Hidden Variables and Uncertainty Quantification: Challenges for the Microbiome

Susan Holmes
@SherlockpHolmes

Pune, December 18, 2019

LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

Part I

Heterogeneity

Heterogeneous data

Homogeneous data are all alike;
all heterogeneous data are
heterogeneous
in their own way.

Heterogeneity of Data

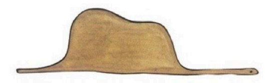
- Status : response/ explanatory.
- Hidden (latent)/measured.
- Types :
  - Continuous, Binary, categorical
  - Graphs/ Trees
  - Images, Spatial Information
- Amounts of dependency: multivariate (co-dependent) features, independent/time series/spatial.
- Different technologies and situations (low biomass, Illumina, MassSpec, RNA-seq, Imaging, CyTOF).

Challenges when working with Longitudinal Multidomain data

- Data Quality : Heterogeneity, unwanted sources of variation.
- Building models from the data.
- Interpretation of analytic output.
- Multiple (dual) dependencies.
- Multidomain, need for registration.
- Uncertainty quantification and inference.
- Reproducibility of results across labs, experimental conditions and users.

Paths in thinking about these heterogeneous systems

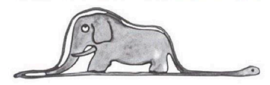
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

Paths in thinking about these heterogeneous systems

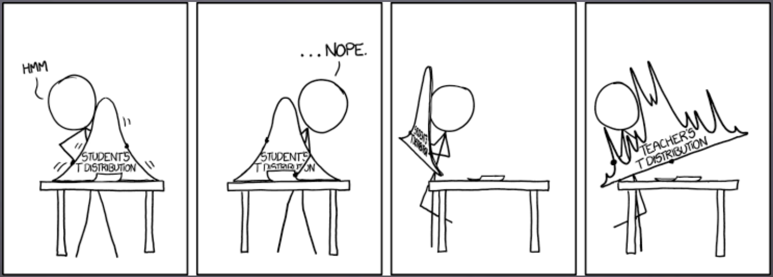
- Think in layers: latent variables or factors enable interpretation.



hidden variables.

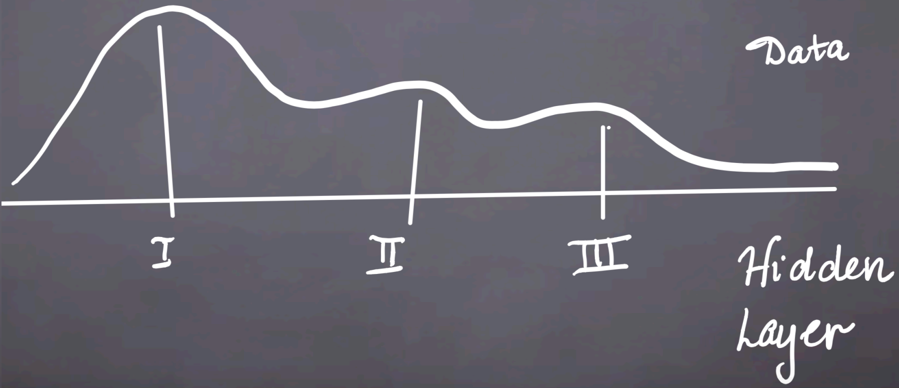
Paths in thinking about these heterogeneous systems

- Think in terms of mixtures (not one parametric population).



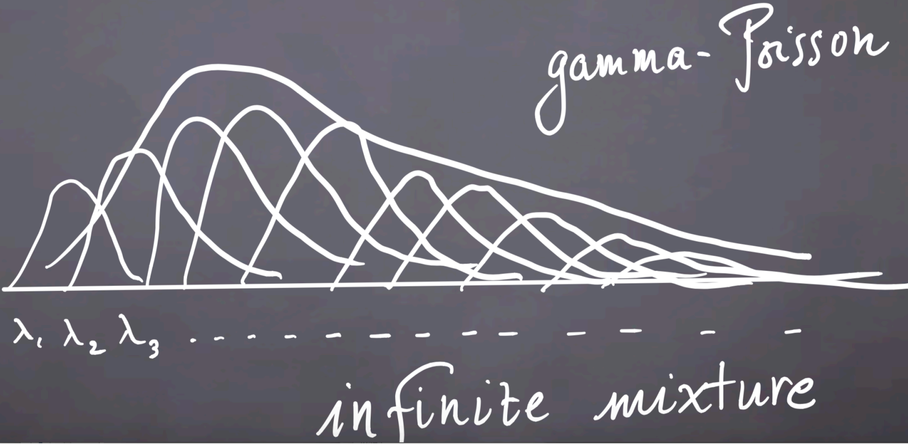
Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.



Paths in thinking about these heterogeneous systems

- Think in layers: latent variables or factors enable interpretation.


gamma-Poisson
λ_1 λ_2 λ_3 . . . - - - - - - - - - - - - - -
infinite mixture

The Yoda of Silicon Valley

“premature optimization is the root of all evil in coding”

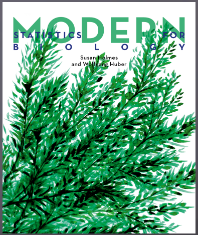
In Statistics

"premature summarization is the root of all evil in statistics"

In Statistics

Statistics: separate the model from the data

$$\begin{tikzcd}
 & \mathcal{F}_{\hat{\theta}} \arrow[dl, bend right] & \\
\mathcal{P}_{\hat{\theta}} & & \hat{\theta} \arrow[ul, bend right] \\
 & \boxed{\mathcal{X}} \arrow[ur, bend right] &
\end{tikzcd}$$



See a complete book:

`http://bios221.stanford.edu/book/`

Read data are counts, the data are not compositional.... *the parameters are!*

- After perturbations amounts of bacteria go up & down.
- Remove contaminants using read numbers (`decontam` and `BARBIE`).
- Estimating depth bias requires read numbers.
- Some bacteria live in symbiosis with others.
- We need the read depths for variability/standard error estimation and uncertainty quantification.
- Data transformations can be used to remove ”multiplicative error” and equalize the variance.

The relative abundances of bacteria and their differences. Different taxa are identified as Amplicon Strain Variant (ASV) generated with **DADA2** (Callahan et al., 2017).

$\theta_{treat} = (p_1, p_2, \dots \, p_J)$ 1 .... J = # ASV's

$\theta_{ctl} = (p_1^c, p_2^c, \dots \, p_J^c)$

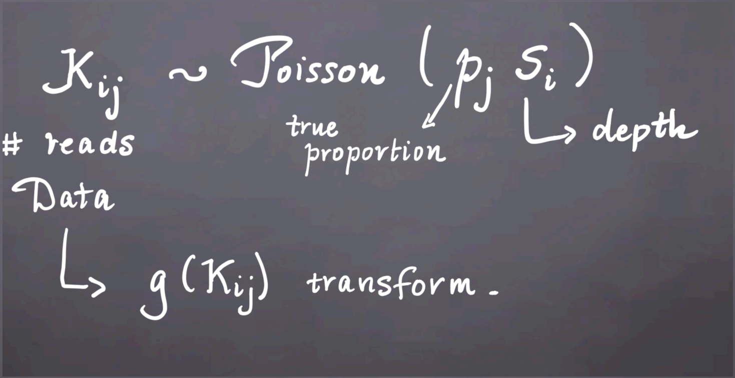
difference $\theta_{treat} - \theta_{ctl}$

$\mathcal{X} \longrightarrow \widehat{\text{diff}} \pm \, ?$ Uncertainty Quantification.

Data

Ecology meets the clinic.

Models for sequencing reads in a microbiome study?



$K_{ij} \sim \text{Poisson}\,(p_j\, s_i)$

$\#$ reads true proportion ↳ depth

Data

↳ $g(K_{ij})$ transform.

Models for sequencing reads in a microbiome study?

β_j: Ben's factor

$$K_{ij} \sim \text{Poisson}(\underbrace{p_j}_{\text{true proportion}} \overset{\beta_j}{\downarrow} \underbrace{S_i}_{\text{depth}})$$

# reads

Data

↳ $g(K_{ij})$ transform.

Errors can be multiplicative and additive

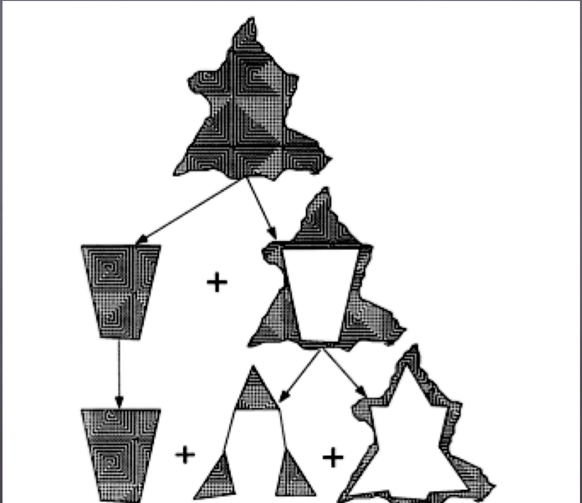
Use a ”hybrid transformation” such as `arcsinh` after removal of library depth factor.

$$\text{asinh}(x) = \log\left(x + \sqrt{x^2 + 1}\right).$$

From this we can see that for large values of x, asinh(x) behaves like the log and is practically equal to log(x) + log(2); for small x the function is close to linear in x.

Part II

Decomposition of Variability



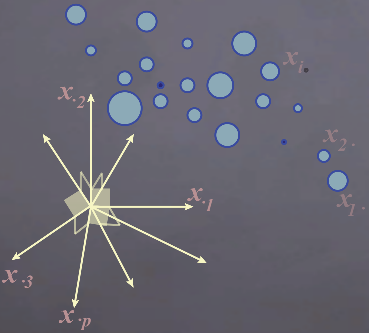
Geometric Approach using Embeddings

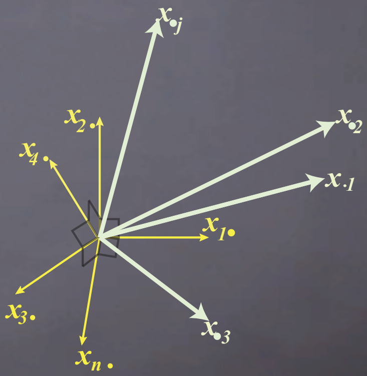
Sample data can often be seen as points in a state space.

$\mathbb{R}^p$

Variables are ‘vectors’ in data point space

$\mathbb{R}^n$





$x^t Q y = < x, y >_Q$

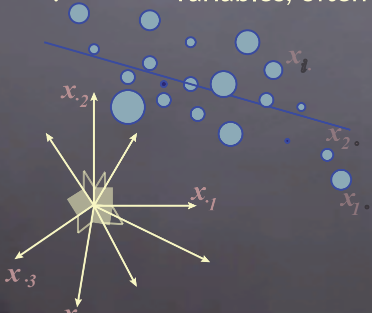
$x^t D y = < x, y >_D$

Duality : Transposable data.

Data Analysis: Geometrical Approach

i. The data are p variables measured on n observations.
ii. X with n rows (the observations) and p columns (the variables).
iii. D is an $n \times n$ matrix of weights on the “observations”, which is most often diagonal but not always.
iv Symmetric definite positive matrix Q, weights on . variables, often

$$Q = \begin{pmatrix} \frac{1}{\sigma_1^2} & 0 & 0 & 0 & \ldots \\ 0 & \frac{1}{\sigma_2^2} & 0 & 0 & \ldots \\ 0 & 0 & \ddots & 0 & \ldots \\ \vdots & \ldots & \ldots & 0 & \frac{1}{\sigma_p^2} \end{pmatrix}.$$



Matrix Decomposition for Dimension Reduction and Embedding

PCA seeks to replace the original (centered) matrix X by a matrix of lower rank, this can be solved using the singular value decomposition of X:

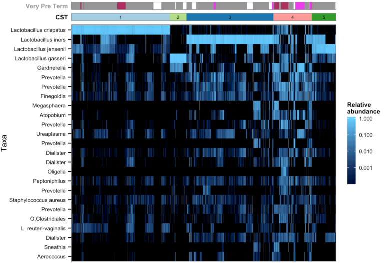
$$X = USV', \text{ with } U'DU = I_n \text{ and } V'QV = I_p \text{ and } S \text{ diagonal}$$

$$XX' = US^2U', \text{ with } U'DU = I_n \text{ and } S^2 = \Lambda$$

PCA is a linear nonparametric multivariate method for dimension reduction. D and Q are the relevant metrics on the dual row and column spaces of n samples and p variables.
Remember: X cannot be of a rank larger than $\min(n, p)$.

Part III

Hidden communities



Pregnancy data: perturbation, stability and preterm

A case-control study of 49 pregnant women:

- 15 delivered preterm.
- From 40 of these women: 3,766 specimens collected weekly during gestation, and monthly after delivery.
- Sites:vagina, distal gut, saliva, and tooth/gum.
- 9 women: validation set collected after the first study was complete.

Methods used: variance stabilization through negative binomial, testing perturbations through linear mixed-effects modeling. Preterm prediction through medoid-based clustering and simple Markov chain. Provided: Simple community temporal trends, community structure, and vaginal community state transitions.

Steps in the analysis

- Variability decomposition.
- Finite State Markov chains.
- Differential abundance testing provides biomarkers for preterm birth.

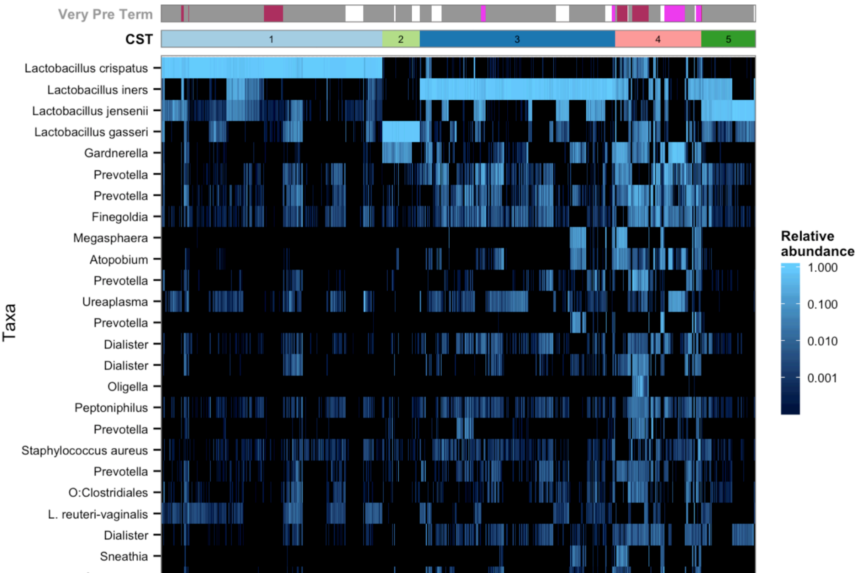
This work involves data collected by D Relman and Dan DiGiulio. Software developed by PJ McMurdie and statistical analyses done jointly with Ben Callahan.

Questions asked?

- Are the community state types the same as seen in previous studies?
- How stable are the communities within each individual during pregnancy?
- What alterations of the vaginal microbiome predict preterm birth?
- How early do these alterations occur?

Previously known Community State Types

Checked clustering of samples into community types.



Markov Chain Model

Transitions between states, as in simple ecological models.

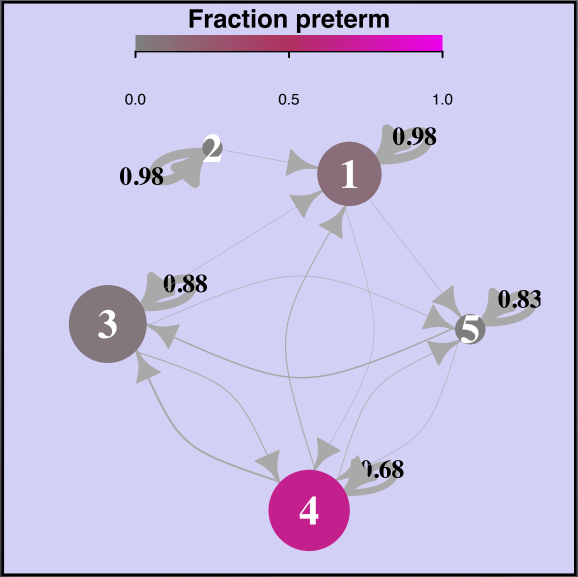
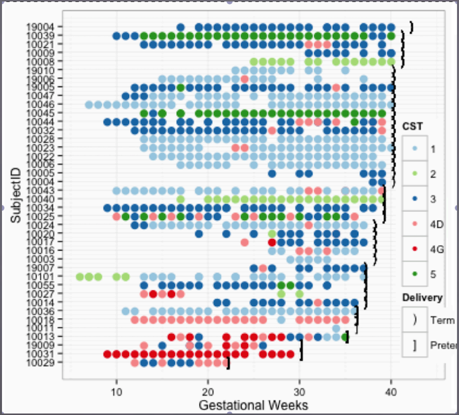


Illustration through Analyses



- Delivery Perturbation
- Preterm Prediction
- Stability

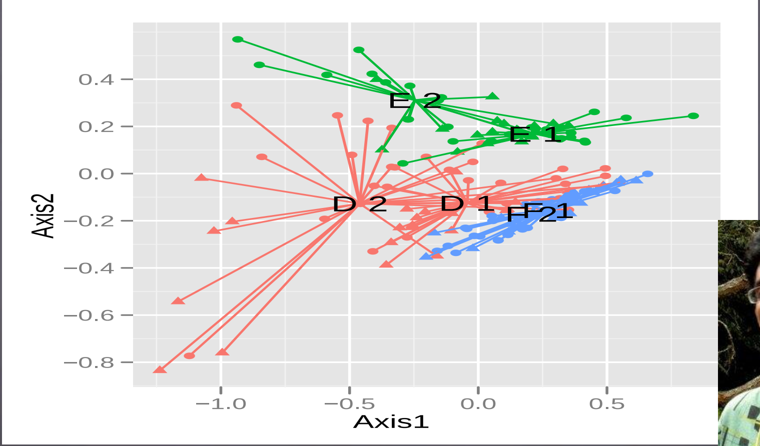
Results of decomposition of heterogeneity

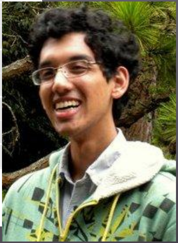
- Prevalence of a Lactobacillus-poor vaginal community state type (CST 4) was inversely correlated with gestational age at delivery. Risk for preterm birth was more pronounced for subjects with CST 4 accompanied by elevated Gardnerella abundances.

Part IV

Latent variables and topic analysis

How to understand the the taxa involved in the perturbation?

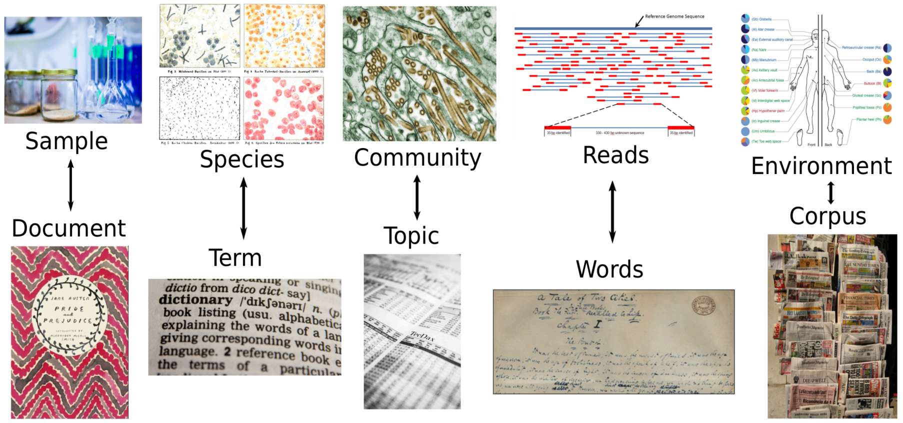




Kris Sankaran

Biostatistics, 2018,
Latent Variable Modeling for the Microbiome.
Kris Sankaran's Topic Page

Parallel between topic and community analyses



Sample ↔ Document

Species ↔ Term

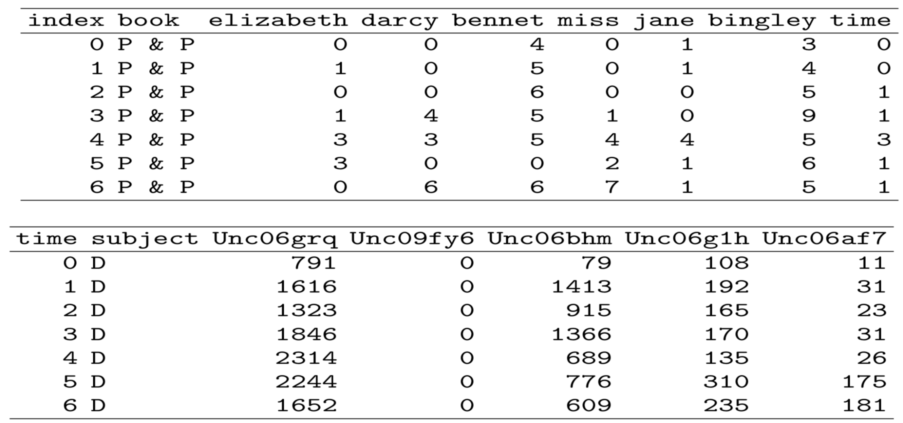
Community ↔ Topic

Reads ↔ Words

Environment ↔ Corpus

Credit: Kris Sankaran

Parallel between topic and community analyses



index	book	elizabeth	darcy	bennet	miss	jane	bingley	time
0	P & P	0	0	4	0	1	3	0
1	P & P	1	0	5	0	1	4	0
2	P & P	0	0	6	0	0	5	1
3	P & P	1	4	5	1	0	9	1
4	P & P	3	3	5	4	4	5	3
5	P & P	3	0	0	2	1	6	1
6	P & P	0	6	6	7	1	5	1

time	subject	Unc06grq	Unc09fy6	Unc06bhm	Unc06g1h	Unc06af7
0	D	791	0	79	108	11
1	D	1616	0	1413	192	31
2	D	1323	0	915	165	23
3	D	1846	0	1366	170	31
4	D	2314	0	689	135	26
5	D	2244	0	776	310	175
6	D	1652	0	609	235	181

Multivariate dependencies

Data depart from a multinomial distribution within each row:

- Some taxa are quasi-exclusive (*Lactobacillus crispatus* and *Gardnerella*).
- Co-occurrence through syntrophy, in which a molecular hydrogen-consuming species (typically a methanogen, like *Methanobrevibacter smithii* in the human gut) enhances the growth of a molecular hydrogen-producing species (any of a number of secondary fermenters in the gut).
- In the mouth (subgingival crevice), where in cases of moderate to severe periodontitis, a methanogen (Methanobrevibacter oralis) is always found with a syntrophic partner.
- There are not a finite number of taxa a priori, taxa evolve, some are sample specific.
- Beware if the number of rows (sample-specimens) is small, the matrix will appear to be low rank.

Statistical Model

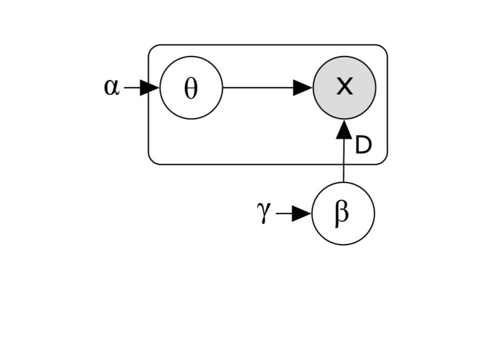
Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling.

It assumes samples have mixed memberships across topics.
(See Pritchard et. al 2000, Blei et. al. 2003)

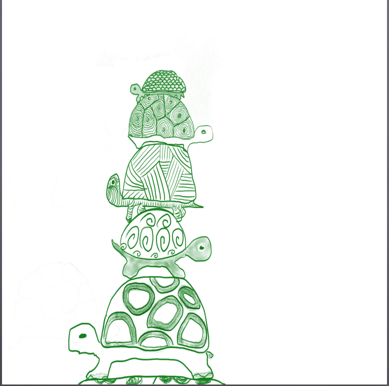
Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling.

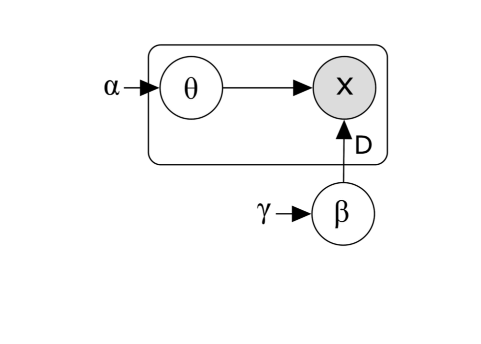
Observed microbiomes $\sim$ mixtures of underlying community types.

Statistical Model



Statistical Model





Statistical Model

samples layer observa.

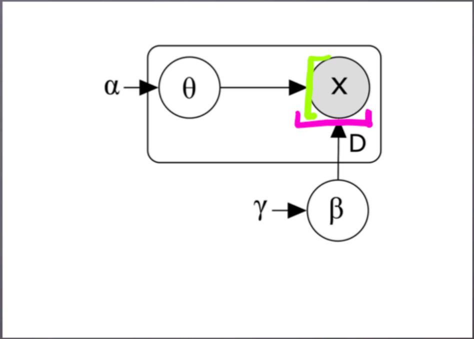
sample 1 ...

⋮

sample n ...

rows(X)

hidden layer for taxa



D Documents

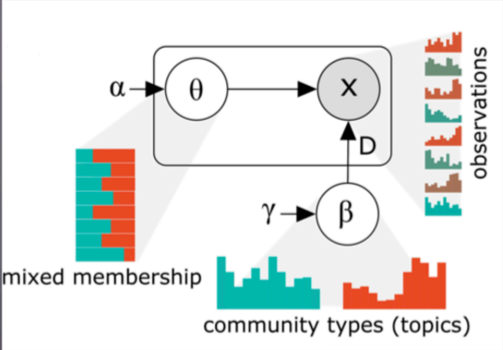
K communities or topics

Statistical Model

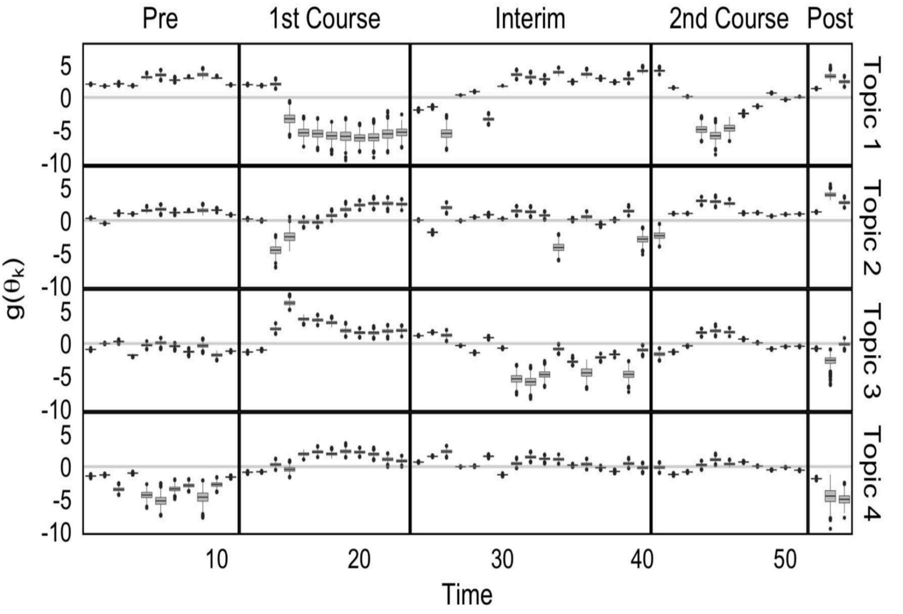
$$x_{d\cdot} \mid \beta \sim \mathrm{Mult}(N_d, B\theta_d)$$

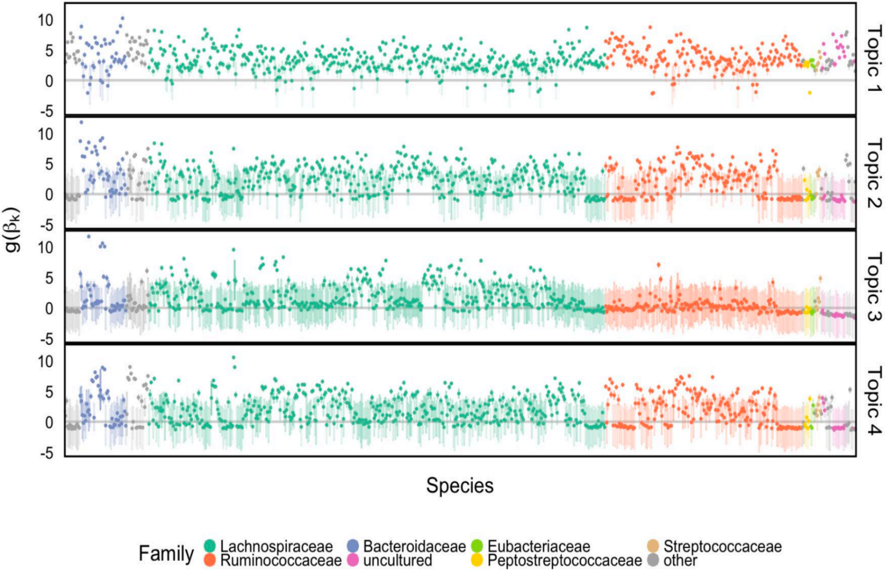
$$\theta_d \sim \mathrm{Dir}(\alpha)$$

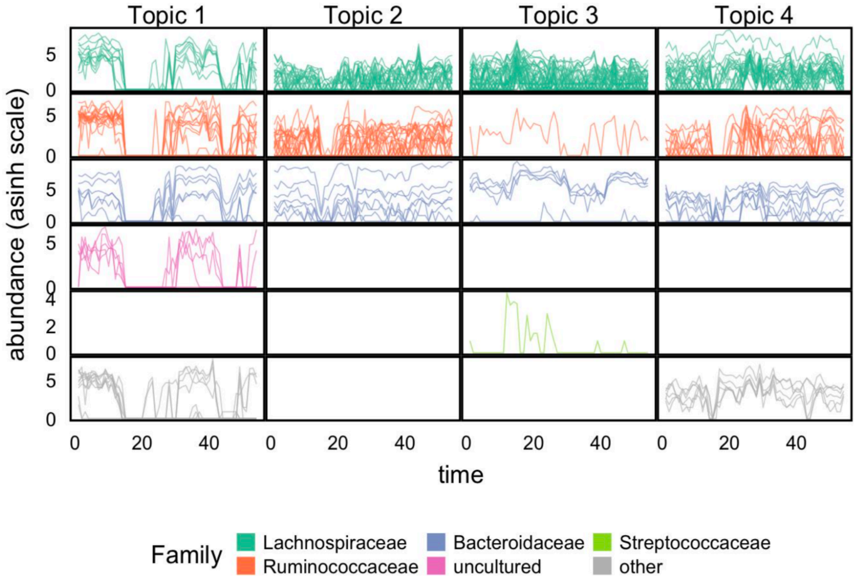
$d = 1 \ldots D$



$$\beta_k \sim \mathrm{Dir}(\gamma) \quad , \; k = 1 \ldots K$$

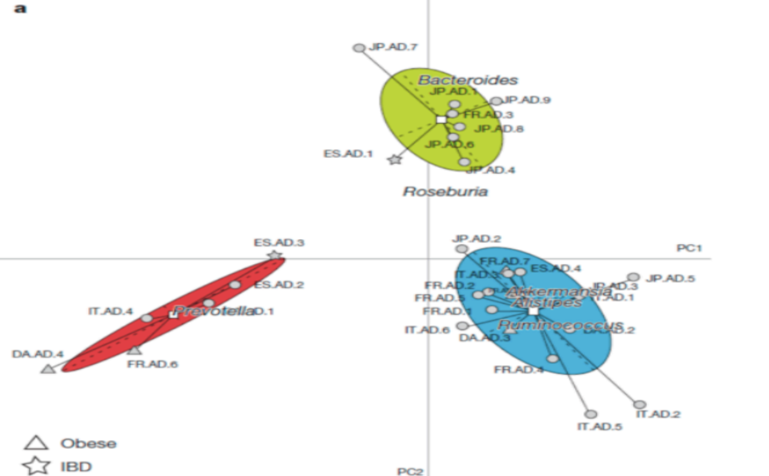

Pre
1st Course
Interim
2nd Course
Post
Topic 1
Topic 2
Topic 3
Topic 4
$g(\theta_k)$
5
0
-5
-10
10
20
30
40
50
Time

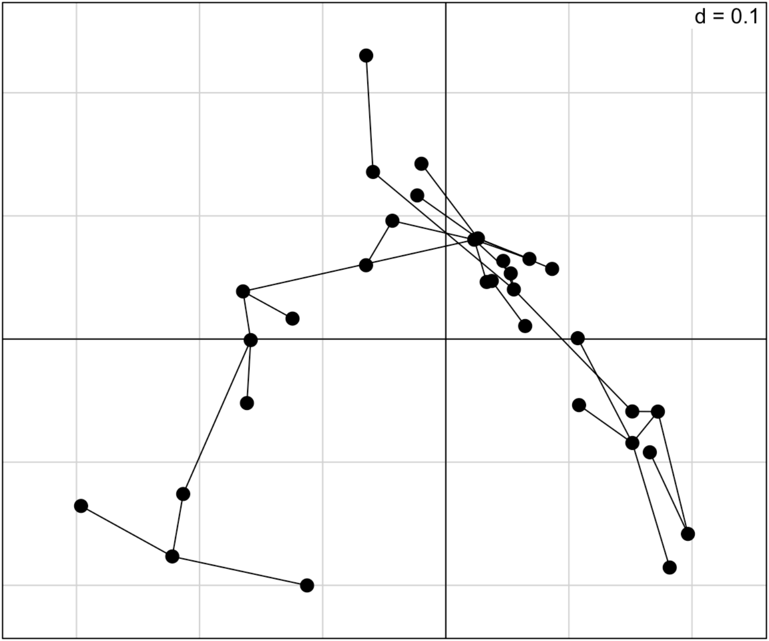

10
5
0
-5
$g(\beta_k)$
Topic 1
Topic 2
Topic 3
Topic 4
Species
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Eubacteriaceae
Peptostreptococcaceae
Streptococcaceae
other

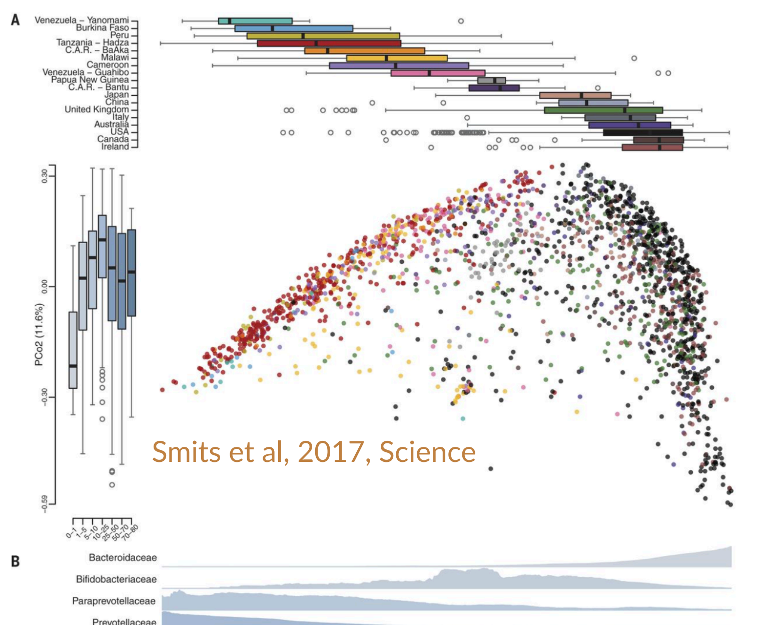

Topic 1
Topic 2
Topic 3
Topic 4
abundance (asinh scale)
5
0
5
0
5
0
5
0
4
2
0
5
0
0
20
40
0
20
40
0
20
40
0
20
40
time
Family
Lachnospiraceae
Ruminococcaceae
Bacteroidaceae
uncultured
Streptococcaceae
other

Part V

Latent (hidden) Gradients for Microbial Communities.

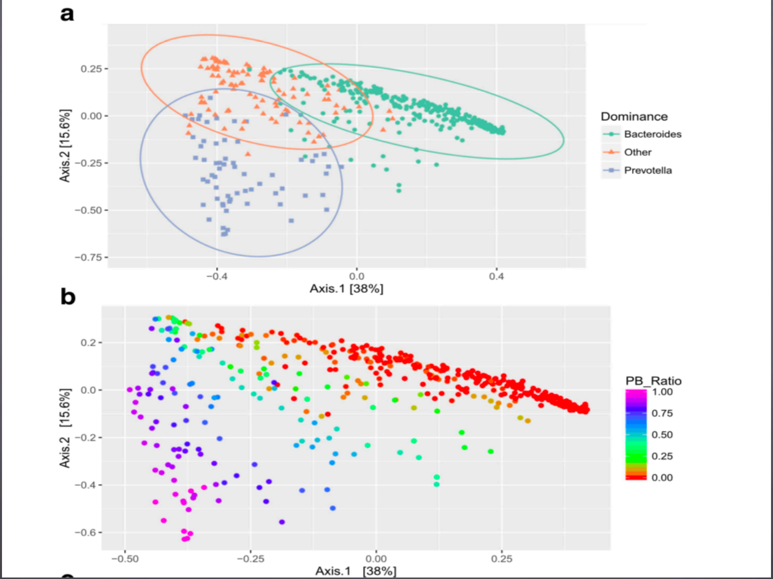



d = 0.1



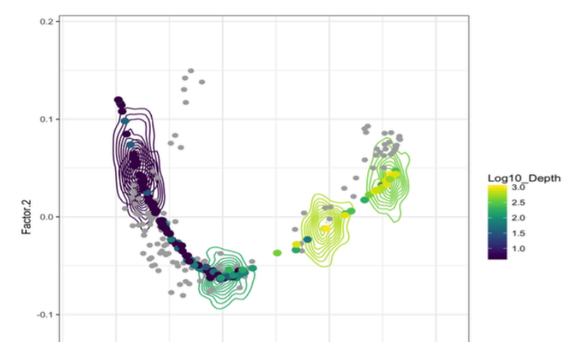
Smits et al, 2017, Science

Gorvitovskaia, Holmes, Huse, 2016, Microbiome



Gradient analysis of Tara Ocean microbiome data (with uncertainties)

Lan Nguyen Huong



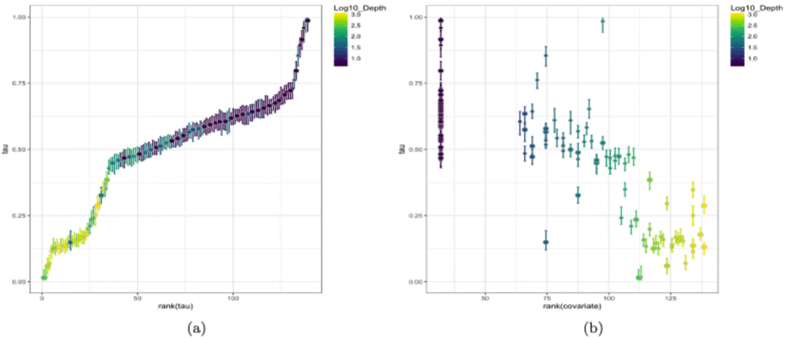


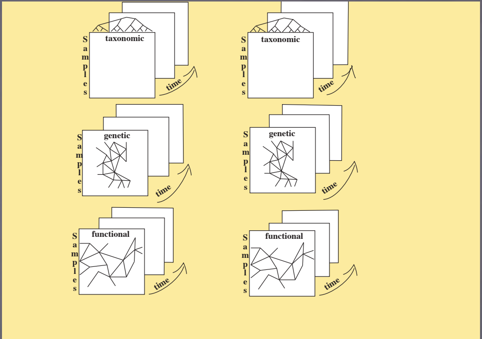
Fig. 4

Latent ordering in TARA Oceans dataset shown with uncertainties. The differences in the slope of plot (**a**) indicate varying data coverage along the underlying gradient. Correlation between the water depth and the latent ordering in microbial composition data is shown in (**b**). Coloring corresponds to log10 of the water depth (in meters) at which the ocean sample was collected

Part VI

Multitable, multidomain methods

Useful first order representation: Many Matrices



- Time series of abundance matrices.
- Bootstrap and Bayesian posterior analyses for many networks.
- Different types of data on same samples (taxa counts, clinical variates, spatial location, mass spec data).
- Networks in longitudinal studies.

Holmes (2005), Duality Diagrams.

Multi-table methods: Inertia/Co-Inertia

Generalize variance and covariance $\longrightarrow$ moments of inertia.
weighted (p_i) sum of distances.

Abundance data in a contingency table $\longrightarrow$ weighted sum of the squares
weighted frequencies (chisquare).

Co-Inertia

When studying two variables measured at the same locations, for instance PH and humidity the standard quantification of covariation is the *covariance*.

$$sum(x1 * y1 + x2 * y2 + x3 * y3)$$

if x and y co-vary –in the same direction this will be big.
A simple generalization to this when the variability is more complicated to measure as above is done through Co-Inertia analysis (CIA).
Co-inertia analysis (CIA) is a multivariate method that identifies trends or co-relationships in multiple datasets which contain the same samples or the same time points.
That is the rows or columns of the matrix have to be weighted similarly and thus must be matchable.

RV coefficient

The global measure of similarity of two data tables as opposed to two vectors can be done by a generalization of covariance provided by an inner product between tables that gives the RV coefficient, a number between 0 and 1, like a correlation coefficient, but for tables.

$$\mathrm{RV}(A, \mathrm{B}) = \frac{\mathrm{Tr}(A'\mathrm{B})}{\sqrt{\mathrm{Tr}(\mathrm{A}'\mathrm{A})}\sqrt{\mathrm{Tr}(\mathrm{B}'\mathrm{B})}}$$

Survey on RV: Josse, Holmes (2017) Statistics Surveys.

Example

Combining different types of data (antibiotic study).

Taxa Read counts (3 patients taking cipro: two time courses) : .

Mass-Spec Positive and Negative ion Mass Spec features and their intensities: .

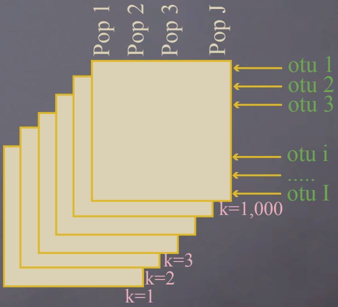
RNA-seq Metagenomic data on genes :.

Here is the RV table of the three array types:

```
> fourtable$RV
          Taxa      Kegg   MassSpec+ MassSpec-
Taxa      1        0.565   0.561      0.670
Kegg      0.565     1      0.686      0.644
MassSpec+ 0.561    0.686   1          0.568
MassSpec- 0.670    0.644   0.568      1
```

Output showing Bayesian posterior uncertainty measures

The methods that we consider here are all related to PCA and use the normalized Gram matrix **S** between biological samples. **S** is the Gram operator matrix of $(Q_{i,1},\ldots,Q_{i,J})$. Based on a single posterior instance of **S**, we can visualize biological samples in a lower dimensional space through PCA, with each biological sample projected once.



Bayesian nonparametric ordination for the analysis of microbial communities
Ren, Bacallado, Favaro, Holmes, Trippa (2017)

Full Bayesian nonparametric model

- We do not know the number of ASVs.
- We suppose underlying low dimensional latent variables for the sample P^j's.
- We use dependent microbial distributions, marginal priors of discrete distributions are built using manipulation of a Gaussian process and then extending this to multiple correlated distributions.

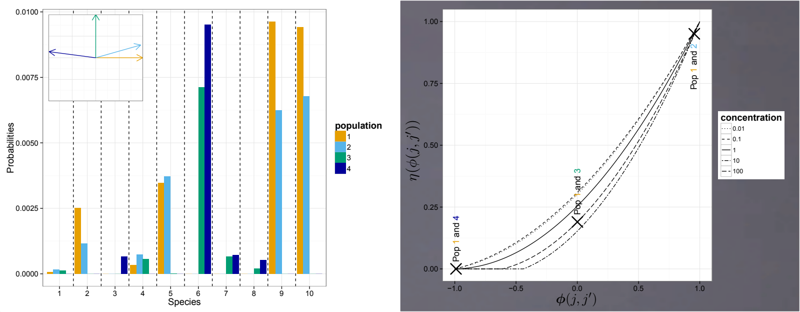


Figure: **Left panel**: realization of 4 microbial distributions from a dependent Dirichlet processes with 10 ASVs **Right panel**: correlation of two random probability measures when the cosine $\phi(j, j')$ between $\mathbf{Y}^j$ and $\mathbf{Y}^{j'}$ varies from -1 to 1. (Ren et al, JASA, 2017).

Parameters for samples

$$\mathbf{Y}^j, j \in \mathcal{J} = \{1, \dots, J\}$$

Define a joint prior on these factors through the Gram matrix

$$(\phi(j_1, j_2))_{j_1, j_2 \in \mathcal{J}}$$

The parameters $\mathbf{Y}^j$ can be interpreted as key characteristics of the biological samples that affect the relative abundance of ASVs.

$$Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle + \epsilon_{i,j}, \tag{1}$$

where the $\epsilon_{i,j}$ are independent Normal variables.

The degree of similarity between the discrete distributions $\{P^j; j \in \mathcal{J}\}$ is summarized by the Gram matrix $(\phi(j, j') = \langle \mathbf{Y}^j, \mathbf{Y}^{j'} \rangle; j, j' \in \mathcal{J})$.
The dependent Dirichlet processes is defined by setting

$$P^j(A) = \frac{\sum_i \mathbb{I}(Z_i \in A) \times \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}{\sum_i \sigma_i \langle \mathbf{X}_i, \mathbf{Y}^j \rangle^{+2}}, \qquad \forall j \in \mathcal{J}, \tag{2}$$

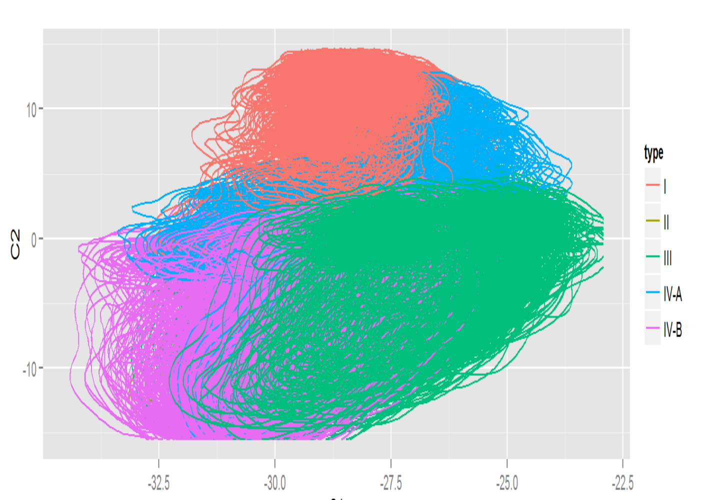
for every $A \in \mathcal{F}$. Here the sequence $(Z_1, Z_2, \ldots)$ and the array $(\mathbf{X}_1, \mathbf{X}_2, \ldots)$, contain independent and identically distributed random variables, while σ is a Poisson process on the unit interval defined by using a prior on $\sigma = (\sigma_1, \sigma_2, \ldots)$, the distribution of ordered points $(\sigma_i > \sigma_{i+1})$ in a Poisson process on $(0, 1)$ with intensity

$$\nu(\sigma) = \alpha\sigma^{-1}(1 - \sigma)^{-1/2}, \tag{3}$$

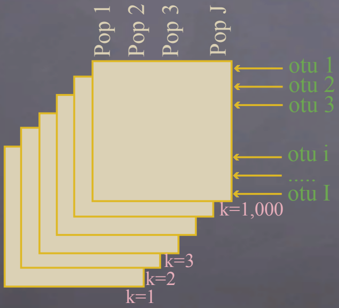
where $\alpha > 0$ is a concentration parameter.
We will use the notation $Q_{i,j} = \langle \mathbf{X}_i, \mathbf{Y}^j \rangle$.

The Naive projection approach

Naively overlaying projections of the principal coordinate loadings generated from different posterior samples of **S** on the same plot *could* show the variability of the projections.

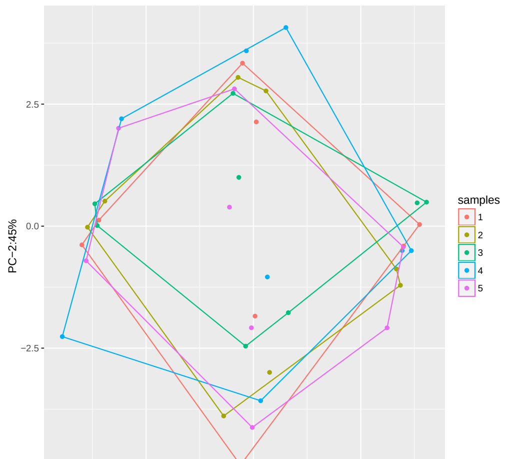


Bootstrap for PCA and MDS of taxa abundance contingency tables



Projection approach for bootstrap and Bayesian MDS

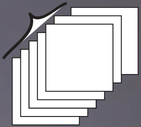
Naively overlaying projections of the principal coordinate loadings generated from different resamples on the same plot *could* show the variability of the projections.



Why?

- Principal coordinate directions are only defined up to a sign.
- Principal coordinates, 1 and 2 or 2 and 3 can be permuted.
- We need to do **registration** first.

Registration: Find $\mathbf{S}_0$



Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.

We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.

We use the **RV** similarity metric between two symmetric square matrices **A** and **B**

$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{AB})/\sqrt{\mathrm{Tr}(\mathbf{AA})\mathrm{Tr}(\mathbf{BB})}$$

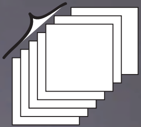
We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

Find lower dimensional consensus space V

For dim 2, $\mathbf{v}_1$ and $\mathbf{v}_2$ of $\mathbf{S}_0$ corresponding to the largest eigenvalues λ_1 and λ_2. All biological samples in V are visualized by projecting rows of $\mathbf{S}_0$ onto V: $(\psi_1^0, \psi_2^0) = \mathbf{S}_0(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$.

Project the rows of posterior sample $\mathbf{S}_k$ onto V by $(\psi_1^k, \psi_2^k) = \mathbf{S}_k(\mathbf{v}_1\lambda_1^{-1/2}, \mathbf{v}_2\lambda_2^{-1/2})$. Overlaying all the ψ^k displays uncertainty of $\mathbf{S}$ in the same linear subspace. Posterior variability of the biological samples' projections is visualized in V by plotting each row of the matrices (ψ_1^k, ψ_2^k), $k = 1, \ldots, K$, in the same figure.

Registration: Find $\mathbf{S}_0$

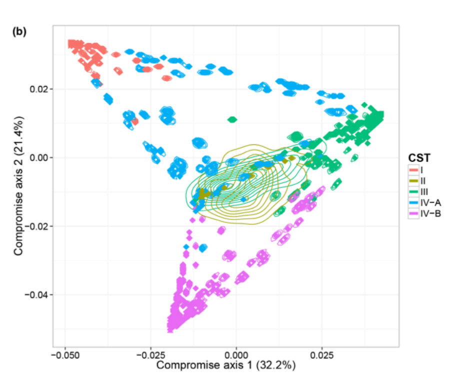


Identify a Gram matrix $\mathbf{S}_0$ that best summarizes K posterior samples' Gram matrix $\mathbf{S}_1, \ldots, \mathbf{S}_K$. Minimizing L_2 loss element-wise leads to $\mathbf{S}_0 = (\sum_i \mathbf{S}_i)/K$.
We prefer to choose $\mathbf{S}_0$, the Gram matrix that maximizes similarity with $\mathbf{S}_1, \ldots, \mathbf{S}_K$.
We use the **RV** similarity metric between two symmetric square matrices **A** and **B**

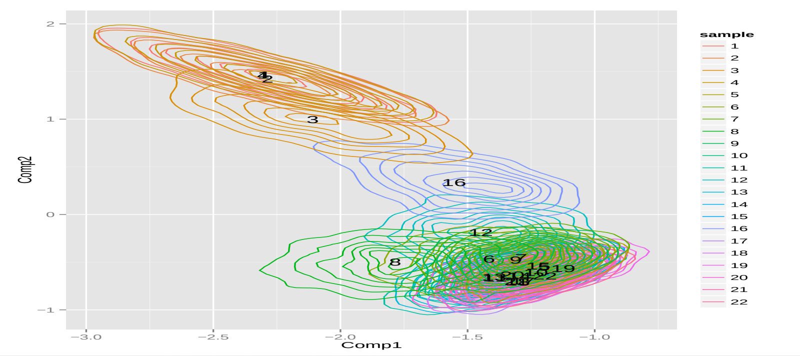
$$\mathrm{RV}(\mathbf{A}, \mathbf{B}) = \mathrm{Tr}(\mathbf{AB})/\sqrt{\mathrm{Tr}(\mathbf{AA})\mathrm{Tr}(\mathbf{BB})}$$

We diagonalize the **RV** matrix to obtain $\mathbf{S}_0$.

We can see the uncertainties



Bayesian Nonparametric Ordination for the Analysis of Microbial Communities, Ren et al, 2017 (JASA).
A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space V.



A contour plot is produced for each biological sample to facilitate visualization of the posterior variability of its position in the consensus space V.

Solutions for microbiome analyses of perturbation studies.

- Maintain all information ⟶ sequences are names.
- Percolate the uncertainty ⟶ contours of uncertainty.
- Interpretation ⟶ latent variables (gradients or clusters).
- Reproducibility ⟶ complete code source.
- Heterogeneity ⟶ multicomponent objects:phyloseq.
- Wait and see approach ⟶ wait to annotate.
- Training and collaboration, learn each others language ⟶ Rmd and html.

Current work and open problems

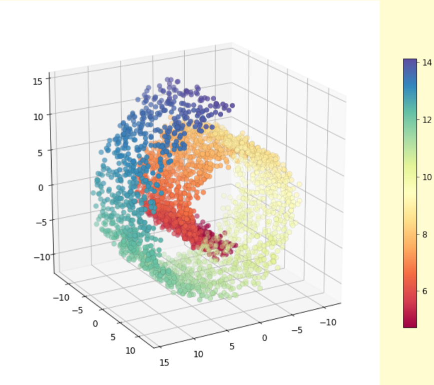
The problem of entanglements.

We call this: non-identifiability, when the model is too rich.

Non linearity: data unfolding

Nonlinear stochastic neighbor embedding

Swiss Roll



Communication between fields and subfields

Blackboxes lead to miscommunication.

- LDA, STRUCTURE, Probabilistic Factor Analysis, Probabilistic PCA, Probabilistic embedding.
- Multiway, multiview, data fusion, data integration, multidomain, multiomics, ...

[1] Benjamin Callahan, Daniel DiGiulio, Daniela Goltsman, Christine Sun, Elizabeth Costello, Pratheepa Jeganathan, Joseph Biggio, Ronald Wong, Maurice Druzin, Gary Shaw, et al. Replication and refinement of a vaginal microbial signature of preterm birth in two racially distinct cohorts of us women. *PNAS*, page 201705899, 2017.

[2] BJ Callahan, PJ McMurdie, MJ Rosen, AW Han, AJ Johnson, and SP Holmes. Dada2: High resolution sample inference from amplicon data. *Nature Methods*, 13(7):581, 2016.

[3] P. Diaconis, S. Goel, and S. Holmes. Horseshoes in multidimensional scaling and kernel methods. *Annals of Applied Statistics*, 2007.

[4] Daniela SA Goltsman, Christine L Sun, Diana M Proctor, Daniel B DiGiulio, Anna Robaczewska, Brian C Thomas, Gary M Shaw, David K Stevenson, Susan P Holmes, Jillian F Banfield, and David A Relman. Metagenomic analysis with strain-level resolution reveals fine-scale variation in the human pregnancy microbiome. *bioRxiv*, page 266700, 2018.

[5] Susan Holmes. Multivariate analysis: The French way. In

D. Nolan and T. P. Speed, editors, *Probability and Statistics: Essays in Honor of David A. Freedman*, volume 56 of *IMS Lecture Notes–Monograph Series*. IMS, Beachwood, OH, 2006.

[6] P. J. McMurdie and S. Holmes. Phyloseq: Reproduible research platform for bacterial census data. *PlosONE*, 2013. April 22,.

[7] P. J. McMurdie and S. Holmes. Waste not, want not: Why rarefying microbiome data is inadmissible. *Plos Computational Biology*, 2014. April 03.

[8] Lan Huong Nguyen and Susan Holmes. Bayesian unidimensional scaling for visualizing uncertainty in high dimensional datasets with latent ordering of observations. *BMC bioinformatics*, 18(10):394, 2017.

[9] Diana M Proctor, Julia A Fukuyama, Peter M Loomer, Gary C Armitage, Stacey A Lee, Nicole M Davis, Mark I Ryder, Susan P Holmes, and David A Relman. A spatial gradient of bacterial diversity in the human oral cavity shaped by salivary flow. *Nature communications*, 9(1):681, 2018.

[10] Boyu Ren, Sergio Bacallado, Stefano Favaro, Susan Holmes, and Lorenzo Trippa. Bayesian nonparametric ordination for the

analysis of microbial communities. *Journal of the American Statistical Association*, (February), 2017.