연세대학교 미래캠퍼스 데이터사이언스학부 교원 채용 공개강의

Kipoong Kim

Department of Statistics at Seoul National University

November 1, 2023

Academic Positions & Experience

- Education
  - 2010–2016 B.S. Dept. of Statistics, Pusan National Univ.
  - 2016–2017 M.S. Dept. of Statistics, Pusan National Univ.
  - 2019–2022 Ph.D. Dept. of Statistics, Pusan National Univ.
    (Advisor: Hokeun Sun)
- Researcher Experience
  - 2017 – 2018 Senior Researcher Korea National Institute of Health
- Teaching Experience
  - Spring & Fall 2021 Part-Time Lecturer
- Academic Positions
  - 2022–Present PostDoc. Dept. of Statistics, Seoul National Univ.
    (Supervisor: Sungkyu Jung)

Research Overview

- Statistical Methodology Papers (SCIE / Total = 6/12)
  - Bioinformatics
  - Low-rank model
- Application Papers (SCIE / Total = 5/6)
  - Biology
  - Plant genomics
  - Medical science

Key Statistical Methodology Papers

Overview

Bioinformatics

Genome-Wide Association Studies (GWAS)

- Suppose that we observed p genetic variants (predictors) and a single phenotype (response) from n individuals, and we denote $\mathbf{X} = (\mathbf{X}_1, \ldots, \mathbf{X}_p) \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^n$.
- Consider a general regression framework

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

  where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^T \in \mathbb{R}^p$.
- In this case, we aim to identify outcome-related variables (that is, variable selection)

$$\mathcal{A} = \{j : \beta_j \neq 0\}.$$

- For this purpose, many statistical methods have been proposed, including the lasso and elastic-net.
- However, we focused on unique features of genomic data to improve statistical power in identification of disease-related variants.

Bioinformatics

Incorporating external information

- **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.
- **Multiple responses:** 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.

Low-rank model

Multi-omics data integration

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

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

- **Multi-omics 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.

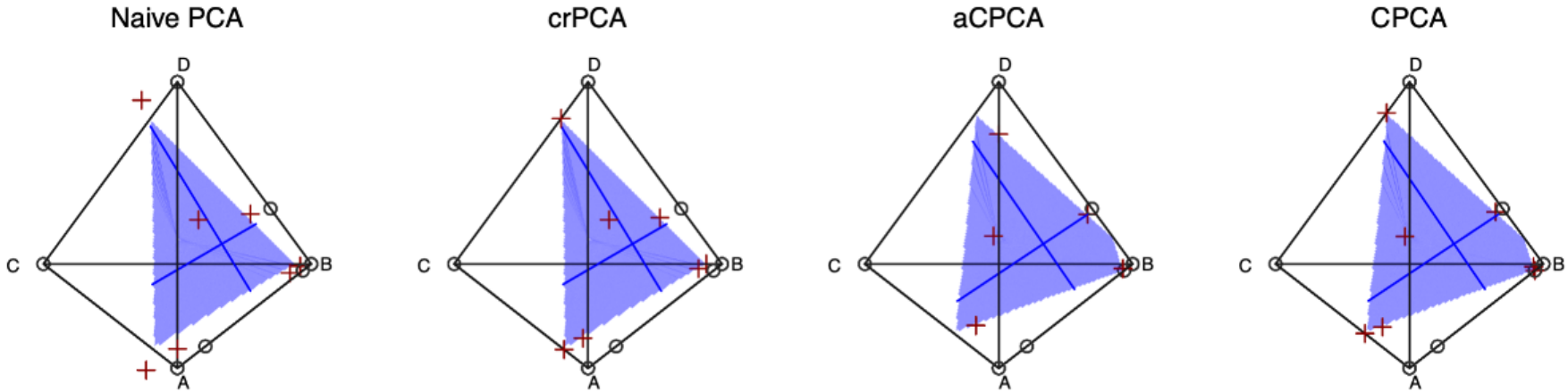
- Goal: Identification of structured association between multi-omics datasets $\mathbf{X}$ and multiple responses $\mathbf{Y}$:

$$[\mathbf{Y}_1, \ldots, \mathbf{Y}_q] = \left[\mathbf{X}_{(1)},\ \mathbf{X}_{(2)},\ \mathbf{X}_{(3)}\right] \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} & \mathbf{0} \\ \mathbf{b}_{21} & \mathbf{b}_{22} & \mathbf{0} \\ \mathbf{b}_{31} & \mathbf{0} & \mathbf{b}_{33} \end{bmatrix} \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \\ 0 & 0 & 0 \end{bmatrix}^T + \mathbf{E}.$$

Low-rank model

Beyond the human genome

- **Microbiome compositional data:** Kipoong Kim, Jaesung Park and Sungkyu Jung (2024). "Principal Component Analysis for zero-inflated compositional data," manuscript in progress.
- We aim to find a principal compositional subspace and the corresponding principal scores minimizing the Euclidean projection error.



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

Key Statistical Methodology Papers

1. Incorporating genetic network into group structured genomic data

Penalized regression with graph-constrained penalty

- Many studies have attempted to incorporate a genetic network in statistical analysis[1,2].
- Penalized regression with the graph-constrained penalty[3]:

$$\underset{\beta \in \mathbb{R}^p}{\arg\min} \; -\ell(\beta) + \lambda_1 \|\beta\|_1 + \frac{\lambda_2}{2} \beta^T L \beta,$$

where $\ell(\beta)$ is a log-likelihood function and $\lambda_1, \lambda_2 > 0$. Here, $\mathbf{L} = \{\ell_{uv}\}_{p \times p}$ is a normalized Laplacian matrix that represents a genetic network among genes.

- Genomic data with a group structure

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

---

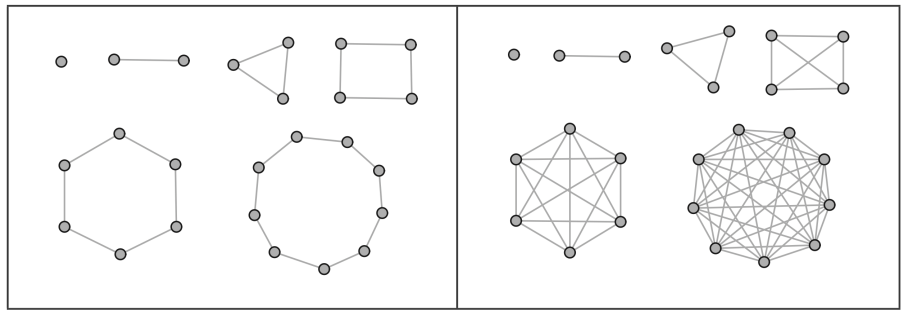
[1] M. Chen *et al.*, *PLoS genetics* **7**, e1001353 (2011).

[2] W. Zhang *et al.*, *PLoS computational biology* **9**, e1002975 (2013).

[3] C. Li, H. Li, *Bioinformatics* **24**, 1175–1182 (2008).

Incorporating genetic networks into group structured data

- (a) Pseudo networks: Ring and Fully connected (F.con)[4]:



- (b) Gene-level dimension reduction[5]:

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

- (c) Group-wise penalties (in progress):

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

[4] H. Sun, S. Wang, *Bioinformatics* **28**, 1368–1375 (2012).

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

Key Statistical Methodology Papers

2. Identification of pleiotropic variants associated with multiple mixed-type responses

Plant genomics with mixed-type responses

- In real application, many genetic studies include a variety of response types such as continuous, ordinal and categorical.
- For example, our cowpea dataset from National Institute of Crop Science, Rural Development Administration:

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

■: qualitative, ■: quantitative

- The goal is to identify genetic variants associated with multiple responses belonging to a specific category.

Variable selection on multiple responses

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

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

  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:

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

Key Statistical Methodology Papers

3. Integrative sparse reduced-rank regression for high-dimensional multi-source data

Reduced-Rank Regression (RRR)

- Multivariate regression model

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

- Reduced-rank regression model[6] as

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

  where $r \leq \min\{n, p, q\}$.

- Advantages:
  - This can effectively take into account the correlation between response variables through the latent variable $\mathbf{XB}$ of rank r.
  - This can dramatically reduce the number of parameters to be estimated, and thus the estimates are more precise;

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

Structural Learning in RRR

- Goal is to identify the structured association between multiple responses and multi-omics 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: $\mathbf{X}_{(3)}\mathbf{b}_{33}$

- A set of parameters is not unique up to an orthogonal matrix; For example, $\mathbf{BA}^T = \mathbf{BQQ}^T\mathbf{A}^T$ for $\mathbf{Q} \in \mathbb{R}^{r\times r}$ such that $\mathbf{QQ}^T = \mathbf{I}_r$.

Identifiability Problem

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

Constrained reduced-rank regression model

- We consider the constrained reduced-rank regression model under the quartimax-simple loading matrix $\mathbf{A}$:

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

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

Identifiability under RE condition

- We need the identifiability of $\mathbf{B}$, not $\mathbf{XB}$.
- Under the restricted eigenvalue condition[7] 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.*

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

Integrative Sparse Reduced-Rank Regression (iSRRR)

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

- Tuning parameters:
  - $\lambda \geq 0$ controls the structured sparsity of $\mathbf{B}$;
  - ν controls the row-wise sparsity of $\mathbf{A}$.

Recent Working Paper

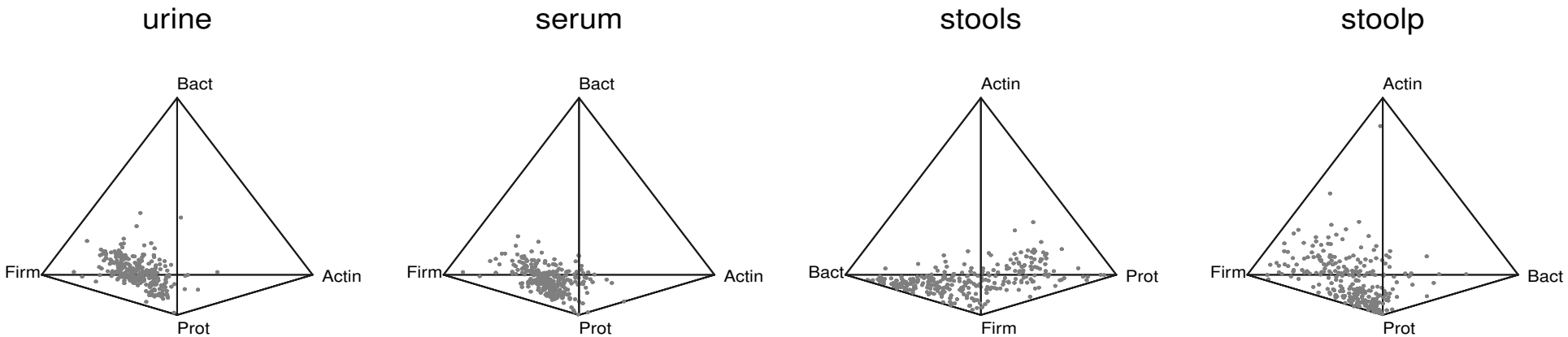
Principal component analysis for zero-inflated compositional data

High-dimensional zero-inflated compositional data

- 16s rRNA microbiome sequencing data
  - (1) Compositionality, (2) High dimensionality, (3) Zero inflation
- Sample space of compositional data is defined as

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

- Real data example with $p = 4$:



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

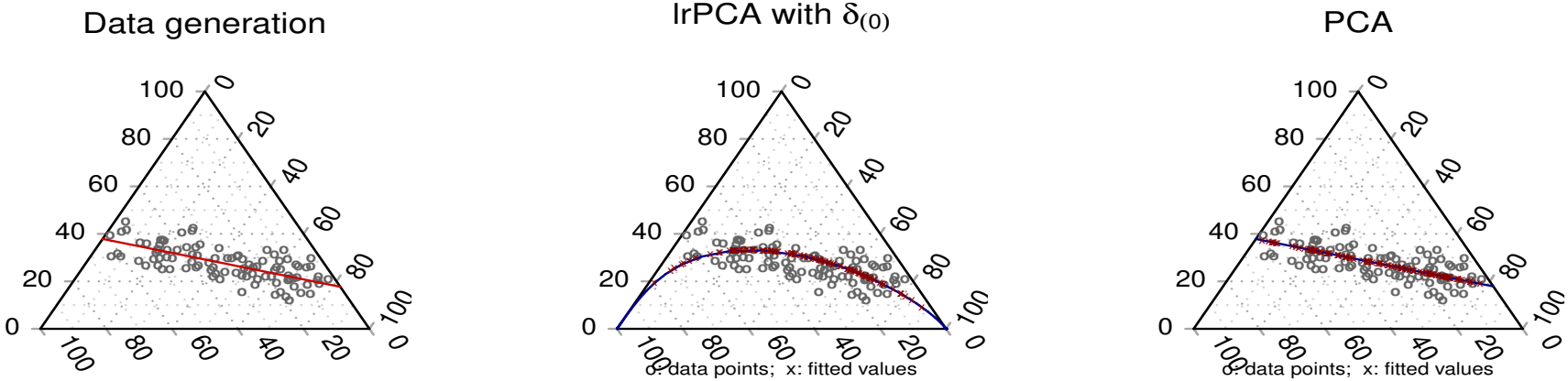
- We aim to find the principal compositional subspace to fit the data

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

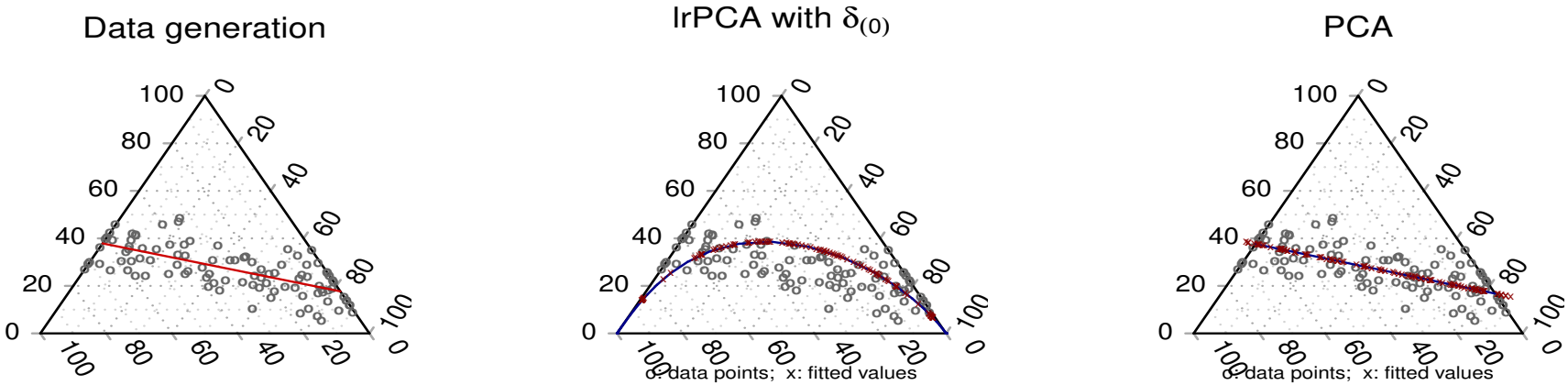
Motivating example: limitation of log-ratio PCA (lrPCA)

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

Without zeros



With 10% zeros



$$\delta_{(0)} = \min\{x_{ij} \in \mathbf{X} : x_{ij} > 0\}$$

The proposed methods: Compositional PCA

- Compositional Reconstructed PCA (crPCA): Given $\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}_1^{PC}, \ldots, \hat{\mathbf{V}}_r^{PC}$,

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

$$\hat{\boldsymbol{\mu}} + u_{i1}\hat{\mathbf{V}}_1^{PC\,T} + \cdots + u_{ir}\hat{\mathbf{V}}_r^{PC\,T} \in \mathbb{C}^p \quad \forall i$$

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

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

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

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

- Compositional PCA (CPCA): Given $\hat{\boldsymbol{\mu}}, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}$,

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

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

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

Theoretical properties

Theorem (Existence)

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

Theorem (Consistency)

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

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

(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 $\mathbb{C}^p$.

Future Plans

Research and Educational

Future Research Plan

- Future research topics
  - Incorporating external information in genome-wide association studies
  - Low-rank model for high-dimensional data
  - Multi-omics data integration
  - Statistical learning model for large-scale cohort data
- Research Grant Plan
  - Focus on interdisciplinary collaboration
  - Current co-workers
    - 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.

Educational Plan

- Available subjects
  - All subjects in Statistics
  - Bioinformatics (or high-dimensional data analysis)
  - Statistical programming language & Visualization
  - Machine learning using Python
- Featured lecture plans
  - Nearly bi-weekly homework assignments
- Student instructional plans
  - Encourage of a wide range of experiences including: research projects, competitions, hackathon and internships.
- All plans follow the rules of the department first

Thank you for your attention