Complex Data Analysis with Application in Neuroimaging

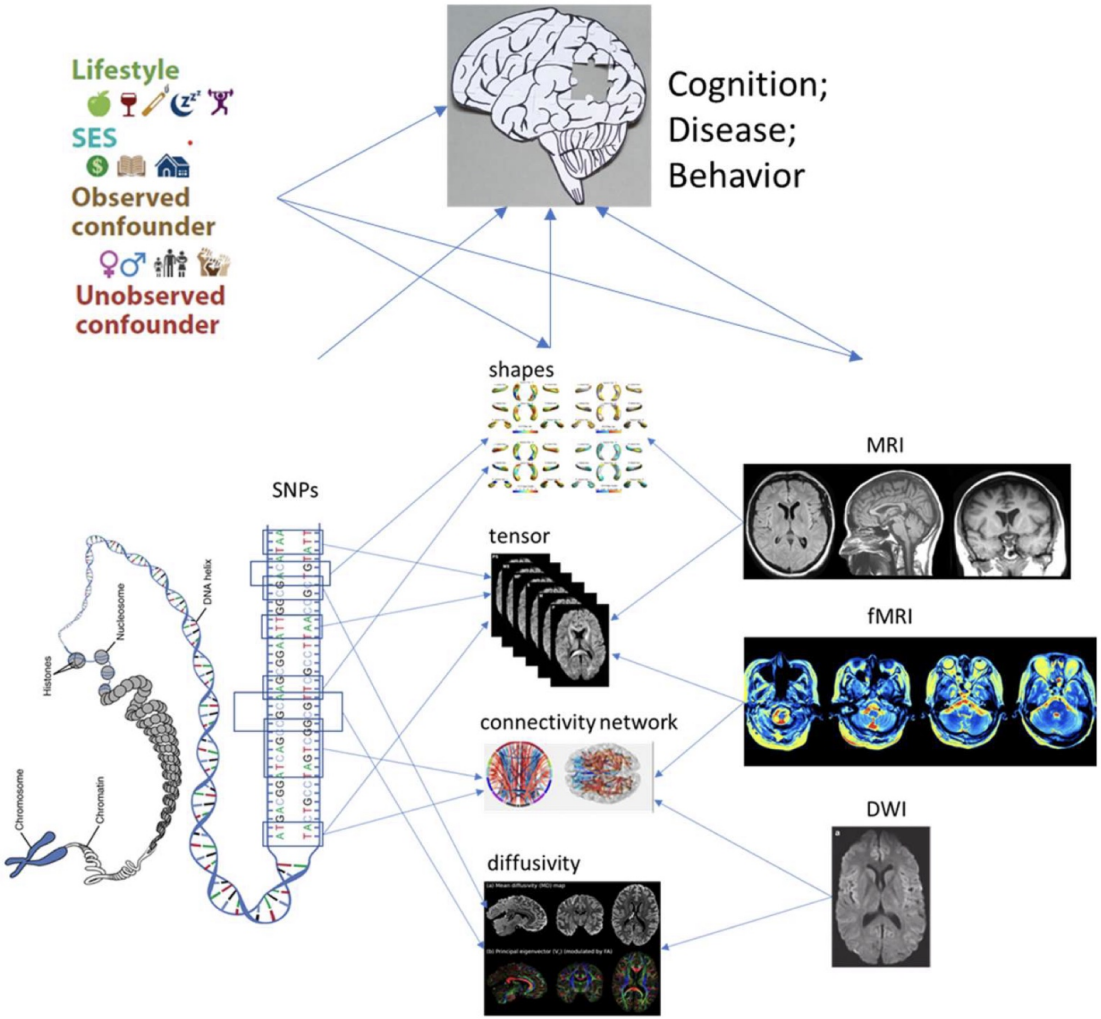
Ting Li

April 29, 2025

Table of Contents

Table of Contents

Motivation



Challenges: (Yang et al., 2024)

1. High dimensionality
2. Interpretability
3. Computational Complexity
4. Multimodal data fusion
5. Non-Euclidean Data

Principal Component Analysis in Geodesic Space

- Discover geometric structure: Identify principal geodesics that capture fundamental data patterns in the metric space;
- Dimensionality reduction: Project high-dimensional data onto a lower-dimensional Euclidean space while preserving essential geometric relationships;
- Downstream task: Enables the efficient application of standard machine learning tasks (regression, classification, clustering) in the reduced Euclidean space.

Literature

Tangent PCA (Extrinsic PCA) [3, 6, 7]

- Key Idea: Projects data onto tangent space at a reference point and applies standard PCA.
- Limitations: Relies on linear approximations, failing to capture intrinsic geometry.

Geodesic PCA (Intrinsic PCA) [9, 10]

- Key Idea: Minimizes intrinsic variance along geodesics on the manifold.
- Limitations:
  - Distance to geodesics is hard to compute.
  - Requires Riemannian metric, limiting adaptability.
  - Difficulty for downstream tasks (e.g., regression, clustering).

Principal Nested Spheres (PNS) [11, 13]

- Key Idea: Iteratively fits nested subspheres for spherical data.
- Limitations: Restricted to sphere and computationally expensive

Literature

Specialized PCA Methods

- **Graph Space PCA**: Computes statistics on unlabeled networks [4, 8].
- **Wasserstein Space PCA**: Minimizes Wasserstein distance for probability distributions [2, 5].
- **Tree PCA**: Analyzes hierarchical structures and phylogenetic trees [1, 12, 15].

Common Limitation: Tailored to specific data structures, lacking general applicability.

- **t-SNE** [14]: Maps high-dimensional data to 2D/3D while preserving local structure.
- Limitation: Randomness and sensitivity to hyperparameters; Interpretability challenges; Inability to add new data.

Summary of the existing literature

- High computational complexity.
- Limited to Riemannian manifolds or specific spaces.
- Unable to produce low-dimensional representations for downstream tasks.
- Lack of statistical theoretical results.

Table of Contents

Geodesic

A geodesic is the shortest path between two points in a metric space.

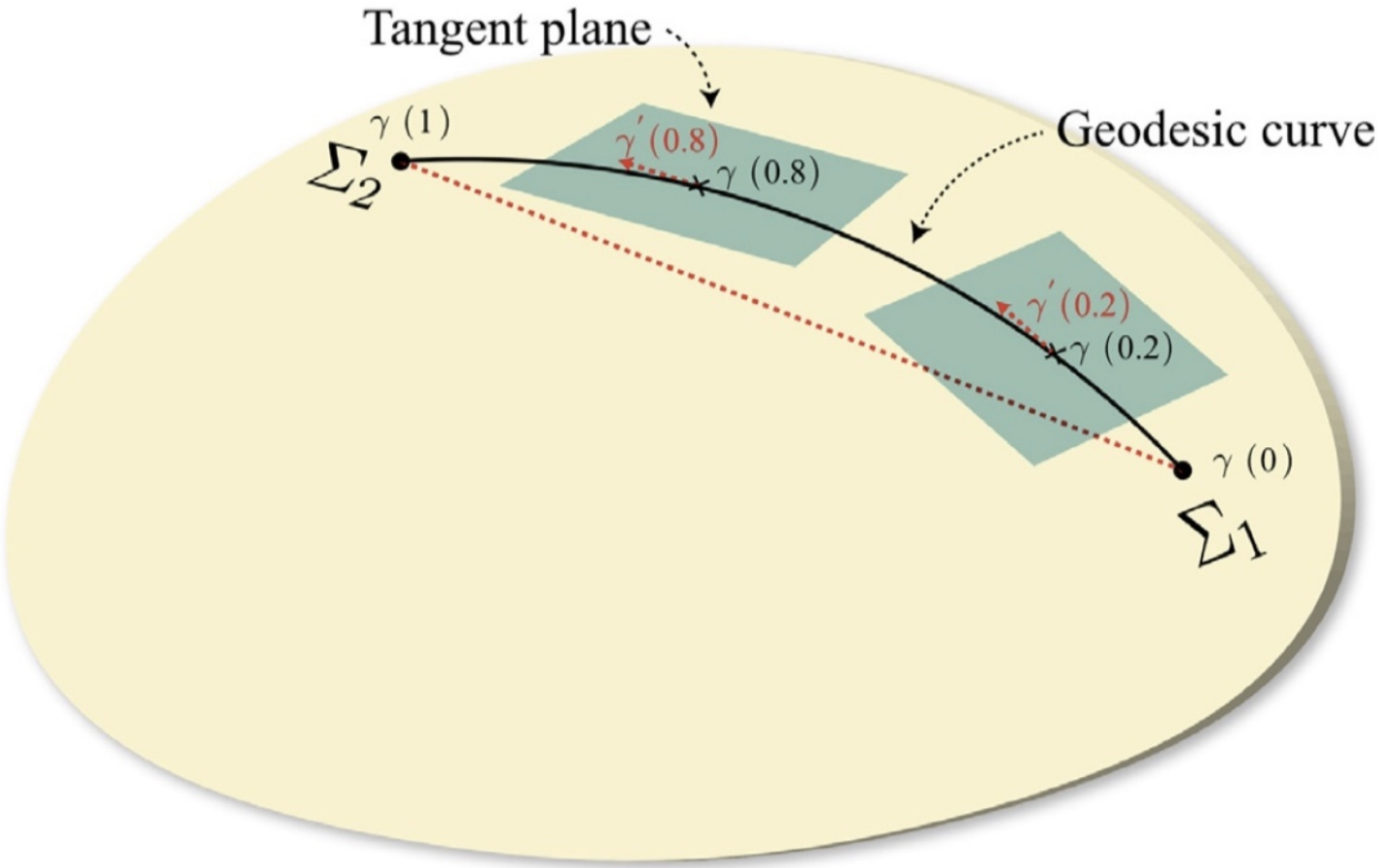


Figure 1: Riemannian manifold and geodesic curve.

Geodesic spaces

A metric space $(\mathcal{M}, d)$ is a geodesic space if for every pair of points $p, q \in \mathcal{M}$, there exits a geodesic connecting them.

Examples of geodesic spaces

- Riemannian manifolds: Euclidean Space, Sphere, Hyperbolic Space.
- CAT(0) space: Hilbert space, Graph-Theoretic Tree.
- Alexandrov space: Convex polyhedron.

Alexandrov angle

For a metric space $(\mathcal{M}, d)$ and three distinct points $p, q, r \in \mathcal{M}$, the comparison angle between q and r at p, denoted by $\bar{\angle}_p(q, r)$, is defined by

$$\bar{\angle}_p(q, r) = \arccos \frac{d^2(p, q) + d^2(p, r) - d^2(q, r)}{2d(p, q)d(p, r)}.$$

The Alexandrov angle between two geodesics γ_p^q and γ_p^r emanating from p to q, r in a uniquely geodesic space, denoted by $\angle_p(q, r)$, is defined by

$$\angle_p(q, r) = \lim_{r \to 0} \sup_{0 < s,t < r} \bar{\angle}_p(\gamma_p^q(t), \gamma_p^r(s)).$$

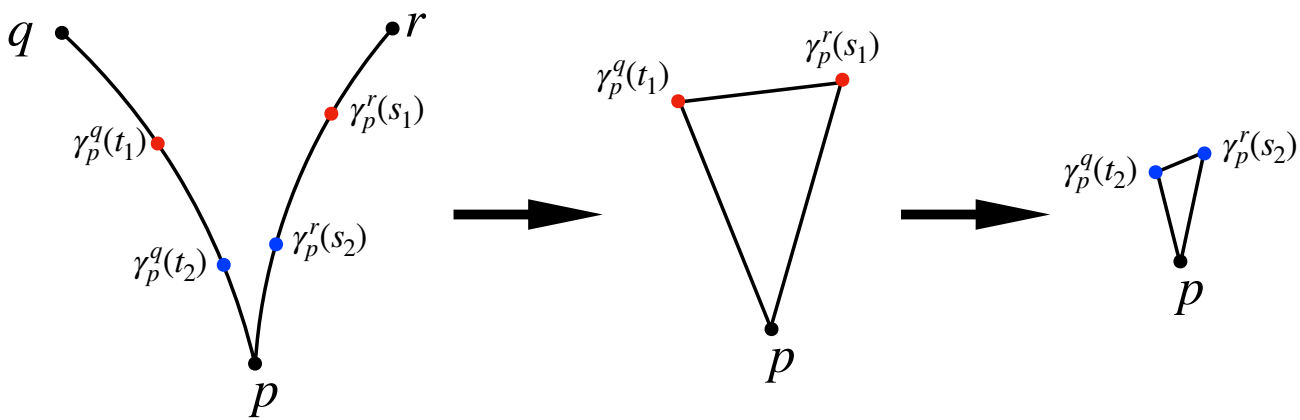


Figure 2: Illustration of Alexandrov angle.

Alexandrov inner product

For three distinct points $p, q, r \in \mathcal{M}$, it is defined as:

$$\langle q, r \rangle_p = d(p, q) d(p, r) \cos(\angle_p(q, r)),$$

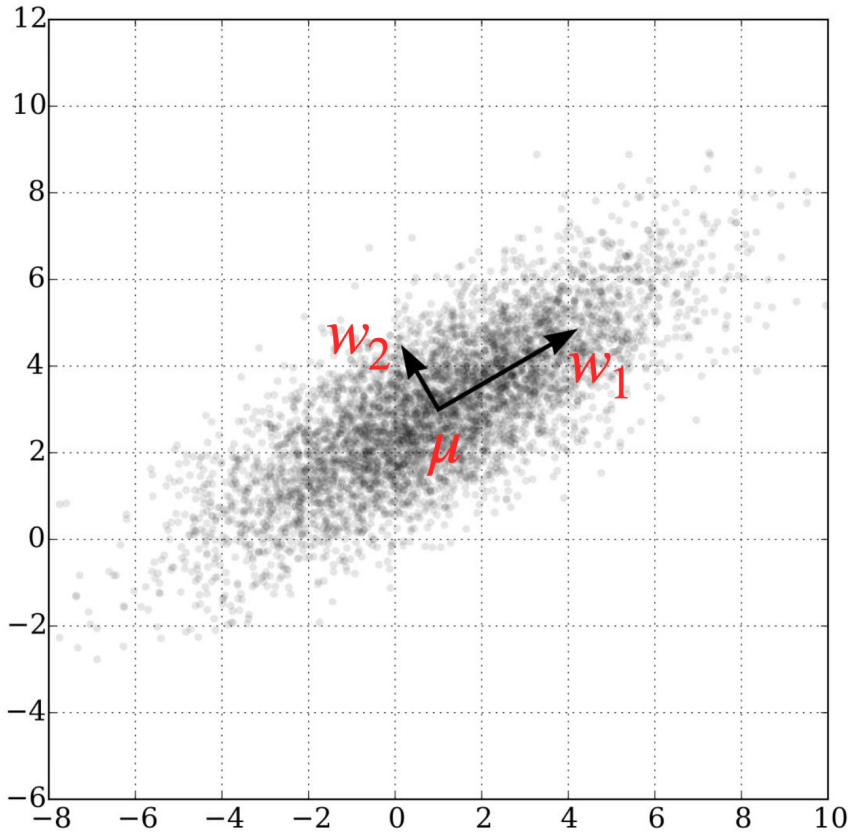
- It can measure the similarity of two points in the space relative to a given point;
- It can be used to define orthogonality: q and r are orthogonal at p if and only if $\langle q, r \rangle_p = 0$.
- It is equivalent to the inner product in Euclidean space,

$$\langle q, r \rangle_p = (q - p)^\top (r - p).$$

PCA in Euclidean space

Assume X is a random vector of $\mathbb{R}^p$, and we seek the direction w that maximizes the variance:

$$\arg\max_{\|w\|=1} \ \mathrm{Var}(w^\top X) = \mathbb{E}\langle w, X - \mu\rangle^2.$$



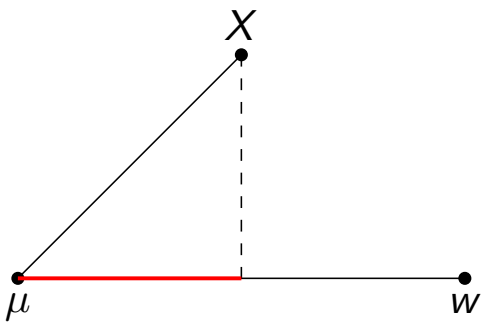
The first PC

By the form of the Alexandrov inner product, we propose the following objective function

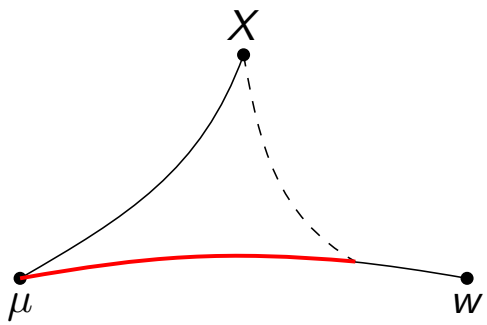
$$\max_{w \in \mathcal{M}} \mathbb{E}\left[\langle X, w\rangle_\mu^2\right], \text{ subject to } d(\mu, w) = r,$$

where $d(\mu, w) = r$ restricts w to the set $\mathcal{M}(r) = \{w \in \mathcal{M} : d(\mu, w) = r\}$. Then the objective function simplifies to:

$$\mathcal{L}(w) = \mathbb{E}\left[d^2(\mu, X)\cos^2(\angle_\mu(X, w))\right], \tag{1}$$



Euclidean Space



Curved Metric Space

The first PC

The first principal component w_1 is defined as:

$$w_1^* = \arg\max_{w \in \mathcal{M}(r)} \mathcal{L}(w).$$

Remark 1

- *Selection of the reference points: Fréchet mean, Geometric median.*
- w_1^* *is not unique even if we constrain* $w \in \mathcal{M}(r)$.
- *We denote that all solutions of* (16) *as* $[w_1^*] := \{w_{1,1}, \ldots, w_{1,s_1}\}$.
- *The distance between the point and a finite set as*

$$\tilde{d}(x, A) = \min_{y \in A} d(x, y),$$

we say the $\hat{w}_1$ *is convergent if it converges to a point in* $[w_1^*]$, *i.e.* $\tilde{d}(\hat{w}_1, [w_1^*]) \xrightarrow{p} 0$.

The first k PCs

Let $\mathcal{M}^k(r) := \{(w_1, \ldots, w_k) : w_i \in \mathcal{M}(r), \forall i = 1, \ldots, k\}$. The parameter space of the PCs is the set of points $(w_1, \ldots, w_k)$ that are orthogonal at μ, which is defined as

$$\Theta^k(r) = \Big\{\theta = (w_1, \ldots, w_k) \in \mathcal{M}^k(r) : (w_1, \ldots, w_k) \text{ is an orthogonal basis;}$$
$$\mathbb{E}d^2(X, \mu)\cos^2(\angle_\mu(X, w_1)) \geq \ldots \geq \mathbb{E}d^2(X, \mu)\cos^2(\angle_\mu(X, w_k))\Big\}.$$

For $\theta_1 = (w_1, \ldots, w_k)$ and $\theta_2 = (w_1', \ldots, w_k')$,

$$d(\theta_1, \theta_2) = \max_{1 \leq s \leq k} d(w_s, w_s').$$

The first k PCs

Our objective of finding the first k PCs is

$$\mathcal{L}(\theta) = \sum_{s=1}^{k} \mathbb{E} d^2(X, \mu) \cos^2(\angle_\mu(X, w_s)).$$

The first k PCs is

$$\theta^* = (w_1^*, \ldots, w_k^*) = \arg\max_{\theta \in \Theta^k(r)} \mathcal{L}(\theta).$$

Illustration

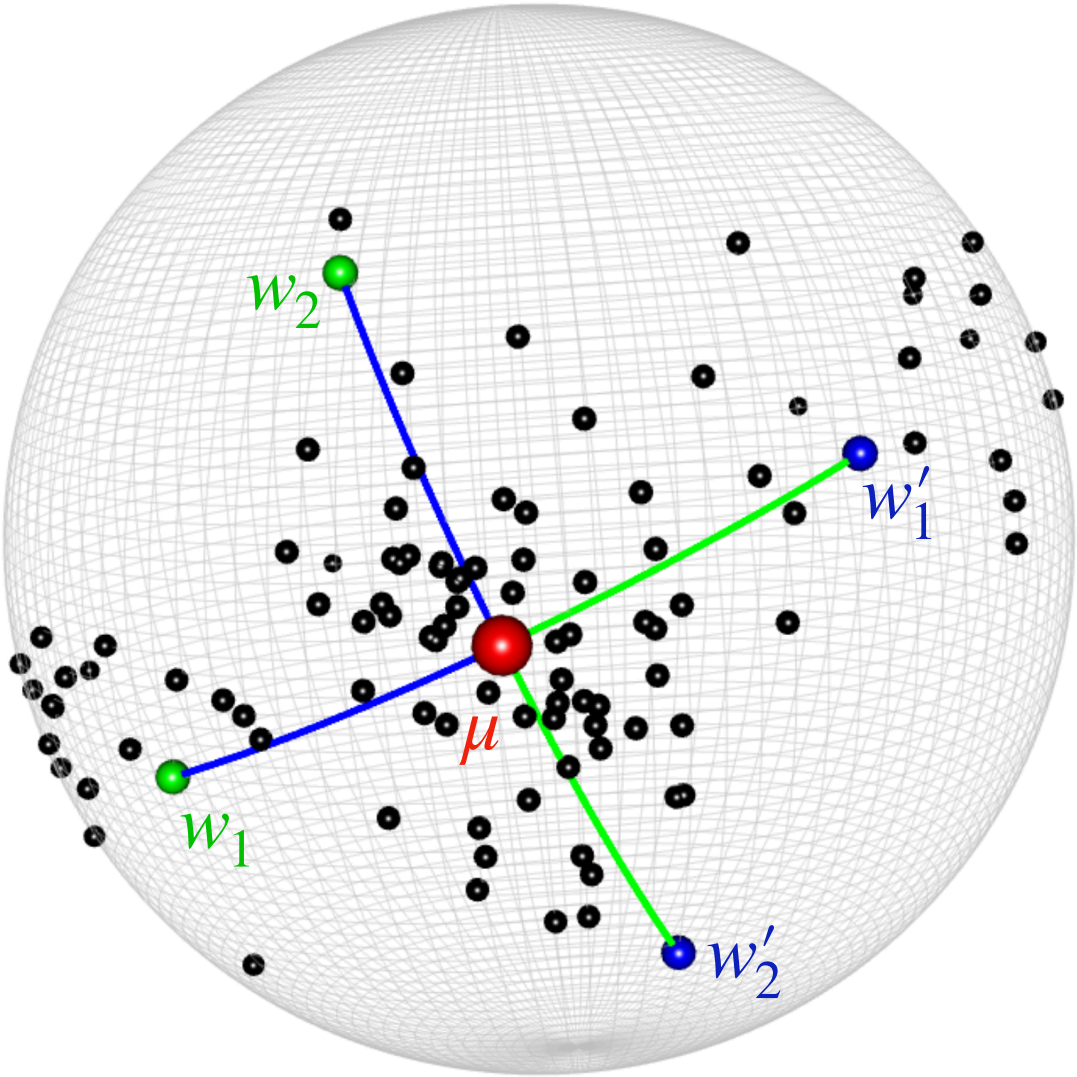


Figure 3: G-PCA on the sphere, w_1, w_2 is a set of PCs, w_1', w_2' is another set of PCs.

PC Score

- Assume we find an orthogonal basis $\{w_1^*, \ldots, w_k^*\}$ that can explains the largest variance of X;

- View $\{w_1^*, \ldots, w_k^*\}$ as the new coordinate system, the coordinates of point X can be expressed as $(\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k$,

$$\lambda_k = d(X, \mu) \cos(\angle_\mu(X, w_k^*)),$$

λ_s is referred to as the sth principal component score.

Estimation

With the sample $X_1, \ldots, X_n$, the sample objective function is

$$\mathcal{L}_n(\theta) = \frac{1}{n} \sum_{s=1}^{k} \sum_{i=1}^{n} d^2(X_i, \mu) \cos^2(\angle_\mu(X_i, w_s)),$$

then the solutions of (18) is

$$\hat{\theta}_n = (\hat{w}_1, \ldots, \hat{w}_k) = \underset{\theta \in \Theta^k(r)}{\arg\max}\, \mathcal{L}_n(\theta).$$

Optimizing the objective function under constraints that $(w_1, \ldots, w_k)$ are orthogonal is still challenging, we adopt the penalty method,

$$\hat{\theta}_\xi = \underset{\theta \in \mathcal{M}^k(r)}{\arg\max}\, \mathcal{L}_n(\theta; \xi) = \mathcal{L}_n(\theta) - \xi \cdot \sum_{1 \le s \ne t \le k} \cos^2(\angle_\mu(w_s, w_t)),$$

as $\xi \to \infty$, the solution $\hat{\theta}_\xi$ would converge to $\hat{\theta}_n$.

Algorithm The generalized LIPO algorithm in geodesic spaces.

Input: Data $\{X_1, \ldots, X_n\}$, reference point μ, PC numbers k;

1: Parameters: penalty parameter ξ, radius $r > 0$, iteration steps T;
2: Set Lipschitz constant $K = 2C_1\{E[d^2(X, \mu)] + \xi k^2\}$;
3: Initialize the initial points $\theta^{(0)} = (w_1^{(0)}, \ldots, w_k^{(0)})$, set $t = 0$;
4: **while** $t \leq T$ **do**
5: Sample θ uniformly from $\mathcal{M}^k(r)$
6: **if** $\min_{j=1,\ldots,t} \left\{\mathcal{L}_n(\theta^{(j)}; \xi) + K \cdot d(\theta, \theta^{(j)})\right\} \geq \max_{j=1,\ldots,t} \mathcal{L}_n(\theta^{(j)}; \xi)$ **then**
7: Set $\theta^{(t+1)} = \theta$, $t = t + 1$;
8: **end if**
9: **end while**

Output: The first k principal components $\hat{\theta}_n = \theta^{(\hat{t})}$, where $\hat{t} = \arg\max_{1 \leq t \leq T} \mathcal{L}_n(\theta^{(t)}; \xi)$.

Algorithm Coordinate Gradient Ascent Algorithm in Riemannian Manifolds.

Input: Data $\{X_1, \ldots, X_n\}$, reference point μ, PC numbers k;

1: Parameters: penalty parameter ξ, radius $r > 0$, tolerance level $\epsilon > 0$, step sizes $\{\gamma(t)\}$, iteration steps T;
2: Initialize the initial points $\theta^{(0)} = \{w_1^{(0)}, \ldots, w_k^{(0)}\}$, set $t = 0$, err $= 2\epsilon$;
3: **while** err $> \epsilon$ **and** $t \leq T$ **do**
4: **for** $s = 1$ to k **do**
5: Compute the gradient of $\mathcal{L}_n(\theta)$ at $w_s^{(t)}$, i.e., $\eta_s(t) = \text{grad } \mathcal{L}_n|_{w_s^{(t)}}$;
6: Update $w_s^{(t)}$: $w_s^{(t+1)} = \text{EXP}_{w_s^{(t)}}\{\gamma(t)\eta_s(t)\}$;
7: Normalize: $w_s^{(t+1)} \leftarrow \gamma_\mu^{w_s^{(t+1)}}(r)$;
8: **end for**
9: Set $\theta^{(t+1)} = (w_1^{(t+1)}, \ldots, w_k^{(t+1)})$;
10: Set err $\leftarrow |\mathcal{L}_n(\theta^{(t+1)}; \xi) - \mathcal{L}_n(\theta^{(t)}; \xi)|$; $t \leftarrow t + 1$;
11: **end while**

Output: The first k principal components $\hat{\theta}_n = \theta^{(t)}$.

Table of Contents

Assumptions

Assumption A

The space $\mathcal{M}$ is a complete, uniquely geodesic metric space.

Assumption B

The second moment of X is finite, that is, $E\{d^2(X, \mu)\} < \infty$, and the reference point μ is an interior point of $\mathcal{M}$.

Assumption C

The set of principal components $[\theta^*]$ is finite, and for any $\epsilon > 0$,

$$\mathcal{L}(\theta^*) > \sup_{d(\theta,[\theta^*])>\epsilon} \mathcal{L}(\theta).$$

Consistency

Theorem 1

Under Assumptions A - C, and fixed k, we have

$$d(\hat{\theta}_n, [\theta^*]) = o_p(1),$$

Theorem 1 establishes the consistency of the estimator $\hat{\theta}$ to the population principal components $[\theta^*]$.

Assumption

We make additional Assumptions to derive the convergence rate of our estimators.

Assumption D

There exist $\eta > 0$, $L > 0$ *and* $\beta > 1$, *for any* θ *satisfying* $d(\theta, [\theta^*]) < \eta$, *we have*

$$\mathcal{L}(\theta^*) - \mathcal{L}(\theta) \geq L d(\theta, [\theta^*])^{\beta}.$$

- $\beta = 2$ when $\mathcal{M}$ is a Riemannian manifold.

Assumption

Assumption E (Local metric entropy condition)

Let $B(\theta_0, \delta) \subset \Theta$ be the ball of radius δ centered at θ_0 and $N(\epsilon, B(\theta, \delta), d)$ be its covering number using balls of size ϵ. There is a constant $C, D > 0$, such that for $\theta_0 \in [\theta^*]$, one of the following holds:

E1: (Log-polynomial metric entropy) $N\left(\epsilon, B\left(\theta_0, \delta\right), d\right) \leq C\left(\frac{\delta}{\epsilon}\right)^{kD}$;

E2: (Polynomial metric entropy) $\log N\left(\epsilon, B\left(\theta_0, \delta\right), d\right) \leq Ck\left(\frac{\delta}{\epsilon}\right)^{D}$.

- Condition E1 typically holds for D-dimensional Riemannian manifolds, such as spaces of symmetric positive definite matrices or spheres.
- Condition E2 accommodates exponential growth in covering numbers, such as in the Wasserstein space $\mathcal{W}_2(\mathbb{R})$ of probability measures on the real line with Wasserstein distance d_W, $\sup_{F \in \mathcal{W}_2(\mathbb{R})} \log N\{\epsilon, B(F, \delta), d_W\} \leq C\delta/\epsilon$ for some positive constant C and all $\delta, \epsilon > 0$.

Convergence rate

Theorem 2

Assume Assumptions A–D. If Assumption E1 also holds, we have

$$d(\hat{\theta}_n, [\theta^*]) = O_p\left\{\left(\frac{n}{D}\right)^{-\frac{1}{2(\beta-1)}}\right\}.$$

If Assumption E2 holds with $D < 2$, we have

$$d(\hat{\theta}_n, [\theta^*]) = O_p\left\{n^{-\frac{1}{2(\beta-1)}}\right\}.$$

If Assumption E2 holds with $D = 2$, we have

$$d(\hat{\theta}_n, [\theta^*]) = O_p\left\{\left(\frac{n}{\log^2 n}\right)^{-\frac{1}{2(\beta-1)}}\right\}.$$

If Assumption E2 holds with $D > 2$, we have

$$d(\hat{\theta}_n, [\theta^*]) = O_p\left\{n^{-\frac{1}{2(D-1)}}\right\}.$$

Theorem 3

Assume the Alexandrov inner product is regular near μ. Then under Assumptions A–D, there is a sufficiently small $r_1 > 0$, such that for any $\tilde{\mu} \in B(\mu, r_1)$,

$$d([\theta^*], [\tilde{\theta}]) \leq C \left\{ L d(\mu, \tilde{\mu}) \right\}^{\frac{1}{\beta}} .$$

where $[\tilde{\theta}]$ is the PCs at $\tilde{\mu}$ and C is a universal constant.

Table of Contents

Data generating process

1. Choose an origin point μ;
2. At the tangent space T_μ, generate k orthogonal tangent vectors $v_1, \ldots, v_k$, and $w_i := \mathrm{Exp}_\mu(v_i)$;
3. For each data point i, apply the Riemannian exponential map at μ to the linear combination of the loadings and the basis vectors, i.e., define $\widetilde{X}_i = \mathrm{Exp}_\mu \left(\sum_{j=1}^k \lambda_{ij} v_j \right)$;
4. For each data point i, generate a random direction E_i at $\widetilde{X}_i$ and a random variable $u_i \sim U(0, \sigma)$, where σ controls the signal-to-noise ratio. Then apply exponential map at $\widetilde{X}_i$ in the direction of E_i, with the magnitude of $u_i \times d(\mu, \widetilde{X}_i)$;
5. Compute the Fréchet mean of samples $X_1, \ldots, X_n$, as $\hat{\mu}$;

6. Evaluation criteria: we consider the angle at the estimated origin $\hat{\mu}$ between w_s and its estimate $\hat{w}_s$,

$$|\cos(\angle_{\hat{\mu}}(w_s^*, \hat{w}_s))|, \quad s = 1, \ldots, k.$$

Symmetric positive definite matrix

We consider the space of SPD matrices, denoted as S_m^+, endowed with the Affine-Invariant and Log-Cholesky metric.

Table 1: Median of $|\cos(\angle_{\hat{\mu}}(w_s^*, \hat{w}_s))|$ over 100 replications in SPD space.

	$n = 50, m = 10$			$n = 100, m = 10$			$n = 50, m = 20$			$n = 100, m = 20$		
σ	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
Affine-Invariant												
0.2	0.988	0.977	0.973	0.994	0.989	0.973	0.989	0.981	0.968	0.995	0.991	0.983
0.4	0.986	0.974	0.968	0.993	0.988	0.973	0.988	0.980	0.965	0.994	0.990	0.980
0.6	0.985	0.973	0.955	0.991	0.987	0.968	0.987	0.979	0.955	0.992	0.986	0.974
0.8	0.981	0.966	0.935	0.989	0.984	0.959	0.982	0.970	0.935	0.991	0.984	0.965
1.0	0.977	0.958	0.905	0.986	0.979	0.946	0.978	0.963	0.909	0.989	0.980	0.956
Log-Cholesky												
0.2	0.987	0.976	0.942	0.993	0.987	0.961	0.987	0.975	0.952	0.994	0.987	0.959
0.4	0.985	0.971	0.942	0.992	0.982	0.954	0.984	0.974	0.942	0.993	0.984	0.956
0.6	0.983	0.966	0.926	0.992	0.981	0.949	0.983	0.975	0.933	0.990	0.982	0.948
0.8	0.981	0.962	0.922	0.989	0.980	0.940	0.981	0.970	0.919	0.989	0.979	0.946
1.0	0.976	0.958	0.882	0.987	0.974	0.921	0.978	0.962	0.886	0.987	0.975	0.934

Kendall's shape space Σ_m^K

Σ_m^K represents shapes formed by K landmarks in m-dimensional space, excluding translation, scaling, and rotation, where each shape is expressed as a $K \times m$ matrix. The geodesic distance between shapes $\pi(X)$ and $\pi(Y)$ is: $\rho(\pi(X), \pi(Y)) = \arccos\left(\max_{A \in SO(m)} \operatorname{tr}(AXY^\top)\right)$.

Table 2: Median of $|\cos(\angle_{\hat{\mu}}(w_s^*, \hat{w}_s))|$ over 100 replications in shape space.

		$k=30, m=2$			$k=100, m=2$			$k=30, m=5$			$k=100, m=5$		
σ	n	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3	w_1	w_2	w_3
0.2	50	0.990	0.982	0.974	0.989	0.980	0.979	0.989	0.983	0.974	0.988	0.981	0.976
0.2	100	0.995	0.992	0.983	0.995	0.991	0.983	0.994	0.990	0.980	0.994	0.991	0.977
0.4	50	0.988	0.981	0.969	0.987	0.978	0.971	0.988	0.980	0.970	0.987	0.980	0.970
0.4	100	0.995	0.991	0.982	0.994	0.989	0.980	0.993	0.990	0.979	0.994	0.990	0.979
0.6	50	0.987	0.980	0.961	0.985	0.976	0.961	0.985	0.977	0.958	0.985	0.976	0.961
0.6	100	0.994	0.989	0.976	0.994	0.988	0.977	0.992	0.988	0.974	0.993	0.989	0.976
0.8	50	0.983	0.974	0.943	0.981	0.970	0.940	0.981	0.972	0.945	0.980	0.970	0.941
0.8	100	0.992	0.986	0.967	0.993	0.987	0.967	0.990	0.986	0.968	0.992	0.985	0.968
1.0	50	0.978	0.966	0.908	0.976	0.963	0.900	0.976	0.967	0.908	0.978	0.964	0.893
1.0	100	0.990	0.982	0.953	0.991	0.983	0.960	0.989	0.983	0.950	0.989	0.983	0.956

Wasserstein space $\mathcal{W}_2(\mathbb{R})$

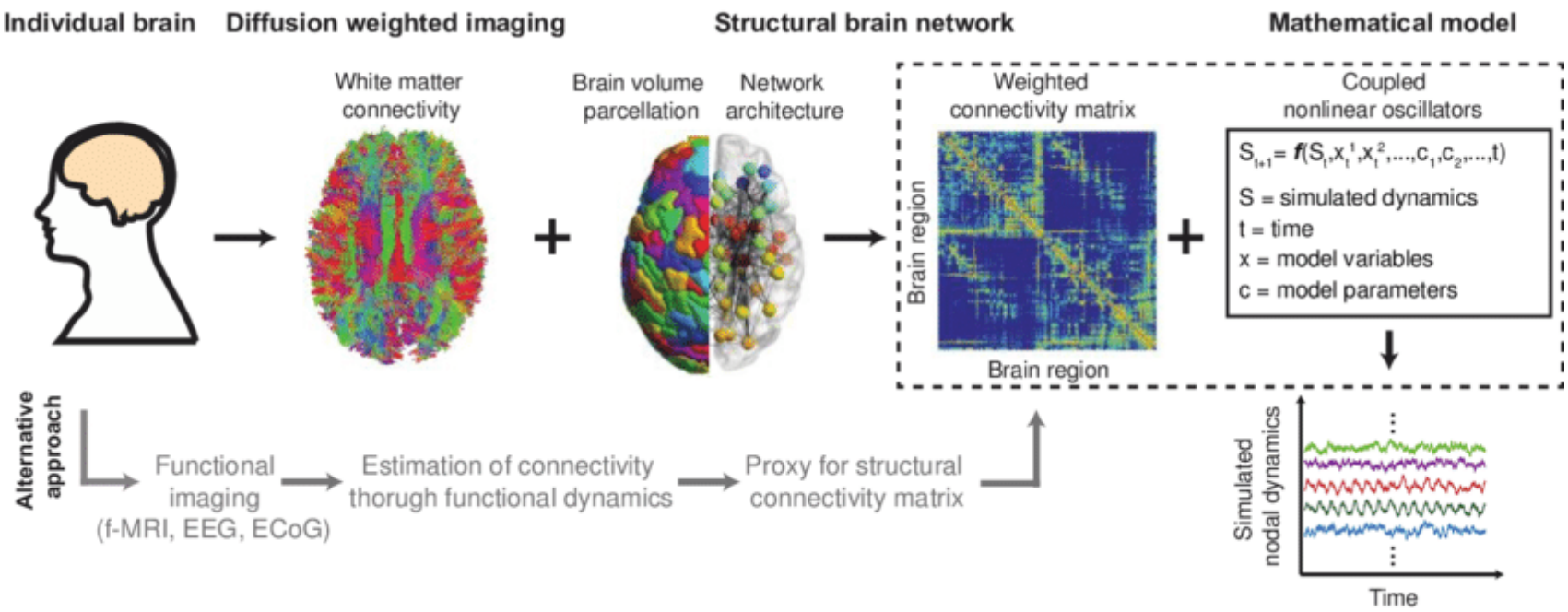
Let $\mathcal{W}_2(\mathbb{R})$ be the space of probability distributions on $\mathbb{R}$ with finite second moments, equipped with the Wasserstein distance, defined for two CDFs F_1 and F_2 as

$$d_W(F_1, F_2) = \left[\int_0^1 \left\{F_1^{-1}(s) - F_2^{-1}(s)\right\}^2 ds\right]^{1/2}.$$

Table 3: Median of $|\cos(\angle_{\hat{\mu}}(w_s^*, \hat{w}_s))|$ over 100 replications in Wasserstein space.

	$n = 50, m = 50$		$n = 100, m = 50$		$n = 50, m = 100$		$n = 100, m = 100$	
σ	w_1	w_2	w_1	w_2	w_1	w_2	w_1	w_2
0.2	0.995	0.987	0.997	0.995	0.996	0.991	0.998	0.994
0.4	0.994	0.983	0.996	0.989	0.995	0.987	0.997	0.991
0.6	0.991	0.977	0.994	0.984	0.993	0.979	0.995	0.984
0.8	0.987	0.963	0.992	0.977	0.989	0.972	0.992	0.979
1.0	0.983	0.956	0.988	0.961	0.985	0.955	0.989	0.969

Brain connectivity networks



Task-fMRI Result

- Source: Emotion task data (Human Connectome Project);
- Scope: 69 regions of interest (ROIs), focusing on 8 task-relevant ROIs;
- Participants: 299, over 176 time points;
- Functional Connectivity: 299 SPD 8×8 covariance matrices;

Applied G-PCA to identify principal components

- Variance Explained by First Three Components: 1st: 36.8%, 2nd: 16.6%, 3rd: 7.3%

Next: focus on first principal component, visualized along first geodesic.

Task-fMRI Result

The matrices were calculated as $\widehat{X}_j = \mathrm{Exp}_{\hat{\mu}}(q_j \hat{v}_1)$, where $\hat{v}_1 = \mathrm{Log}_{\hat{\mu}}(\hat{w}_1)$, $q_1, \ldots, q_8$ is the eight of $\{\hat{\lambda}_{11}, \ldots, \hat{\lambda}_{n1}\}$.

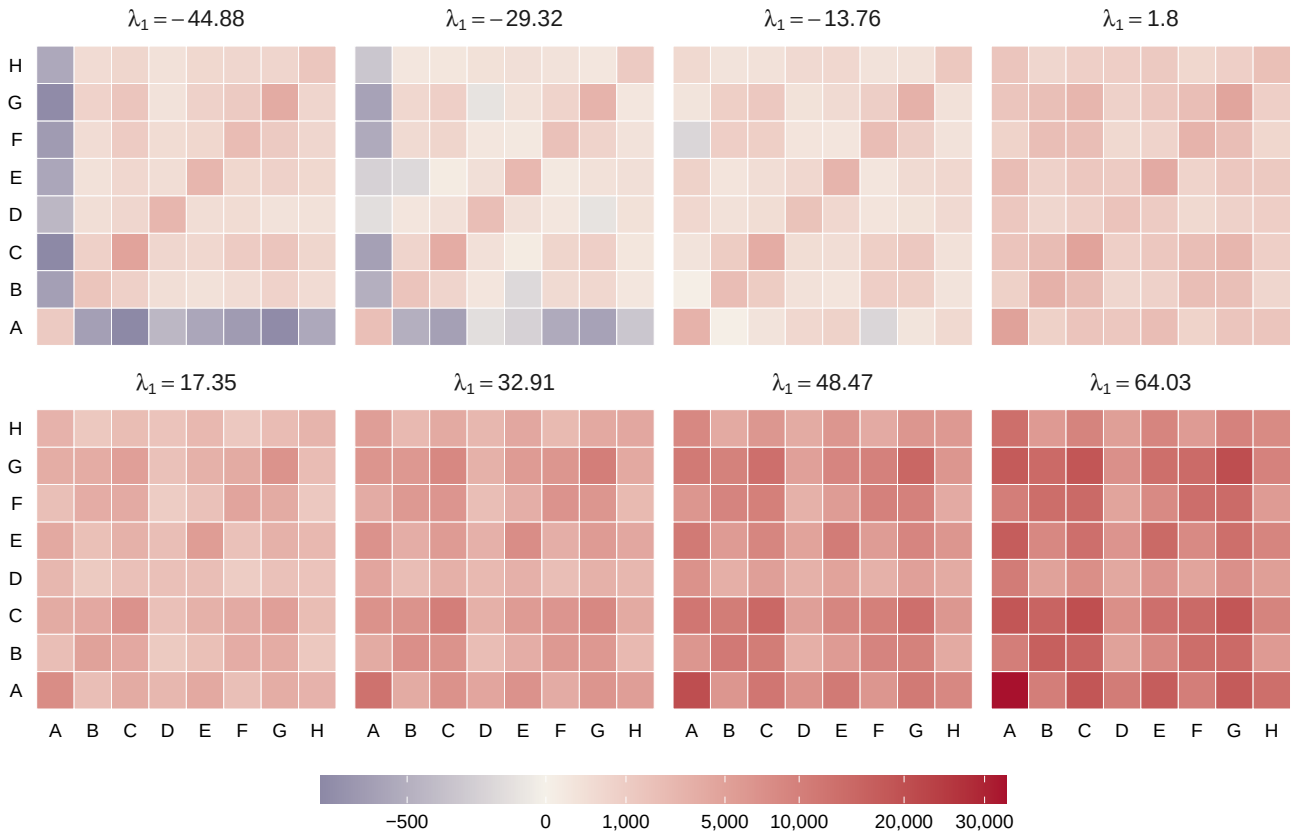


Figure 4: Variation in brain connectivity matrices along the first principal component directions.

Task-fMRI Result

- Older, lower scores: age-related neural decline or greater emotional processing maturity
- Females, lower scores: higher sensitivity and intuitive efficiency in emotion recognition

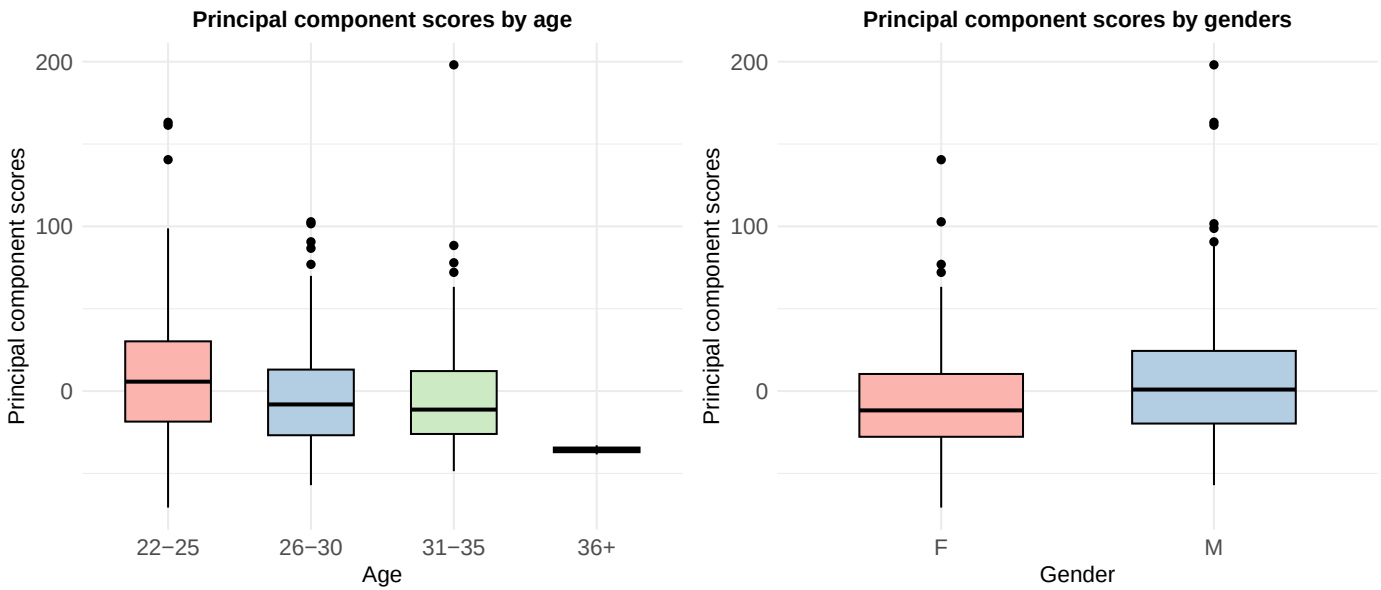


Figure 5: Principal component scores by age and genders.

Task-fMRI Result

Table 4: Pairwise correlations between first principal component scores and task-related variables.

Variable	Correlation	p-value
Age	-0.1519	0.0085(**)
GenderM	0.2113	0.0002(***)
Emotion_Task_Acc	-0.1559	0.0069(**)
Emotion_Task_Median_RT	0.0495	0.3941
Emotion_Task_Face_Acc	-0.1742	0.0025(**)
Emotion_Task_Face_Median_RT	0.0642	0.2682
Emotion_Task_Shape_Acc	-0.1065	0.0658
Emotion_Task_Shape_Median_RT	0.1181	0.0412(*)

The p-values are indicated in parentheses: (*) for $p < 0.05$, (**) for $p < 0.01$, and (***) for $p < 0.001$.

Corpus Callosum

- 2D shape data by 50 landmarks, totally 409 individuals.
- Covariates:
  - Alzheimer's Disease (186) or Normal Control (223).
  - Gender: 214 male and 195 female.
  - Age: ranges from 55 to 92.
  - Marital Status (Widowed/Divorced/Never married).
  - Education length: range from 4 to 20.
  - Retirement.

Results

- The first three principal components explain 53.1%, 13.1%, and 9.2% of the total variations.

Table 5: Regression results on the first three PC scores

Covariate	w_1	w_2	w_3
(Intercept)	-14.3074(***)	8.4402(***)	0.7046
GenderM	1.1287(*)	0.0989	-0.1379
Handness	-0.8531	-0.1964	0.2715
Widowed	-0.7857	-0.5813	0.6298(*)
Divorced	-0.0558	0.1396	-0.7051
Never_married	0.7205	-0.5701	0.0544
Edu_length	0.2025(**)	-0.0603	0.0492
Retirement	0.1755	-0.4701	-0.0924
Age	0.1426(***)	-0.0821(***)	-0.0228
DiagnosisAD	1.3712(**)	-1.3558(***)	0.1705

The p-values are indicated in parentheses: (*) for $p < 0.05$, (**) for $p < 0.01$, and (***) for $p < 0.001$.

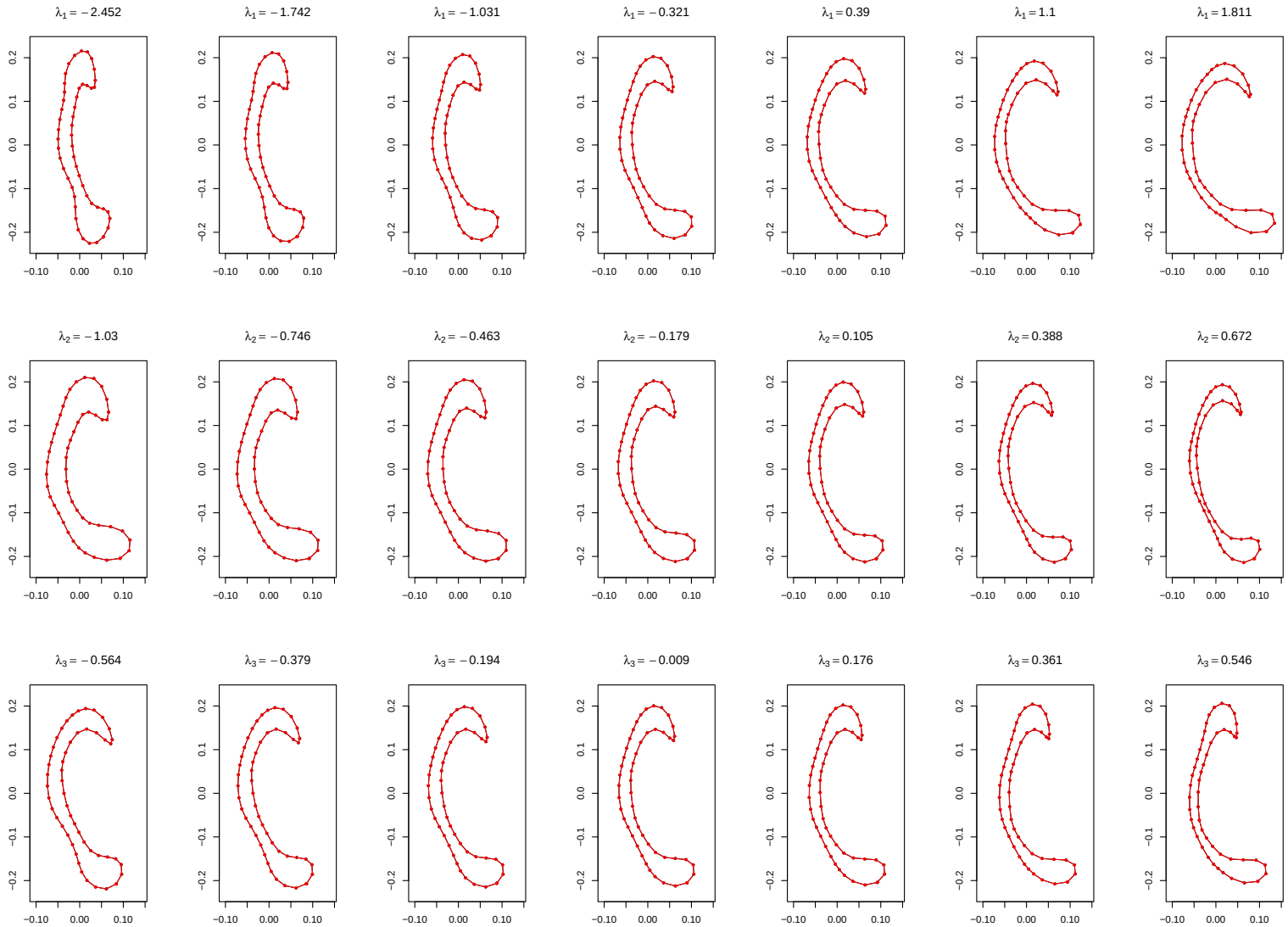
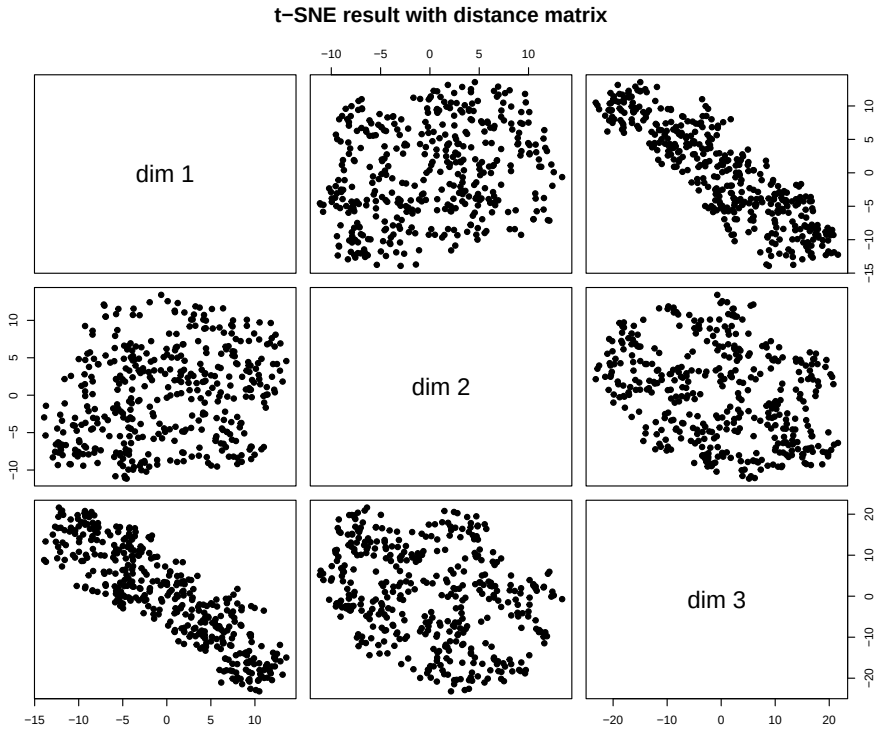


Figure 6: Variation in corpus callosum shapes corresponds to the first three principal component directions.

Compare with t-SNE



- Unstable output: the results of t-SNE are random.
- Lack of interpretability: the coordinates obtained by t-SNE dimensionality reduction do not have clear meanings.

Table of Contents

Conclusion and discussion

- Propose a unified G-PCA method in geodesic spaces.
- Provide efficient algorithms for computations.
- Establish the theoretical results of PCA in geodesic spaces.

References I

[1] Burcu Aydın, Gábor Pataki, Haonan Wang, Elizabeth Bullitt, and J. S. Marron. A principal component analysis for trees. *The Annals of Applied Statistics*, 3(4):1597 – 1615, 2009.

[2] Jérémie Bigot, Raúl Gouet, Thierry Klein, and Alfredo López. Geodesic PCA in the Wasserstein space by convex PCA. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(1):1 – 26, 2017.

[3] Jonathan Boisvert, Xavier Pennec, Hubert Labelle, Farida Cheriet, and Nicholas Ayache. Principal spine shape deformation modes using riemannian geometry and articulated models. In *Articulated Motion and Deformable Objects: 4th International Conference, AMDO 2006, Port d'Andratx, Mallorca, Spain, July 11-14, 2006. Proceedings 4*, pages 346–355. Springer, 2006.

References II

[4] Anna Calissano, Aasa Feragen, and Simone Vantini. Populations of unlabelled networks: graph space geometry and generalized geodesic principal components. *Biometrika*, page asad024, 04 2023. ISSN 1464-3510.

[5] Steven Campbell and Ting-Kam Leonard Wong. Efficient convex pca with applications to wasserstein gpca and ranked data. *Journal of Computational and Graphical Statistics*, pages 1–12, 2024.

[6] P Thomas Fletcher and Sarang Joshi. Principal geodesic analysis on symmetric spaces: Statistics of diffusion tensors. In *International Workshop on Mathematical Methods in Medical and Biomedical Image Analysis*, pages 87–98. Springer, 2004.

[7] 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(8):995–1005, 2004.

References III

[8] Xiaoyang Guo, Aditi Basu Bal, Tom Needham, and Anuj Srivastava. Statistical shape analysis of brain arterial networks (ban). *The Annals of Applied Statistics*, 16(2):1130–1150, 2022.

[9] Stephan Huckemann and Herbert Ziezold. Principal component analysis for riemannian manifolds, with an application to triangular shape spaces. *Advances in Applied Probability*, 38(2):299–319, 2006.

[10] Stephan Huckemann, Thomas Hotz, and Axel Munk. Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, pages 1–58, 2010.

[11] Sungkyu Jung, Ian L. Dryden, and J. S. Marron. Analysis of principal nested spheres. *Biometrika*, 99(3):551–568, 07 2012. ISSN 0006-3444.

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

References IV

[13] Xavier Pennec. Barycentric subspace analysis on manifolds. *The Annals of Statistics*, 46(6A):2711 – 2746, 2018.

[14] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

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