Hidden Structure in Data

Object Data and Keys to Designing Parameter-Efficient Models

Chun-Hao Yang

National Taiwan University

Statistical Science Camp

Aug. 24, 2023

Outline

Motivation

- Modern data are often of high dimension and complex.
- A typical approach is to represent the data as vectors or matrices.
- Unconstrained vector/matrix representation is not able to reveal the hidden structure in the data.
- We need to figure out how to decompose data into different “modes” in order to have a deeper understanding of the data.
- **Modes of Variation**: data variation in different modes, i.e., variation in length, direction, etc.

Example

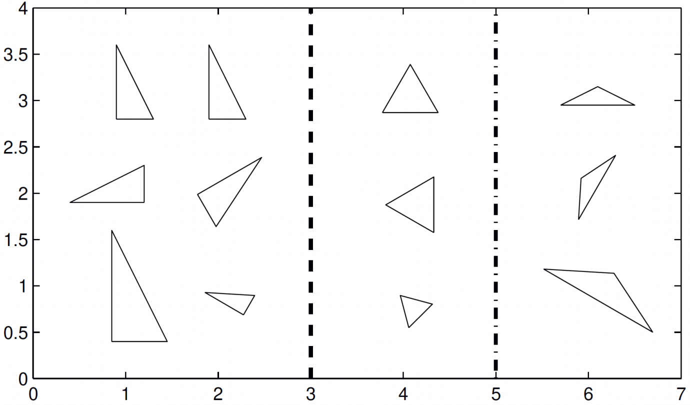


Figure: Orbits in the space of all triangles [1, Fig. 1.9].

Triangle Example

- If we use a 3×2 matrix to represent a triangle, we are not able to easily tell whether two triangles are different in size, orientation, position, or shape.
- Size: $a \in \mathbb{R}_+$ (1 degree of freedom)
- Position: $o \in \mathbb{R}^2$ (2 degrees of freedom)
- Orientation: $\theta \in [0, 2\pi]$ (1 degree of freedom)
- The remaining degrees $6 - 1 - 2 - 1 = 2$ are responsible for shape.
- What is the space for shapes?

What is Object Data?

- Traditionally, the samples are represented as vectors or matrices.
- Constrained vectors/matrices are able to represent the data more accurately.
- Example: vector with unity length $\Rightarrow$ direction
- Usually, the constraints make the sample space non-Euclidean.
- Object data are those residing on a non-Euclidean space, e.g., a curved space.

Connectivity Matrix

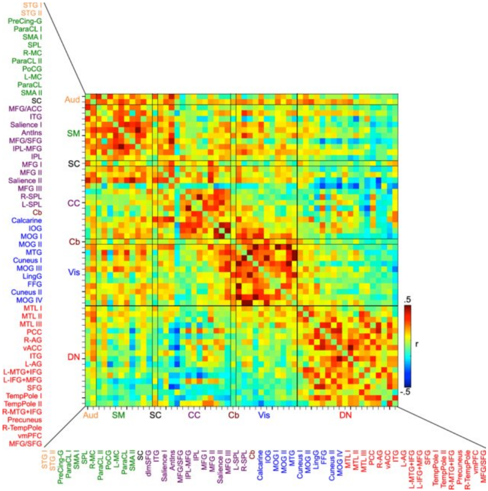


Figure: Functional Connectivity Matrix [2, Fig. 3].

Phylogenetic Tree

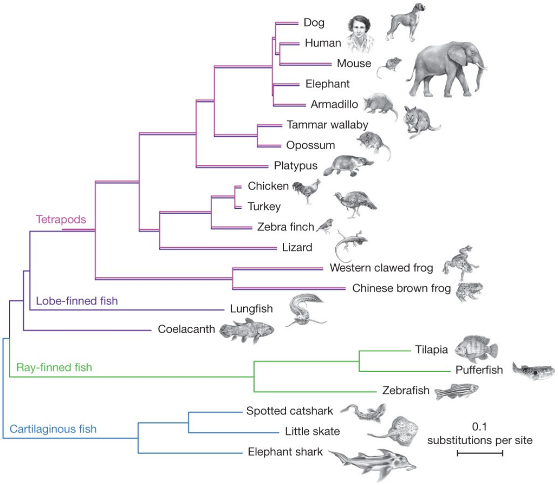


Figure: Phylogenetic Tree [3].

Shape

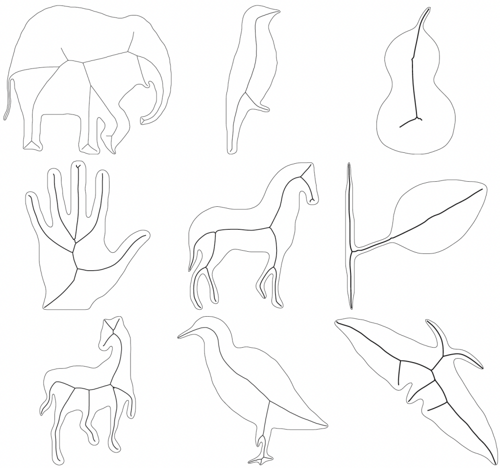


Figure: Example of planar shapes [4, Fig. 4.5].

Sample Space of Object Data

- The sample spaces of the object data are all non-Euclidean: the tree space, the space of SPD matrices, the shape space, etc.
- Vector operations, e.g., addition and scalar multiplication, are no longer valid.
- How can we compute some simple statistics, e.g., mean?

Invariance and Equivariance

- **Invariance**: When the samples are transformed, the inference remains unchanged.
- **Equivariance**: When the samples are transformed, the inference changes accordingly.
- Example:
  - Sample mean is equivariant to translation and scale
  - Variance is invariant to translation but equivariant to scale
- Invariance/Equivariance allow us to transform the data to make inference easier.

Location-Scale Invariance and Equivariance

- The t-test is invariant to location-scale transformations.
  - $\Rightarrow$ We can standardize the data without changing the conclusion.
- Linear regression is also invariant/equivariant to location-scale transformations.
- What invariance/equivariance to object data have?
  - $\Rightarrow$ It depends on the sample space of the object data.

Invariance/Equivariance for Object Data

- Scale/rotation/translation invariance for shapes.
- Antipodal invariance for directions.
- Rotation/affine invariance for SPD matrices.

An Example: The Shape of the Corpus Callosum

- The shape of the CC varies with age, sex, intellectual ability, etc.
- Its size and shape are also associated with disease progression of some neurodevelopmental disorders, such as autism and Schizophrenia.

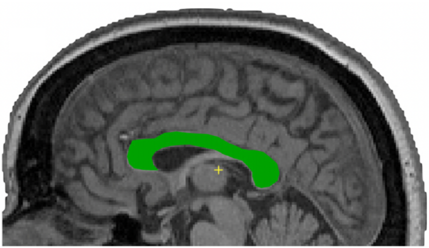


Figure: The shape of a corpus callosum [5].

Statistical Questions about Shapes

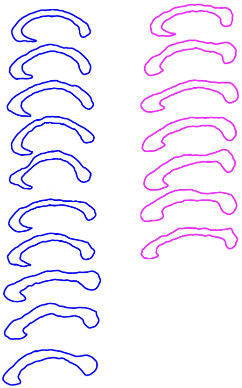


Figure: The CC shapes of male (blue) and female (magenta).

How can we answer some statistical questions about shapes? For example,

- Test H_0: Shape of Female = Shape of Male.
- What variations in shape are associated with sex?
- What is the relationship between age and the shape of CC?

What do we need to analyze object data?

- Suppose now we have some observations $X_1, \ldots, X_n$ from the sample space $\mathcal{X}$.
- What is the most important notion we need for $\mathcal{X}$ in order to perform statistical analysis?
  - probability distribution?
  - sample mean?
  - expectation/variance?
- The most fundamental one is a **distance**, or any measure of dissimilarity.

Recall: what is a distance?

A *distance* $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a function such that

1. $d(x, y) \geq 0$ and $d(x, y) = 0$ iff $x = y$,
2. $d(x, y) = d(y, x)$, and
3. $d(x, y) \leq d(x, z) + d(y, z)$ for any $x, y, z \in \mathcal{X}$.

- The pair $(\mathcal{X}, d)$ is called a *metric space*.
- A distance has more than what we need; in many cases, a *divergence* also works.
- We will see what we can achieve with only a distance function.

Fréchet Mean

- Let $(\mathcal{X}, d)$ be a metric space and $x_1, \ldots, x_n \in \mathcal{X}$.
- Define $F : \mathcal{X} \to [0, \infty)$ by

$$F(m) = \sum_{i=1}^{n} d^2(x_i, m).$$

- This is called the Fréchet variance at m.
- The set of minimizers of F is called the Fréchet mean set of $x_1, \ldots, x_n$.

Example: $\mathcal{X} = \mathbb{R}$

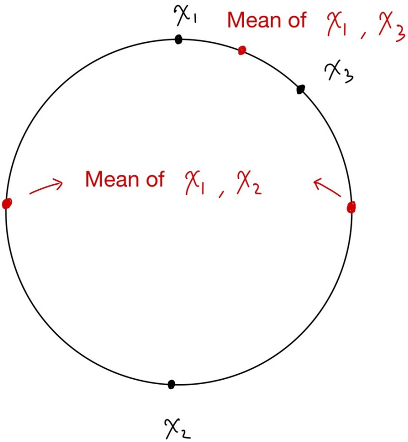
- Let $\mathcal{X} = \mathbb{R}$ and $d(x, y) = |x - y|$.
- Then $F(m) = \sum_{i=1}^{n}(x_i - m)^2$.
- The Fréchet mean of $x_1, \ldots, x_n$ is

$$\arg\min_{m \in \mathbb{R}} \sum_{i=1}^{n}(x_i - m)^2 = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}.$$

- The Fréchet mean is a generalization of the arithmetic mean to general metric spaces.

Example: Circle

- Let $\mathcal{X} = S^1 = \{(x, y) : x^2 + y^2 = 1\} \subseteq \mathbb{R}^2$ and $d(x_1, x_2) = \arccos(x_1^T x_2)$, i.e., the angle between x_1 and x_2.



Fréchet Mean

- In general, there is no unique sample mean for $x_1, \ldots, x_n \in \mathcal{X}$.
- However, if
  1. $\mathcal{X}$ has non-positive sectional curvatures, or
  2. $x_1, \ldots, x_n$ are "not far from each other",

  the FM of $x_1, \ldots, x_n$ is unique.
- Examples of non-positively curved spaces: $\mathbb{R}^d$, SPD(d), etc.
- Examples of positively curved spaces: sphere, etc.
- For circles/spheres, if all the samples are in the same hemisphere, the FM is unique.

What can we do with FM?

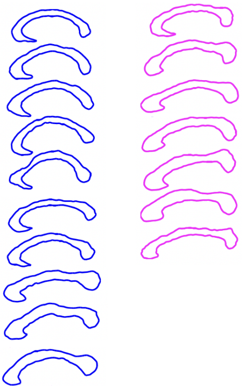


Figure: The CC shapes of male (blue) and female (magenta).

Test H_0: Shape of Female = Shape of Male.

1. Compute the FM of shapes for two groups: $\bar{X}_M$ and $\bar{X}_F$.
2. Compute the distance $d_{\text{obs}} = d(\bar{X}_M, \bar{X}_F)$.
3. Use permutation test to obtain a p-value.

Geodesic

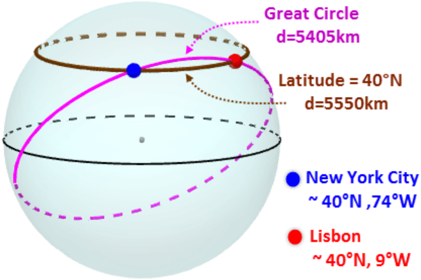
- For $x_0, x_1 \in \mathcal{X}$, a geodesic $\gamma(t)$ with $\gamma(0) = x_0$ and $\gamma(1) = x_1$ is a curve such that $\text{Length}_{t\in[0,1]}(\gamma) = d(x_0, x_1)$.
- On a manifold $\mathcal{M}$, a geodesic can also be determined by a point $x \in \mathcal{M}$ and a tangent vector $v \in T_x\mathcal{M}$.
- Given $x \in \mathcal{M}$ and $v \in T_x\mathcal{M}$, the geodesic is the solution to the differential equation $\gamma'(0) = v$ with the initial condition $\gamma(0) = x$.

Example: Geodesics on a Sphere

- Let $S^n = \{x \in \mathbb{R}^{n+1} : \|x\| = 1\}$. The geodesic for $x \in S^n$ and $\mathbf{v} \in T_x S^n$ is

$$\gamma(t) = \cos(\|\mathbf{v}\|t)x + \sin(\|\mathbf{v}\|t)\frac{\mathbf{v}}{\|\mathbf{v}\|}.$$

- For a sphere, the geodesic is a segment of a great circle.



Exp/Log Map for a Manifold

- Let $\mathcal{M}$ be a manifold and $T_x\mathcal{M}$ be the tangent space of $\mathcal{M}$ at x.
- $T_x\mathcal{M}$ is a vector space.
- $\gamma_v(t)$ is a geodesic starting at x with direction v.

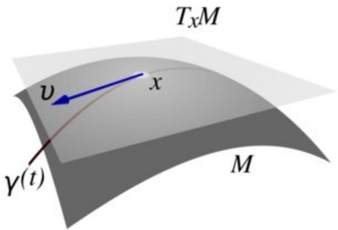


Figure: The tangent space at x.

Exp/Log Map for a Manifold

- Exponential map: $\text{Exp}_x : T_x\mathcal{M} \to \mathcal{M}$ (always exists).
- Log Map: $\text{Log}_x : \mathcal{M} \to T_x\mathcal{M}$ (exists only on a neighborhood of x).
- In fact, $d(x_i, x_j) = \|\text{Log}_{x_i}(x_j)\|$ provided that the log map exists.

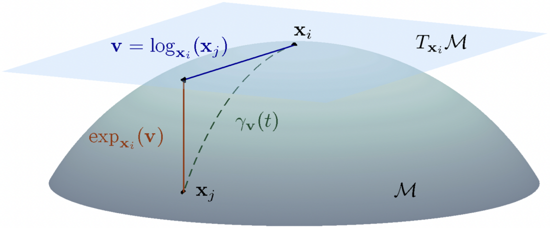


Figure: The Log/Exp map [6, Fig. 1].

Principal Geodesic Analysis

- Find a FM μ.
- Project all the data onto $T_\mu \mathcal{M}$ using the Log map.
- Perform PCA on the tangent space.

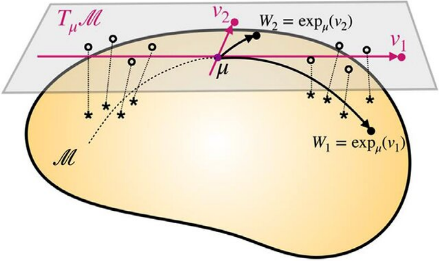


Figure: Principal Geodesic Analysis [7].

Geodesic Regression

- Suppose now we have $\{x_i, y_i\}_{i=1}^n$ where $x_i \in \mathbb{R}$ and $y_i \in \mathcal{M}$.
- We want to model the relationship between x_i and y_i.

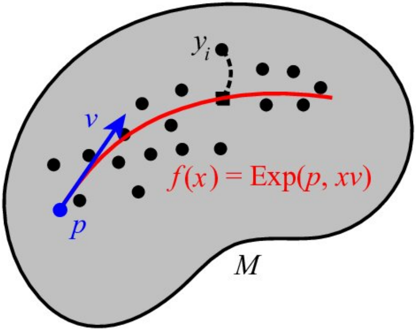


Figure: Geodesic Regression [8].

Geodesic Regression

- Model:

$$y = \mathrm{Exp}_p(xv + \epsilon), \epsilon \sim N(0, \sigma^2 I) \tag{1}$$

- The point $p \in \mathcal{M}$ is the “intercept” and $v \in T_p\mathcal{M}$ is the “slope”.
- Suppose Log_p exists. Model (1) is equivalent to

$$\mathrm{Log}_p y = xv + \epsilon,$$

  that is, it is the linear regression on the tangent space $T_p\mathcal{M}$.

Linearization

- When the sample space is a manifold, we can use the Exp/Log map to map the samples to a vector space back and forth.
- The vector space is often the tangent space at an FM.
- On the tangent space, we can apply the usual statistical methods.
- Example: PCA and linear regression on the tangent space.

Problems with linearization

- The linearization technique works well only when the samples are clustered.
- There is no natural coordinate system on the tangent space.
- Linearization relies on the FM, which might not be unique.
- Linearization loses the geometrical information of the sample space.

Extrinsic Methods

- In many cases, the sample space is embedded in a higher dimensional Euclidean space.
- For example, $x_1, \ldots, x_n \in S^2 \subseteq \mathbb{R}^3$ are represented as 3-dim vectors but they are actually on a 2-dim manifold (sphere).
- What if we consider

$$\tilde{x} = \frac{\bar{x}}{\|\bar{x}\|}, \quad \text{where} \quad \bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \ ?$$

- Will it be the same as $\mathrm{FM}(x_1, \ldots, x_n)$? In general, no.
- $\mathrm{FM}(x_1, \ldots, x_n)$ is also called an **intrinsic mean**, and $\tilde{x}$ is called an **extrinsic mean**.

Extrinsic Methods

- Suppose we have $X_1, \ldots, X_n \in \mathcal{M} \subseteq \mathbb{R}^d$ where $\mathcal{M}$ is a k-dim manifold and $k \ll d$.
- We can simply treat the X_i's as d-dim vectors.
- However, d is larger than the actual dimension of X_i's and hence we might need a larger model.

Conclusion

- For object data, we can do some basic statistical analysis with only the notion of a distance.
- Geodesics and other more advanced geometric concepts are also helpful.
- Linearization and extrinsic approaches are good first steps. However, they work well only when the data are clustered.

References I

[1] James Stephen Marron and Ian L Dryden. *Object oriented data analysis*. CRC Press, 2021.

[2] R Matthew Hutchison and J Bruce Morton. Tracking the brain's functional coupling dynamics over development. *Journal of Neuroscience*, 35(17):6849–6859, 2015.

[3] Matt Haber and Joel Velasco. Phylogenetic Inference. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition, 2022.

[4] Kaleem Siddiqi and Stephen Pizer. *Medial representations: mathematics, algorithms and applications*, volume 37. Springer Science & Business Media, 2008.

[5] Shantanu H Joshi, Katherine L Narr, Owen R Philips, Keith H Nuechterlein, Robert F Asarnow, Arthur W Toga, and Roger P Woods. Statistical shape analysis of the corpus callosum in Schizophrenia. *Neuroimage*, 64:547–559, 2013.

References II

[6] Alvina Goh and René Vidal. Clustering and dimensionality reduction on riemannian manifolds. In *2008 IEEE Conference on computer vision and pattern recognition*, pages 1–7. IEEE, 2008.

[7] Kisung You and Hae-Jeong Park. Re-visiting riemannian geometry of symmetric positive definite matrices for the analysis of functional connectivity. *NeuroImage*, 225:117464, 2021.

[8] P Thomas Fletcher and Miaomiao Zhang. Probabilistic geodesic models for regression and dimensionality reduction on riemannian manifolds. In *Riemannian Computing in Computer Vision*, pages 101–121. Springer, 2016.