Principal component analysis for zero-inflated compositional data

Kipoong Kim, Jaesung Park, and Sungkyu Jung

Department of Statistics, Seoul National University

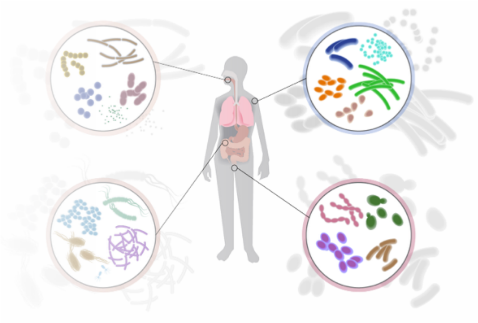
February 2, 2024

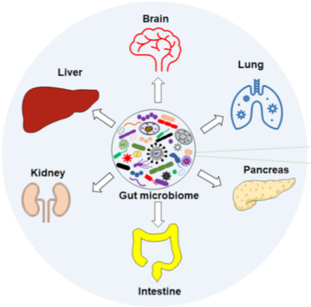
1. Introduction
    - Microbiome
    - Compositional data
2. The proposed method
    - Global compositional PCA
    - Compositional PCA
    - Approximated compositional PCA
3. Theoretical properties
    - Existence
    - Consistency
4. Simulation studies
5. Real data analysis

Our motivating data

16s rRNA microbiome sequencing data

- Formation of the Human Microbiome
  - Initial Colonization: Begins at birth, influenced by delivery method (vaginal vs. C-section) and breastfeeding.
  - Early Life ($\sim$ 1000 days): Shaped by diet transition and environmental exposure, including family and pets.
  - Adulthood: Continuously influenced by diet, lifestyle, and medication.
  - Other Factors: Genetics, geography, health status, and age also play roles.

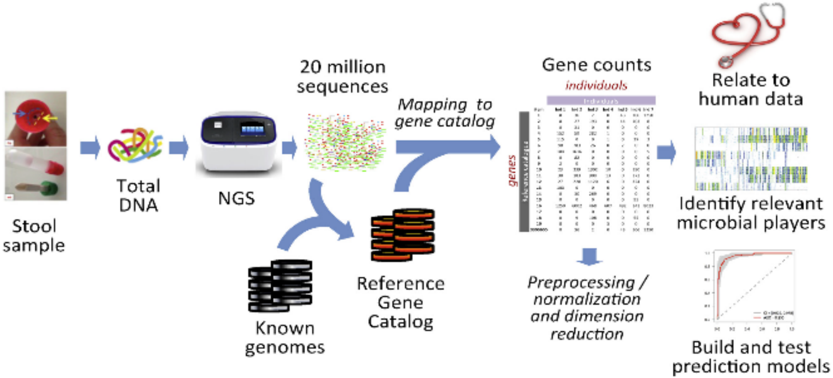




16s rRNA microbiome sequencing data

NGS Technologies

- Samples → DNA extraction → PCR+Library prep. → Sequencing & Mapping → Microbiome count data



16s rRNA microbiome sequencing data

Sampling examples

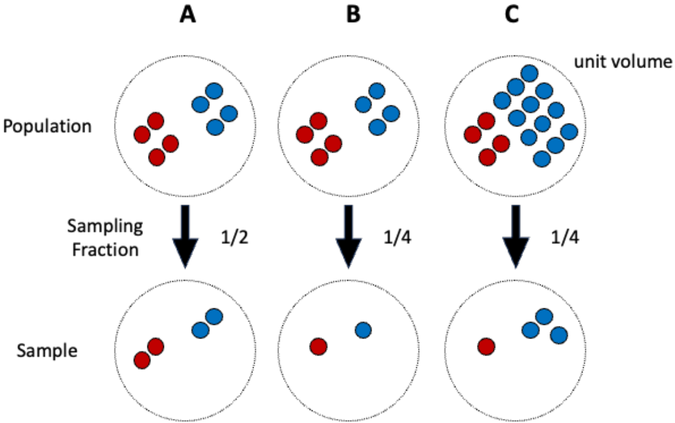


Figure: A vs B: different library size; A vs C: different sampling fraction

16s rRNA microbiome sequencing data

Challenges

- In analyzing such microbiome count data, many researchers encounter several challenges
  - **Variability in Library Size**: Differences in sequencing platforms and technical issues can cause significant variability in the number of reads across samples (Gloor et al., 2017).
    ⇒ Normalization to compositional data
  - **High Dimensionality**: The vast number of microbial taxa in samples adds complexity to data analysis.
    ⇒ Dimension reduction method
  - **Zero Inflation**: Insufficient sampling or specific sampling designs may lead to underrepresentation of rare taxa, resulting in data sparsity (Martíin-Fernaíndez et al., 2015).
    ⇒ Dealing with the zero inflation.
- In this work, we aim to develop a new dimension reduction method for zero-inflated compositional data.

Microbiome compositional data

- Compositional space (the sample space of compositional data):

$$\mathbb{C}^p = \left\{ (x_1, \ldots, x_p) : \sum_{j=1}^{p} x_j = 1; x_j \geq 0 \text{ for all } j \right\}.$$

- The compositional space can be thought of as $p-1$ dimensional `convex hull` embedded in $\mathbb{R}^p$.

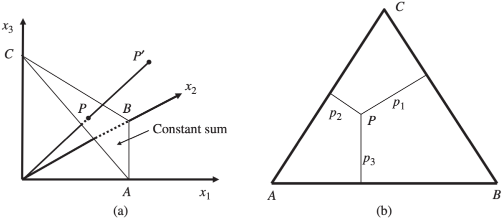


Figure: (a) Simplex embedded in the positive orthant of $\mathbb{R}^3$. (b) Ternary diagram.

- Compositional data that reside on a simplex does not admit the standard Euclidean geometry
  - e.g., not closed under addition and scalar multiplication

Microbiome compositional data

Existing methods

- There have been developments on compositional data analysis based on the so-called *Aitchison geometry*, which is based on the log-ratio transformation.
  - Additive log-ratio: $\text{alr}(\boldsymbol{x}) = \log x_j - \log x_J,\ J \in \{1, \ldots, p\}$
  - Centered log-ratio: $\text{clr}(\boldsymbol{x}) = \log x_j - \frac{1}{p}\sum_{j=1}^{p} \log x_j$
  - etc.
- Log-ratio PCA (Aitchison, 1983) to cope with both linear and curved data patterns.

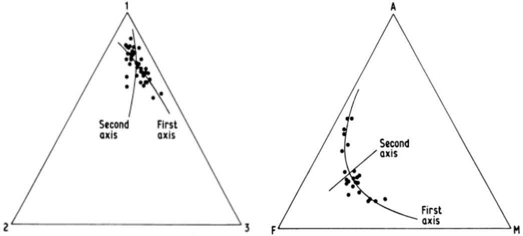


Figure: Ternary diagram with log-ratio principal axes

Log-ratio PCA

Dealing with zeros in log-ratio transformation

- Zero replacement strategies
  - Simple replacement

$$r_j = \begin{cases} \frac{1}{1+\sum_{k:x_k=0}\delta}\delta_j, & \text{if } x_j = 0, \\ \frac{1}{1+\sum_{k:x_k=0}\delta}x_j, & \text{if } x_j > 0, \end{cases}$$

  - Additive, Multiplicative, and etc. replacements.

  where δ is a small zero-replacement value.

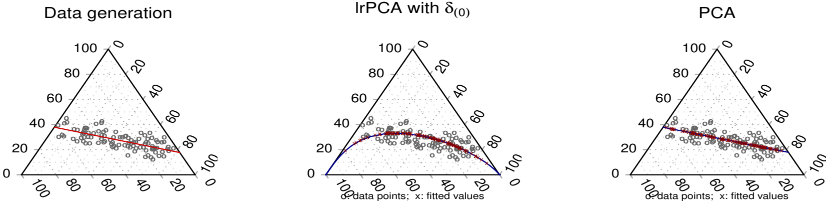
- Determination of δ
  - Count-level data: $\delta = 0.5$
  - Compositional-level data: $\delta = \frac{1}{2}\min\{x_j : x_j > 0\}$.

Limitation of log-ratio PCA

Sensitivity analysis for the zero replacement

- However, the zero inflation may result in the distortion.

Without zeros

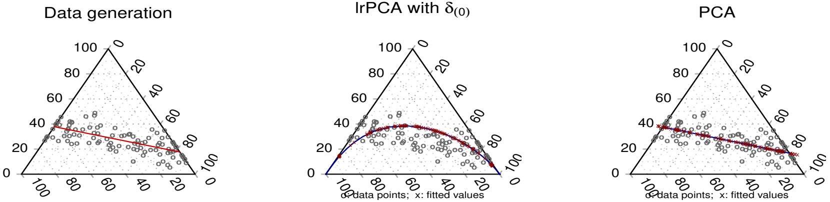


Data generation

lrPCA with $\delta_{(0)}$

PCA

With 10% zeros



Data generation

lrPCA with $\delta_{(0)}$

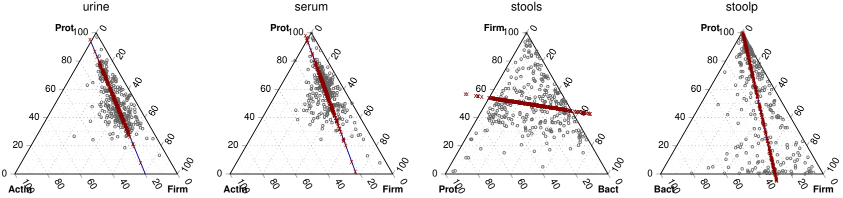
PCA

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

PCA for zero-inflated compositional data

Compositional reconstruction

- We want to propose a new dimension reduction method that prevents its low-rank reconstructions from being out of the composition space.
- Intuitive approach: compositional reconstruction PCA (crPCA)
  - Find the principal directions (classical PCA)
  - Project the principal scores into the compositional space



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

Real data example

Rank-1 reconstruction

- Composition plots of the rank-1 reconstruction in urine dataset (in the order of log-ratio PCA, crPCA, and CPCA).

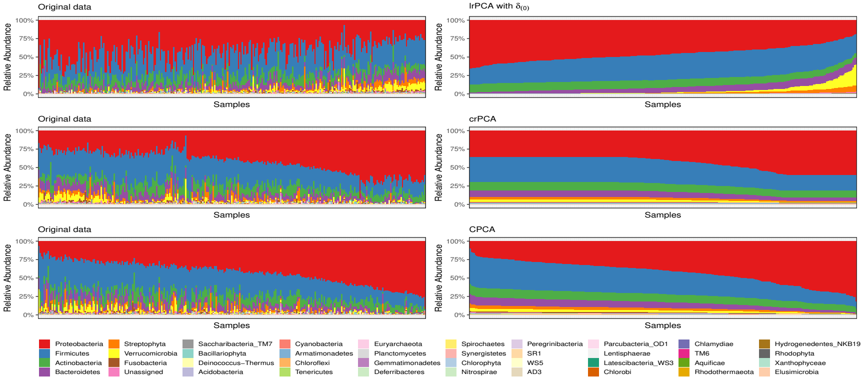


Figure: Left: the original data. Right: reconstructed data. The same sample orders were maintained between left and right panels for each method, based on its estimated first score.

Global Compositional PCA

Main goal

- Denote the i-th row of $\mathbf{A}$ by $\boldsymbol{a}_i$ and the k-th column of $\mathbf{A}$ by $\mathbf{A}_k$.

- Global compositional PCA (global CPCA) problem:

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

  subject to

  - $\mathbf{U}$ and $\mathbf{V}$ have `orthogonal` and `orthonormal` columns
  - $\boldsymbol{\mu} \in \mathbb{C}^p$
  - $\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i \in \mathbb{C}^p$ for all $i = 1, \ldots, n$.

- The mean vector $\boldsymbol{\mu}$ was set to the sample mean $\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$ for simplicity.

Global Compositional PCA

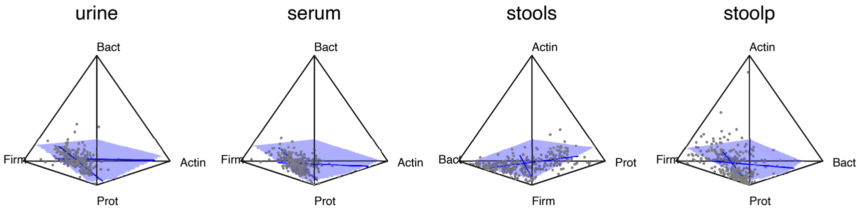
Main goal

- Compositional subspace, spanned by $\{\mathbf{V}_1, \dots, \mathbf{V}_r\}$ at $\boldsymbol{\mu}$:

$$\mathbb{CS}_{(\boldsymbol{\mu};\{\mathbf{V}_1,\dots,\mathbf{V}_r\})} := \mathbb{C}^p \cap \{\boldsymbol{\mu} + c_1\mathbf{V}_1 + \cdots + c_r\mathbf{V}_r : c_1, \dots, c_r \in \mathbb{R}\}$$

- Alternatively, global CPCA finds an r-dimensional compositional subspace, where
  - the data is best approximated and
  - the low-rank reconstruction lies within the compositional subspace.

- e.g.



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

Compositional PCA

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:

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

$$\vdots$$

  - Rank-k case:

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

for $k = 1, \ldots, r$.

Compositional PCA

The proposed methods

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

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

  subject to

  - $\boldsymbol{\mu} + \sum_{h=1}^{k-1} u_{i\,h}\hat{\mathbf{V}}_h + u_{ik}\mathbf{V}_k \in \mathbb{C}^p \quad \forall i$
  - $\mathbf{V}_k \perp \mathbf{1}_p, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}$ and $\|\mathbf{V}_k\|_2 = 1$

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

$$\underset{\mathbf{U}_k,\mathbf{V}_k}{\arg\min} \ \|\mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \hat{\mathbf{U}}_1\hat{\mathbf{V}}_1^T - \cdots - \hat{\mathbf{U}}_{k-1}\hat{\mathbf{V}}_{k-1}^T - \mathbf{U}_k\mathbf{V}_k^T\|_F^2, \qquad (3)$$

  subject to

  - $\boldsymbol{\mu} + \sum_{h=1}^{k-1} \hat{u}_{ih}\hat{\mathbf{V}}_h + u_{ik}\mathbf{V}_k \in \mathbb{C}^p \quad \forall i$
  - $\mathbf{V}_k \perp \mathbf{1}_p, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \ \|\mathbf{V}_k\|_2 = 1$

- In each method, we use the alternating algorithm to estimate both scores and directions.

Computational Algorithm

Sub-problems: U-update for CPCA

- The problem can be expressed as the individual problem for the i-th sample, $i = 1, \ldots, n$:

$$\begin{aligned} \underset{u_{i1},\ldots,u_{ik}}{\arg\min} \; \|\boldsymbol{x}_i - \boldsymbol{\mu} - u_{i1}\hat{\mathbf{V}}_1 - \cdots - u_{i,k-1}\hat{\mathbf{V}}_{k-1} - u_{ik}\mathbf{V}_k\|_2^2 \\ \text{subject to } \boldsymbol{\mu} + \textstyle\sum_{h=1}^{k-1} u_{i,h}\hat{\mathbf{V}}_h + u_{ik}\mathbf{V}_k \in \mathbb{C}^p, \end{aligned} \tag{4}$$

  where $\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \mathbf{V}_k \perp \mathbf{1}_p$ are fixed.

Proposition 1

The problem (4) can be expressed as a quadratic programming problem:

$$\hat{\boldsymbol{u}}_i = \underset{\boldsymbol{u}_i \in \mathbb{R}^n}{\arg\min} \; \boldsymbol{u}_i^T(\tilde{\mathbf{V}}^T\tilde{\mathbf{V}})\boldsymbol{u}_i - 2\left\{(\boldsymbol{x}_i - \boldsymbol{\mu})^T\tilde{\mathbf{V}}\right\}\boldsymbol{u}_i \;\; \textit{subject to } \tilde{\mathbf{V}}\boldsymbol{u}_i \geq -\boldsymbol{\mu},$$

where $\tilde{\mathbf{V}} = (\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \mathbf{V}_k)$.

Computational Algorithm

Sub-problems: U-update for aCPCA

- The problem for the i-th sample can be expressed as:

$$\underset{u_{ik}\in\mathbb{R}}{\arg\min}\ \|\boldsymbol{x}_i - \boldsymbol{c}_i - u_{ik}\mathbf{V}_k\|_2^2 \ \text{ subject to } \boldsymbol{c}_i + u_{ik}\mathbf{V}_k \in \mathbb{C}^p, \tag{5}$$

where $\boldsymbol{c}_i = \boldsymbol{\mu} + \sum_{h=1}^{k-1} \hat{u}_{ih}\hat{\mathbf{V}}_h \in \mathbb{C}^p$ and $\mathbf{V}_k \perp \mathbf{1}_p, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}$ are fixed.

Proposition 2

The solution of (5) is given by

$$\hat{u}_{ik} = \begin{cases} m_k & \text{if } \langle \boldsymbol{x}_i - \boldsymbol{c}_i, \mathbf{V}_k\rangle \le m_k, \\ \langle \boldsymbol{x}_i - \boldsymbol{c}_i, \mathbf{V}_k\rangle & \text{if } m_k \le \langle \boldsymbol{x}_i - \boldsymbol{c}_i, \mathbf{V}_k\rangle \le M_k, \\ M_k & \text{if } \langle \boldsymbol{x}_i - \boldsymbol{c}_i, \mathbf{V}_k\rangle \ge M_k, \end{cases}$$

where $m_k = \max_j \{-\mu_j/v_{jk}\}$ *and* $M_k = \min_j \{-\mu_j/v_{jk}\}$.

Computational Algorithm

Sub-problems: V-update

- Both aCPCA and CPCA problems for $\mathbf{V}_k$ can be expressed as follows. For fixed $(\mathbf{U}_1, \ldots, \mathbf{U}_k)$,

$$\operatorname*{arg\,min}_{\mathbf{V}_k \in \mathbb{R}^p} \left\| \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{U}_1 \hat{\mathbf{V}}_1^T - \cdots - \mathbf{U}_{k-1} \hat{\mathbf{V}}_{k-1}^T - \mathbf{U}_k \mathbf{V}_k^T \right\|_F^2 \tag{6}$$

subject to the appropriate constraints.

Proposition 3

The problem (6) is equivalent to a quadratic programming problem given by

$$\operatorname*{arg\,min}_{\mathbf{V}_k \in \mathbb{R}^p} \mathbf{V}_k^T (\mathbf{U}_k^T \mathbf{U}_k) \mathbf{V}_k - 2 \left(\sum_{i=1}^{n} u_{ik} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right)^T \mathbf{V}_k$$

$$\textit{subject to } (\mathbf{1}, \hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1})^T \mathbf{V}_k = \mathbf{0} \textit{ and } u_{ik} \mathbf{V}_k \geq -\boldsymbol{c}_i \ \forall i,$$

where $\boldsymbol{c}_i = \boldsymbol{\mu} + u_{i1} \hat{\mathbf{V}}_1 + \cdots + u_{i,k-1} \hat{\mathbf{V}}_{k-1}$.

Computational Algorithm

CPCA

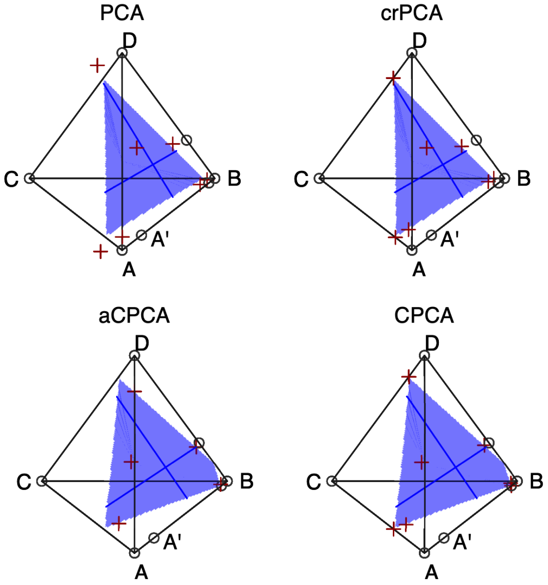
```
Algorithm 1: Rank-$k$ approximation for CPCA
Input: $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$ and $(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1})$.
Initialize $\mathbf{V}_k^{(0)} \perp \mathbf{1}_p$.
Repeat for $t = 0, 1, 2, \ldots$:
    1 U-update: obtain $\boldsymbol{u}_i^{(t+1)}$ by (4) with $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ and $\mathbf{V}_k = \mathbf{V}_k^{(t)}$ $\forall i$.
    2 U-shrinkage: $\boldsymbol{u}_i^{(t+1)} \leftarrow (1 - \frac{\gamma}{t+1})\boldsymbol{u}_i^{(t+1)}$.
    3 V-update: obtain $\mathbf{V}_k^{(t+1)}$ by (6) with $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ and $\mathbf{U} = (\mathbf{U}_1^{(t+1)}, \ldots, \mathbf{U}_k^{(t+1)})$
    4 V-scaling: $\mathbf{V}_k^{(t+1)} \leftarrow \mathbf{V}_k^{(t+1)} / \|\mathbf{V}_k^{(t+1)}\|_2$.
until convergence: $\|\mathbf{V}_k^{(t+1)} - \mathbf{V}_k^{(t)}\|_F^2 < \epsilon$.
Re-estimation of $\mathbf{U}$: estimate $\boldsymbol{u}_i^{(t+1)}$ without the shrinkage $\forall i$.
Output: $(\mathbf{U}_1^{(t+1)}, \ldots, \mathbf{U}_k^{(t+1)})$ and $(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \mathbf{V}_k^{(t+1)})$.
```

Computational Algorithm

Approximated CPCA

```
Algorithm 2: Rank-$k$ approximation for aCPCA
Input: $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $(\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_{k-1})$ and $(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1})$.
Initialize $\mathbf{V}_k^{(0)}$.
Repeat for $t = 0, 1, 2, \ldots$:
    1 U-update: obtain $u_{ik}^{(t+1)}$ by (5) with $\boldsymbol{c}_i = \bar{\boldsymbol{x}} + \sum_{h=1}^{k-1} \hat{u}_{ih} \hat{\mathbf{V}}_h$ and $\mathbf{V}_k = \mathbf{V}_k^{(t)}$ $\forall i$.
    2 U-shrinkage: $u_{ik}^{(t+1)} \leftarrow (1 - \frac{\gamma}{t+1}) u_{ik}^{(t+1)}$.
    3 V-update: obtain $\mathbf{V}_k^{(t+1)}$ by (6) with $\boldsymbol{\mu} = \bar{\boldsymbol{x}}$ and $\mathbf{U} = (\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_{k-1}, \mathbf{U}_k^{(t+1)})$.
    4 V-scaling: $\mathbf{V}_k^{(t+1)} \leftarrow \mathbf{V}_k^{(t+1)} / \|\mathbf{V}_k^{(t+1)}\|_2$.
until convergence: $\|\mathbf{V}_k^{(t+1)} - \mathbf{V}_k^{(t)}\|_F^2 < \epsilon$.
Re-estimation of $\mathbf{U}$: estimate $u_{ik}^{(t+1)}$ without the shrinkage $\forall i$.
Output: $(\hat{\mathbf{U}}_1, \ldots, \hat{\mathbf{U}}_{k-1}, \mathbf{U}_k^{(t+1)})$ and $(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1}, \mathbf{V}_k^{(t+1)})$.
```

An illustrative comparison



Theoretical properties

- Consider that X is a $\mathbb{C}^p$-valued random element defined on the probability space $(\Omega, \mathcal{A}, \mathcal{P})$. We denote i.i.d. copies of X by $\mathcal{X}_n := (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)$.
- Let $\Pi_{\mathcal{Z}}(\boldsymbol{x}) = \arg\min_{\boldsymbol{z} \in \mathcal{Z}} \|\boldsymbol{x} - \boldsymbol{z}\|_2^2$ be an Euclidean projection of $\boldsymbol{x} \in \mathbb{R}^p$ onto a nonempty closed convex subset $\mathcal{Z} \subset \mathbb{C}^p$.
    - This projection is unique because $\mathcal{Z}$ is a nonempty closed convex subset and the norm is strictly convex and differentiable.
- For a nonempty closed convex subset $\mathsf{CS} \subset \mathbb{C}^p$, let us define the population risk and empirical risk with respect to CS by

$$R(\mathsf{CS}) := \mathbb{E}\|X - \Pi_{\mathsf{CS}}(X)\|_2^2 \quad \text{and} \quad R_n(\mathsf{CS}; \mathcal{X}_n) := \frac{1}{n}\sum_{i=1}^{n} \|\boldsymbol{x}_i - \Pi_{\mathsf{CS}}(\boldsymbol{x}_i)\|_2^2,$$

respectively.

Theoretical properties

- Let us denote by $\mathbb{CS}_{k,\mathcal{Z}}$ the set of all k-dimensional compositional subspaces containing a subset $\mathcal{Z}$ of $\mathbb{C}^p$ with dim(span($\mathcal{Z}$)) $< k$.

- Then, we can write the proposed CPCA problem as

$$\hat{F}_k = \operatorname*{arg\,min}_{\mathsf{CS} \in \mathbb{CS}_{k,\hat{F}_{k-1}}} R_n(\mathsf{CS}; \mathcal{X}_n) = \frac{1}{n} \textstyle\sum_{i=1}^{n} \|\boldsymbol{x}_i - \Pi_{\mathsf{CS}}(\boldsymbol{x}_i)\|_2^2$$

  for a given $\mathcal{X}_n = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)$, and its population version as

$$F_k = \operatorname*{arg\,min}_{\mathsf{CS} \in \mathbb{CS}_{k,F_{k-1}}} R(\mathsf{CS}) = \mathbb{E}\|X - \Pi_{\mathsf{CS}}(X)\|_2^2$$

  : Forward Principal Compositional Subspace

- Using this framework, we will show
  - the **existence** of F_k, $\hat{F}_k$ and their directions $(V_k, \hat{V}_k)$,
  - the **consistency** of $\hat{F}_k$ to F_k and $\hat{V}_k$ to V_k in an adequate topology using the generalized Fréchet mean framework introduced in Park and Jung (2023+).

Theoretical properties: Existence

- To show the existence of F_k, we utilize the fact that the continuous function on a compact set has a minimizer.
  - Recall that F_k is a minimizer of $R(\cdot)$ among $\mathbb{CS}_{k,F_{k-1}}$.
  - We first give a topology to $\mathbb{CS}_{k,F_{k-1}}$ so that $R(\cdot) : \mathbb{CS}_{k,F_{k-1}} \to [0, \infty)$ is a continuous function on a compact set.
  - We utilize the Hausdorff distance (Beer, 1993) to measure a discrepancy between two compositional subspaces, defined by

    $$h(A, B) := \max \left(\sup_{a \in A} \inf_{b \in B} \|a - b\|_2, \ \sup_{b \in B} \inf_{a \in A} \|a - b\|_2 \right)$$

    for nonempty closed subsets A and B of $\mathbb{C}^p$.
  - Let us denote by $\mathcal{H}(\mathbb{C}^p)$ the collection of all nonempty closed subsets of $\mathbb{C}^p$ endowed with Hausdorff distance h.

Theoretical properties: Existence

- Then, we can check the continuity of the risk function $R(\cdot)$ with respect to the Hausdorff distance h through the following lemma.

Lemma 1

For any distribution of X, $R(\cdot) : \mathcal{H}(\mathbb{C}^p) \to [0, \infty)$ is continuous with respect to h.

- Next, we can show the compactness of the minimizing domain $\mathbb{CS}_{k,F_{k-1}} \subset \mathcal{H}(\mathbb{C}^p)$, except that the subspace F_{k-1} lies completely on a simplex boundary.

Lemma 2

Let us denote the simplex boundary by $\partial\mathbb{C}^p := \{\boldsymbol{x} \in \mathbb{C}^p : x_j = 0$ for some $j\}$. Then, the following holds:

(1) $\mathcal{H}(\mathbb{C}^p)$ *is compact.*

(2) $\mathbb{CS}_{k,\mathcal{Z}}$ *is compact for all $k = 1, \ldots, p$ and for $\mathcal{Z} \in \mathbb{CS}_{k-1}$ such that $\mathcal{Z} \not\subset \partial\mathbb{C}^p$.*

Theoretical properties: Existence

- Together with Lemma 1 and 2, the existence of F_k and V_k is immediately established as described in Theorem 3.

Theorem 3

For any distribution of X with $P(X_j = 0) < 1$ for all $j = 1, \dots, p$, the forward principal compositional subspace F_k and its direction V_k exist for all $k = 1, \dots, p$.

- The empricial estimators $\hat{F}_k$ and $\hat{V}_k$ can be regarded as a special case of F_k and V_k derived from a distribution that assigns probability $1/n$ to each of $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_n$, for a fixed $\mathcal{X}_n$.
- Since each $\boldsymbol{x}_i \in \mathbb{C}^p \backslash \partial \mathbb{C}^p$ with probability 1, we have $\bar{\boldsymbol{x}} \in \mathbb{C}^p \backslash \partial \mathbb{C}^p$ almost surely. Thus, for any distribution of X with $P(X_j = 0) < 1 \ \forall j$, $\hat{F}_k$ and $\hat{V}_k$ exist almost surely.

Theoretical properties: Consistency

The generalized Fréchet mean framework (Park and Jung, 2023+)

- By applying the generalized Fréchet mean framework to our CPCA problem, we will show that $\hat{F}_k$ converges almost surely to F_k under an assumption.
  - In the generalized Fréchet mean framework, a nonempty closed subset defined on a more general metric space (M, d) is considered.
  - Let $\mathcal{H}(M)$ be the collection of all nonempty closed subsets of M endowed with Hausdorff distance h.
  - Subsequently, the minimizers of population risk and empirical risk with different minimizing domains are defined as

$$E_0 = \operatorname*{arg\,min}_{m \in M_0} \mathbb{E}[\mathfrak{c}(X, m)]$$

and

$$\hat{E}_n = \operatorname*{arg\,min}_{m \in M_n} \frac{1}{n} \sum_{i=1}^{n} \mathfrak{c}(X_i, m),$$

where $\mathfrak{c} : T \times M \to \mathbb{R}$ is a loss function (T is the data space), $M_0 \in \mathcal{H}(M)$, and $M_n : (\Omega, \mathcal{A}, \mathcal{P}) \to \mathcal{H}(M)$ is a sequence of closed random subsets of M.

Theoretical properties: Consistency

Regularity conditions

For almost all $\omega \in \Omega$, $M_n(\omega)$ converges to M_0 in the following sense of Kuratowski (Beer, 1993):
For $P_0 \in \mathcal{H}(M)$ and $P_n \in \mathcal{H}(M)$, we say that P_n converges to P_0 in the sense of Kuratowski if P_n and P_0 satisfies the following:

(i) If an arbitrary sequence $m_n \in P_n$ has an accumulation point, then that point is in P_0.

(ii) For an arbitrary $m_0 \in P_0$, there exist a sequence $m_n \in P_n$ that converges to m_0.

(M, d) is a separable and complete metric space. In other words, M is a polish metric space.

$\mathfrak{c}(t, \cdot) : M \to \mathbb{R}$ is continuous for each $t \in T$.

Let $B(m, r)$ be the open ball in M with center m and radius $r > 0$. For every $m \in M$, there exists $r = r_m > 0$ such that $\pi_{m,r}(X)$ and $\Pi_{m,r}(X)$ are integrable, where $\pi_{m,r}(t) := \inf_{m' \in B(m,r)} \mathfrak{c}(t, m')$ and $\Pi_{m,r}(t) := \sup_{m' \in B(m,r)} \mathfrak{c}(t, m')$ for $m \in M, r > 0, t \in T$ are the local infimum and supremum of the loss function $\mathfrak{c}$ in the neighborhood of m.

$\overline{\cup_{n \geq N} \hat{E}_n}$ is compact for some $N \in \mathbb{N}$ almost surely, where $\overline{E}$ indicates the closure of a set E.

Theoretical properties: Consistency

Main theorem of Park and Jung (2023+)

Theorem 4

Suppose $E_0 = \{m_0\}$ is a singleton set. Under regularity conditions on the loss function and on the metric space M, all sequences $m_n \in \hat{E}_n$ converge almost surely to m_0.

- Our CPCA problem is a special case of the generalized Fréchet mean framework with

$$T = \mathbb{C}^p, \quad \mathfrak{c}(\cdot, \bullet) = \ell(\cdot, \bullet), \quad E_0 = F_k, \quad \hat{E}_n = \hat{F}_k,$$
$$M = \mathcal{H}(\mathbb{C}^p), \quad M_0 = \mathbb{CS}_{k,F_{k-1}}, \text{ and } M_n = \mathbb{CS}_{k,\hat{F}_{k-1}},$$

where $\ell(\boldsymbol{x}, \mathsf{CS}) = \|\boldsymbol{x} - \Pi_{\mathsf{CS}}(\boldsymbol{x})\|_2^2$.

In addition, the regularity conditions are all satisfied.

Theoretical properties: Consistency

Consistency for the principal compositional subspace

- By Theorem 4, we can show the almost sure convergence of the principal compositional subspace to its population counterpart.

Assumption 1

F_k uniquely exists for all $k = 1, \ldots, p$.

Corollary 5

Under Assumption 1, for any distribution of X with $P(X_j = 0) < 1 \ \forall j$, the following holds for all $k = 1, \ldots, p$ almost surely:

$$\lim_{n \to \infty} h(\hat{F}_k, F_k) = 0.$$

Theoretical properties: Consistency

Consistency for the principal compositional direction

- This result also leads to the almost sure convergence of $\hat{V}_k$ to V_k, since the k-th principal compositional direction is uniquely determined (up to sign changes) for a given sequence of principal compositional subspaces.
- Finally, we obtain the following result on the consistency of the proposed principal compositional direction.

Corollary 6

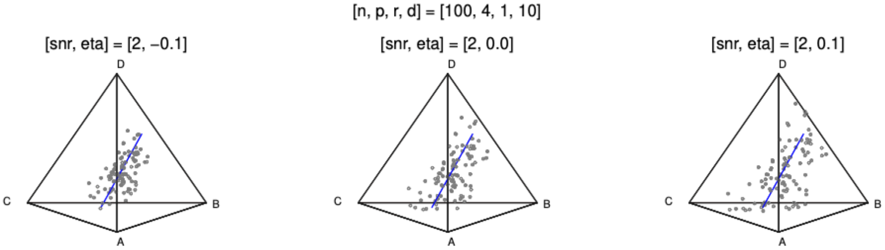
Under Assumption 1, for any distribution of X with $P(X_j = 0) < 1 \ \forall j$, the direction $\hat{V}_k(\mathcal{X}_n)$ converges almost surely to V_k for all $k = 1, \dots, p$ in the following sense:

$$\lim_{n\to\infty} \|\hat{V}_k(\mathcal{X}_n) - V_k\|_2 = 0.$$

Simulation studies

Scenario 1: Linear pattern

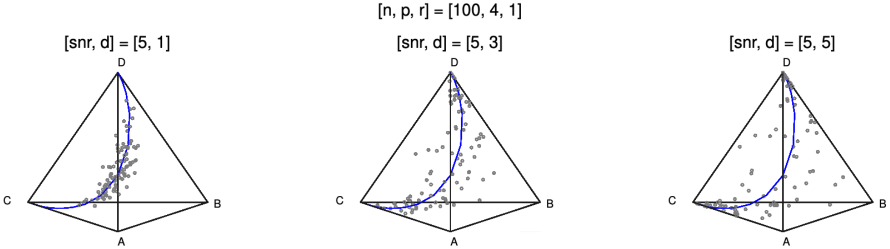
- We consider the model $\boldsymbol{x}_i = \Pi_{\mathbb{C}^p}(\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i + \boldsymbol{e}_i) \in \mathbb{C}^p$ for $i = 1, \ldots, n$.
  - The mean vector $\boldsymbol{\mu} \sim$ Dir$(10, \ldots, 10)$.
  - The directions $\mathbf{V} =$ Orth$(\mathbf{V}^*)$ with $\mathbf{V}^* = \{v_{jk}^*\}$ and $v_{jk}^* \sim N(0,1)$ such that $[\mathbf{1}_p, \mathbf{V}]^T[\mathbf{1}_p, \mathbf{V}] = \mathbf{I}_{r+1}$, for $j = 1, \ldots, p$ and $k = 1, \ldots, r$.
  - The scores $u_{ik} \sim TN(0, (d/k)^2; a_k - \dfrac{\eta}{\log(p)}, b_k + \dfrac{\eta}{\log(p)})$, where $[a_k, b_k]$ is the confined support which ensures any vectors within $[\boldsymbol{\mu} + a_k\mathbf{V}_k, \boldsymbol{\mu} + b_k\mathbf{V}_k]$ to be inside $\mathbb{C}^p$.
  - The error term $\boldsymbol{e}_i = (\mathbf{I}_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^T)\boldsymbol{e}_i^*$ with $e_{ij}^* \sim U(-\delta, \delta)$.



Simulation studies

Scenario 2: Curved pattern

- We consider the following log-normal model $\boldsymbol{x}_i^* = \mathcal{C}\left[\exp(\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i + \boldsymbol{e}_i)\right] \in \mathbb{C}^p$ for $i = 1, \ldots, n$, where $\mathcal{C}(\cdot)$ is a closure operator.
    - The mean vector $\boldsymbol{\mu}$ was set to $(0, \ldots, 0)$.
    - The directions $\mathbf{V}$ were generated in the same way to Scenario 1.
    - The scores $u_{ik} \sim N(0, (d/k)^2)$
    - The errors $e_{ij} \sim N(0, \sigma_e^2)$.
    - We apply the hard-thresholding and closure operators again to $\boldsymbol{x}_i^*$, with threshold of $0.01/\log(p)$ ($0.01/\log(p)$ is 0.00217 for $p = 100$).



Simulation studies

Parameters

- $n \in \{50, 100, 500, 1000\}$, $p = 100$, and $r = 5$
- Scenario 1
  - SNR $= 2$
  - $\eta = 0.1$
    - Prop. of 0's $=$ 14 – 16% empirically.
- Scenario 2
  - SNR $= 5$ in a centered log-ratio scale
  - $d = 3$
    - Prop. of 0's $=$ 0.2 – 0.6% empirically.

Simulation studies

Evaluation criterion

- The out-of-sample reconstruction error on an independent test data was calculated as

$$\sqrt{\frac{1}{n_{\text{test}}p}\sum_{i=1}^{n_{\text{test}}}\|\boldsymbol{x}_i^{\text{test}} - \text{Proj}_{(\bar{\boldsymbol{x}};\{\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_r\})}(\boldsymbol{x}_i^{\text{test}})\|_2^2}$$

  where $n_{\text{test}} = 1000$, $\boldsymbol{x}_i^{\text{test}}$ is the i-th observation vector of the test data.

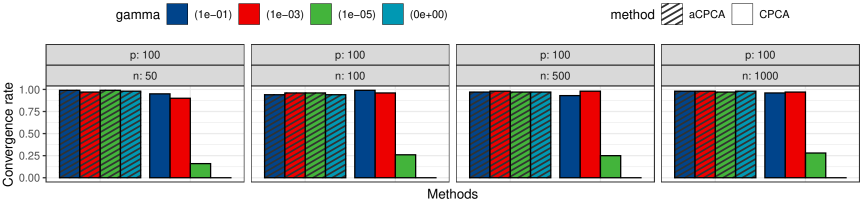
- All results are presented as averages over 100 simulated replicates

Simulation results

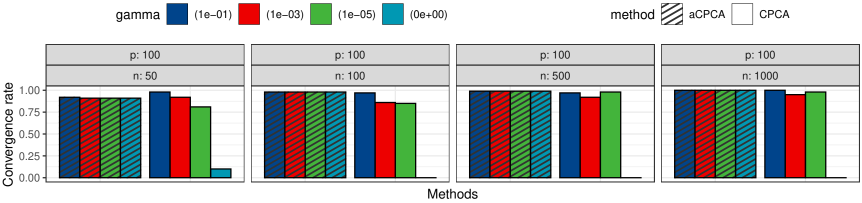
Convergence

- The proportion of cases that converged over 100 simulation replicates
  - We choose the shrinkage parameter $\gamma = 0.1$ as an optimal.

Scenario 1: Linear pattern

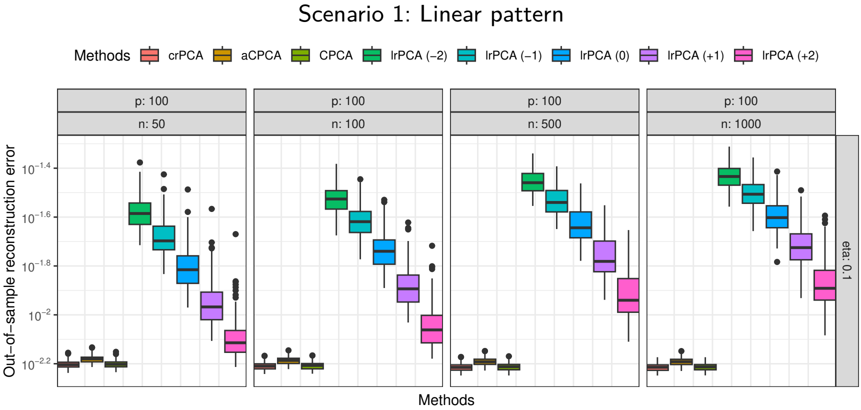


Scenario 2: Curved pattern



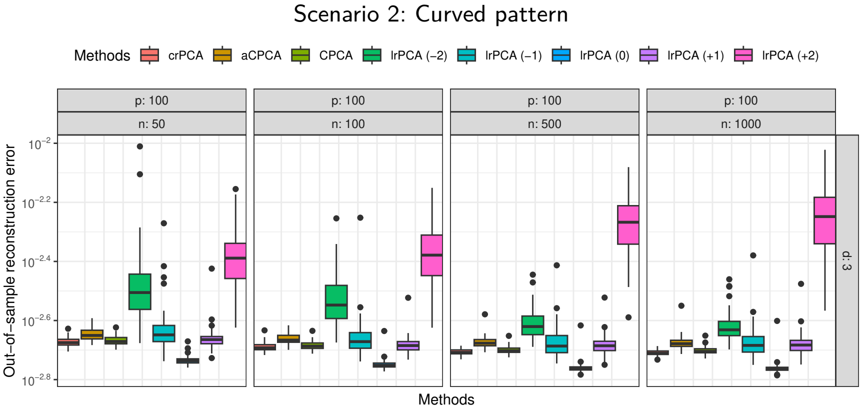
Simulation results

Out-of-sample reconstruction error



Simulation results

Out-of-sample reconstruction error



Real data analysis: microbiome data

- Microbiome counts of reads were measured at four different body sites (urine, serum, stool-s, stool-p) for $n = 293$ individuals.
- The counts of reads were amalgamated to the phylum level, resulting in data dimensions of $p = 40, 44, 46, \text{ and } 32$, respectively.

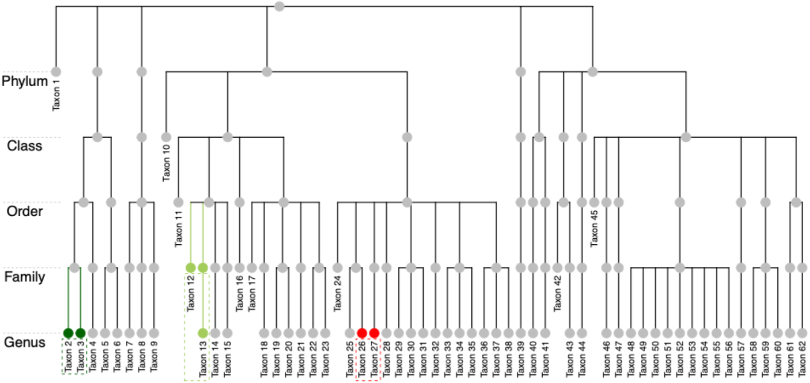
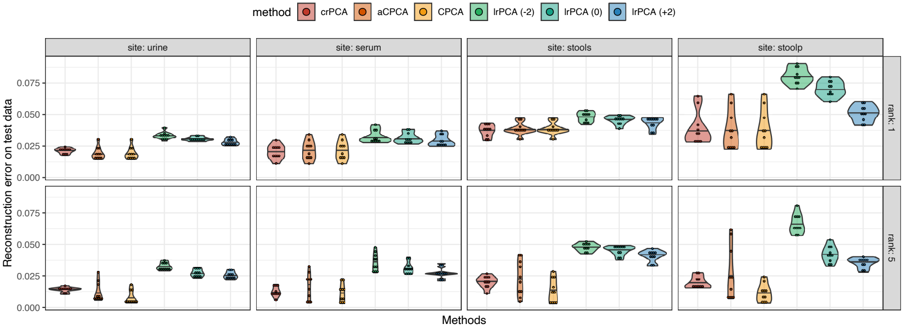


Figure: Taxonomic hierarchy

Real data anlysis

Rank-1 and rank-5 results

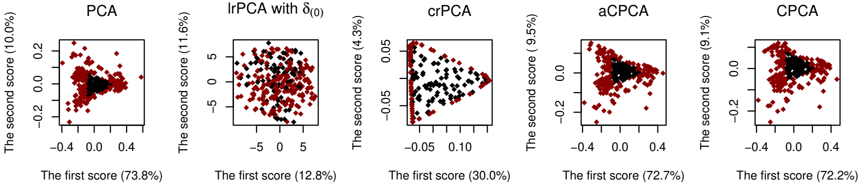
- 10-fold cross-validated (CV) reconstruction error
  - Top: $r = 1$; Bottom: $r = 5$:
  - Methods: crPCA, aCPCA, CPCA, and lrPCA with $\delta_{(-2)}, \delta_{(0)}, \delta_{(+2)}$.



Real data analysis

Rank-2 result: PC scores

- The first two PC scores estimated in the urine dataset:



  - The red points is the samples out of a simplex in PCA reconstruction.
  - In log-ratio PCA, the proportion of variance explained was calculated in a centered log-ratio scale.

Real data analysis

Rank-2 result: PC directions

- The first two PC directions estimated in the urine dataset:

Taxa	$\hat{v}_{j1}$				$\hat{v}_{j2}$			
	lrPCA	crPCA	aCPCA	CPCA	lrPCA	crPCA	aCPCA	CPCA
Proteobacteria	-0.130	0.865	0.863	0.863	-0.063	-0.266	0.275	0.275
Bacillariophyta	0.060	0.010	0.005	0.005	-0.562	0.013	-0.011	-0.020
Elusimicrobia	-0.080	$<$0.001	$<$0.001	$<$0.001	-0.029	$<$0.001	$<$0.001	$<$0.001
Xanthophyceae	-0.072	$<$0.001	$<$0.001	$<$0.001	-0.030	$<$0.001	$<$0.001	$<$0.001
Rhodophyta	-0.076	$<$0.001	$<$0.001	$<$0.001	-0.028	$<$0.001	$<$0.001	$<$0.001
⋮			⋮				⋮	
Verrucomicrobia	0.546	-0.092	-0.055	-0.055	-0.210	0.246	-0.277	-0.255
Streptophyta	0.321	-0.074	-0.065	-0.065	0.037	0.391	-0.328	-0.302
Actinobacteria	-0.094	-0.066	-0.072	-0.072	0.038	-0.141	0.197	0.220
Bacteroidetes	0.047	-0.145	-0.150	-0.150	-0.002	0.414	-0.453	-0.476
Firmicutes	-0.045	-0.461	-0.469	-0.469	-0.010	-0.723	0.702	0.699

The direction of lrPCA is in a centered log-ratio scale.

Summary

- In this work, we proposed three types of compositional PCA based on the Euclidean projection onto the principal compositional subspace.
- These methods outperformed the existing log-ratio PCA when linear patterns are present in zero-inflated data, and they demonstrated comparable performance in scenarios with curved patterns.
- Although the proposed optimization problems are inherently non-convex, we empirically guaranteed their convergence by utilizing the shrinkage parameter.
- We also established the existence and consistency of the forward principal compositional subspace and its direction.
    - We are also interested in robust compositional PCA as future research.

Thank you for your attention ! ☺