Compositional PCA based on projection onto simplicial subspace with application to microbiome data

Kipoong Kim, Jaesung Park, and Sungkyu Jung

Department of Statistics, Seoul National University

September 1, 2023

Compositional data

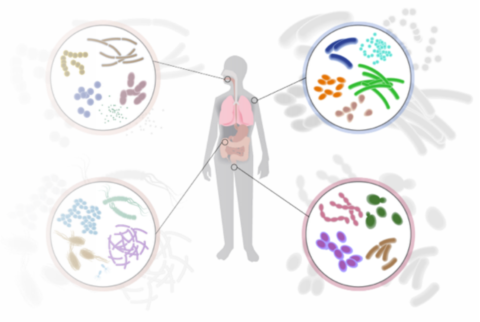
- Compositional data consists of vectors of proportions summing to one: e.g.
  - Geology, a rock composed of different minerals $[0.1, 0.3, 0.6]$;
  - Demography, a town or country
  - Epidemiology, 24-hour time-use data

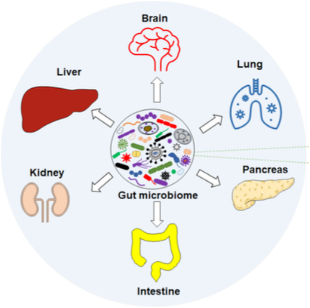
Table: Average monthly expenses per household

Type	ID	Housing	Foodstuff	Transport	Commun.	Sum
Absolute information	1	269	430	287	128	1114
	2	403	645	431	192	1671
	3	592	946	631	282	2450
Information expressed in %	1	24	39	26	11	100
	2	24	39	26	11	100
	3	24	39	26	11	100

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.





Our motivating data

- 16s rRNA microbiome sequencing data
  - The data is collected by **counts of reads**, which can vary significantly between samples due to the DNA extraction process, the concentration of microbial cells, and technical problem.
  - For example,

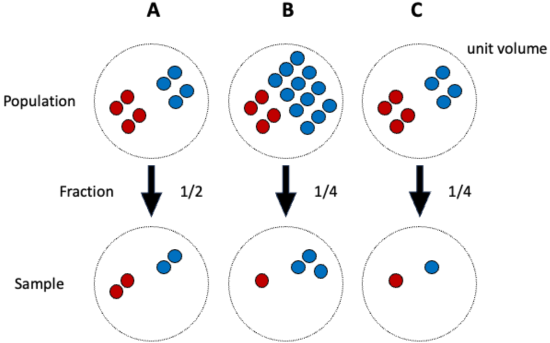


Figure: A vs B, C: different sampling fraction; A, B vs C: different library size

Compositional microbiome data

- Normalization to compositional data to create equal library sizes:

The number of counts

	x	y	z
A	1	4	6
B	5	15	30

⇒

The proportion of counts

	x	y	z
A	0.09	0.36	0.54
B	0.1	0.3	0.6

- Sample space of compositional data is a simplex defined as

$$\mathbb{S}^p = \{(x_1, \dots, x_p) : x_1 \geq 0, \dots, x_p \geq 0; x_1 + \cdots + x_p = 1\}.$$

- Closure operation $\mathcal{C} : \mathbb{R}^p_+ \to \mathbb{S}^p$ is defined as

$$\mathcal{C}(\boldsymbol{x}) = \left[\frac{x_1}{\sum_{j=1}^p x_j}, \dots, \frac{x_p}{\sum_{j=1}^p x_j} \right]$$

- In microbiome data,
  - Counts of reads: $\boldsymbol{x}^* \in \mathbb{R}^p_+$
  - Relative abundance: $\boldsymbol{x} = \mathcal{C}(\boldsymbol{x}^*) \in \mathbb{S}^p$

Compositional data

- Subcomposition and Amalgamation
  - Given a composition $\boldsymbol{x}$ and a selection of interest $\mathcal{A} = \{j_1, \ldots, j_a\}$, a `subcomposition` $\boldsymbol{x}_{\mathcal{A}}$, with a parts, can be written as

    $$\boldsymbol{x}_{\mathcal{A}} = \mathcal{C}[x_{j_1}, \ldots, x_{j_a}] = \left[\frac{x_{j_1}}{\sum x_{j_\ell}}, \ldots, \frac{x_{j_a}}{\sum x_{j_\ell}} \right],$$

    and the value $\sum_{j \in \mathcal{A}} x_j$ is called `amalgamated component`.
- Spurious correlation (bias towards negative correlation):

  $$\begin{aligned} 0 &= cov(x_1, x_1 + \cdots + x_p) \\ &= var(x_1) + cov(x_1, x_2) + \cdots + cov(x_1, x_p) \\ -var(x_1) &= cov(x_1, x_2) + \cdots + cov(x_1, x_p) \end{aligned}$$

  At least one of the covariances on the right must be negative.

Compositional data

- Simplex can be thought of as $p-1$ dimensional `convex hull` embedded in p-dimensional real space.

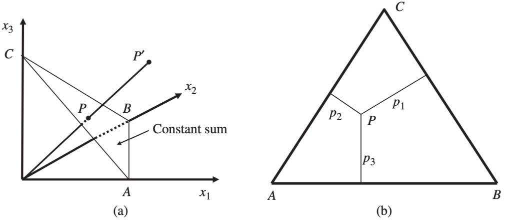


Figure: (a) Simplex embedded in the positive orthant of $\mathbb{R}^3$. (b) Ternary diagram.

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

Aitchison geometry

- There have been developments on compositional data analysis based on the so-called *Aitchison geometry*[1], which is based on the log-ratio transformation.

- The most common log-ratio transformation is *centered log-ratio (clr) transformation*.
  - Additive log-ratio operator: $alr(\boldsymbol{x}) = \log x_j - \log x_J,\ J \in \{1, \dots, p\}$
  - Centered log-ratio operator: $clr(\boldsymbol{x}) = \log x_j - \frac{1}{p} \sum_{j=1}^{p} \log x_j$
  - Isometric log-ratio operator: $ilr(\boldsymbol{x}) = \log x_j - \frac{1}{p} \sum_{j=1}^{p} \log x_j$

- The inverse operator:

$$inv(\boldsymbol{z}) = \mathcal{C}(\exp \boldsymbol{z}) = \left[\frac{\exp z_1}{\sum_{j=1}^{p} \exp z_j}, \dots, \frac{\exp z_p}{\sum_{j=1}^{p} \exp z_j} \right].$$

[1] J. Aitchison, *Journal of the Royal Statistical Society: Series B* **44**, 139–160 (1982).

Existing PCA methods for compositional data

- Log-ratio PCA[2]: copes with both linear and curved data patterns.

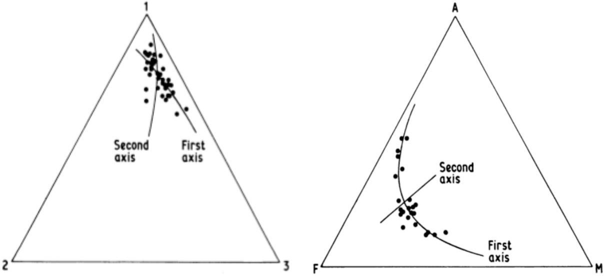


Figure: Ternary diagram with log-ratio principal axes

- Limitation
  - the log-ratio transformation could be inadequate to accommodate the *distinctive features of microbiome data* such as zero inflation, over dispersion and the presence of taxonomic tree structure among microbes.

[2] J. Aitchison, *Biometrika* **70**, 57–65 (1983).

Dealing with zeros in log-ratio transformation

- Zero replacement strategies
  - Simple replacement

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

  - Additive replacement

$$r_j = \begin{cases} \frac{\delta_j(Z+1)N}{(N+Z)^2}, & \text{if } x_j = 0, \\ x_j - \frac{\delta_j(Z+1)Z}{(N+Z)^2}, & \text{if } x_j > 0, \end{cases}$$

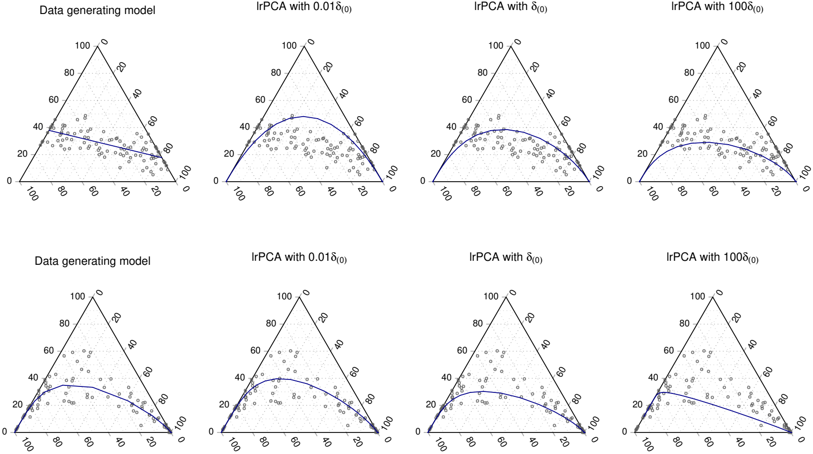
  - Multiplicative replacement

$$r_j = \begin{cases} \delta_j, & \text{if } x_j = 0, \\ (1 - \frac{\sum_{k:x_k=0} \delta_k}{1})x_j, & \text{if } x_j > 0, \end{cases}$$

where δ_j is a small value (e.g. $\min\{x_j : x_j > 0\}$), Z is the number of zeros, and N is the number of nonzeros $(i.e., N + Z = p)$.

Sensitivity analysis for the zero replacement

- Based on $\delta_{(0)} = \min\{x_j : x_j > 0\}$ (minimum of nonzero compositions),



Data generating model

lrPCA with $0.01\delta_{(0)}$

lrPCA with $\delta_{(0)}$

lrPCA with $100\delta_{(0)}$

Data generating model

lrPCA with $0.01\delta_{(0)}$

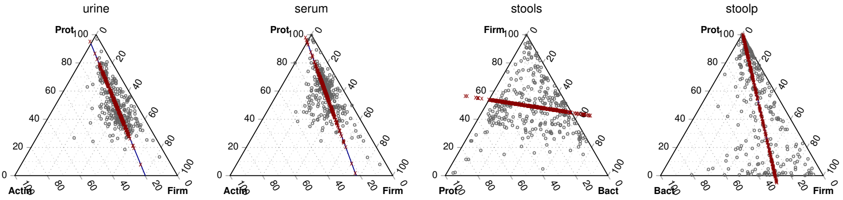
lrPCA with $\delta_{(0)}$

lrPCA with $100\delta_{(0)}$

- The log-ratio PCA result highly depends on the zero-replacement value.

Subcompositional plot: linearity pattern

- Subcomposition plot for the three most abundant microbes with naive PCA axes



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

- Low-rank approximation of compositional data do not belong to a simplex.
    - New statistical method (Compositional PCA).

Main goal

- We denote the transpose of the i-th row vector by $\boldsymbol{a}_i$ and the k-th column vector by $\mathbf{A}_k$ for a matrix $\mathbf{A}$.
- We want to solve the following 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
  - $\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i^T \in \mathbb{S}^p \ \forall i$ for $\boldsymbol{\mu} \in \mathbb{S}^p$

- Simplicial subspace

$$\mathbb{S}^p_{\mathbf{V}_1,\ldots,\mathbf{V}_k} := \mathbb{S}^p \cap span(\{\mathbf{V}_1, \ldots, \mathbf{V}_k\})$$

  (intersection of affine subspace spanned by $\mathbf{V}_1, \ldots, \mathbf{V}_k$ and $\mathbb{S}^p$)

- There is no relationship between $(\mathbf{U}_{k-1}, \mathbf{V}_{k-1})$ and $(\mathbf{U}_k, \mathbf{V}_k)$: e.g. $\mathbb{S}^p_{\mathbf{V}_1,\ldots,\mathbf{V}_{k-1}} \not\subseteq \mathbb{S}^p_{\mathbf{V}_1,\ldots,\mathbf{V}_{k-1},\mathbf{V}_k}$.

Main ideas

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

$$\mathbb{S}^p_{\mathbf{V}_1} \subset \mathbb{S}^p_{\mathbf{V}_1,\mathbf{V}_2} \subset \cdots \subset \mathbb{S}^p_{\mathbf{V}_1,\ldots,\mathbf{V}_k} \subset \cdots .$$

- (Score) Projection onto principal simplicial subspace:

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

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

Two types of projection approaches

- One-dimensional projection

$$\Pi_{\boldsymbol{v}}^{one}(\boldsymbol{x}_i;\boldsymbol{\mu}) = \mathop{\arg\min}_{u_i \in \mathbb{R}} \|\boldsymbol{x}_i - \boldsymbol{\mu} - u_i\boldsymbol{v}\|_2^2 \quad \text{s.t. } \boldsymbol{\mu} + u_i\boldsymbol{v} \in \mathbb{S}_{\boldsymbol{v}}^p$$

- Multi-dimensional projection

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

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

- Example with 2-dimensional simplicial subspace embedded in $\mathbb{S}^4$, where the blue cross is out of the subspace and the red cross is the projected point.

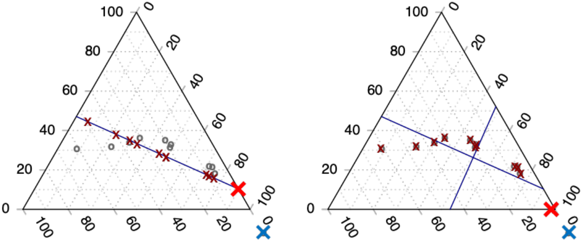


Figure: Left: one-dimensional projection; Right: multi-dimensional projection

Three types of compositional PCA

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

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

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

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

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

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

  under the appropriate compositional constraints.

- Sequential alternating minimization:
  - update $\mathbf{U}_k$ and $\mathbf{V}_k$

    (a) sequentially for $k = 2, \ldots, r$ and (b) alternately by fixing another.

Alternating algorithm: Rank-1 estimation

- Repeat the followings for $t = 0, 1, \ldots$:
    - U-update: Given $\mathbf{V}_1^{(t)}$,

$$u_{i1}^{(t+1)} = \Pi^{one}_{\mathbf{V}_1^{(t)}}(\boldsymbol{x}_i; \boldsymbol{\mu}) = \Pi^{mult}_{\mathbf{V}_1^{(t)}}(\boldsymbol{x}_i; \boldsymbol{\mu}) \quad \forall i$$

    - U-shrinkage: $\mathbf{U}_1^{(t+1)} \leftarrow (1 - \frac{\gamma}{t+1})\mathbf{U}_1^{(t+1)}$.
    - V-update: Given $\mathbf{U}_1^{(t+1)}$,

$$\mathbf{V}_1^{(t+1)} = \underset{\mathbf{V}_1 : \mathbf{V}_1 \perp \mathbf{1}_p}{\arg\min} \left\| \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \mathbf{U}_1^{(t+1)} \mathbf{V}_1^T \right\|_F^2 \quad \text{s.t. } \boldsymbol{\mu} + u_{i1}^{(t+1)} \mathbf{V}_1 \in \mathbb{S}^p \ \forall i.$$

    - V-scaling: $\mathbf{V}_1^{(t+1)} \leftarrow \mathbf{V}_1^{(t+1)} / \|\mathbf{V}_1^{(t+1)}\|_2$

  until convergence, $\|\mathbf{V}_1^{(t+1)} - \mathbf{V}_1^{(t)}\|_F^2 < \epsilon = 10^{-10}$.

- Re-estimation of $\mathbf{U}_1$:

$$u_{i1}^{(t+1)} \leftarrow \Pi^{one}_{\mathbf{V}_1^{(t+1)}}(\boldsymbol{x}_i; \boldsymbol{\mu}) = \Pi^{mult}_{\mathbf{V}_1^{(t+1)}}(\boldsymbol{x}_i; \boldsymbol{\mu}) \quad \forall i \quad \text{for a given } \mathbf{V}_1^{(t+1)}.$$

Alternating algorithm: Rank-k estimation in aCPCA

- For $(\hat{\mathbf{U}}_1, \hat{\mathbf{V}}_1, \dots, \hat{\mathbf{U}}_{k-1}, \hat{\mathbf{V}}_{k-1})$ fixed, repeat the followings for $t = 0, 1, \dots$:
  - Let $\hat{\mathbf{C}}_{k-1} = \mathbf{1}\boldsymbol{\mu}^T + \sum_{h=1}^{k-1} \hat{\mathbf{U}}_h \hat{\mathbf{V}}_h^T$.
  - U-update: Given $\mathbf{V}_k^{(t)}$,

$$u_{ik}^{(t+1)} = \Pi^{one}_{\mathbf{V}_k^{(t)}}(\boldsymbol{x}_i; \hat{\boldsymbol{c}}_{i,k-1}) \ \ \forall i.$$

  - U-shrinkage: $\mathbf{U}_k^{(t+1)} \leftarrow (1 - \frac{\gamma}{t+1})\mathbf{U}_k^{(t+1)}$.
  - V-update: Given $\mathbf{U}_k^{(t+1)}$,

$$\mathbf{V}_k^{(t+1)} = \underset{\mathbf{V}_k : \mathbf{V}_k \perp \mathbf{1}_p}{\arg\min} \left\| \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \sum_{h=1}^{k-1} \hat{\mathbf{U}}_h \hat{\mathbf{V}}_h^T - \mathbf{U}_k^{(t+1)} \mathbf{V}_k^T \right\|_F^2$$

$$\text{s.t. } \boldsymbol{\mu} + \sum_{h=1}^{k-1} \hat{u}_{ih} \hat{\mathbf{V}}_h + u_{ik}^{(t+1)} \mathbf{V}_k \in \mathbb{S}^p \ \forall i.$$

  - V-scaling: $\mathbf{V}_k^{(t+1)} \leftarrow \mathbf{V}_k^{(t+1)} / \|\mathbf{V}_k^{(t+1)}\|_2$

  until convergence, $\|\mathbf{V}_k^{(t+1)} - \mathbf{V}_k^{(t)}\|_F^2 < \epsilon = 10^{-10}$.

- Re-estimation of $\mathbf{U}_k$:

$$u_{ik}^{(t+1)} \leftarrow \Pi^{one}_{\mathbf{V}_k^{(t+1)}}(\boldsymbol{x}_i; \hat{\boldsymbol{c}}_{i,k-1}) \ \ \forall i \ \text{ for a given } \mathbf{V}_k^{(t+1)}.$$

Alternating algorithm: Rank-k estimation in CPCA

- For $(\hat{\mathbf{V}}_1, \ldots, \hat{\mathbf{V}}_{k-1})$ fixed, repeat the followings for $t = 0, 1, \ldots$:
  - U-update: Given $\mathbf{V}_k^{(t)}$,

$$\boldsymbol{u}_i^{(t+1)} = \Pi^{mult}_{\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_{k-1},\mathbf{V}_k}(\boldsymbol{x}_i; \boldsymbol{\mu}) \ \ \forall i$$

  - U-shrinkage: $[\mathbf{U}_1^{(t+1)}, \ldots, \mathbf{U}_k^{(t+1)}] \leftarrow (1 - \frac{\gamma}{t+1})[\mathbf{U}_1^{(t+1)}, \ldots, \mathbf{U}_k^{(t+1)}]$.
  - V-update: Given $\mathbf{U}_1^{(t+1)}, \ldots, \mathbf{U}_k^{(t+1)}$,

$$\mathbf{V}_k^{(t+1)} = \underset{\mathbf{V}_k : \mathbf{V}_k \perp \mathbf{1}_p}{\arg\min} \left\| \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \sum_{h=1}^{k-1} \mathbf{U}_h^{(t+1)} \hat{\mathbf{V}}_h^T - \mathbf{U}_k^{(t+1)} \mathbf{V}_k^T \right\|_F^2$$

$$\text{s.t. } \boldsymbol{\mu} + \sum_{h=1}^{k-1} u_{ih}^{(t+1)} \hat{\mathbf{V}}_h + u_{ik}^{(t+1)} \mathbf{V}_k \in \mathbb{S}^p \ \forall i.$$

  - V-scaling: $\mathbf{V}_k^{(t+1)} \leftarrow \mathbf{V}_k^{(t+1)} / \|\mathbf{V}_k^{(t+1)}\|_2$

  until convergence, $\|\mathbf{V}_k^{(t+1)} - \mathbf{V}_k^{(t)}\|_F^2 < \epsilon = 10^{-10}$.

- Re-estimation of $\mathbf{U}_1, \ldots, \mathbf{U}_k$:

$$\boldsymbol{u}_i^{(t+1)} \leftarrow \Pi^{mult}_{\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_{k-1},\mathbf{V}_k^{(t+1)}}(\boldsymbol{x}_i; \boldsymbol{\mu}) \ \ \forall i \ \ \text{for a given } \mathbf{V}_k^{(t+1)}.$$

Optimization problems

- One-dimensional projection problem

$$\arg\min_{u_i \in \mathbb{R}} \|\boldsymbol{x}_i - \boldsymbol{\mu} - u_i \boldsymbol{v}\|_2^2$$

: Closed form solution

- Multi-dimensional projection problem

$$\arg\min_{u_{i1},\ldots,u_{ik} \in \mathbb{R}^p} \|\boldsymbol{x}_i - \boldsymbol{\mu} - u_{i1}\mathbf{V}_1 - \cdots - u_{ik}\mathbf{V}_k\|_2^2$$

$$\text{subject to } \boldsymbol{\mu} + u_i 1\mathbf{V}_1 + \cdots + u_{ik}\mathbf{V}_k \in \mathbb{S}^p_{\mathbf{V}_1,\ldots,\mathbf{V}_k}$$

: Quadratic Programming (QP)

- Update of $\mathbf{V}_k$

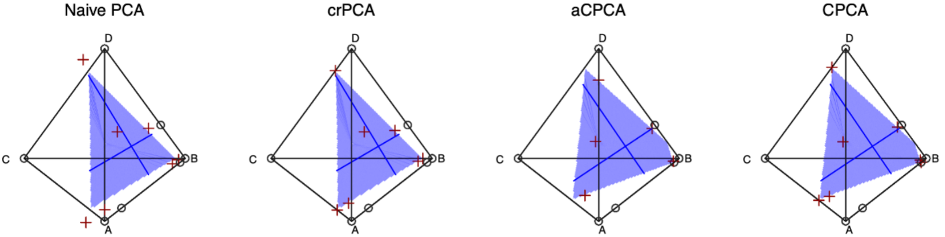
$$\mathbf{V}_k = \arg\min_{\mathbf{V}_k} \left\| \mathbf{X} - \mathbf{1}\boldsymbol{\mu}^T - \sum_{h=1}^{k-1} \mathbf{U}_h \hat{\mathbf{V}}_h^T - \mathbf{U}_k \mathbf{V}_k^T \right\|_F^2$$

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

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

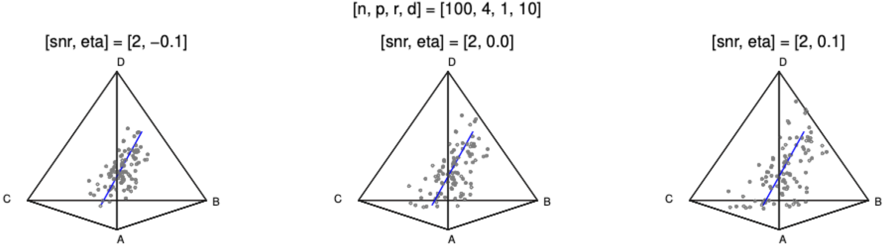
: Quadratic Programming (QP)

Comparative illustration



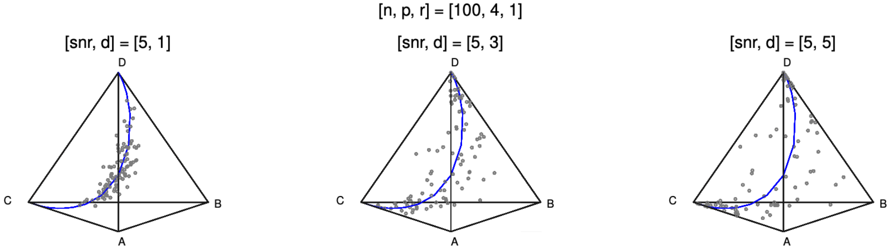
Simulation design: Linear pattern

- Centroid: $\boldsymbol{\mu} \sim Dir(10, \ldots, 10)$
- Loadings: $\mathbf{V} = Orth(\mathbf{V}^*)$ where $v^*_{jk} \sim N(0, 1)$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$, and $\mathbf{V}_1, \ldots, \mathbf{V}_r \perp \mathbf{1}_p \;\; (r = 5)$.
- Scores: $\mathbf{U} = \{u_{ik}\}$ with $u_{ik} \sim TN(0, (d/k)^2; a_k - \frac{\eta}{\log(p)}, b_k + \frac{\eta}{\log(p)})$, 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 $\mathbb{S}^p$ $(d = 10 \;\&\; \eta = 0.1)$.
- Simulated data: $\boldsymbol{x}_i = Proj_{\,\mathbb{S}^p}\left[\boldsymbol{\mu}^T + \mathbf{V}\boldsymbol{u}_i + (\mathbf{I}_p - \frac{1}{p}\mathbf{1}_p\mathbf{1}_p^T)\boldsymbol{e}_i\right]$, where $e_{ij} \sim U(-\delta, \delta)$, $Proj_{\,\mathbb{S}^p}$ is a projection operator onto a simplex, and δ was set to achive a specified SNR.



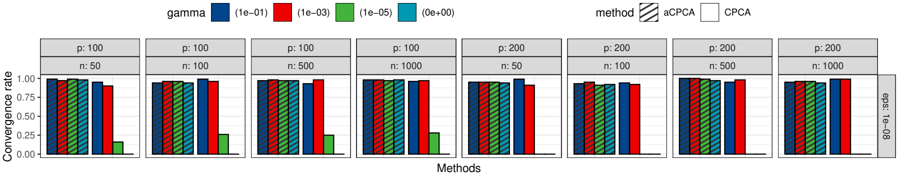
Simulation design: Curved pattern

- Centroid: $\boldsymbol{\mu} = (0, \ldots, 0)$
- Loadings: $\mathbf{V} = Orth(\mathbf{V}^*)$ where $v^*_{jk} \sim N(0,1)$, $\mathbf{V}^T\mathbf{V} = \mathbf{I}_r$, and $\mathbf{V}_1, \ldots, \mathbf{V}_r \perp \mathbf{1}_p \;\; (r = 5)$.
- Scores: $\mathbf{U} = \{u_{ik}\}$ with $u_{ik} \sim N(0, (d/k)^2)$ $(d = 3)$
- Simulated data: $\boldsymbol{x}_i = \mathcal{C}\left[\exp(\boldsymbol{\mu} + \mathbf{V}\boldsymbol{u}_i + \boldsymbol{e}_i)\right]$, where $\boldsymbol{e}_i = (e_{i1}, \ldots, e_{ip})$ with $e_{ij} \sim N(0, \sigma_e^2)$.

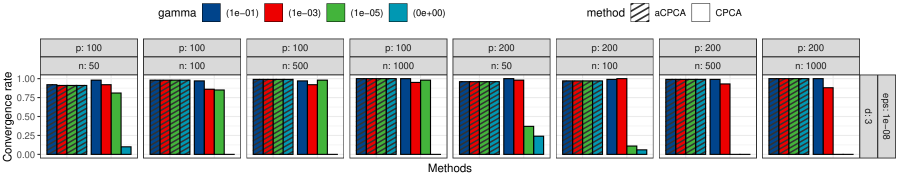


Simulation result: Convergence rate

- The proportion of cases that converged over 100 simulation replicates



Linear

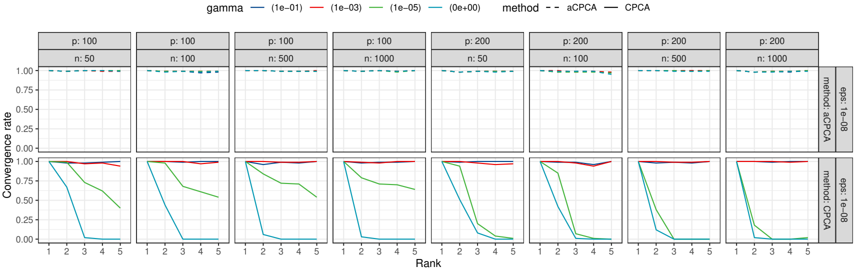


Curved

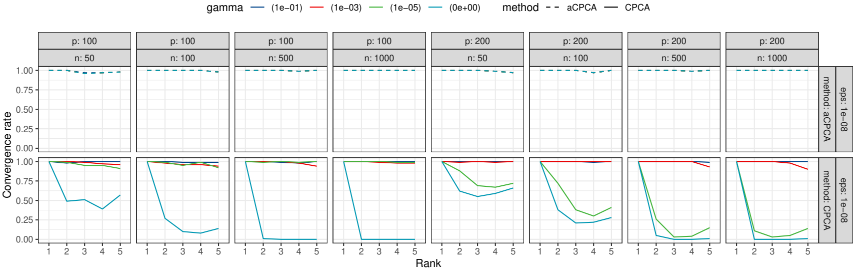
- We choose the shrinkage parameter $\gamma = 0.1$ as an optimal.

Simulation result: Convergence rate for each rank

Linear



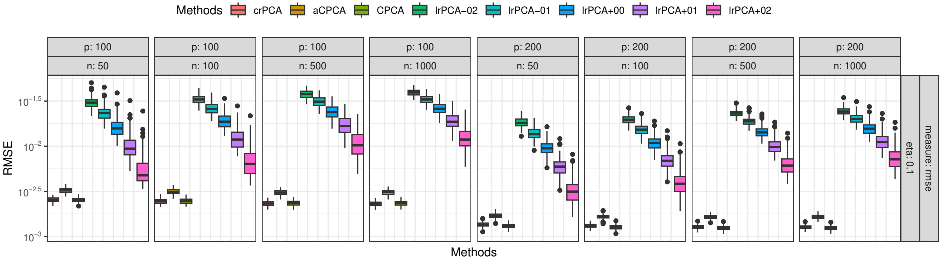
Curved



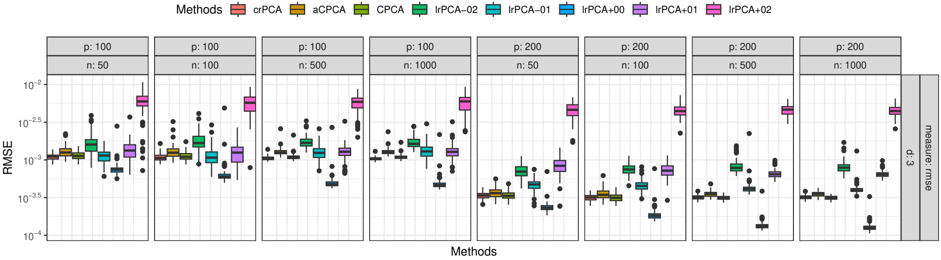
Simulation result: Estimation performance

- RMSE: $\|(\mathbf{1}\boldsymbol{\mu}^T + \mathbf{U}\mathbf{V}^T) - (\mathbf{1}\hat{\boldsymbol{\mu}}^T + \hat{\mathbf{U}}\hat{\mathbf{V}}^T)\|_F / \sqrt{np}$

Linear

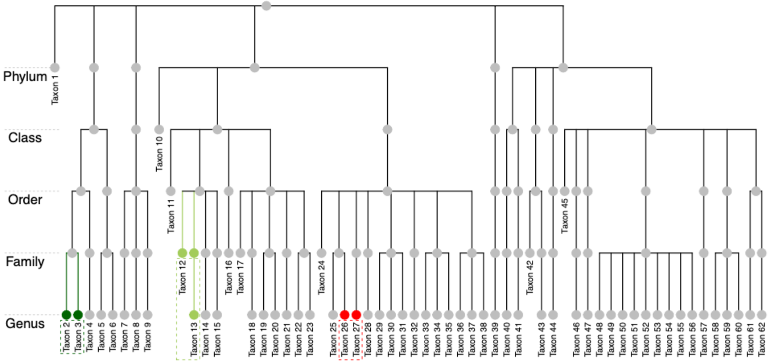


Curved



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 $p = 40, 44, 46,$ and 32, respectively.
- Microbiome data is highly sparse so that 70-76% of elements are zero.
- Data dimension (p) can vary according to the taxonomic level.



Cross-validated reconstruction error

- Reconstruction error on test: $\sqrt{\frac{1}{n_{\text{test}}p}\sum_i \|\boldsymbol{x}_i^{\text{test}} - u_{i1}^{\text{test}}\hat{\mathbf{V}}_1 - \cdots - u_{ir}^{\text{test}}\hat{\mathbf{V}}_r\|_2^2}$ where $\boldsymbol{u}_i^{\text{test}} = \Pi_{\hat{\mathbf{V}}_1,\ldots,\hat{\mathbf{V}}_r}(\boldsymbol{x}_i^{\text{test}};\hat{\boldsymbol{\mu}})$.
- 10-fold CV reconstruction error with rank=1 and rank=5.

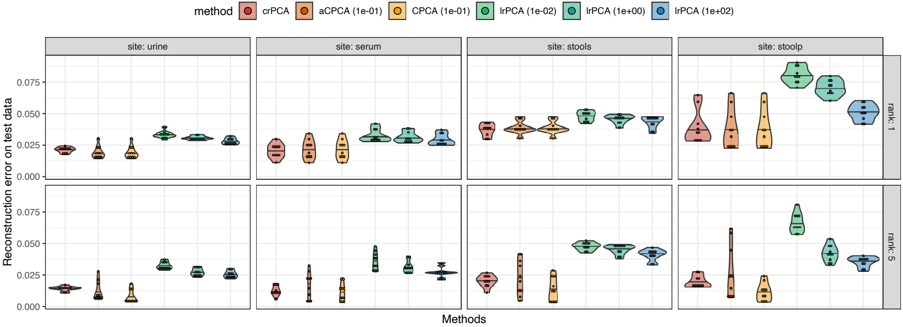
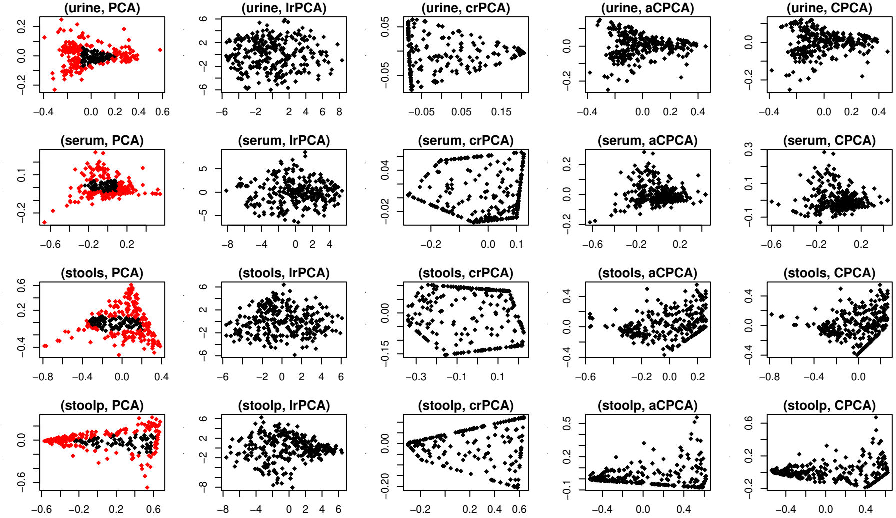


Figure: Top: rank-1; Bottom: rank-5. The lrPCA methods with zero-replacement value $\frac{1}{100}\delta_{(0)}, \delta_{(0)}, 100\delta_{(0)}$ are denoted by lrPCA (1e-02), lrPCA (1e+00), lrPCA (1e+02), respectively.

Real data analysis: The first two PC scores



- The red points represents the samples out of a simplex.

Real data analysis: Compositional plot

- Compositional plot for the rank-1 reconstructed data.

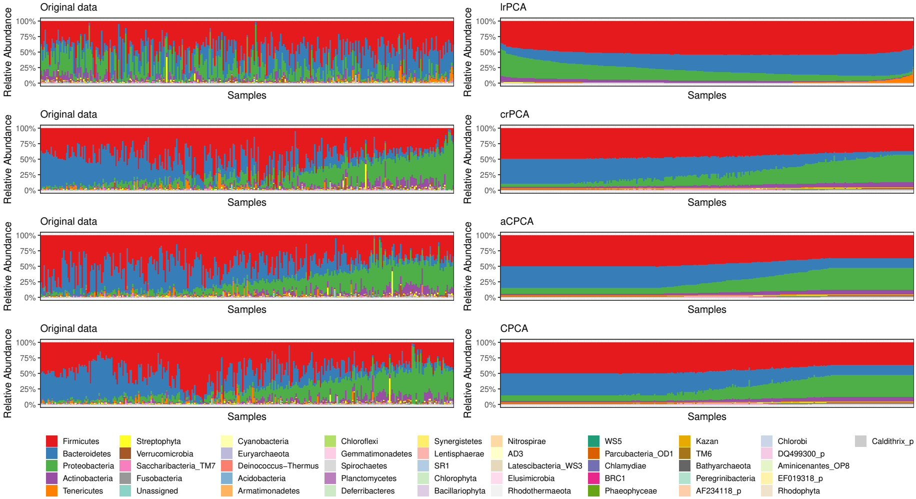


Figure: Right: reconstructed data by PCA; Left: original data. The samples of both the right and left panels are sorted in the same order based on the first PC score.

Conclusion

- In this work, we proposed three types of compositional PCA based on the projection onto the simplicial subspace.
- They performed better than the existing log-ratio PCA in the presence of linear pattern in zero-inflated data.
- Although the proposed optimization problem is clearly non-convex, the convergence is empirically guaranteed by a proper shrinkage parameter.
- We will show the existence and consistency of the simplicial subspace. Furthermore, we are also interested in the robust compositional PCA for future research.

Thank you for your attention ! ☺