강릉원주대학교 데이터사이언스학과 전임교원 신규채용 공개강의

Kipoong Kim

Department of Statistics, Seoul National University

December 21, 2023

Academic Positions & Experiences

- 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)

1. 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 $\mathbf{Y} \in \mathbb{R}^{n \times q}$.

1. Variable selection

- We aim to identify disease-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 unique features of genomic data to improve statistical power in discovering disease-associated genetic variants.
  - Genetic network
  - Correlation structure among multiple phenotypes

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]:

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

  - 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:

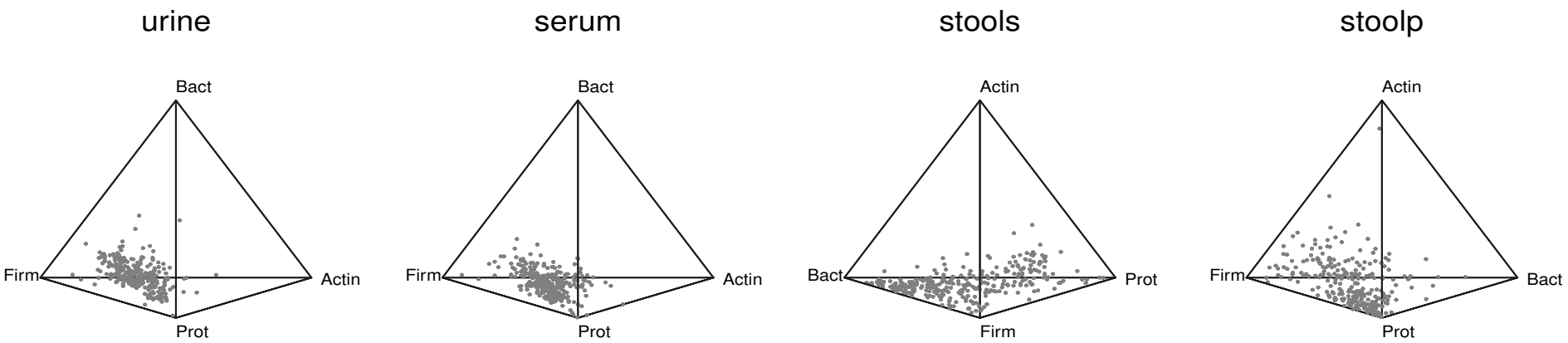
$$\hat{\Pi}_j(\Lambda; \mathbf{X}, \mathbf{Y}) = \mathbb{E}^*[\hat{\pi}_j(\Lambda; \mathbf{X}, \mathbf{Y})].$$

2. Non-Euclidean data analysis: Compositional data

- Sample space of compositional data is defined as

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

- Real data example with $p = 4$:

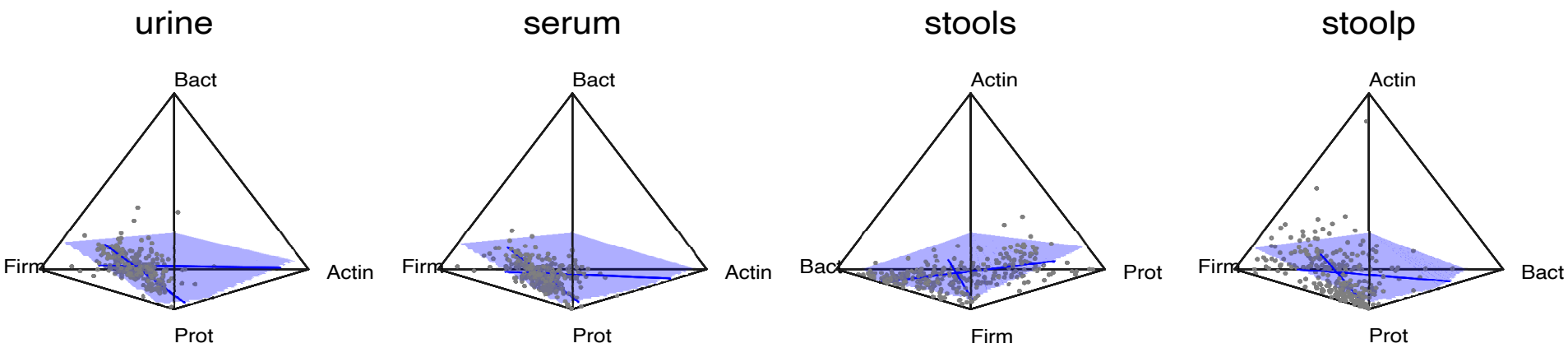


*Prot: Proteobacteria; Firm: Firmicutes; Actin: Actinobacteria; Bact: Bacteroidetes

- 16s rRNA microbiome sequencing data
  - (1) Compositionality, (2) High dimensionality, (3) Zero inflation

2. Non-Euclidean data analysis: Compositional data

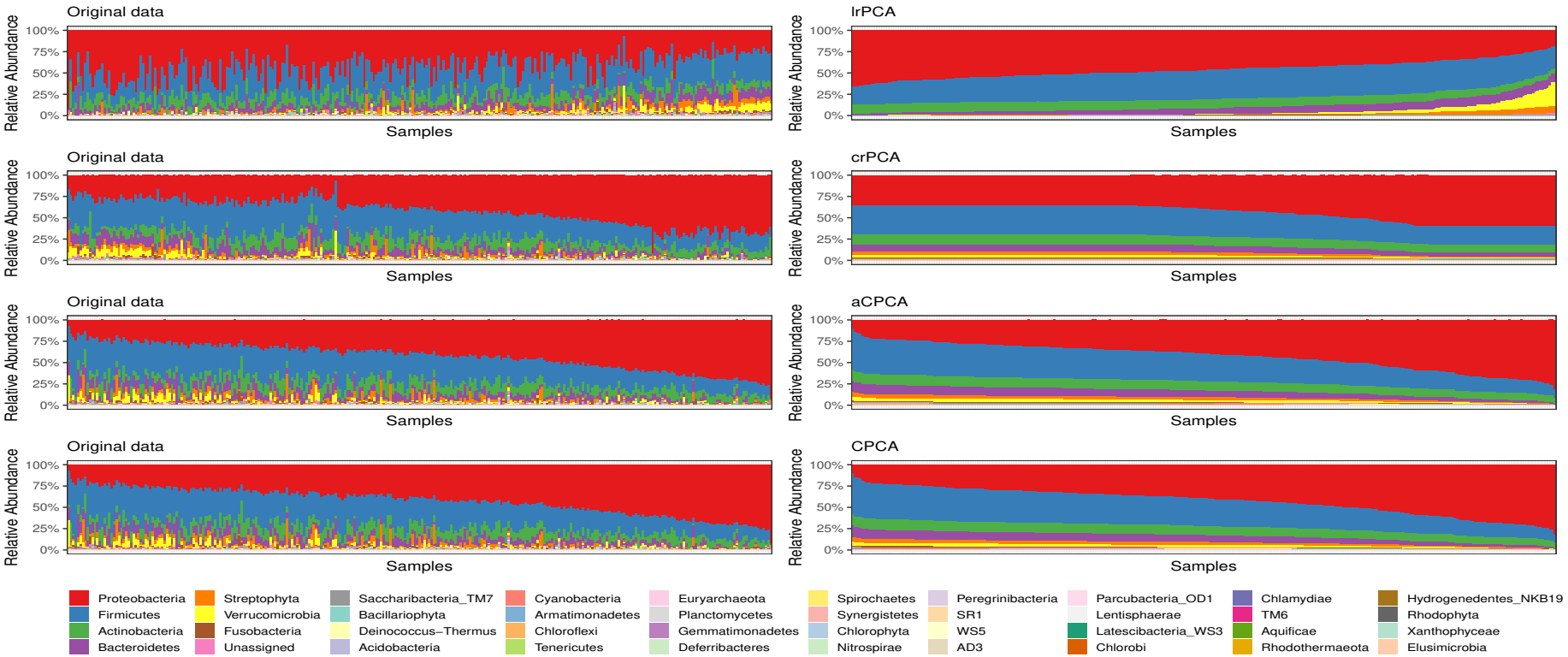
- Kim et al (2023+) proposed a new dimension reduction method for zero-inflated compositional data
- It aims to find a principal compositional subspace and the corresponding principal scores minimizing the Euclidean projection error.



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

2. Non-Euclidean data analysis: Compositional data

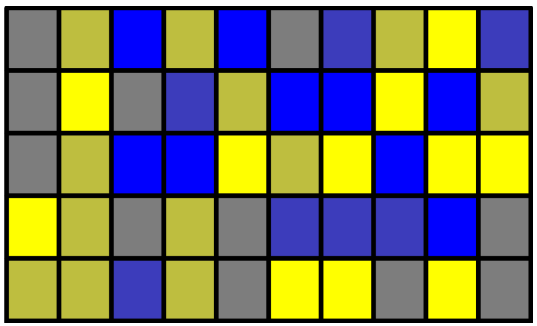
- The existing log-ratio PCA can have distortion in its reconstruction.



3. Multi-source data integration

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

- 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 $\text{rank}(\mathbf{C}) \leq r$ 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

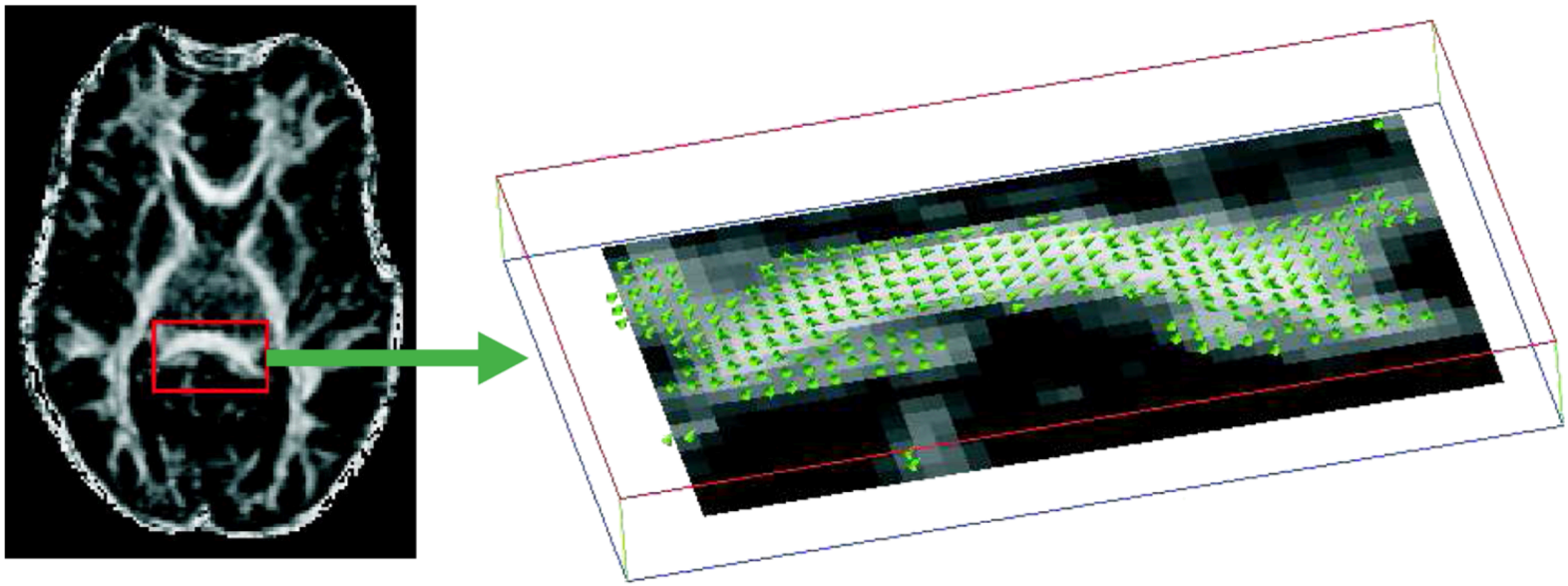
- Current research areas
  - Variable selection in high-dimensional genomic data
  - Low-rank model for non-Euclidean data
  - Multi-source data integration
- Future research topics of interest
  - **Large-scale cohort data analysis** (in the long term)

    - 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



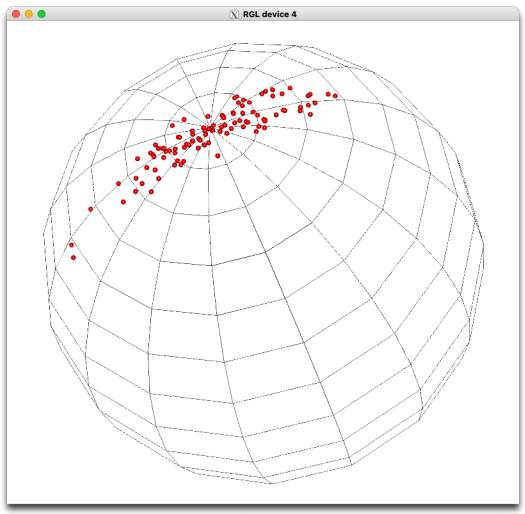
[6]

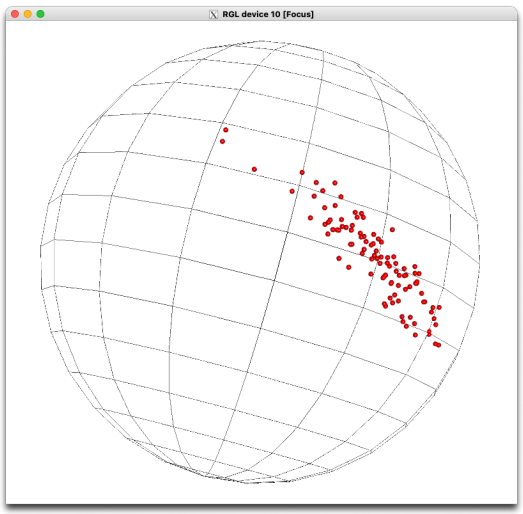
- DTI data collects the directions for the movements of water molecules at multiple brain regions.

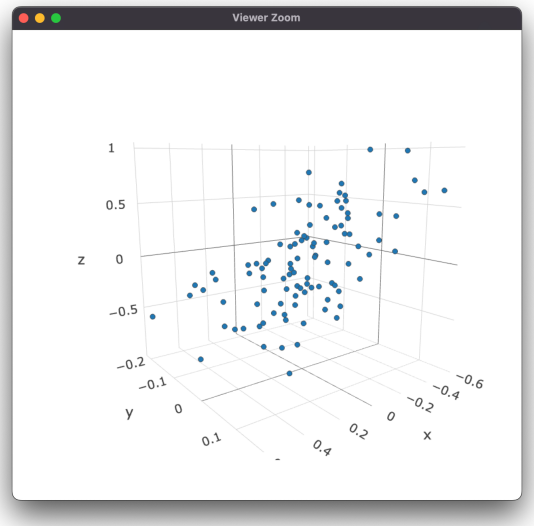
[6] V. Koltchinskii *et al.* (2007).

Non-Euclidean data integration: Spherical data

- Structural decomposition for multiple data sets (ongoing):







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