A framework for multi-view analysis of microbiome data

Jing Ma

Fred Hutchinson Cancer Research Center
Public Health Sciences Division

2 July 2019

Collaborators

FRED HUTCH
CURES START HERE™

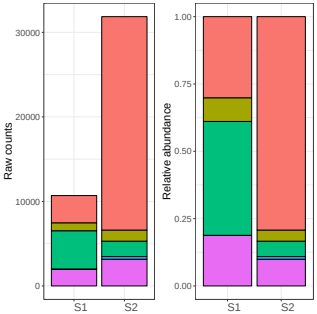
Yue Wang

Ali Shojaie

Tim Randolph

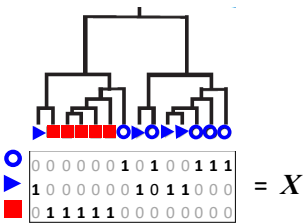
Microbiome Data

- $\boldsymbol{X} = (x_{ij})_{n \times p}$ matrix of microbiome data for n samples and p taxa
- Due to sample differences, often work with relative abundances



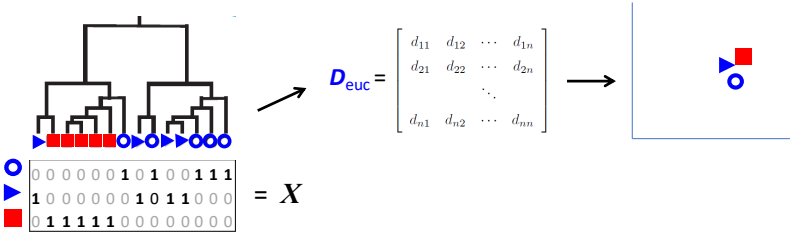
Microbiome Data

- Similarities among samples better captured by phylogenetic tree
- Many methods for capturing phylogenetic distances, e.g. UniFrac dist.



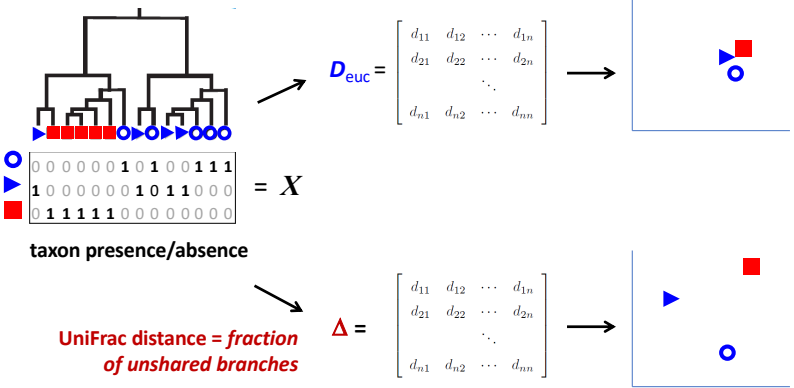
Microbiome Data

- Similarities among samples better captured by phylogenetic tree
- Many methods for capturing phylogenetic distances, e.g. UniFrac dist.



Microbiome Data

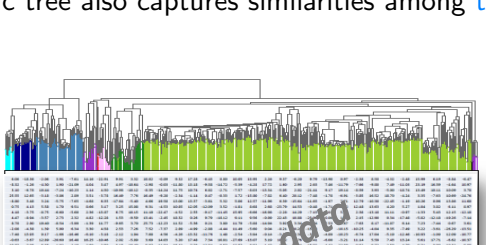
- Similarities among samples better captured by phylogenetic tree
- Many methods for capturing phylogenetic distances, e.g. UniFrac dist.



Microbiome Data

- Similarities among samples better captured by phylogenetic tree
- The phylogenetic tree also captures similarities among taxa.

$\boldsymbol{Q}$ =

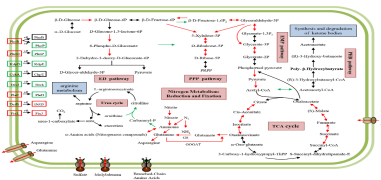


$\boldsymbol{X}$ =

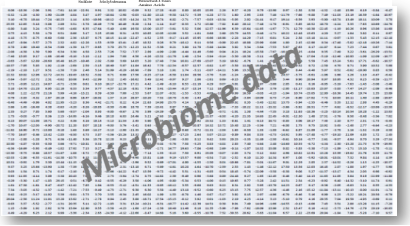
Microbiome Data

- Similarities among samples better captured by phylogenetic tree
- Alternatively, can consider information from metabolic pathways.

$\boldsymbol{Q}$ =



$\boldsymbol{X}$ =



Microbiome Data Analysis

FRED HUTCH
CURES START HERE™

- Need to capture
  - Similarities among samples – non-Euclidean (e.g. UniFrac distance)
  - Similarities among taxa – phylogenetic tree, pathway information, etc.

Microbiome Data Analysis

FRED HUTCH
CURES START HERE™

- Need to capture
  - Similarities among samples – non-Euclidean (e.g. UniFrac distance)
  - Similarities among taxa – phylogenetic tree, pathway information, etc.
- Often use exploratory data analysis tools
  - PCoA (aka MDS)
  - DPCoA (Double PCoA)

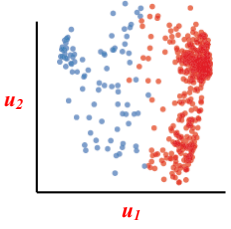
Exploratory Analysis using PCoA

First recall PCA

$$X = \begin{bmatrix} x_1^1 & x_1^2 & \cdots & x_1^p \\ x_2^1 & x_2^2 & \cdots & x_2^p \\ x_3^1 & x_3^2 & \cdots & x_3^p \\ & & \ddots & \\ x_n^1 & x_n^2 & \cdots & x_n^p \end{bmatrix}$$

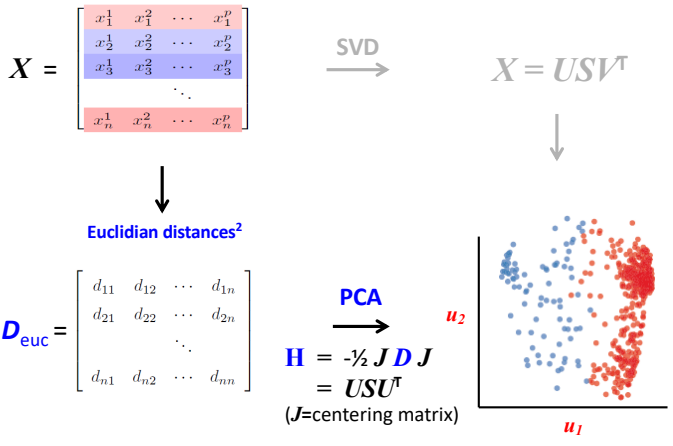
$\xrightarrow{\text{SVD}}$ $X = USV^T$

n samples (rows)
p variables (cols)



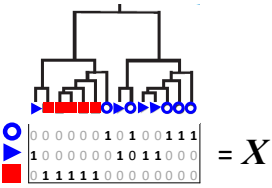
Exploratory Analysis using PCoA

First recall PCA



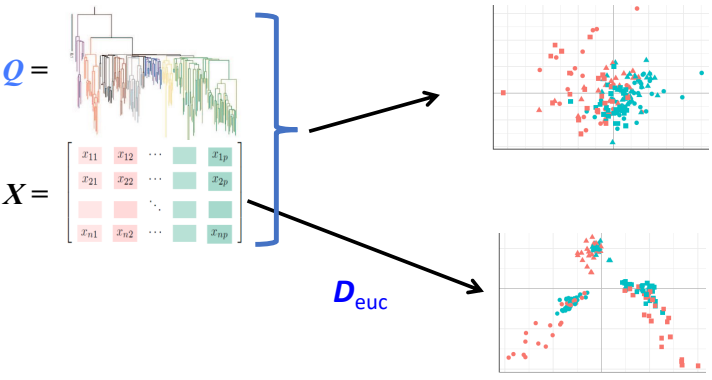
Exploratory Analysis using PCoA

UniFrac PCoA



Exploratory Analysis using DPCoA

DPCoA[1] based on phylogenetic tree



[1] Pavoine et al. 2004; Purdom. AOAS, 2011

Beyond Exploratory Analysis: Important Variables

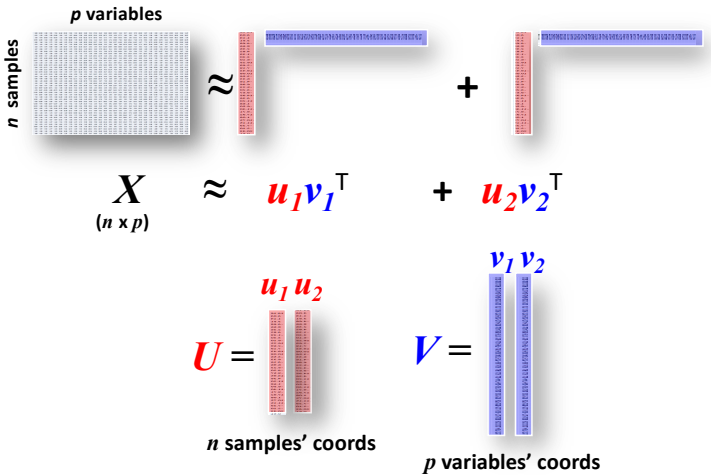
FRED HUTCH
CURES START HERE™

Biplots provide simultaneous visualization of samples and variables

Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

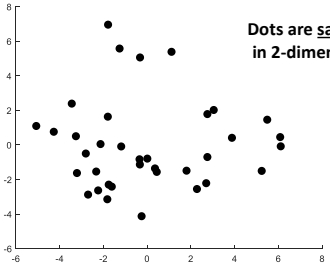
Biplots provide simultaneous visualization of samples and variables



Beyond Exploratory Analysis: Important Variables

FRED HUTCH CURES START HERE™

Biplots provide simultaneous visualization of samples and variables

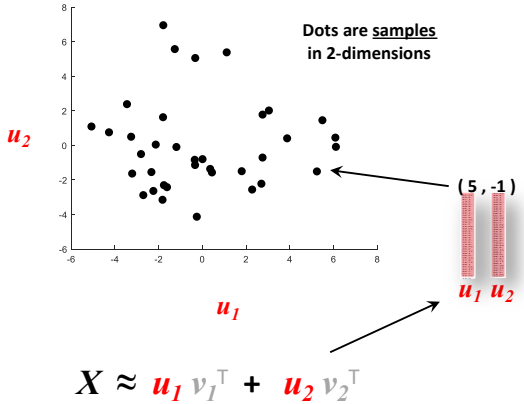


$$X \approx u_1 v_1^\top + u_2 v_2^\top$$

Beyond Exploratory Analysis: Important Variables

FRED HUTCH CURES START HERE™

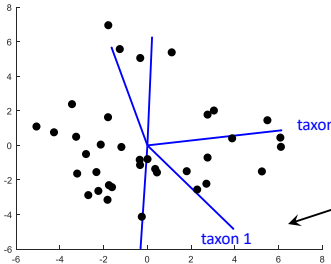
Biplots provide simultaneous visualization of samples and variables



Beyond Exploratory Analysis: Important Variables

FRED HUTCH CURES START HERE™

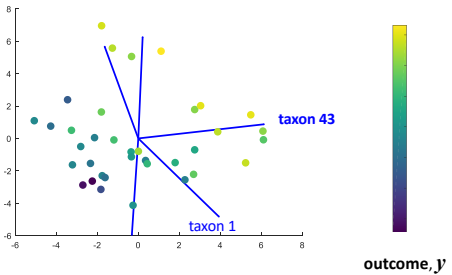
Biplots provide simultaneous visualization of samples and variables



Beyond Exploratory Analysis: Important Variables

Biplots provide simultaneous visualization of samples and variables

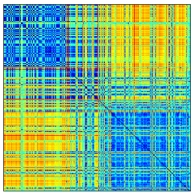
- Can also overlay outcome



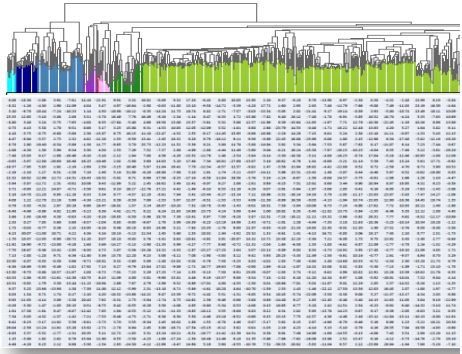
$$X \approx u_1 v_1^\top + u_2 v_2^\top$$

Microbiome Data are Doubly Structured

$$\mathbf{H} = -\tfrac{1}{2}\, \boldsymbol{J}\, \boldsymbol{\Delta}\, \boldsymbol{J}$$



$$\boldsymbol{Q} =$$



Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

Can we draw a biplot that accounts for the external structures?

[2]Satten et al. PLOS One, 2017

Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

Can we draw a biplot that accounts for the external structures?

- Unfortunately, there is no biplot with PCoA because
  - PCA uses $\boldsymbol{X}$ and gives: $\boldsymbol{X} = \boldsymbol{USV}^\intercal$ (SVD)
  - PCoA uses $\boldsymbol{\Delta}$ and only gives: $\boldsymbol{US}^2\boldsymbol{U}^\intercal$ (no $\boldsymbol{V}$)

[2]Satten et al. PLOS One, 2017

Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

Can we draw a biplot that accounts for the external structures?

- Unfortunately, there is no biplot with PCoA because
  - PCA uses $\boldsymbol{X}$ and gives: $\boldsymbol{X} = \boldsymbol{USV}^\intercal$ (SVD)
  - PCoA uses $\boldsymbol{\Delta}$ and only gives: $\boldsymbol{US}^2\boldsymbol{U}^\intercal$ (no $\boldsymbol{V}$)
- Existing approaches[2] in the field are approximate/add hoc.

[2] Satten et al. PLOS One, 2017

Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

Can we draw a biplot that accounts for the external structures?

- SVD gives $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}$ by solving

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\mathsf{T}\|_F$$

where $\|A\|_F = \text{trace}(A^\mathsf{T}A)$.

[3] Allen et al. JASA, 2014

Beyond Exploratory Analysis: Important Variables

FRED HUTCH
CURES START HERE™

Can we draw a biplot that accounts for the external structures?

- SVD gives $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top$ by solving

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\|_F$$

  where $\|A\|_F = \text{trace}(A^\top A)$.

- Consider instead a general norm to incorporate $\boldsymbol{H}$ and $\boldsymbol{Q}$:

$$\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\top\|_{\boldsymbol{H},\boldsymbol{Q}}$$

  where $\|A\|_{\boldsymbol{H},\boldsymbol{Q}} = \text{trace}(A^\top \boldsymbol{H} A \boldsymbol{Q})$.

[3] Allen et al. JASA, 2014

Beyond Exploratory Analysis: Important Variables

FRED HUTCH CURES START HERE™

Can we draw a biplot that accounts for the external structures?

- SVD gives $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\intercal$ by solving

$$\underset{\boldsymbol{U},\boldsymbol{S},\boldsymbol{V}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\intercal\|_F$$

  where $\|A\|_F = \text{trace}(A^\intercal A)$.

- Consider instead a general norm to incorporate $\boldsymbol{H}$ and $\boldsymbol{Q}$:

$$\|\boldsymbol{X} - \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^\intercal\|_{\boldsymbol{H},\boldsymbol{Q}}$$

  where $\|A\|_{\boldsymbol{H},\boldsymbol{Q}} = \text{trace}(A^\intercal \boldsymbol{H} A \boldsymbol{Q})$.

- The GMD (Gen'zd Matrix Decomp[3]) gives $\boldsymbol{X} = \mathcal{U}\mathcal{S}\mathcal{V}^\intercal$ such that $\mathcal{U}^\intercal \boldsymbol{H} \mathcal{U} = \mathcal{V}^\intercal \boldsymbol{Q} \mathcal{V} = I_K$, and $\mathcal{S}$ is the diagonal matrix of GMD values.

[3] Allen et al. JASA, 2014

Variance Explained

SVD gives

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}$$

GMD gives

$$\boldsymbol{X} = \mathcal{U}\mathcal{S}\mathcal{V}^{\mathsf{T}}$$

$\boldsymbol{S}$ and $\mathcal{S}$ are diagonal matrices of weights:

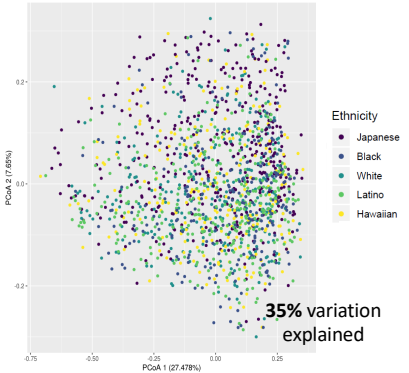
$$\boldsymbol{S} = \begin{pmatrix} \sigma_1 & & \\ & \sigma_2 & \\ & & \sigma_3 \end{pmatrix} \quad \text{(decreasing: } \sigma_1 > \sigma_2 > \sigma_3\text{)}$$

$\Rightarrow$ These quantify "variance explained" by components 1, 2, 3, ...

GMD Biplot

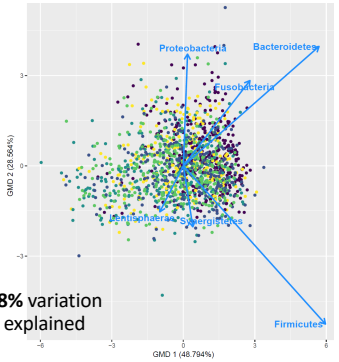
The GMD-biplot displays samples and variables using columns of $\mathcal{U}$ and $\mathcal{V}$

PCoA using UniFrac $\mathbf{\Delta}$



35% variation explained

PCoA using $\boldsymbol{X}$ and UniFrac $\mathbf{\Delta}$; phyla (arrows) from **GMD biplot**



78% variation explained

Supervised Learning with GMD: GMDR

FRED HUTCH
CURES START HERE™

- GMD generalizes SVD for doubly structured data
- Can thus use GMD for supervised learning, similar to PCR

Principal Component Regression

FRED HUTCH
CURES START HERE™

- Linear model $y = \boldsymbol{X}\beta + \varepsilon$
- Using two dimensions

$$y \approx \gamma_1 u_1 + \gamma_2 u_2 + \varepsilon = \boldsymbol{U}_{(2)} \boldsymbol{S}_{(2)} \boldsymbol{V}_{(2)}^{\mathsf{T}} \beta + \varepsilon$$

- Coefficient

$$\hat{\beta}_{PCR} = \boldsymbol{V}_{(2)} \boldsymbol{S}_{(2)}^{-1} \boldsymbol{U}_{(2)}^{\mathsf{T}} y$$

Ridge Regression

- Linear model $y = \boldsymbol{X}\beta + \varepsilon$
- Using all p dimensions

$$y = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathsf{T}}\beta + \varepsilon$$

- Coefficient

$$\hat{\beta}_{Ridge} = \boldsymbol{V}W_{\lambda}\boldsymbol{S}^{-1}\boldsymbol{U}^{\mathsf{T}}y,$$

where W_{λ} is a diagonal matrix of weights and λ a tuning parameter.

GMD Regression

- Linear model $y = \boldsymbol{X}\beta + \varepsilon$
- Incorporating $\boldsymbol{H}$ and $\boldsymbol{Q}$

$$y = \mathcal{U}\mathcal{S}\mathcal{V}^{\mathsf{T}}\beta + \varepsilon$$

- Coefficient

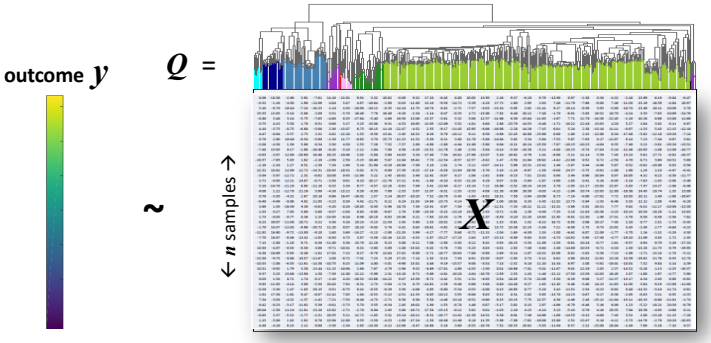
$$\hat{\beta}_{GMD} = \boldsymbol{Q}\mathcal{V}\mathcal{W}\mathcal{S}^{-1}\mathcal{U}^{\mathsf{T}}\boldsymbol{H}y,$$

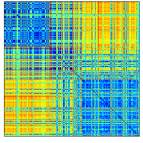
where $\mathcal{W}$ is a diagonal matrix of weights:

- $\mathcal{W}_j = \mathbf{1}_{j\in\mathcal{J}} \rightarrow \hat{\beta}_{GMDR}(\mathcal{J}), \mathcal{J} \subset \{1, \ldots, p\}$
- $\mathcal{W} = \mathcal{S}^2(\mathcal{S}^2 + \lambda I_n)^{-1} \rightarrow \hat{\beta}_{KPR} = \arg\min_\beta\{\|y - \boldsymbol{X}\beta\|^2_{\boldsymbol{H}} + \lambda\|\beta\|^2_{\boldsymbol{Q}^{-1}}\}$[4]

[4] Randolph et al. AOAS, 2018

Summary

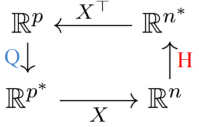




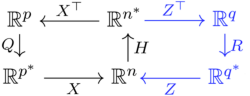
$$\mathbf{H} = -\tfrac{1}{2}\, \boldsymbol{J}\, \boldsymbol{\Delta}\, \boldsymbol{J}$$

Summary

- Discussed visualization and regression based on GMD
- GMD is closely related to the duality[5] between viewing the data from the perspective of samples and variables



- Our framework
  - encompasses classical methods, both unsupervised (PCA, PCoA/MDS, biplots) and supervised (ridge, GLS) methods
  - extends them to non-standard settings (multi-view data)



[5] Escoufier. 1977; de la Cruz and Holmes. AOAS, 2011