창원대학교 통계학과
전임교원 신규채용 공개강의

Kipoong Kim

Department of Statistics, Seoul National University

January 10, 2024

Introduction

History: Kipoong Kim

- I was born in Busan and lived there for 30 years
- Education
  - 2010–2016 B.S. in Statistics, Pusan National Univ.
  - 2016–2017 M.S. in Statistics, Pusan National Univ.
  - 2019–2022 Ph.D. in Statistics, Pusan National Univ.
- Teaching Experience
  - Spring & Fall 2021 Part-Time Lecturer
    - ★ (a) Statistical Programming Language, (b) Biostatistics,
    - ★ (c) Introduction to Statistics, (d) Mathematical Statistics
- Academic Positions
  - 2022–Present PostDoc. in Statistics, Seoul National Univ.

Research Overview

- Main interest is to develop new statistical methodologies to better understand the data produced in various fields.
- My research areas include:
  - Variable selection in high-dimensional genomic data
  - Low-rank models for non-Euclidean data (e.g. compositional, spherical)
  - Multi-source data integration
- I am also interested in collaborating with researchers in other fields such as psychology, biology, plant genetics, and medicine.
- As a result, we have published a total of 16 papers in the last 5 years
  - Statistical methodology: 10 papers (SCIE=6)
  - Application: 6 papers (SCIE=5)

Research Overview

Variable selection

- Suppose that we observed p genetic variants (predictors) and q phenotypes (responses) from n individuals.

- Then, we can consider the following frameworks:
  - For $q = 1$, univariate linear regression

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

    where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^n$.
  - For $q > 1$, multivariate regression

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \mathbf{E},$$

    where $\mathbf{B} = \{\beta_{jk}\} \in \mathbb{R}^{p \times q}$ and $\mathbf{Y} \in \mathbb{R}^{n \times q}$.

Research Overview

Variable selection

- We aim to identify outcome-related variables (i.e., variable selection)

$$\{j : \beta_j \neq 0\} \text{ for } q = 1 \quad \text{or} \quad \{(j,k) : \beta_{jk} \neq 0\} \text{ for } q > 1$$

- To this end, many statistical methodologies have been developed over time, including the lasso and elastic-net.
- Here, we focused on the external information of genomic data to improve statistical power in variable selection.
  - Genetic network
  - Multi-response information

Research Overview

Incorporating external information in variable selection

- **Genetic network:** Kipoong Kim†, and Hokeun Sun (2019). "Incorporating Genetic networks into case-control association studies with high-dimensional DNA methylation data". *BMC Bioinformatics*, 20, 510.
- **Multi-response information:** Kipoong Kim†, Taehwan Jun, Bokeun Ha, Shuang Wang and Hokeun Sun (2023). "New statistical selection method for pleiotropic variants associated with both quantitative and qualitative traits," *BMC Bioinformatics*, 24, 381.
- For variable selection, an appropriate threshold π_{thr} is required

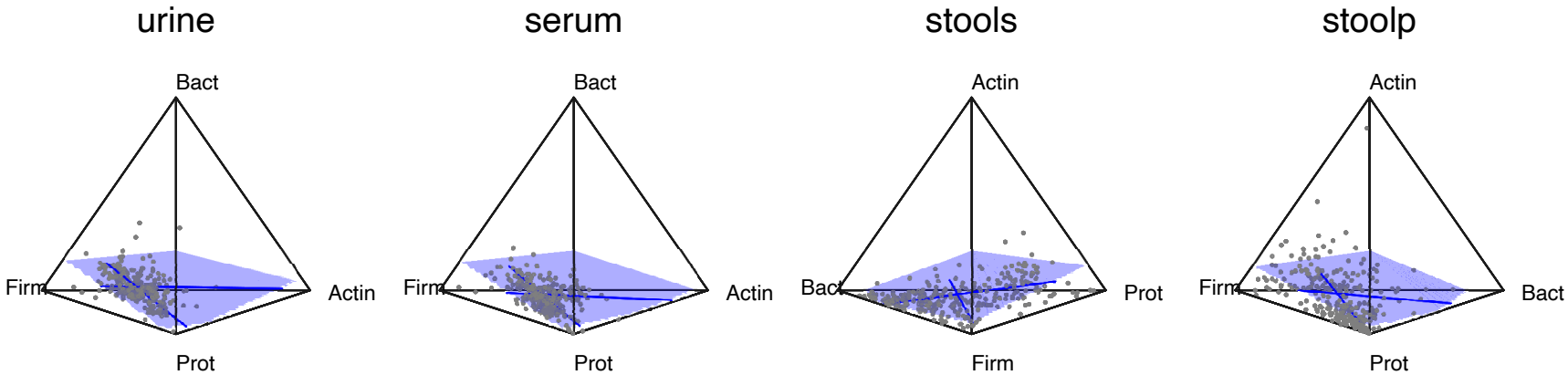
$$\hat{\mathcal{A}} = \{j : \Pi_j \geq \pi_{\text{thr}}\}.$$

- **Error control:** Kipoong Kim†, Jajoon Koo, and Hokeun Sun (2020). "An Empirical threshold of selection probability for analysis of high-dimensional correlated data," *Journal of Statistical Computation and Simulation*, 90(9), 1606–1617.

Research Overview

Low-rank models for non-Euclidean data

- **Compositional data:** Kipoong Kim†, Jaesung Park† and Sungkyu Jung (2024). "Principal Component Analysis for zero-inflated compositional data," submitted to *Computational Statistics and Data Analysis*.
- We aim to estimate a principal compositional subspace that best approximates the given compositional data.



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

- We also investigated theoretical properties of the principal compositional subspace including its existence and consistency.

Research Overview

Multi-source data integration

- Multi-source data can be thought of as a set of datasets produced from different multiple sources:

  $$\{ \text{Gene expression, DNA methylation, RNA sequencing, } \dots \}$$

  - Each can be thought of as a design matrix, so we have

    $$\mathbf{X} = [\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \mathbf{X}_{(3)}, \dots].$$

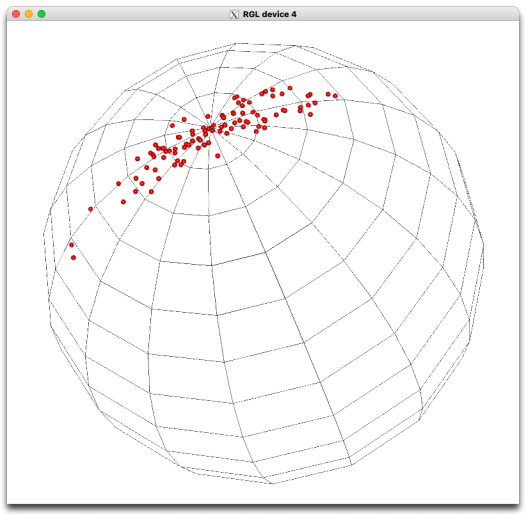
- The goal is to estimate the structural relationship between multi-source data $\mathbf{X}$ and multiple responses $\mathbf{Y}$:

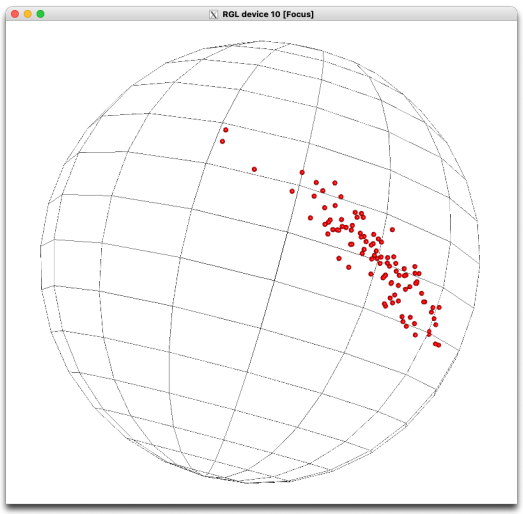
- **Multi-source data:** Kipoong Kim† and Sungkyu Jung (2024). "Integrative sparse reduced-rank regression via orthogonal rotation for analysis of high-dimensional multi-source data," *Statistics and Computing*, 34, 2.

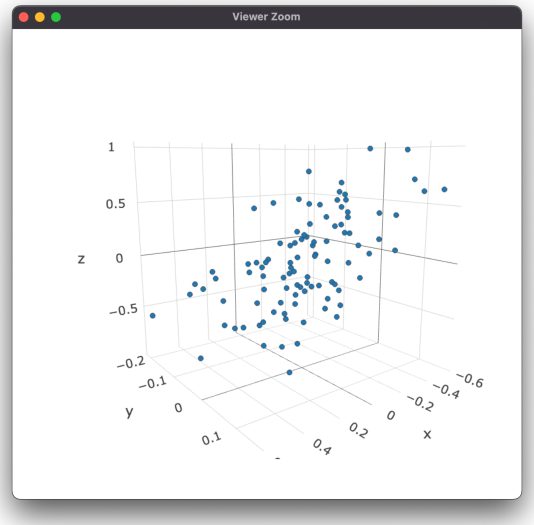
Research Overview

Non-Euclidean data integration

- Spherical data refers to data that is distributed on the surface of a hyper-sphere.
- Structural decomposition for multiple data sets (ongoing):







Key Statistical Methodology Papers

1. Variable selection: Incorporating genetic network

- Genomic data with a group structure: e.g. SNP, DNA-methylation.

$$\mathbf{X} = (\underbrace{\mathbf{X}_1, \ldots, \mathbf{X}_{p_1}}_{\text{1st gene}} \mid \underbrace{\mathbf{X}_{p_1+1}, \ldots, \mathbf{X}_{p_2}}_{\text{2nd gene}} \mid \cdots \mid \underbrace{\mathbf{X}_{p_{m-1}+1}, \ldots, \mathbf{X}_{p_m}}_{m\text{-th gene}})$$

  - Gene-level dimension reduction (2019)[1]:

$$(\mathbf{X}_1, \ldots, \mathbf{X}_{p_1} \mid \mathbf{X}_{p_1+1}, \ldots, \mathbf{X}_{p_2} \mid \cdots \mid \mathbf{X}_{p_{m-1}+1}, \ldots, \mathbf{X}_{p_m}) \\ \downarrow \qquad\qquad \downarrow \qquad\qquad \downarrow \\ \tilde{\mathbf{X}}_1 \qquad\quad \tilde{\mathbf{X}}_2 \qquad\quad \tilde{\mathbf{X}}_m$$

  - Group-wise penalties (2023+):

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \; -\ell(\boldsymbol{\beta}) + \lambda_1 \sum_{k=1}^{m} \|\boldsymbol{\beta}_k\|_2 + \frac{\lambda_2}{2} \sum_{u \sim v} \left(\frac{\|\boldsymbol{\beta}_u\|_2}{\sqrt{d_u}} - \frac{\|\boldsymbol{\beta}_v\|_2}{\sqrt{d_v}} \right)^2.$$

[1] K. Kim, H. Sun, *BMC bioinformatics* **20**, 1–15 (2019).

1. Variable selection: Multiple mixed-type responses

- In real application, many genetic studies include response variables with various data types such as continuous, ordinal and categorical:
  - e.g., cowpea dataset from Rural Development Administration:

Categories	**Phenotypes**		
Seed	Seed coat color	Seed coat pattern	Seed shape
	Seed coat gloss	100-seed weight	
Flowering	Flower color	Days for flowering	Days for ripening
Pod	Pod color	Pod curve	Seed density
	Shattering	Pod length	Seed numbers

■: categorical, ■: continuous

- Kim et al. (2023)[2] proposed a statistical method based on penalized regression to identify genetic variants associated with multiple mixed-type responses belonging to a specific category.

[2] K. Kim *et al.*, *BMC bioinformatics* **24**, 381 (2023).

1. Variable selection: Multiple mixed-type responses

- Consider a penalized regression with a sparsity-inducing penalty on the k-th response, $k = 1, \ldots, q$:

$$\hat{\boldsymbol{\beta}}_k^{\lambda_k}(\mathbf{X}, \mathbf{Y}_k) = \underset{\boldsymbol{\beta}_k \in \mathbb{R}^p}{\arg\min} \; -\ell_k(\boldsymbol{\beta}_k; \mathbf{X}, \mathbf{Y}_k) + P_{\lambda_k}(\boldsymbol{\beta}_k),$$

  where $\ell_k(\cdot)$ is the log-likelihood function corresponding to the k-th response.

- We define the number of associated responses with the j-th predictor as

$$\hat{\pi}_j(\Lambda; \mathbf{X}, \mathbf{Y}) = \sum_{k=1}^{q} \mathbb{I}\left(\hat{\beta}_{jk}^{\lambda_k}(\mathbf{X}, \mathbf{Y}_k) \neq 0\right),$$

  where $\Lambda = (\lambda_1, \ldots, \lambda_q)$ is a set of penalty parameters.

- We propose the selection score defined by its bootstrap expectation:

$$\hat{\Pi}_j(\Lambda; \mathbf{X}, \mathbf{Y}) = \mathbb{E}^*[\hat{\pi}_j(\Lambda; \mathbf{X}, \mathbf{Y})].$$

1. Variable selection: Multiple mixed-type responses

- Manhattan plot

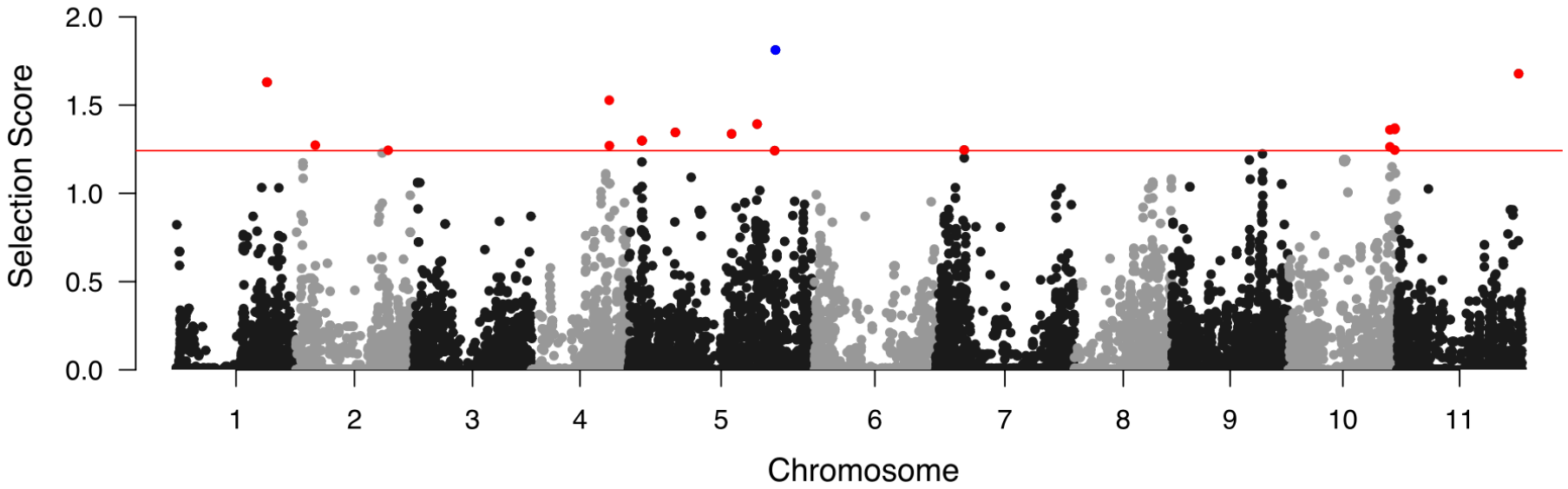


Figure: Top 20 ranked predictors are colored by red or blue.

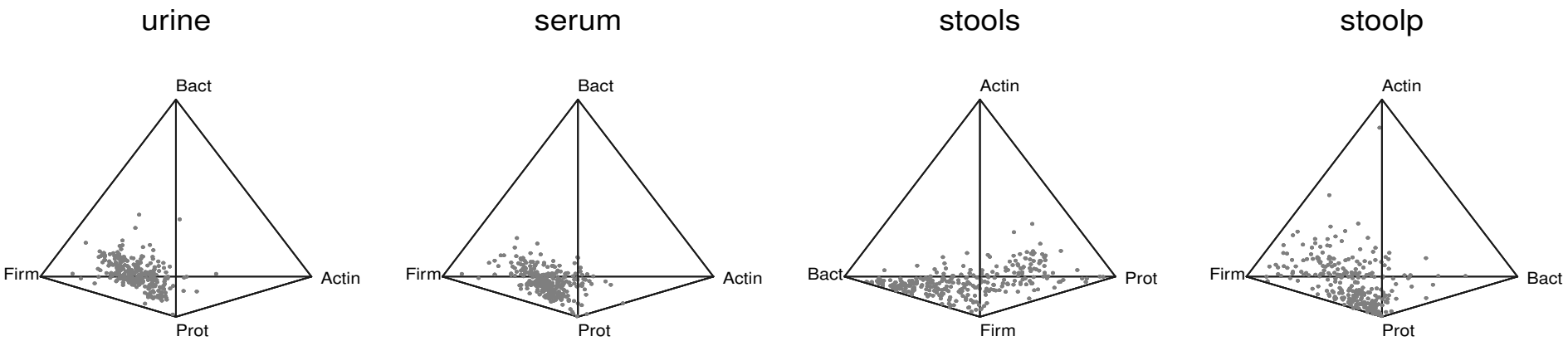
2. Non-Euclidean data analysis

Compositional data

- 16s rRNA microbiome sequencing data
  - (1) Compositionality, (2) High dimensionality, (3) Zero inflation
- The sample space is limited to a compositional space (simplex):

$$\mathbb{C}^p = \{(x_1, \ldots, x_p) \in \mathbb{R}_+^p : x_1 \geq 0, \ldots, x_p \geq 0; \sum_{j=1}^{p} x_j = 1\}.$$

- Real data example with $p = 4$:



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

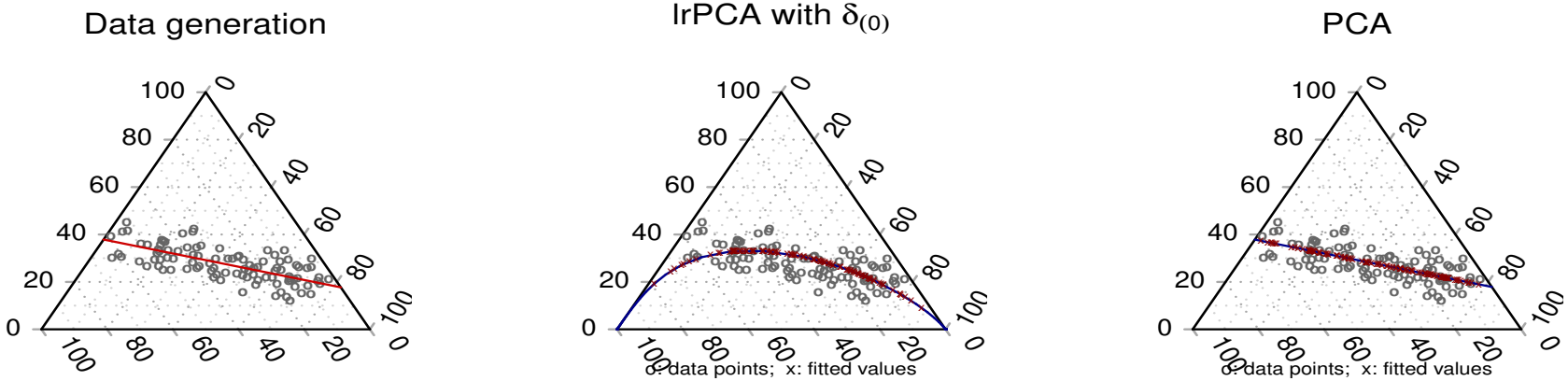
- In this work, we aim to find a new dimension reduction method for zero-inflated compositional data.

2. Non-Euclidean data analysis

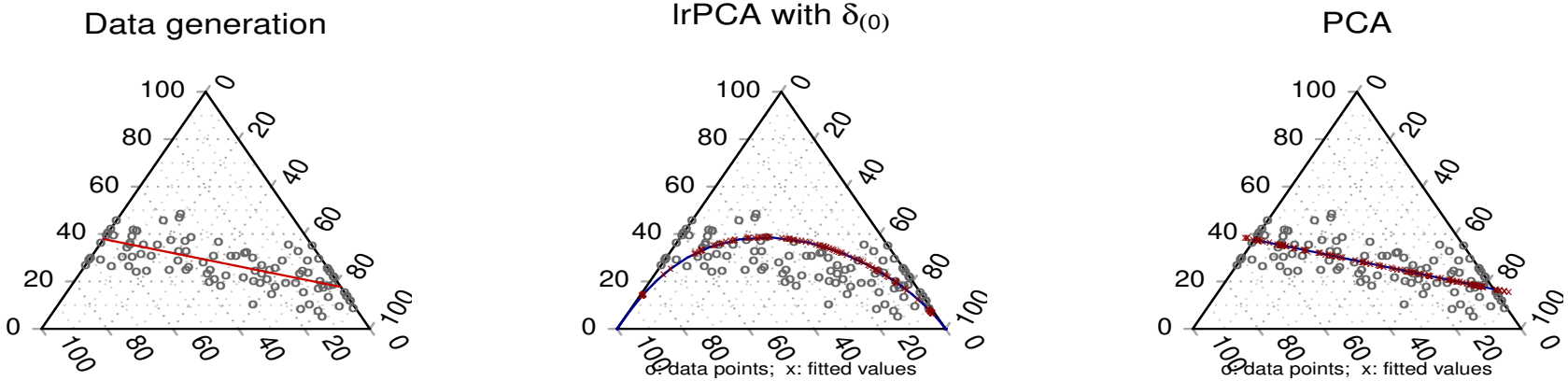
Motivating example: Limitation of log-ratio PCA

- Log-ratio PCA (lrPCA) is to apply the classical PCA after the log-ratio transformation. For dealing with zeros, some zero replacement strategies are applied. However, the zero inflation may result in the distortion.

Without zeros



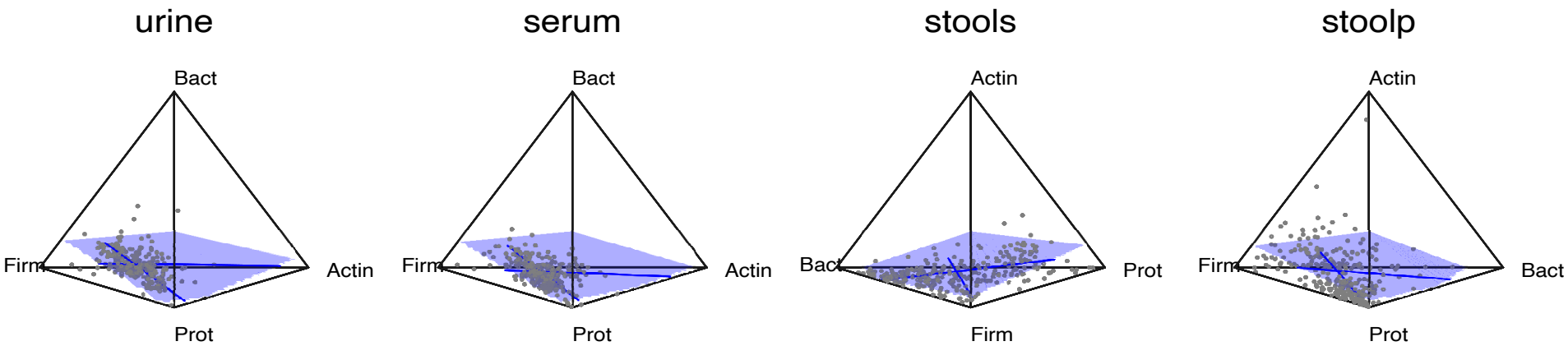
With 10% zeros



2. Non-Euclidean data analysis

Principal compositional subspace

- Kim et al (2023+) aim to find a principal compositional subspace that best approximates the data, by minimizing the Euclidean projection error.



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

- Compositional subspace spanned by $\mathbf{V}_1, \ldots, \mathbf{V}_k$ at $\boldsymbol{\mu}$

$$\mathbb{CS}_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_k} := \mathbb{C}^p \cap \{\boldsymbol{\mu} + c_1\mathbf{V}_1, \ldots, c_k\mathbf{V}_k : c_1, \ldots, c_k \in \mathbb{R}\}$$

2. Non-Euclidean data analysis

Main goal

- We denote the transpose of the i-th row vector by $\mathbf{a}_i$ and the k-th column vector by $\mathbf{A}_k$ for a matrix $\mathbf{A}$.
- We want to solve the following global compositional PCA (global CPCA) problem:

$$\underset{\mathbf{U}\in\mathbb{R}^{n\times r},\ \mathbf{V}\in\mathbb{R}^{p\times r}}{\arg\min} \|\mathbf{X}-\mathbf{1}\boldsymbol{\mu}^T-\mathbf{U}\mathbf{V}^T\|_F^2,$$

  subject to
  - $\mathbf{U}$ and $\mathbf{V}$ have `orthogonal` and `orthonormal` columns
  - $\boldsymbol{\mu}+\mathbf{V}\mathbf{u}_i^T\in\mathbb{C}^p\ \forall i$ for $\boldsymbol{\mu}\in\mathbb{C}^p$
- A sequence of compositional subspaces by global CPCA is not nested: e.g.

$$\mathbb{CS}_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_{k-1}} \not\subseteq \mathbb{CS}_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_{k-1},\mathbf{V}_k}.$$

2. Non-Euclidean data analysis

Estimation techniques

- (Direction) Construction of a nested sequence of principal compositional subspaces:

$$\{\boldsymbol{\mu}\} \subset \mathbb{CS}^p_{\boldsymbol{\mu};\mathbf{V}_1} \subset \mathbb{CS}^p_{\boldsymbol{\mu};\mathbf{V}_1,\mathbf{V}_2} \subset \cdots \subset \mathbb{CS}^p_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_k} \subset \cdots$$

- (Score) Projection onto principal compositional subspace:

$$\mathbf{u}_i = \Pi_{\mathbb{CS}_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_k}}(\mathbf{x}_i;\boldsymbol{\mu})$$
$$\text{such that } \boldsymbol{\mu} + u_{i1}\mathbf{V}_1 + \cdots + u_{ik}\mathbf{V}_k \in \mathbb{CS}^p_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_k}$$

for $\mathbf{x}_i \in \mathbb{C}^p$ and $\mathbf{V}_1, \ldots, \mathbf{V}_k \perp \mathbf{1}_p$.

2. Non-Euclidean data analysis

Sequential estimation procedure

- With the appropriate constraints,
  - Rank-1 case:

$$(\hat{\mathbf{U}}, \hat{\mathbf{V}}_1) = \underset{\mathbf{U}_1, \mathbf{V}_1}{\arg\min} \, \|\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{U}_1\mathbf{V}_1^T\|_F^2,$$

  - Rank-2 case:

$$(\hat{\mathbf{U}}, \hat{\mathbf{V}}_2) = \underset{(\mathbf{U}_1, \mathbf{U}_2), \mathbf{V}_2}{\arg\min} \, \|\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{U}_1\hat{\mathbf{V}}_1^T - \mathbf{U}_2\mathbf{V}_2^T\|_F^2,$$

  - $\vdots$
  - Rank-k case:

$$(\hat{\mathbf{U}}, \hat{\mathbf{V}}_k) = \underset{(\mathbf{U}_1, \ldots, \mathbf{U}_k), \mathbf{V}_k}{\arg\min} \, \|\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{U}_1\hat{\mathbf{V}}_1^T - \cdots - \mathbf{U}_k\mathbf{V}_k^T\|_F^2,$$

- We use the alternating algorithm to estimate $\mathbf{U}$ and $\mathbf{V}_k$.

2. Non-Euclidean data analysis

Two types of projection approaches

- One-dimensional projection

$$\Pi^{one}_{\mathbf{v}}(\mathbf{x}_i;\boldsymbol{\mu}) = \underset{u_i\in\mathbb{R}}{\arg\min}\ \|\mathbf{x}_i - \boldsymbol{\mu} - u_i\mathbf{v}\|_2^2 \quad \text{subject to } \boldsymbol{\mu} + u_i\mathbf{v} \in \mathbb{CS}_{\boldsymbol{\mu};\mathbf{v}}$$

- Multi-dimensional projection

$$\Pi^{mult}_{\mathbf{V}_1,\ldots,\mathbf{V}_k}(\mathbf{x}_i;\boldsymbol{\mu}) = \underset{u_{i1},\ldots,u_{ik}\in\mathbb{R}}{\arg\min}\ \|\mathbf{x}_i - \boldsymbol{\mu} - u_{i1}\mathbf{V}_1 - \cdots - u_{ik}\mathbf{V}_k\|_2^2$$

$$\text{subject to } \boldsymbol{\mu} + u_{i1}\mathbf{V}_1 + \cdots + u_{ik}\mathbf{V}_k \in \mathbb{CS}_{\boldsymbol{\mu};\mathbf{V}_1,\ldots,\mathbf{V}_k}$$

2. Non-Euclidean data analysis

Three proposed methods

- Compositional PCA (CPCA):

Given $\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}$,

$$\underset{\mathbf{U}_1,\ldots,\mathbf{U}_k,\mathbf{V}_k}{\arg\min} \|\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}^T - \mathbf{U}_1\hat{\mathbf{V}}_1^T - \cdots - \mathbf{U}_{k-1}\hat{\mathbf{V}}_{k-1}^T - \mathbf{U}_k\mathbf{V}_k^T\|_F^2,$$

subject to

$$\hat{\boldsymbol{\mu}} + u_{i1}\hat{\mathbf{V}}_1 + \cdots + u_{i\,k-1}\hat{\mathbf{V}}_{k-1} + u_{ik}\mathbf{V}_k \in \mathbb{CS}_{\hat{\boldsymbol{\mu}};\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_{k-1},\mathbf{V}_k} \quad \forall i$$

$$\mathbf{V}_k \perp \mathbf{1}_p, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \ \|\mathbf{V}_k\|_2 = 1$$

2. Non-Euclidean data analysis

Three proposed methods

- Approximated CPCA (aCPCA):

Given $\hat{\boldsymbol{\mu}}, (\hat{\mathbf{U}}_1, \hat{\mathbf{V}}_1), \ldots, (\hat{\mathbf{U}}_{k-1}, \hat{\mathbf{V}}_{k-1})$,

$$\underset{\mathbf{U}_k, \mathbf{V}_k}{\arg\min} \; \|\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}^T - \hat{\mathbf{U}}_1\hat{\mathbf{V}}_1^T - \cdots - \hat{\mathbf{U}}_{k-1}\hat{\mathbf{V}}_{k-1}^T - \mathbf{U}_k\mathbf{V}_k^T\|_F^2,$$

subject to

$$\hat{\boldsymbol{\mu}} + \hat{u}_{i1}\hat{\mathbf{V}}_1 + \cdots + \hat{u}_{i\,k-1}\hat{\mathbf{V}}_{k-1} + u_{ik}\mathbf{V}_k \in \mathbb{CS}_{\hat{\boldsymbol{\mu}};\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_{k-1},\mathbf{V}_k} \quad \forall i$$

$$\mathbf{V}_k \perp \mathbf{1}_p, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \; \|\mathbf{V}_k\|_2 = 1$$

2. Non-Euclidean data analysis

Three proposed methods

- Compositional Reconstructed PCA (crPCA):

  Given $\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}_1^{PC}, \ldots, \hat{\mathbf{V}}_r^{PC}$,

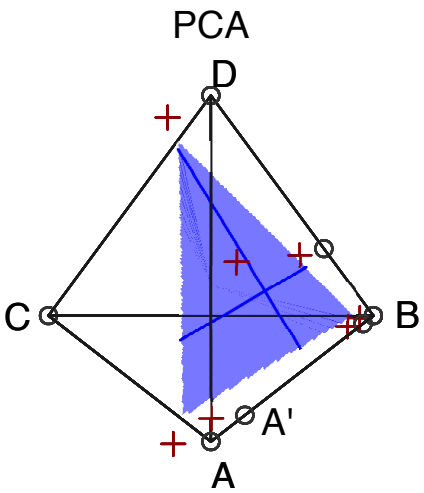
$$\underset{\mathbf{U}_1,\ldots,\mathbf{U}_r}{\arg\min} \|\mathbf{X} - \mathbf{1}\hat{\boldsymbol{\mu}}^T - \mathbf{U}_1\hat{\mathbf{V}}_1^{PC^T} - \cdots - \mathbf{U}_r\hat{\mathbf{V}}_r^{PC^T}\|_F^2,$$

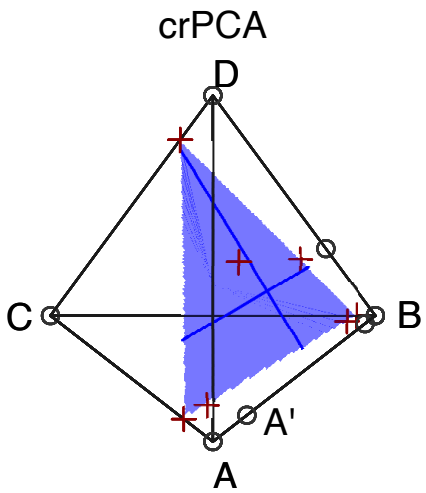
$$\text{subject to}$$

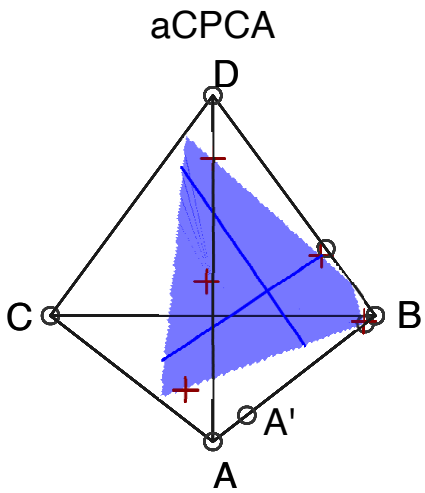
$$\hat{\boldsymbol{\mu}} + u_{i1}\hat{\mathbf{V}}_1^{PC} + \cdots + u_{ir}\hat{\mathbf{V}}_r^{PC} \in \mathbb{CS}_{\hat{\boldsymbol{\mu}};\hat{\mathbf{V}}_1^{PC},\ldots,\hat{\mathbf{V}}_r^{PC}} \quad \forall i$$

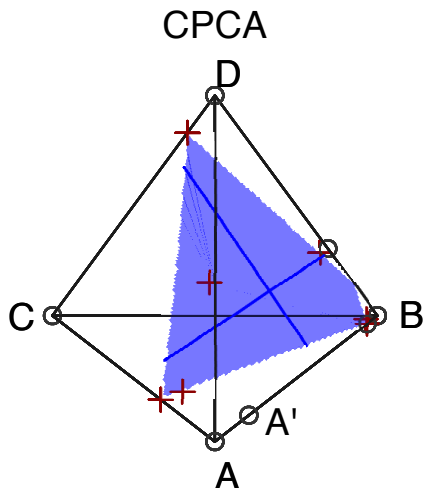
2. Non-Euclidean data analysis

An illustrative comparison





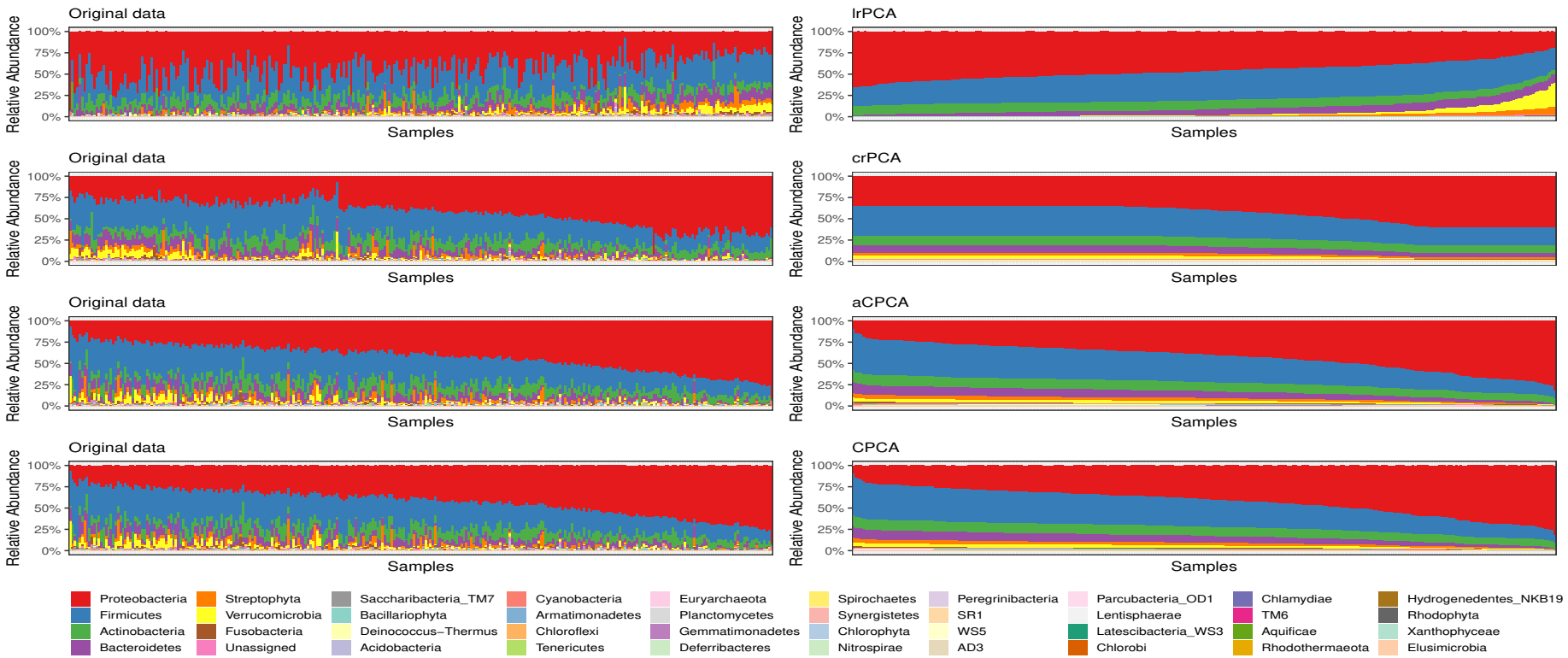




2. Non-Euclidean data analysis

Rank-1 approximation

- The existing log-ratio PCA can have distortion in its reconstruction.



2. Non-Euclidean data analysis

Theoretical properties

Theorem (Existence)

The principal compositional subspaces and principal compositional directions ($\mathbb{CS}_{\boldsymbol{\mu};\mathbf{v}_1,\ldots,\mathbf{v}_k}$, V_k, $\mathbb{CS}_{\hat{\boldsymbol{\mu}};\hat{\mathbf{v}}_1,\ldots,\hat{\mathbf{v}}_k}$, and $\hat{V}_k$) exist for all $k = 1, \ldots, p$.

Theorem (Consistency)

Assume $\mathbb{CS}_{\boldsymbol{\mu};\mathbf{v}_1,\ldots,\mathbf{v}_k}$ uniquely exists for all $k = 1, \ldots, p$. Then, the followings hold almost surely.

(a) $\lim_{n\to 0} \; h\left(\mathbb{CS}_{\hat{\boldsymbol{\mu}};\hat{\mathbf{v}}_1,\ldots,\hat{\mathbf{v}}_k}, \mathbb{CS}_{\boldsymbol{\mu};\mathbf{v}_1,\ldots,\mathbf{v}_k}\right) = 0.$

(b) $\lim_{n\to 0} \; \|\hat{V}_k(\mathcal{X}_n) - V_k\| = 0.$

where h is the Hausdorff distance defined by $h(A, B) := \max\{\sup_{a\in A} \inf_{b\in B} \|a - b\|_2, \sup_{b\in B} \inf_{a\in A} \|a - b\|_2\}$ for nonempty closed subsets A and B of $\mathbb{C}^p$.

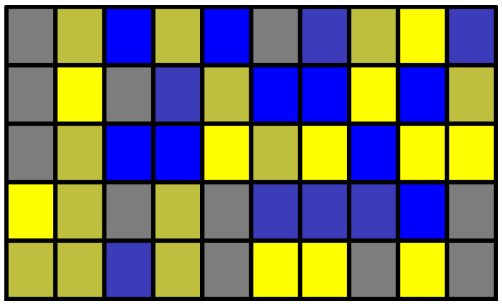
3. Multi-source data integration

- Multi-source data is defined as a set of datasets produced from different multiple sources:
  $\{$ Gene expression, DNA methylation, RNA sequencing, $\ldots\}$

  - Each can be thought of as a design matrix, so we have

$$\mathbf{X} = [\mathbf{X}_{(1)}, \mathbf{X}_{(2)}, \mathbf{X}_{(3)}, \ldots].$$

- Our goal is to estimate the structural relationship between multi-source data and drug responses $(q > 50)$.



3. Multi-source data integration

Reduced-rank regression (RRR)

- Multivariate regression with low-rank assumption

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\mathbf{C}_{p\times q} + \mathbf{E}_{n\times q}$$

  where $\text{rank}(\mathbf{C}) = r$ and $r \leq \min\{n, p, q\}$.

- This leads to the reduced-rank regression model[3]:

$$\mathbf{Y}_{n\times q} = \mathbf{X}_{n\times p}\mathbf{B}_{p\times r}\mathbf{A}^T_{q\times r} + \mathbf{E}_{n\times q}.$$

  - This can dramatically reduce the number of parameters to be estimated $(pq \to (p+q)r)$.
  - Thus, the estimates are more precise

[3] A. J. Izenman, *Journal of Multivariate Analysis* **5**, 248–264 (1975).

3. Multi-source data integration

Structural Learning in RRR

- Goal is to identify the structured association between multiple responses and multi-source datasets

$$\mathbf{Y} = \left[\mathbf{X}_{(1)},\ \mathbf{X}_{(2)},\ \mathbf{X}_{(3)}\right] \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} & \mathbf{0} \\ \mathbf{b}_{21} & \mathbf{b}_{22} & \mathbf{0} \\ \mathbf{b}_{31} & \mathbf{0} & \mathbf{b}_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ 0 & 0 & 0 \end{bmatrix}^{\top} + \mathbf{E},$$

- Structural relationship between $\mathbf{X}$ to $\mathbf{Y}$ through $\mathbf{XB}$:
  - The first column is *joint* structure: $\mathbf{X}_{(1)}\mathbf{b}_{11} + \mathbf{X}_{(2)}\mathbf{b}_{21} + \mathbf{X}_{(3)}\mathbf{b}_{31}$
  - The second column is *partially-joint* structure: $\mathbf{X}_{(1)}\mathbf{b}_{12} + \mathbf{X}_{(2)}\mathbf{b}_{22}$
  - The third column is *individual* structure: $\mathbf{X}_{(3)}\mathbf{b}_{33}$

3. Multi-source data integration

Identifiability Problem

- To this end, we can consider the following penalized optimization problem:

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\mathbf{A}^T\|_F^2 + \lambda\sum_{i=1}^{d}\sum_{k=1}^{r}\sqrt{p_i}\|\mathbf{b}_{ik}\|_2 + \nu^\star\sum_{h=1}^{q}\|\mathbf{a}_h\|_2,$$

  where $\lambda \geq 0$ controls the structured sparsity of $\mathbf{B}$ and $\nu^\star$ controls the row-wise sparsity of $\mathbf{A}$.

- However, the parameters are not unique up to an orthogonal matrix; For example, $\mathbf{B}\mathbf{A}^T = \mathbf{B}\mathbf{Q}\mathbf{Q}^T\mathbf{A}^T$ for $\mathbf{Q} \in \mathbb{R}^{r\times r}$ such that $\mathbf{Q}\mathbf{Q}^T = \mathbf{I}_r$.

3. Multi-source data integration

Constraint for the rotational indeterminacy

- Quartimax criterion: $\mathcal{F}(\mathbf{A}) = \sum_{j=1}^{q} \sum_{k=1}^{r} A_{jk}^4$ for a generic matrix $\mathbf{A}$.

Definition (Quartimax-simple structure)

Given $\mathbf{A} \in \mathbb{R}^{q \times r}$, the rotated matrix $\mathbf{AQ}$ is said to have a *quartimax-simple structure* if $\mathbf{Q}$ maximizes the quartimax criterion $\mathcal{F}(\mathbf{AQ})$ over all $\mathbf{Q} \in \mathcal{O}(r)$. Also, a set of semi-orthogonal matrices with simple structure is defined as

$$\mathcal{O}_{\mathcal{S}}(q,r) = \left\{ \mathbf{A}\hat{\mathbf{Q}} : \hat{\mathbf{Q}} = \arg \max_{\mathbf{Q} \in \mathcal{O}(r)} \mathcal{F}(\mathbf{AQ}),\ \mathbf{A} \in \mathcal{O}(q,r) \right\}.$$

where $\mathcal{O}(r) = \left\{ \mathbf{Q} \in \mathbb{R}^{r \times r} : \mathbf{Q}^T\mathbf{Q} = \mathbf{QQ}^T = \mathbf{I}_r \right\}$ and $\mathcal{O}(q,r) = \left\{ \mathbf{A} \in \mathbb{R}^{q \times r} : \mathbf{A}^T\mathbf{A} = \mathbf{I}_r \right\}$.

3. Multi-source data integration

Constrained reduced-rank regression model

- We propose the constrained reduced-rank regression model with the condition $\mathbf{A} \in \mathcal{O}_S(q, r)$:

$$\mathbf{Y} = \mathbf{XBA}^T + \mathbf{E}, \quad \mathbf{A} \in \mathcal{O}_S(q, r), \tag{1}$$

  where $\mathbf{E} = (\mathbf{e}_1, \ldots, \mathbf{e}_n)^T$ with $\mathbf{e}_l \sim \mathcal{N}_q(\mathbf{0}, \sigma^2\mathbf{I})$, $l = 1, \ldots, n$.

- The following proposition illustrates the identifiability of (1).

Proposition

In model (1), if $\mathbf{B}^T\mathbf{X}^T\mathbf{XB}$ has r distinct positive eigenvalues for the fixed design matrix $\mathbf{X}$, then the parameter set $(\mathbf{A}, \mathbf{XB}, \sigma^2)$ is identifiable up to simultaneous signed permutations of the columns of $\mathbf{A}$ and $\mathbf{XB}$.

3. Multi-source data integration

Identifiability under RE condition

- We need the identifiability of $\mathbf{B}$, not $\mathbf{XB}$.
- Under the restricted eigenvalue condition[4] on $\mathbf{X}$, we have the following corollary.

Corollary

Assume that $\mathbf{B}$ has at most s nonzero elements. If the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$ satisfies the RE condition over $\mathbb{C}(2s, \xi)$ for some $\xi > 0$, the set of parameters $(\mathbf{A}, \mathbf{B}, \sigma^2)$ is identifiable up to simultaneous signed permutations of the columns.

[4] P. J. Bickel *et al.*, *The Annals of Statistics* **37**, 1705–1732 (2009).

3. Multi-source data integration

Integrative Sparse Reduced-Rank Regression (iSRRR)

- Kim and Jung (2024)[5] propose to estimate $\mathbf{A}$ and $\mathbf{B}$ for integrative sparse reduced-rank regression (iSRRR) by solving the constrained optimization problem

$$\min_{\mathbf{A},\mathbf{B}} \frac{1}{2n}\|\mathbf{Y}-\mathbf{X}\mathbf{B}\mathbf{A}^T\|_F^2 + \lambda\sum_{i=1}^{d}\sum_{k=1}^{r}\sqrt{p_i}\|\mathbf{b}_{ik}\|_2$$

$$\text{subject to } \mathbf{A}\in\mathcal{O}_S(q,r) \text{ and } \mathbf{A}\in\mathcal{T}(\nu),$$

where $\mathcal{T}(\nu) = \left\{\mathbf{A}\in\mathcal{O}(q,r) : \min_{j:\,\mathbf{a}_j\neq\mathbf{0}}\|\mathbf{a}_j\|_2 \geq \nu\right\}$.

[5] K. Kim, S. Jung, *Statistics and Computing* **34**, 2 (2024).

Future Research Plan

Future Research Plan

- Current research areas
  - Variable selection in high-dimensional genomic data
  - Low-rank model for non-Euclidean data
  - Multi-source data integration
- Future research topics of interest
  - **Large-scale cohort data analysis** (in the long term)

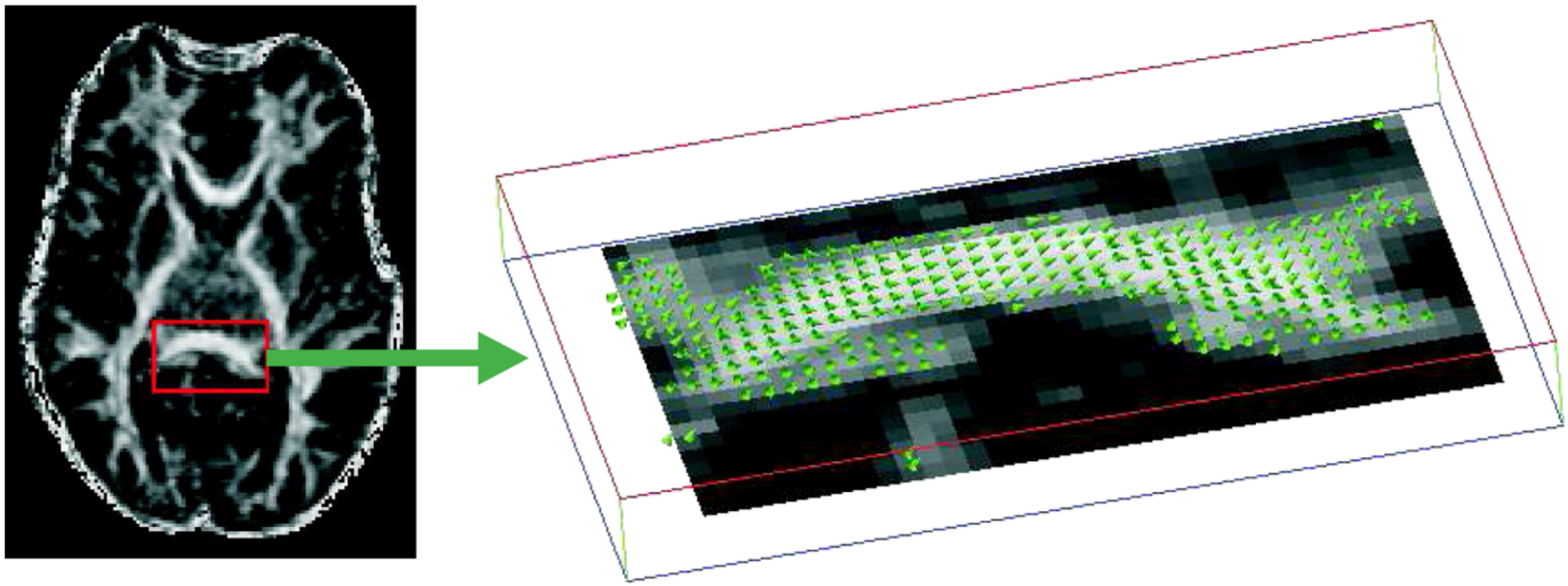
    - e.g. UK biobank data with 500,000 individuals

    - Transfer learning from large-scale data to smaller data of interest.
  - **Non-Euclidean data integration**

Non-Euclidean data integration

Spherical data

- Diffusion Tensor Imaging (DTI) data



[6]

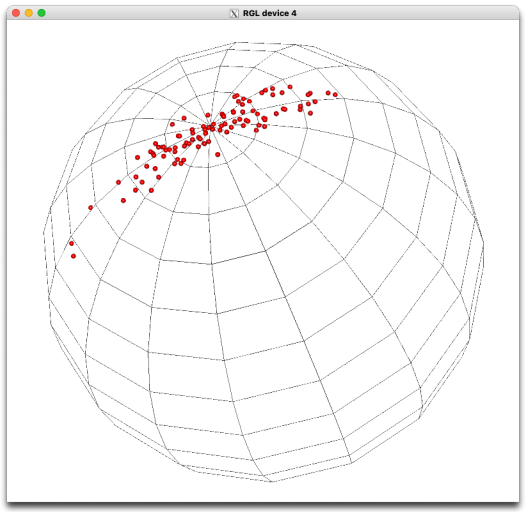
- DTI data collects the directions for the movements of water molecules at multiple brain regions.
- The direction can be represented by a vector with unit norm.

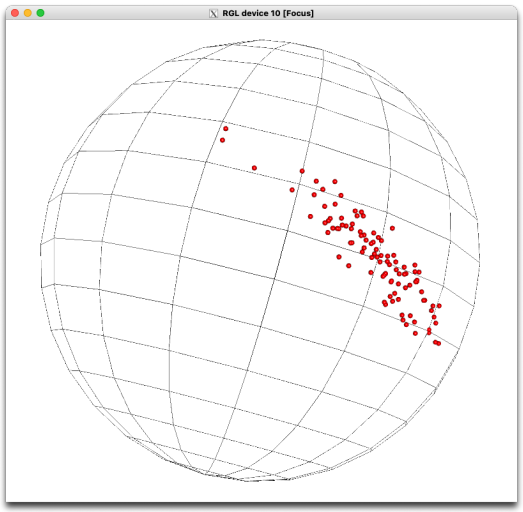
[6] V. Koltchinskii *et al.* (2007).

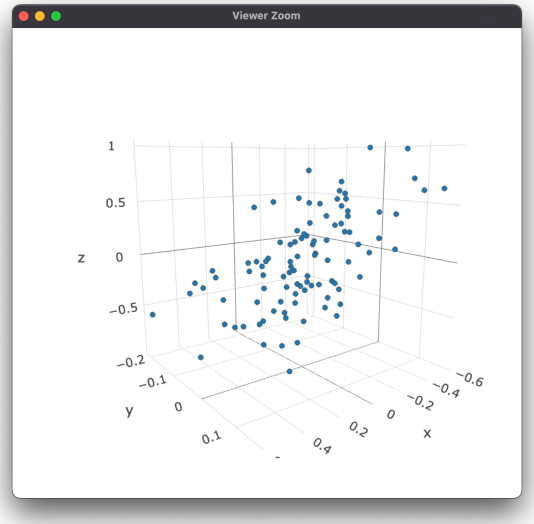
Non-Euclidean data integration

Spherical data

- Spherical data refers to data that is distributed on the surface of a hyper-sphere.
- Structural decomposition for multiple datasets (ongoing):







- We expect to be able to correlate functional differences in the brain with other biological data.

Non-Euclidean data integration

Joint and Individual Variation Explained (JIVE) model

- JIVE model for gaussian-valued data

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

- For the d-th source case:

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

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

Non-Euclidean data integration

JIVE for non-Euclidean case

- $X_{ij}^{(d)} \sim$ Exponential family with the natural parameter $\theta_{ij}^{(d)}$
- For spherical data, we use the von Mises-Fisher (vMF) distribution:

$$f_{\text{vmf}}(\mathbf{x}; \boldsymbol{\theta}, \kappa) = C_p(\kappa) \exp(\kappa \cdot \boldsymbol{\theta}^T \mathbf{x})$$

- $g(\mathbb{E}X_{ij}^{(d)}) = \theta_{ij}^{(d)} = \boldsymbol{\mu}_j^{(d)} + \mathbf{u}_i^{(0)^T} \mathbf{v}_j^{(d)} + \mathbf{u}_i^{(d)^T} \mathbf{a}_j^{(d)}$
- Matrix-version

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

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

- Estimate each of $\mathbf{U}_{(0)}$, $\mathbf{V}_{(d)}$, $\mathbf{U}_{(d)}$ and $\mathbf{A}_{(d)}$ with others fixed

Future Research Plan

- Collaboration with many researchers in various fields:
  - Dept. of Statistics, Pusan/Seoul National Univ.
  - Data Discovery Science Institute, Seoul National Univ.
  - Korea National Institute of Health (KNIH)
  - Center for Happiness Studies, Seoul National Univ.
  - School of Medicine, Pusan National Univ.

- By leveraging these collaborative relationships, we aim to successfully secure research grants and publish good results in the future.

Thank you for your attention