAP-2022-008-EN

Snapshots of modern mathematics from Oberwolfach provide exciting insights into current mathematical research. They are written by participants in the scientific program of the Mathematisches Forschungsinstitut Oberwolfach (MFO). The snapshot project is designed to promote the understanding and appreciation of modern mathematics and mathematical research in the interested public worldwide. All snapshots are published in cooperation with the IMAGINARY platform and can be found on www.imaginary.org/snapshots and on www.mfo.de/snapshots.

ISSN 2626-1995

Mathematisches Forschungsinstitut
Oberwolfach gGmbH
Schwarzwaldstr. 9 – 11
77709 Oberwolfach
Germany

Director
Gerhard Huisken