Learning from Human Microbiome

Jing Ma

Statistics, Texas A&M

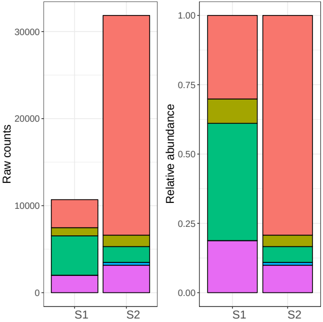
7 February 2020

Human Microbiome

Credit: Antoine Doré

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



Scientific Questions

Exploratory analysis

- Dimension reduction (Ordination)
- Microbial interactions
- Controlling batch effects
- ...

Supervised learning

- Is the microbiota associated with an outcome?
- Which taxa are associated with an outcome?
- ...

Scientific Questions

Exploratory analysis

- Dimension reduction (Ordination)
- Microbial interactions
- Controlling batch effects
- ...

Supervised learning

- Is the microbiota associated with an outcome?
- Which taxa are associated with an outcome?
- ...

Dimension Reduction

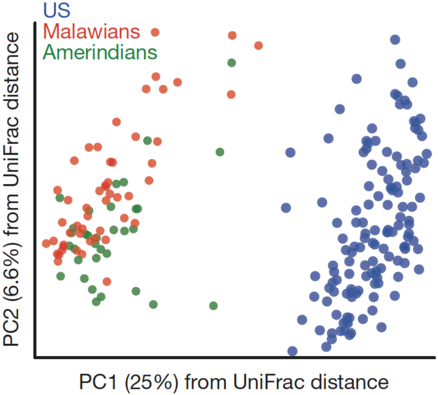


Figure: PCoA of unweighted unifrac distances for the fecal microbiota of adults[1]

[1] Yatsunenko et al. Nature, 2012

Dimension Reduction

Existing methods:

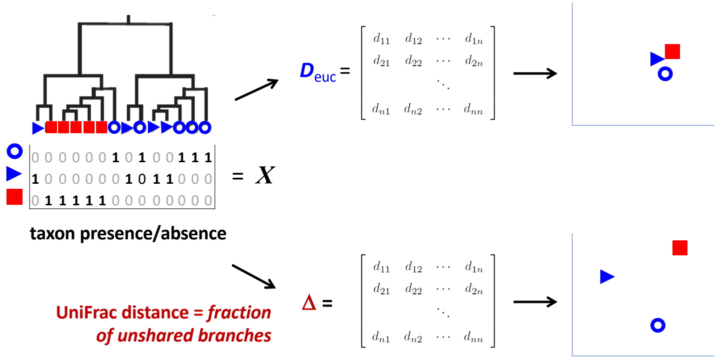
- PCoA (aka MDS), DPCoA

Limitations:

- unable to visualize both samples and variables.
- unable to account for two-way structures.

Two-way Structures

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



○	0	0	0	0	0	0	1	0	1	0	0	1	1	1
▶	1	0	0	0	0	0	0	1	0	1	1	0	0	0
■	0	1	1	1	1	1	0	0	0	0	0	0	0	0

= X

taxon presence/absence

$$D_{euc} = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ & & \ddots & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

UniFrac distance = *fraction of unshared branches*

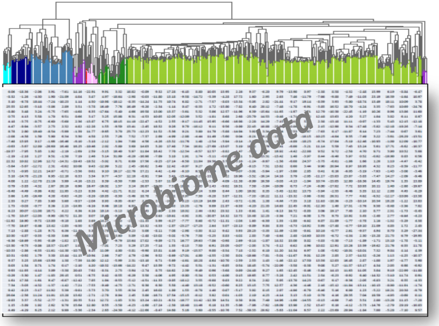
$$\Delta = \begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1n} \\ d_{21} & d_{22} & \cdots & d_{2n} \\ & & \ddots & \\ d_{n1} & d_{n2} & \cdots & d_{nn} \end{bmatrix}$$

Two-way Structures

- The phylogenetic tree also captures similarities among taxa. Alternatively, can consider information from metabolic pathways.

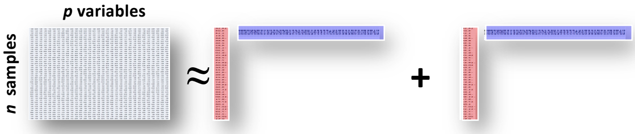
Q =

X =

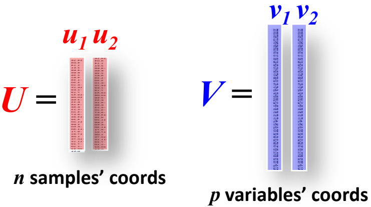


GMD Biplot

First recall PCA biplot

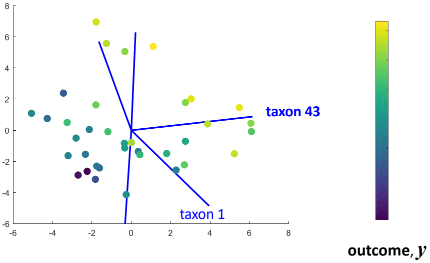


$$\underset{(n \times p)}{X} \approx u_1 v_1^{\mathsf{T}} + u_2 v_2^{\mathsf{T}}$$



GMD Biplot

First recall PCA biplot



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

GMD Biplot

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

[2] Allen et al. JASA, 2014

GMD Biplot

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

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

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

[2] Allen et al. JASA, 2014

GMD Biplot

- 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 $\|A\|_F = \mathrm{trace}(A^{\intercal}A)$.

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

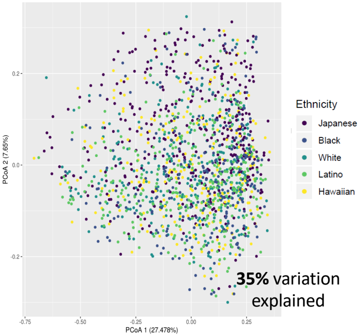
- The GMD (Gen'zd Matrix Decomp[2]) gives $\boldsymbol{X} = \mathcal{U}\mathcal{S}\mathcal{V}^{\intercal}$ such that $\mathcal{U}^{\intercal}\boldsymbol{H}\mathcal{U} = \mathcal{V}^{\intercal}\boldsymbol{Q}\mathcal{V} = I_K$, and $\mathcal{S}$ is the diagonal matrix of GMD values.

[2] Allen et al. JASA, 2014

GMD Biplot

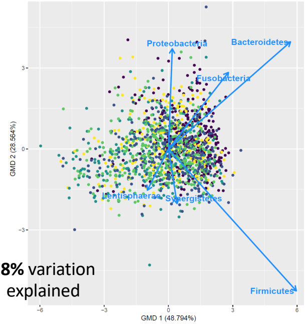
The GMD-biplot[3] displays samples and variables using columns of $\mathcal{U}$ and $\mathcal{V}$

PCoA using UniFrac $\boldsymbol{\Delta}$



35% variation explained

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



78% variation explained

[3] Yue et al. mSystems, 2019

Supervised Learning with GMD

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

GMD Regression and Inference

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

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

- Coefficient

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

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

[5] Yue et al. Submitted, 2020

GMD Regression and Inference

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

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

- Coefficient

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

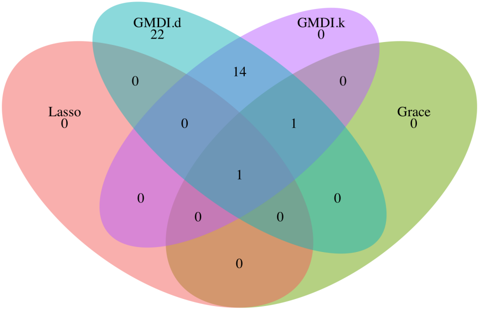
  where $\mathcal{W}$ is a diagonal matrix of weights:
  - $\mathcal{W}_j = \mathbf{1}_{j \in \mathcal{J}} \rightarrow \hat{\beta}_{GMDR}(\mathcal{J}), \mathcal{J} \subset \{1, \ldots, p\}$
  - $\mathcal{W} = \mathcal{S}^2(\mathcal{S}^2 + \lambda I_n)^{-1} \rightarrow \hat{\beta}_{KPR} = \arg\min_{\beta}\{\|y - \boldsymbol{X}\beta\|^2_{\boldsymbol{H}} + \lambda\|\beta\|^2_{\boldsymbol{Q}^{-1}}\}^4$
- GMD inference (GMDI[5]) $H_{0,j} : \beta^{\star}_j = 0$.

[4] Randolph et al. AOAS, 2018

[5] Yue et al. Submitted, 2020

Application to Yatsunenko Data

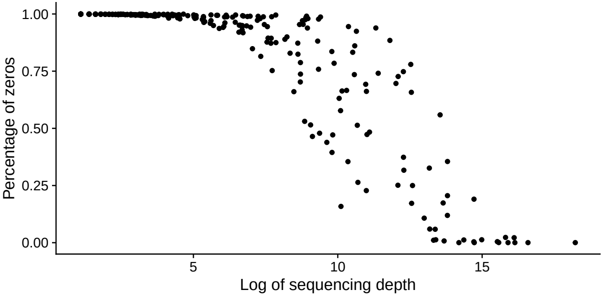
- Which bacteria are associated with age?
- Significant associations from multivariate methods[6] (FDR=0.1)



[6] Ridge test by Bühlmann (2013) returns 0 sig association.

Open Questions: Missing Data

- Microbiome data are zero-inflated.
- Zeros are not missing at random.



Open Questions: Microbial Network Analysis

We previously worked on constructing microbial co-occurrence network from presence/absence data[7].

[7] Cai et al. Biometrika, 2019

Open Questions: Microbial Network Analysis

We previously worked on constructing microbial co-occurrence network from presence/absence data[7].

- How to define dependence between two taxa?
- Marginal vs. conditional?
- How to jointly analyze microbiome and metabolomic data?

[7] Cai et al. Biometrika, 2019

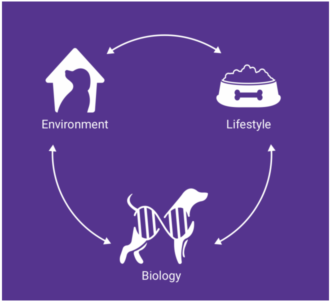
Open Questions: Interaction Testing

Microbiome by environment interaction

$$y_i = \alpha_0 + \alpha' E_i + \beta' G_i + \gamma E_i G_i + \epsilon_i,$$

- E_i: low-dim covariates.
- G_i: high-dim genetic markers.
- Interest in testing whether the interaction $\gamma = 0$.
- Existing variance components test fails to control type I error if G_i is high-dimensional.

The Dog Aging Project



- Co-led by University of Washington and Texas A&M University.
- To understand how genes, lifestyle, and environment influence aging.
- Multiple data types: survey data, electronic medical records, omics data, etc.

References

- Differential Markov random field analysis with an application to detecting differential microbial community networks. *Biometrika.* 2019
- The GMD-biplot and its application to microbiome data. *mSystems.* 2019
- Generalized matrix decomposition: estimation and inference for two-way structured data. 2020+

Thank You!

GitHub: drjingma / Website: drjingma.com