Non-Euclidean data analysis with applications to multi-omics data

Kipoong Kim

Department of Statistics, Changwon National University

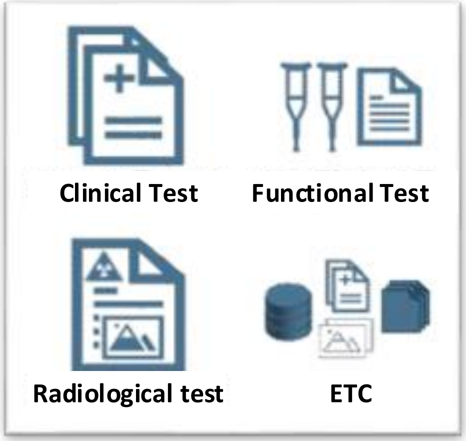
July 17, 2025

Multi-omics datasets from an individual

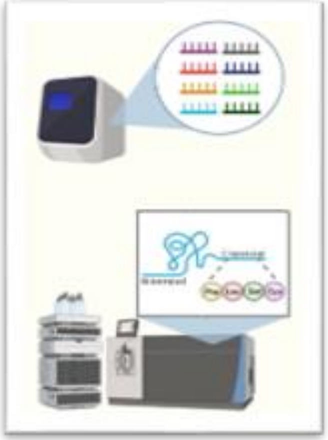
Clinical Outcomes

DATE	DAY	BRAIN	LUNGS	HEART	SYSTOLIC	DIASTOLIC	CELSIUS	PULSE
11/1/2020	Sunday	5	5	5	123	82	36.6	172
11/2/2020	Monday	5	5	5	119	78	36.6	179
11/3/2020	Tuesday	5	5	5	111	80	36.6	84
11/4/2020	Wednesday	5	5	5	120	80	36.6	162
11/5/2020	Thursday	5	4	5	120	80	36.6	52
11/6/2020	Friday	5	5	5	125	81	36.6	80
11/7/2020	Saturday	2	4	5	90	56	37.2	95
11/8/2020	Sunday	2	2	3	101	68	37.4	171
11/9/2020	Monday	5	4	4	147	95	37.6	76
11/10/2020	Tuesday	5	3	4	199	133	37.7	151
11/11/2020	Wednesday	4	2	3	97	70	37.8	154
11/12/2020	Thursday	4	3	4	193	125	38.3	140
11/13/2020	Friday	2	1	2	114	74	38.4	134
11/14/2020	Saturday	2	1	3	207	151	38.5	102

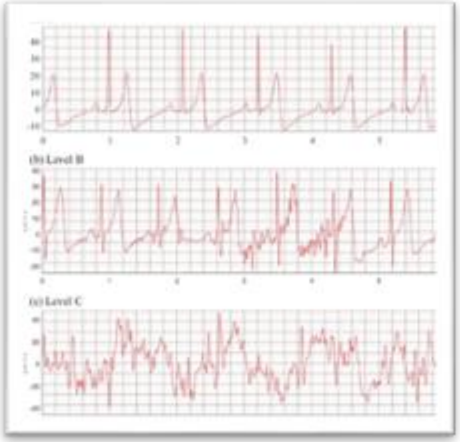
Electronic Health Records



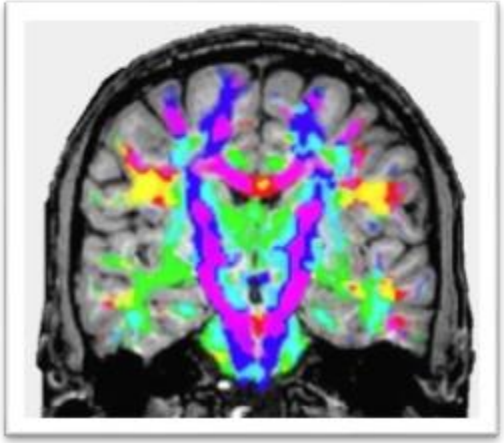
Sequencing data



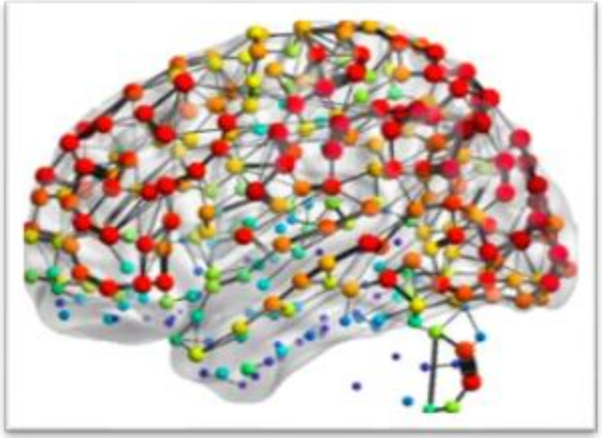
Electrocardiogram (ECG)



Medical Imaging data



Connectome



What is the non-Euclidean data?

- Non-Euclidean data:
  - Data whose underlying structure is not Euclidean

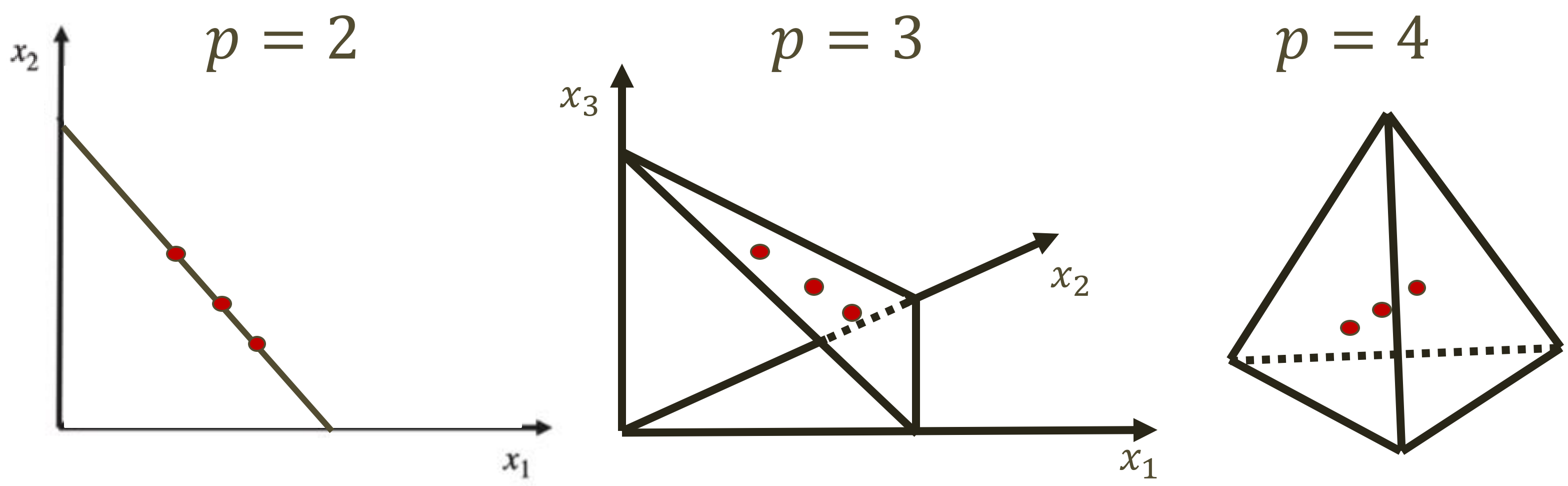
- Some characteristics of $\mathbb{R}^p$:
  - It is a vector space closed under the vector addition $+$ and scalar multiplication $\cdot$

- Non-Euclidean data lying on the following spaces will be introduced:
  - Finite-dimensional vector spaces which are not closed under $+$ and $\cdot$

[1] Microbiome compositional data

- Compositional space

$$\mathbb{C}^{p-1} := \left\{ \boldsymbol{x} = \left[x_1, \dots, x_p \right] \in \mathbb{R}^p :\ x_1 + \cdots x_p = 1,\ x_j > 0\ \ \forall j \right\}$$



- For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{C}^{p-1}$ and $c \in \mathbb{R} \setminus \{1\}$,
  - (i) $\boldsymbol{x} + \boldsymbol{y} \notin \mathbb{C}^{p-1}$, (ii) $c \cdot \boldsymbol{x} \notin \mathbb{C}^{p-1}$, (iii) $\boldsymbol{x} - \boldsymbol{y} \notin \mathbb{C}^{p-1}$

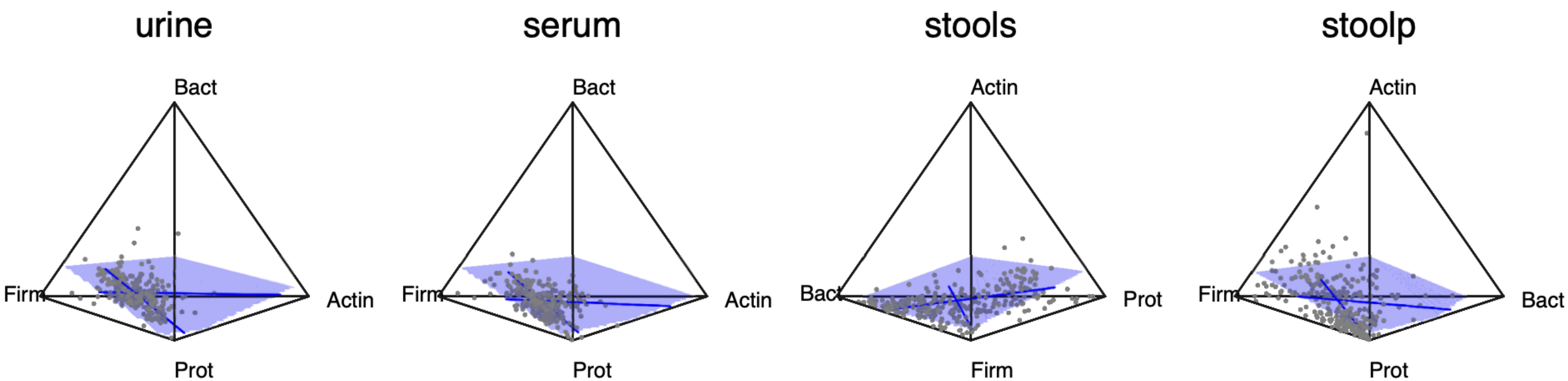
PCA for zero-inflated compositional data[1]

- Global compositional PCA solves the following optimization problem:

$$\underset{\boldsymbol{\mu}\in\mathbb{C}^{p-1},\,\mathbf{U},\,\mathbf{V}}{\text{minimize}} \sum_{i=1}^{n} \|\mathbf{x_i} - \boldsymbol{\mu} - \mathbf{V}\boldsymbol{u}_\mathbf{i}\|_2^2$$

subject to

- $\mathbf{U} = [\mathbf{u_1}, \dots, \boldsymbol{u}_\boldsymbol{n}]^\mathbf{T}$ and $\mathbf{V}$ have orthogonal and orthonormal columns
- $\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i \in \mathbb{C}^{p-1}$ for all $i = 1, \dots, n$

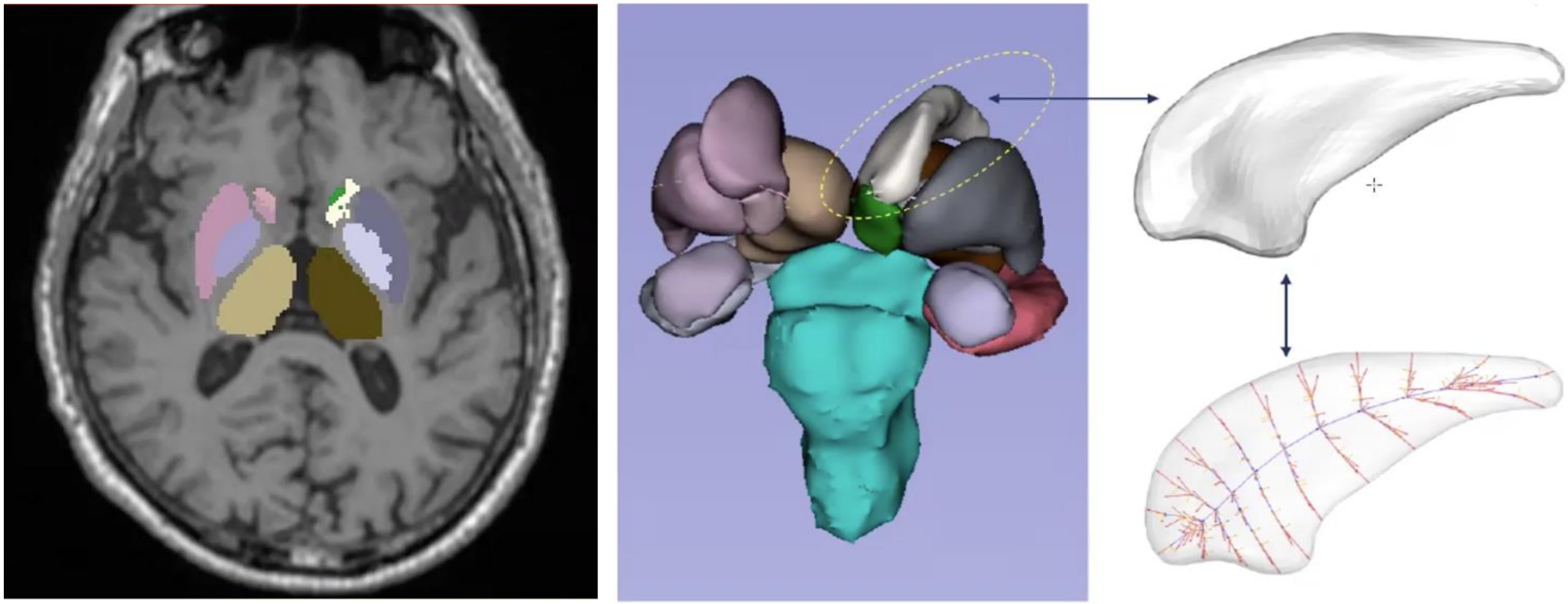


*Prot: Proteobacteria; Firm: Firmicutes; Actin: Actinobacteria; Bact: Bacteroidetes

[1] Kim, K., Park, J., and Jung, S. (2024). Principal component analysis for zero-inflated compositional data, Computational Statistics and Data Analysis, 198, 107989.

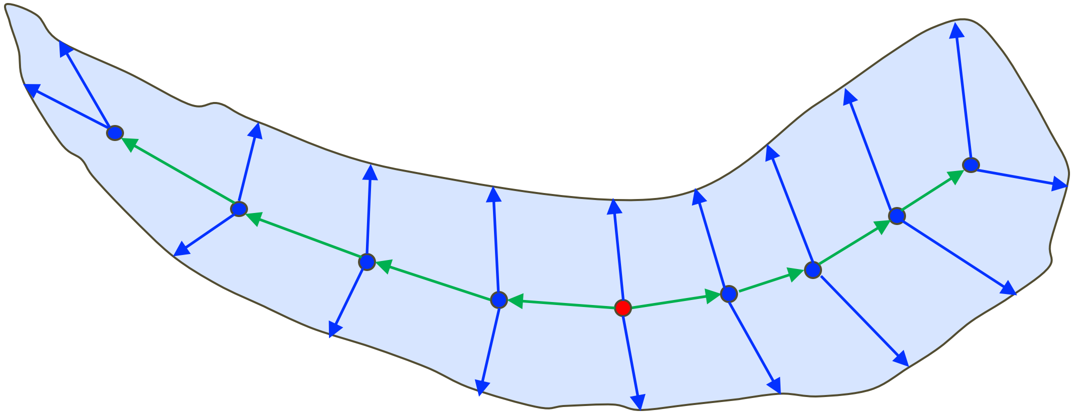
[2] Shape data from a brain MRI

- This study tests for shape differences in the hippocampus, derived from brain MRI, between a Parkinson's disease (PD) group and a normal control group.



Discrete-skeletal representation of a shape

- For simplicity, we use a discrete-skeletal representation (ds-rep). This represents a shape with:
  - spine centroid (red point)
  - n_c connection directions (green arrows) and their lengths
  - n_s spoke directions (blue arrows) and their lengths



The shape space of ds-reps

- Then, from an individual, the following variable is collected:

$$\underbrace{\mathbb{R}^3}_{\text{centroid}} \times \underbrace{(\mathbb{S}^2)^{n_c} \times \mathbb{R}_+^{n_c}}_{\text{connections}} \times \underbrace{(\mathbb{S}^2)^{n_s} \times \mathbb{R}_+^{n_s}}_{\text{spokes}}$$

where $\mathbb{S}^{q-1} := \{\boldsymbol{x} \in \mathbb{R}^q : x_1^2 + \cdots + x_q^2 = 1\}$, $\mathbb{R}_+ = \{\mathrm{x} \in \mathbb{R} : \mathrm{x} > 0\}$

- Here, $\mathbb{S}^{q-1}$ is also clearly non-Euclidean space.

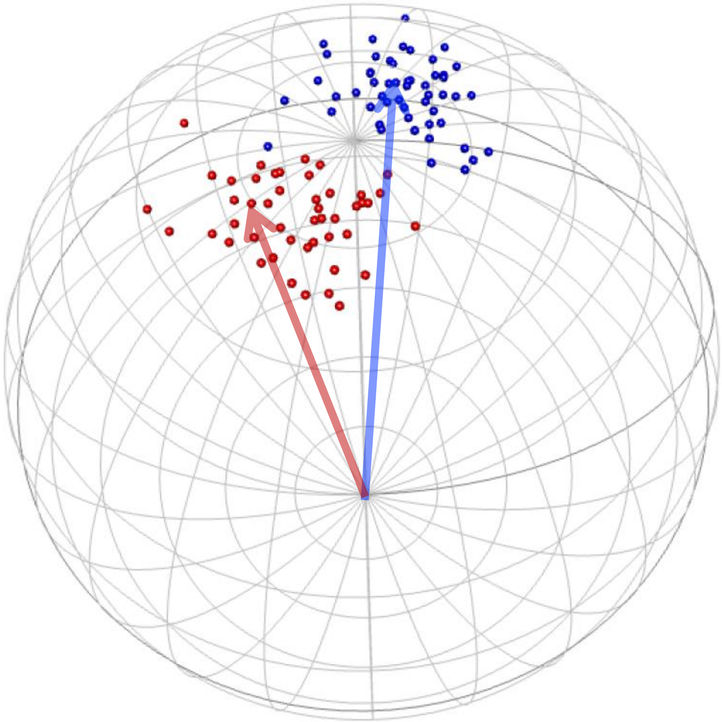
- We focus on a direction at a single spoke or connection.
  - It lies on a two-dimensional sphere ($\mathbb{S}^2$).

Two-sample testing for spherical responses[2]

- PD group $\sim vMF(\boldsymbol{\mu}_1, \kappa_1)$ vs. Control group $\sim vMF(\boldsymbol{\mu}_2, \kappa_2)$

  with a pdf $f_{vMF}(\boldsymbol{y}: \boldsymbol{\zeta}, \kappa) = \mathrm{C}_q(\kappa) \cdot \exp(\kappa \cdot \boldsymbol{\zeta}^T \boldsymbol{y})$

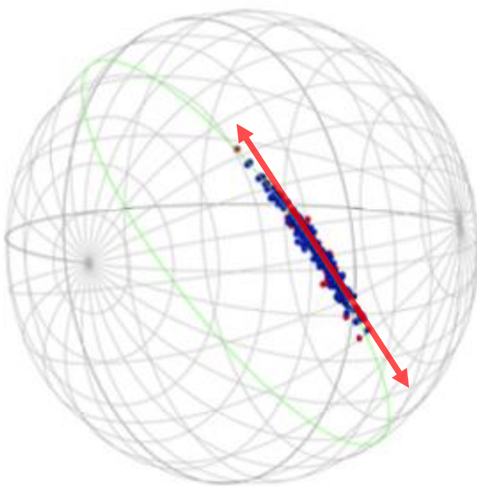
  - $\boldsymbol{\zeta}$: mean direction
  - κ : concentration parameter

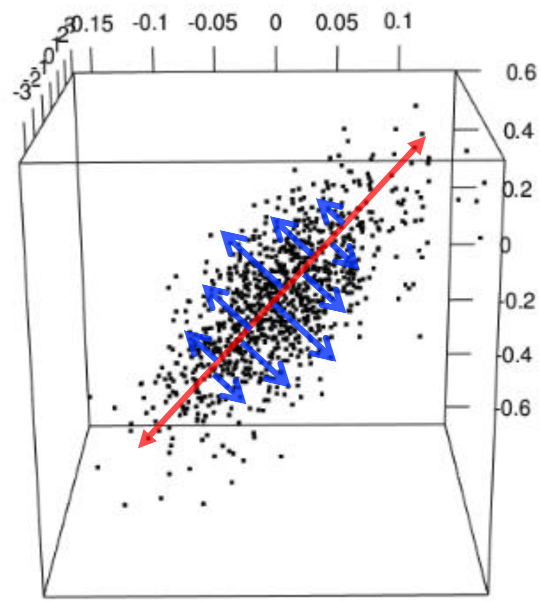


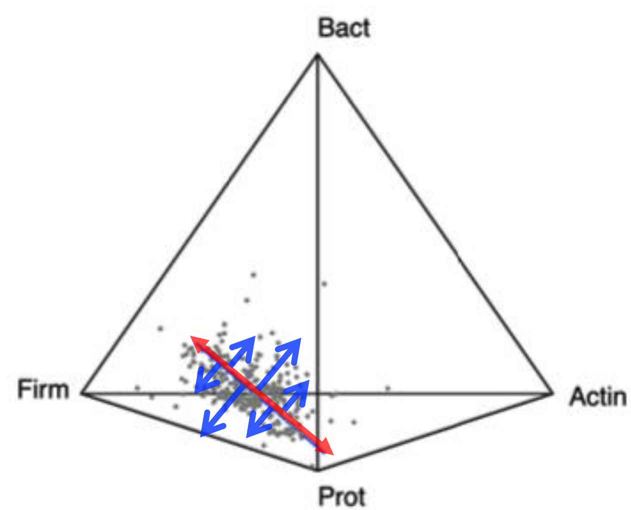
[2] Kipoong Kim and Sungkyu Jung (2025+). Generalized linear model for spherical response with projection-based inference

[3] Non-Euclidean data integration

- Structural decomposition for multiple (non-)Euclidean datasets:
  - Joint variation (red arrow)
  - Individual variation (blue arrow)







Multi-source data integration

- Joint and Individual Variation Explained (JIVE) model

$$\begin{bmatrix} \mathbf{X}_{(1)}^T \\ \vdots \\ \mathbf{X}_{(D)}^T \end{bmatrix} = \underbrace{\begin{bmatrix} \boldsymbol{\mu}_{(1)} \\ \vdots \\ \boldsymbol{\mu}_{(D)} \end{bmatrix} \mathbf{1}_n^T}_{\text{Intercept}} + \underbrace{\begin{bmatrix} \mathbf{V}_{(1)} \\ \vdots \\ \mathbf{V}_{(D)} \end{bmatrix} \mathbf{U}_{(0)}^T}_{\text{Joint}} + \underbrace{\begin{bmatrix} \mathbf{A}_{(1)} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{A}_{(D)} \end{bmatrix} \begin{bmatrix} \mathbf{U}_{(1)}^T \\ \vdots \\ \mathbf{U}_{(D)}^T \end{bmatrix}}_{\text{Individual}} + \underbrace{\begin{bmatrix} \mathbf{E}_{(1)} \\ \vdots \\ \mathbf{E}_{(D)} \end{bmatrix}}_{\text{Error}}$$

- For the d-th source case:

$$\mathbf{X}_{(d)} = \mathbf{1}\boldsymbol{\mu}_{(d)}^T + \mathbf{U}_{(0)}\mathbf{V}_{(d)}^T + \mathbf{U}_{(d)}\mathbf{A}_{(d)}^T + \mathbf{E}_{(d)},$$

Where $\mathbf{U}_{(0)} \in \mathbb{R}^{n \times r_0}$, $\mathbf{V}_{(d)} \in \mathbb{R}^{p_d \times r_0}$, $\mathbf{U}_{(d)} \in \mathbb{R}^{\mathrm{n} \times r_d}$, $\mathbf{A}_{(\mathrm{d})} \in \mathbb{R}^{p_d \times r_d}$

- We will extend this JIVE model to non-Euclidean case

Thank you for your attention !