Discovery and Visualization of Latent Structure
with Applications to the Microbiome

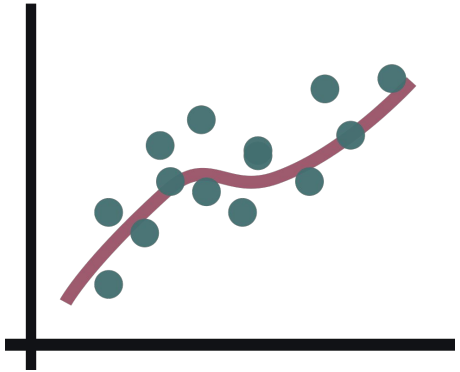
Kris Sankaran
Department of Statistics

March 9, 2018

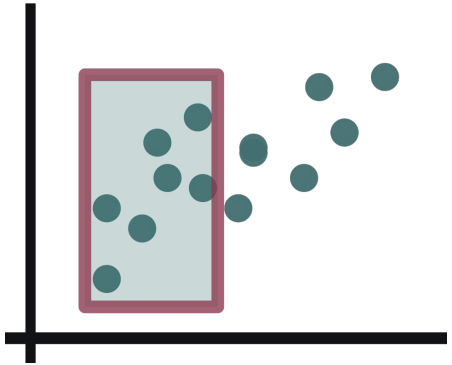
Follow along
https://bit.ly/2HffOsS

Approaches

Modeling Workflows

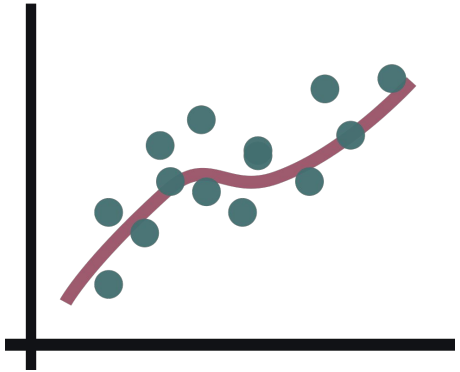


Visualization Packages



Approaches

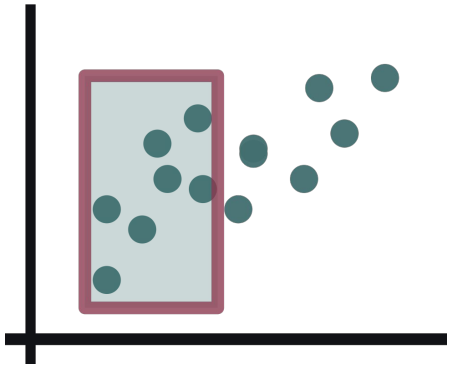
Modeling Workflows



- Compute reductions in data
- Characterize uncertainty in reductions
- Sufficiency provides summaries

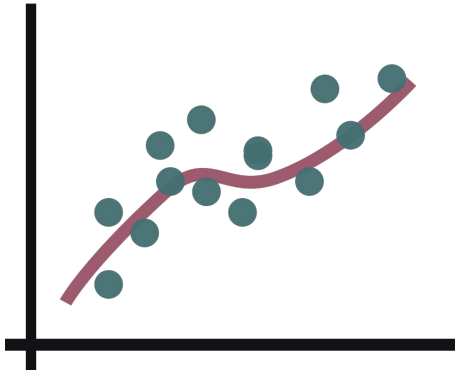
$$X \rightarrow \hat{\theta}$$

Visualization Packages



Approaches

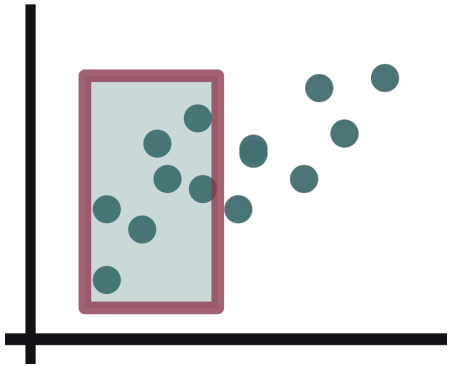
Modeling Workflows



- Compute reductions in data
- Characterize uncertainty in reductions
- Sufficiency provides summaries

$$X \rightarrow \hat{\theta}$$

Visualization Packages

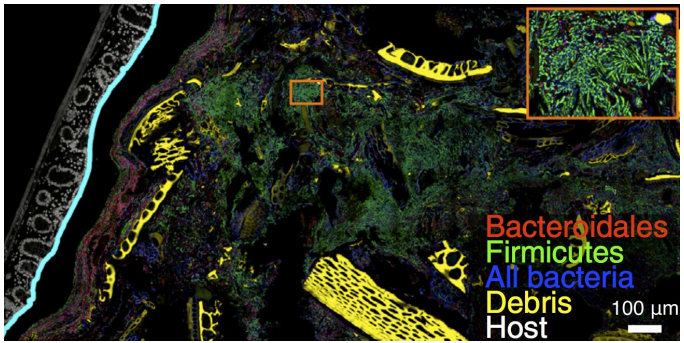


- Simplify navigation across data
- Suggest simple models + hypotheses
  - Motivate reductions

$$\Theta?$$

Microbiome Studies

- **Microbiome**: All the bacteria living together in some place (e.g., the human gut)
  - Quantification enabled by advances in 16S rRNA and metagenomics sequencing technologies
- **Typical goal**: Characterize variation in community profiles across conditions
- Intersection of several sciences
  - **Ecology + Medicine**: Bacterial communities can have medical implications
  - **Evolution + Microbiology**: A data-rich setting for studying evolutionary principles



Imaging the microbiome. Figure from [Earle et. al 2015].

Data Sources

- 16S Sequencing: Marker gene allows quantification of variants
- Size: Typically 10 - 1000 samples, 500 - 2000 bacteria

Data Sources

- 16S Sequencing: Marker gene allows quantification of variants
- Size: Typically 10 - 1000 samples, 500 - 2000 bacteria
- Contextual data are available
  - Samples → subject characteristics

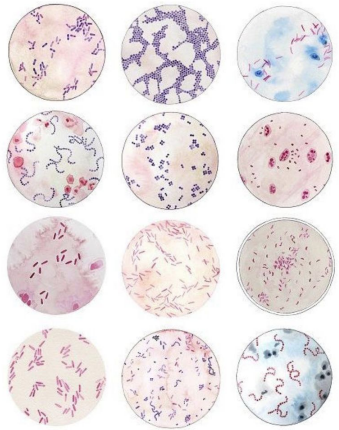
Data Sources

- 16S Sequencing: Marker gene allows quantification of variants
- Size: Typically 10 - 1000 samples, 500 - 2000 bacteria
- Contextual data are available
  - Samples → subject characteristics
  - Bacteria → taxonomic information

Probabilistic Latent Variable Modeling for the Microbiome

Motivating Idea

- Schloss and Handelsman [2007] noted parallels between text and microbiome data
  - Limited application of text models in microbiome studies
  - Probability models are useful in studies with complex structure
- **Contribution**: In the spirit of [Callahan 2016, Fukuyama 2017], describe a workflow, provide code, and compare existing methods, focusing on interpretation, usability, and visualization



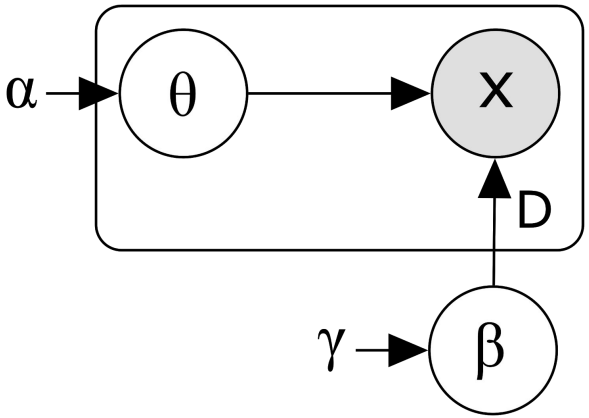
Motivating Idea

- Schloss and Handelsman [2007] noted parallels between text and microbiome data
  - Limited application of text models in microbiome studies
  - Probability models are useful in studies with complex structure
- **Contribution**: In the spirit of [Callahan 2016, Fukuyama 2017], describe a workflow, provide code, and compare existing methods, focusing on interpretation, usability, and visualization

index	book	eliza.	darcy	bennet	miss	jane
0	P&P	0	0	4	0	1
1	P&P	1	0	5	0	1
2	P&P	0	0	6	0	0
3	P&P	1	4	5	1	0
4	P&P	3	3	5	4	4

time	subject	sp_1	sp_2	sp_3	sp_4	sp_5
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

Models Considered: Latent Dirichlet Allocation

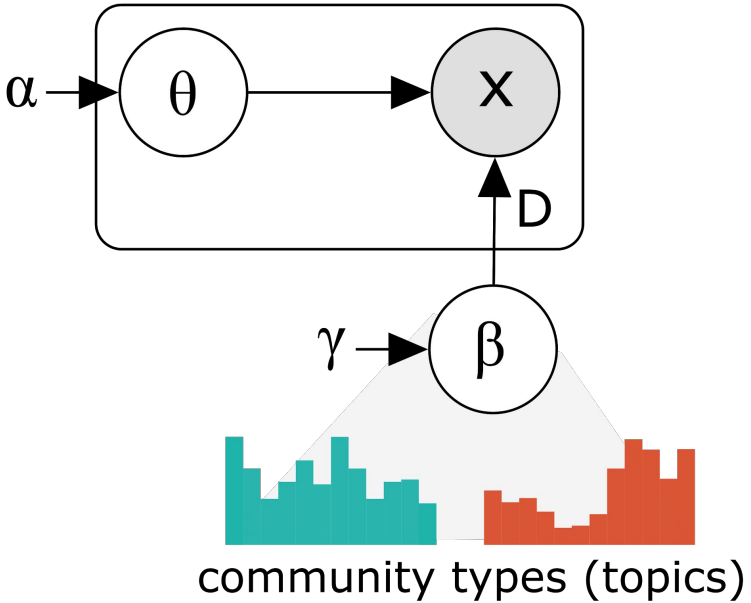


- Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling that assumes samples have mixed memberships across topics [Pritchard et. al 2000, Blei et. al. 2003]
  - Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling
- Observed microbiomes ≈ mixtures of underlying community types

Models Considered: Latent Dirichlet Allocation

$$\beta_k \overset{iid}{\sim} \text{Dir}(\gamma) \text{ for } k = 1, \ldots, K$$

- Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling that assumes samples have mixed memberships across topics [Pritchard et. al 2000, Blei et. al. 2003]
  - Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling
- Observed microbiomes ≈ mixtures of underlying community types

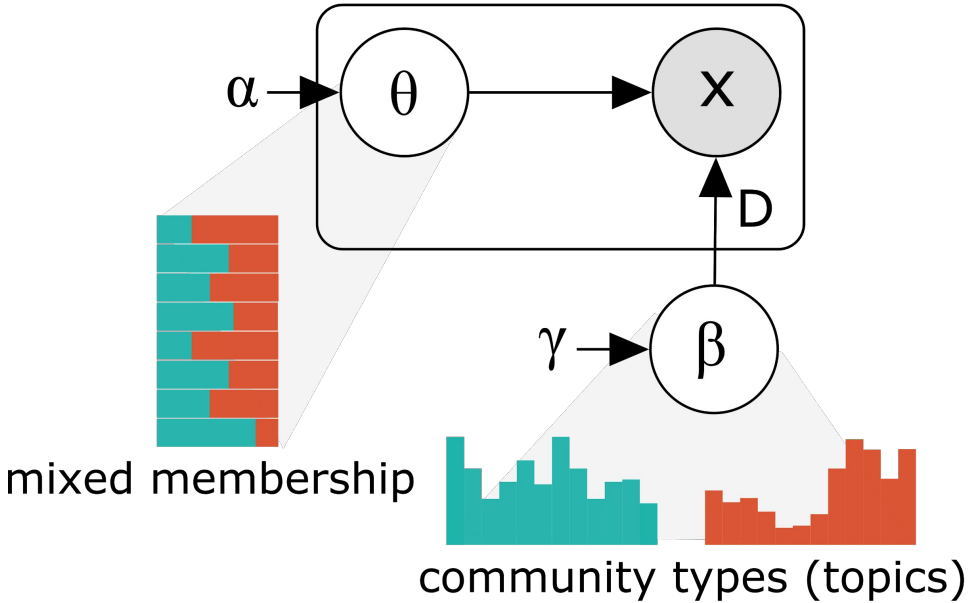


Models Considered: Latent Dirichlet Allocation

$$\theta_d \overset{iid}{\sim} \text{Dir}(\alpha) \text{ for } d = 1, \ldots, D$$

$$\beta_k \overset{iid}{\sim} \text{Dir}(\gamma) \text{ for } k = 1, \ldots, K$$

- Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling that assumes samples have mixed memberships across topics [Pritchard et. al 2000, Blei et. al. 2003]
  - Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling
- Observed microbiomes ≈ mixtures of underlying community types



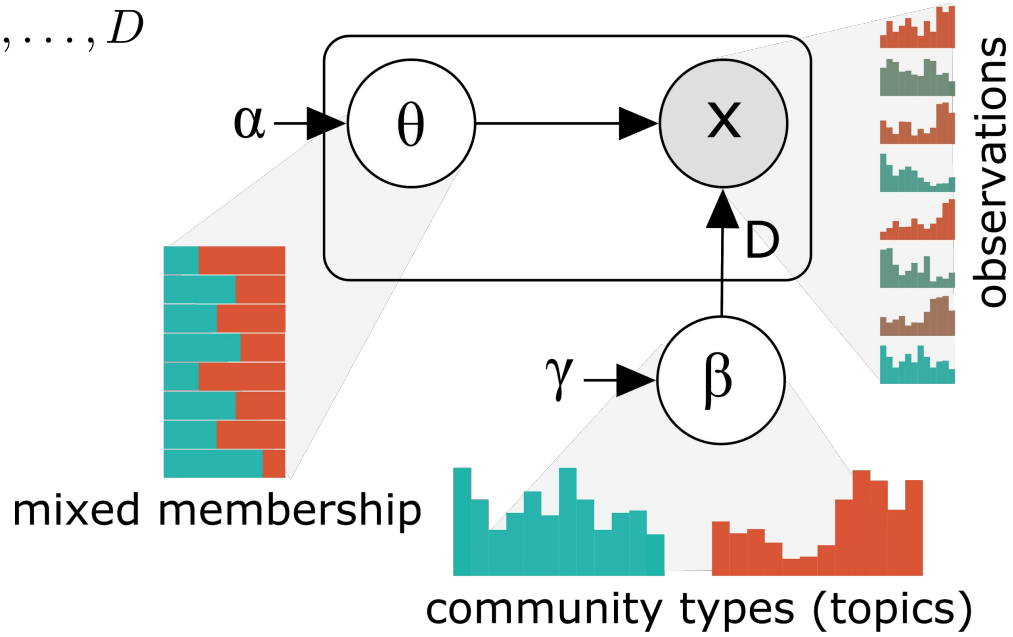
Models Considered: Latent Dirichlet Allocation

$$x_{d\cdot} \,|\, (\beta_k)_1^K \overset{iid}{\sim} \text{Mult}\,(N_d, B\theta_d) \;\text{ for } d = 1, \ldots, D$$

$$\theta_d \overset{iid}{\sim} \text{Dir}\,(\alpha) \;\text{ for } d = 1, \ldots, D$$

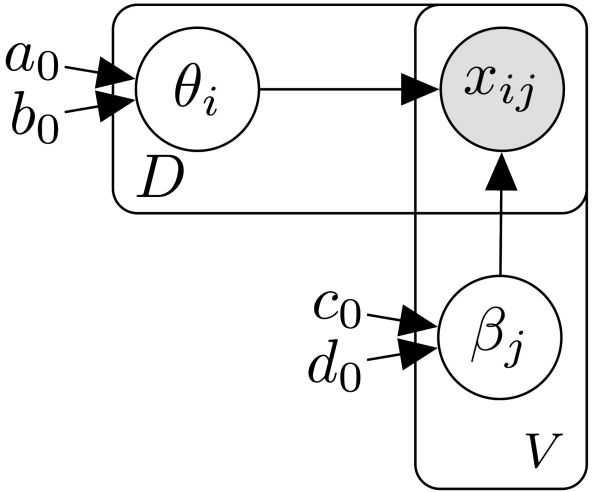
$$\beta_k \overset{iid}{\sim} \text{Dir}\,(\gamma) \;\text{ for } k = 1, \ldots, K$$

- Latent Dirichlet Allocation (LDA) is an alternative to Multinomial Mixture Modeling that assumes samples have mixed memberships across topics [Pritchard et. al 2000, Blei et. al. 2003]
  - Posterior inference can be done with variational approximations or (collapsed) Gibbs sampling
- Observed microbiomes ≈ mixtures of underlying community types



Models Considered: Gamma-Poisson Factorization

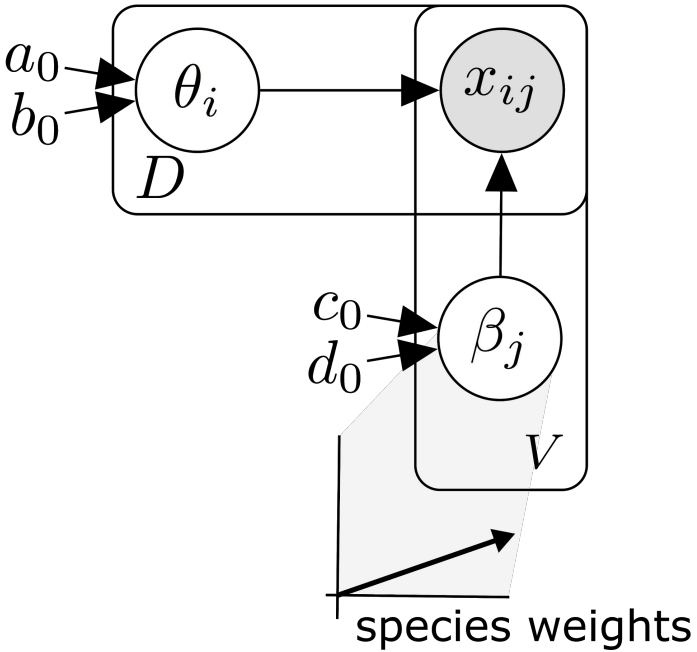
- Nonnegative Matrix Factorization using a low-rank Gamma-Poisson model [Canny 2004]
- Mixture and topic vectors are no longer required to sum to one
- The expected size of an entry is set by the angle between associated mixture and topic vectors



Models Considered: Gamma-Poisson Factorization

$$\beta_v \overset{iid}{\sim} \text{Gam}\,(c_0, d_0) \;\text{ for } v = 1, \ldots, V$$

- Nonnegative Matrix Factorization using a low-rank Gamma-Poisson model [Canny 2004]
- Mixture and topic vectors are no longer required to sum to one
- The expected size of an entry is set by the angle between associated mixture and topic vectors

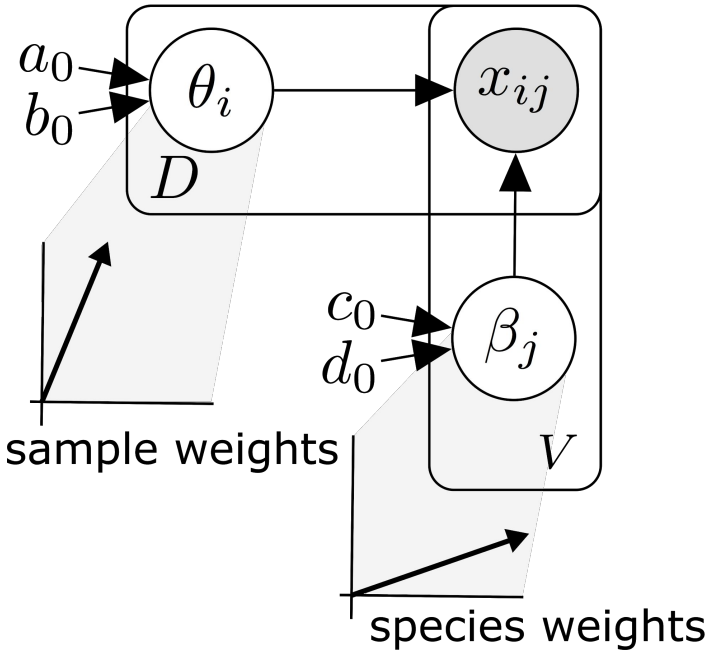


Models Considered: Gamma-Poisson Factorization

$$\theta_d \overset{iid}{\sim} \mathrm{Gam}\,(a_0, b_0) \;\text{ for } d = 1, \ldots, D$$

$$\beta_v \overset{iid}{\sim} \mathrm{Gam}\,(c_0, d_0) \;\text{ for } v = 1, \ldots, V$$

- Nonnegative Matrix Factorization using a low-rank Gamma-Poisson model [Canny 2004]
- Mixture and topic vectors are no longer required to sum to one
- The expected size of an entry is set by the angle between associated mixture and topic vectors



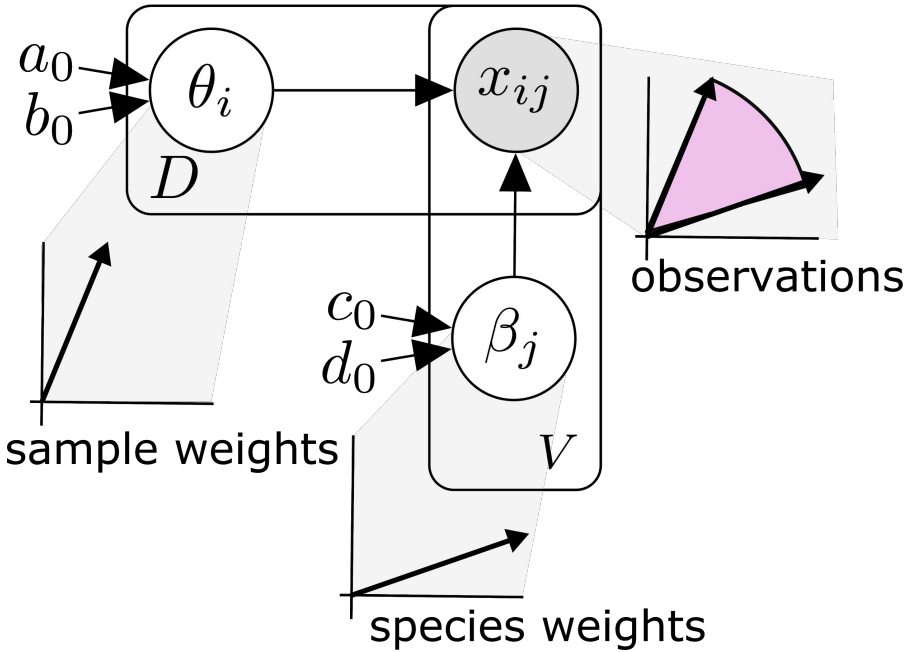
Models Considered: Gamma-Poisson Factorization

$$x_{dv} \overset{iid}{\sim} \text{Poi}\left(\theta_d^T \beta_v\right) \text{ for } d = 1, \ldots, D \text{ and } v = 1, \ldots V$$

$$\theta_d \overset{iid}{\sim} \text{Gam}\left(a_0, b_0\right) \text{ for } d = 1, \ldots, D$$

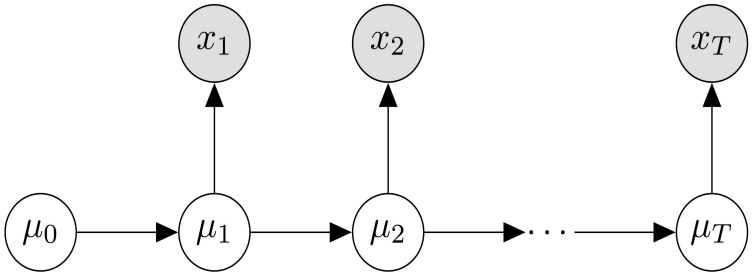
$$\beta_v \overset{iid}{\sim} \text{Gam}\left(c_0, d_0\right) \text{ for } v = 1, \ldots, V$$

- Nonnegative Matrix Factorization using a low-rank Gamma-Poisson model [Canny 2004]
- Mixture and topic vectors are no longer required to sum to one
- The expected size of an entry is set by the angle between associated mixture and topic vectors



Models Considered: Dynamic Unigrams

- The Dynamic Unigram Model supposes the distribution of species within a sample evolves smoothly over time
- This is accomplished by passing a Gaussian random walk through a multilogit link
- The factorized version is the Dynamic Topic Model [Blei and Lafferty 2006]

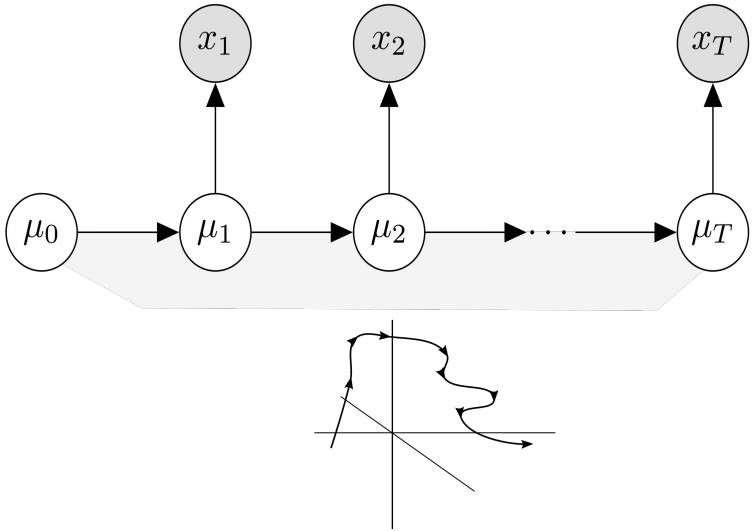


Models Considered: Dynamic Unigrams

$$\mu_t|\mu_{t-1} \overset{iid}{\sim} \mathcal{N}\left(\mu_{t-1}, \sigma^2 I_V\right) \text{ for } t = 1, \ldots, T$$

$$\mu_0 \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2 I_V\right)$$

- The Dynamic Unigram Model supposes the distribution of species within a sample evolves smoothly over time
- This is accomplished by passing a Gaussian random walk through a multilogit link
- The factorized version is the Dynamic Topic Model [Blei and Lafferty 2006]



Models Considered: Dynamic Unigrams

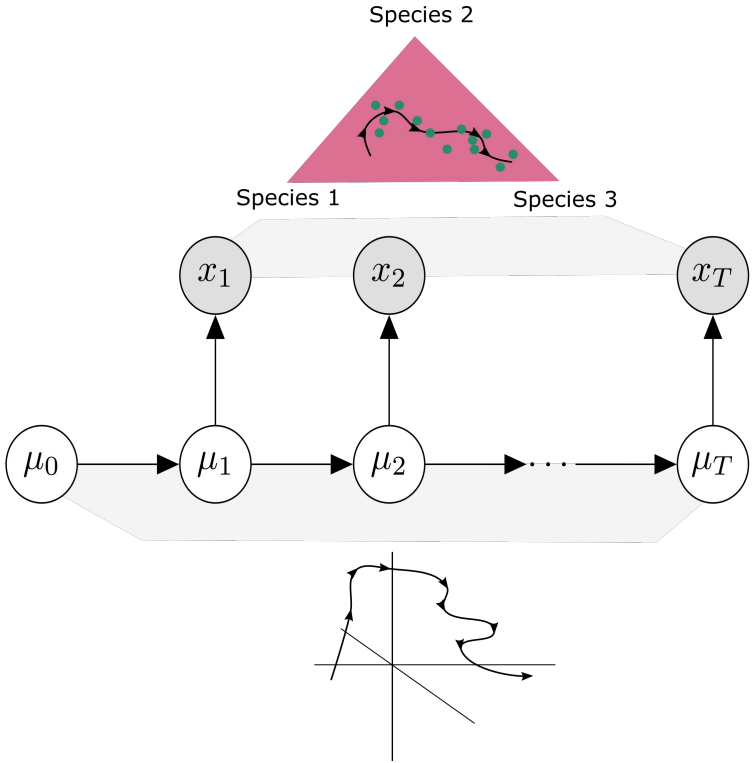
$$x_{d\cdot}|\mu_{t(d)} \overset{iid}{\sim} \text{Mult}\left(N_d, S\left(\mu_{t(d)}\right)\right) \text{ for } d = 1, \ldots, D$$

$$\mu_t|\mu_{t-1} \overset{iid}{\sim} \mathcal{N}\left(\mu_{t-1}, \sigma^2 I_V\right) \text{ for } t = 1, \ldots, T$$

$$\mu_0 \overset{iid}{\sim} \mathcal{N}\left(0, \sigma^2 I_V\right)$$

$$[S(\mu)]_v = \frac{\exp \mu_v}{\sum_{v'} \exp \mu_{v'}}$$

- The Dynamic Unigram Model supposes the distribution of species within a sample evolves smoothly over time
- This is accomplished by passing a Gaussian random walk through a multilogit link
- The factorized version is the Dynamic Topic Model [Blei and Lafferty 2006]



Bacterial Dynamics and Antibiotic Time Courses

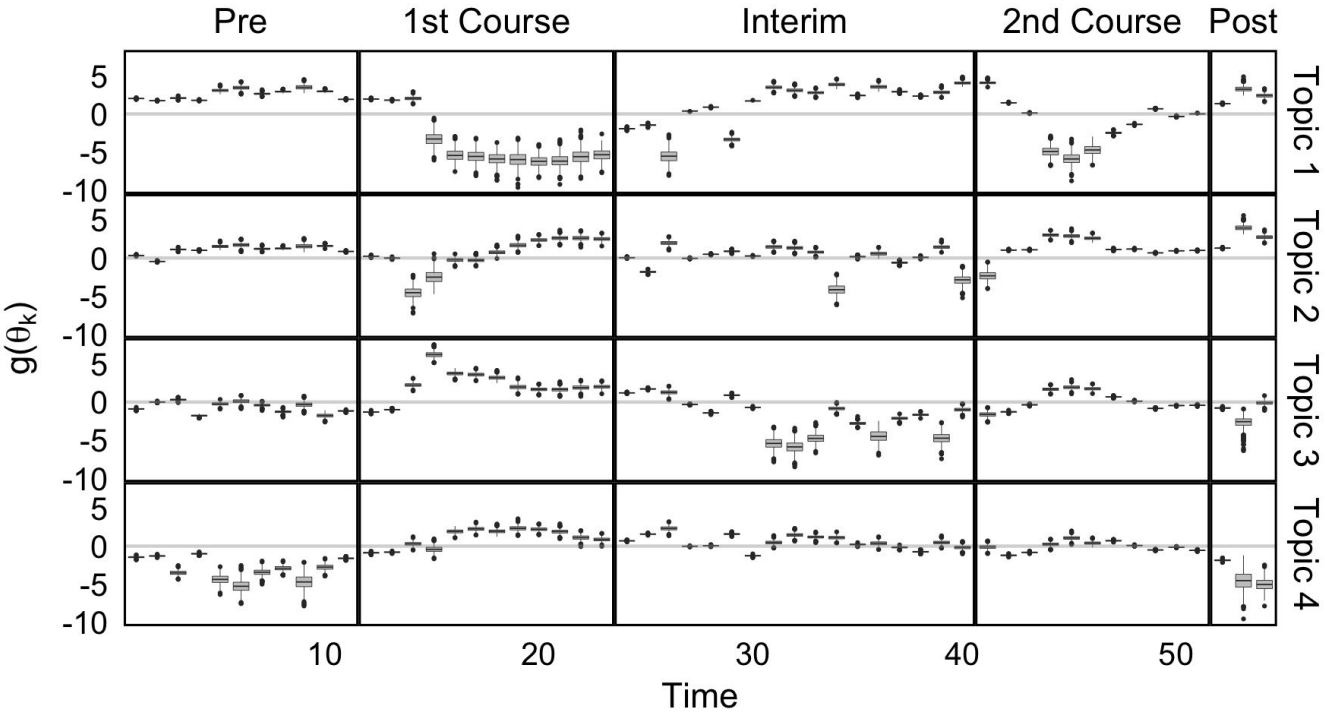
- Revisiting the study [Dethlefsen and Relman, 2011]
- Study Goal: How does the microbiome respond to antibiotics, keeping in mind typical day-to-day variation? ["Wildfire" analogy]

Bacterial Dynamics and Antibiotic Time Courses

- Revisiting the study [Dethlefsen and Relman, 2011]
- Study Goal: How does the microbiome respond to antibiotics, keeping in mind typical day-to-day variation? ["Wildfire" analogy]
- Data: 3 subjects, across ~ 50 timepoints, with two antibiotic time courses introduced in-between
- We study one subject at a time

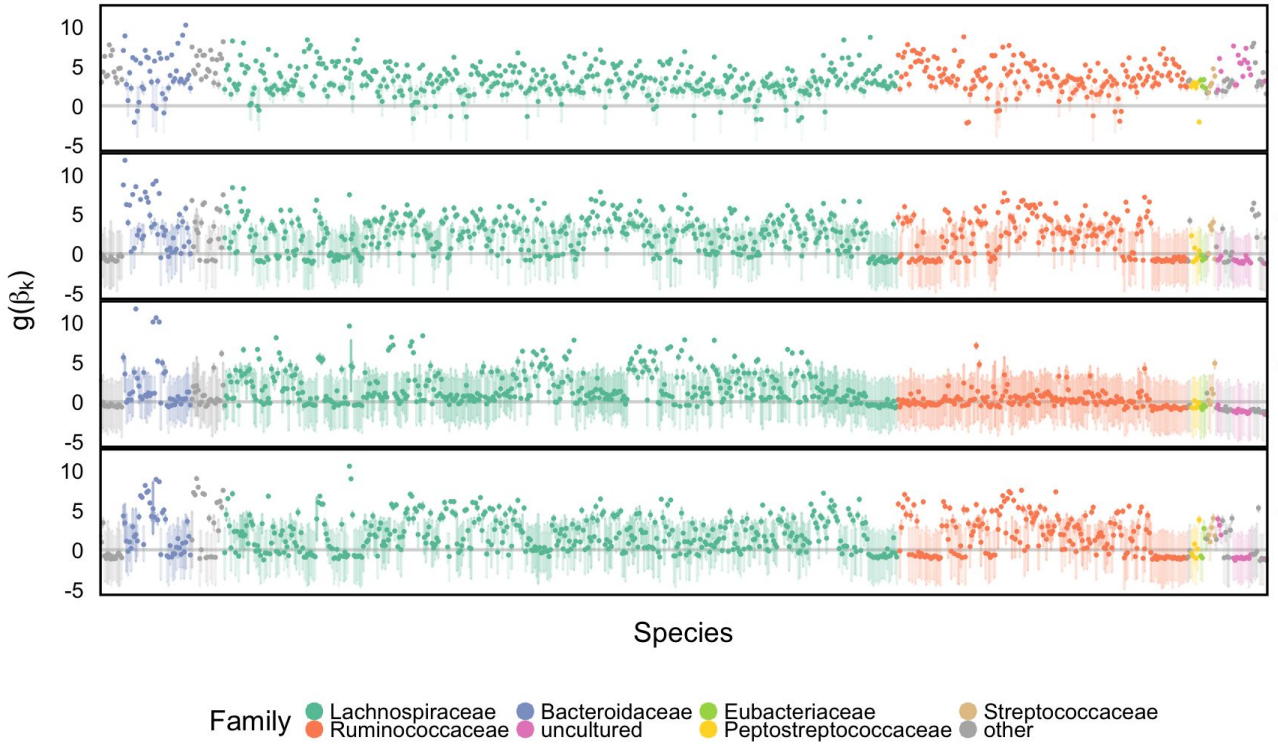
Estimated Mixed-Memberships: θ_{dk}

- Boxplots represent approximate posteriors for θ_{dk} and their evolution
- A centered log transform has been applied
- Interpretations
  - Slow recovery
  - Fast recovery
  - Relatively increase
  - No effect



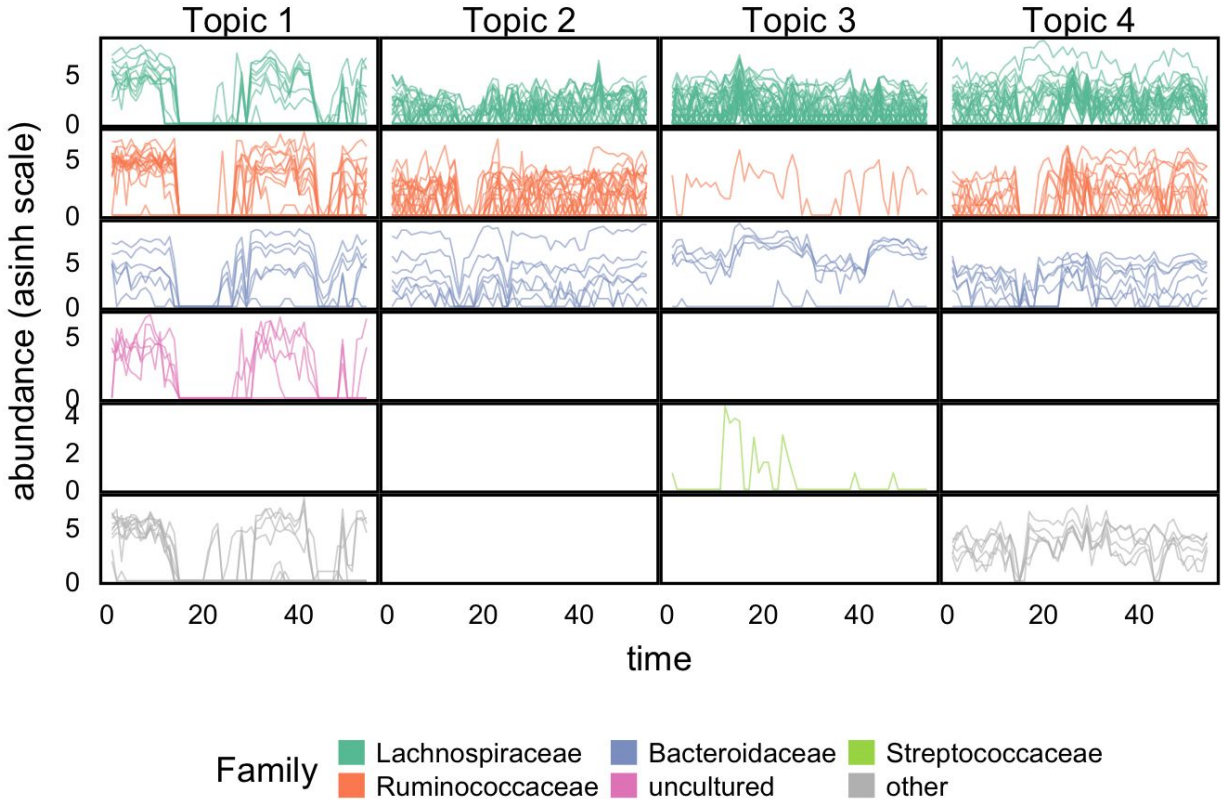
Estimated Topics: β_{vk}

- Intervals represent approximate β_{vk} posteriors
- Each row is a topic, each interval is a species, colors are taxonomic groups
- A centered log transform has been applied
- The third topic is noticeably less diverse, corresponding to the community during antibiotics time courses



Topic Prototypes

- Prototypes are species representative of individual topics
  - 50 species with large $\beta_{kv} - \sum_{k' \neq k} \beta_{k'v}$
- Rows are taxonomic groups, columns are topics
- Each trajectory gives the abundance of a single species over time

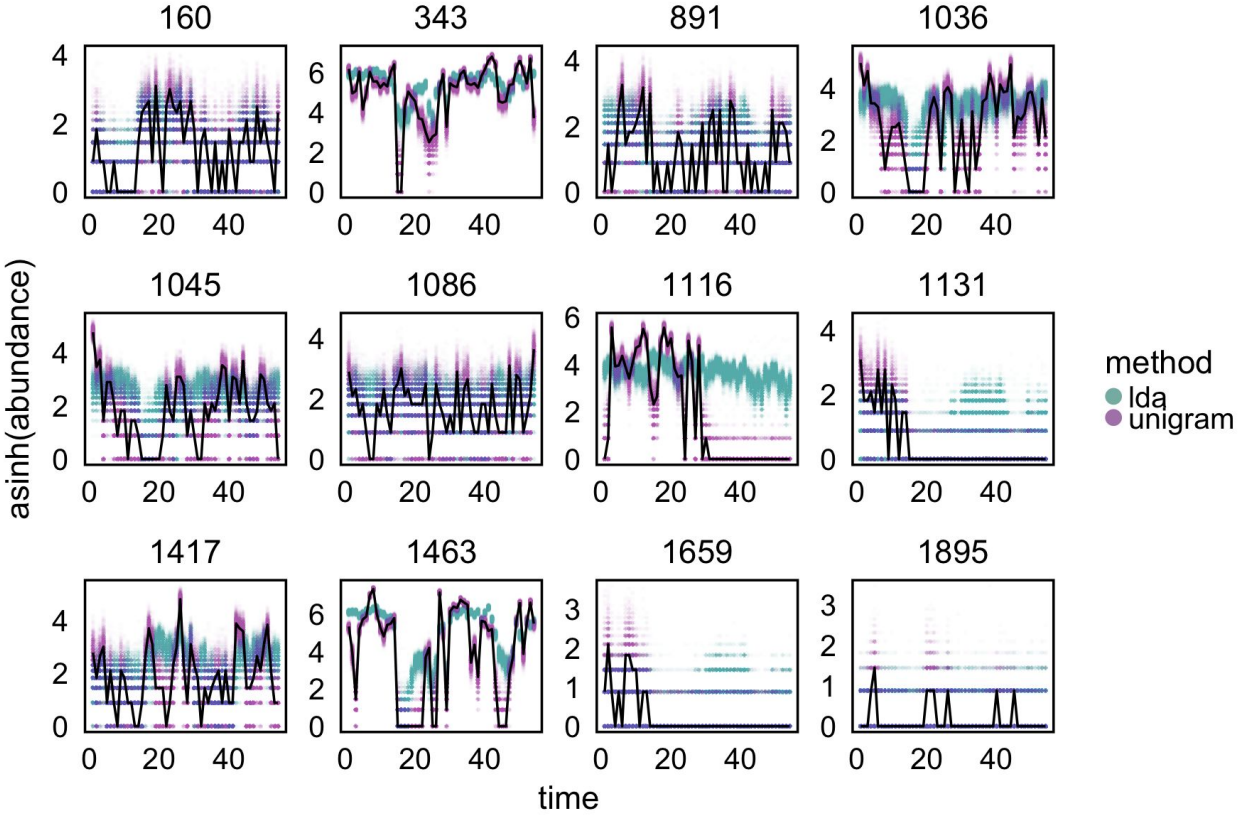


Model Criticism

- We simulate data from the posterior predictive,

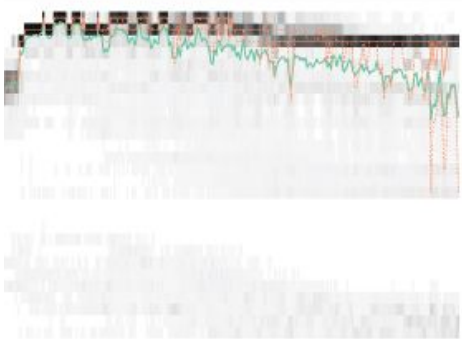
$$p\left(x^{*}|x\right)=\int p\left(x^{*}|\theta\right)p\left(\theta|x\right)d\theta$$

- Each panel represents one species, the black lines are observed species trajectories
- The blue and purple lines are LDA and unigram posterior predictive samples
  - The unigram model overfits
  - Species 1116 is an interesting outlier



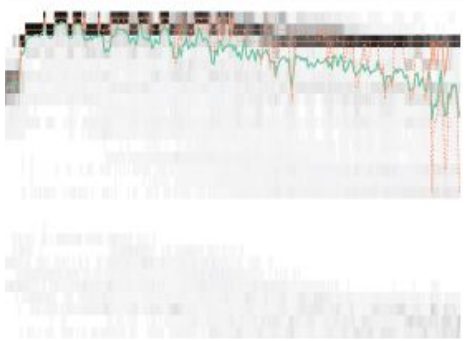
Other Workflow Guides

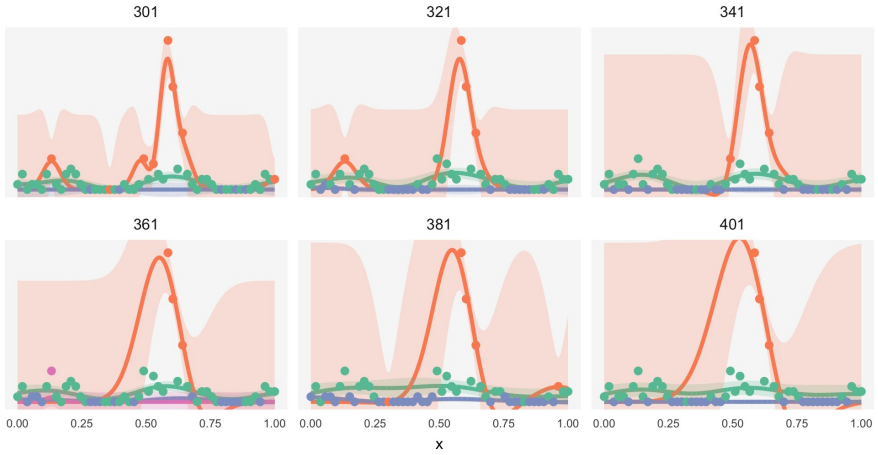
- Distill, critique, and compare existing approaches
- Guide for turning raw reads into abundance matrices



Other Workflow Guides

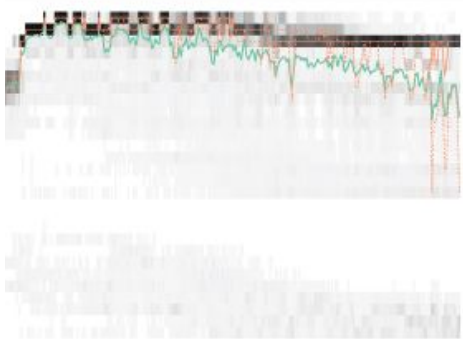
- Distill, critique, and compare existing approaches
- Guide for turning raw reads into abundance matrices
- Methods for segmenting dynamic regimes

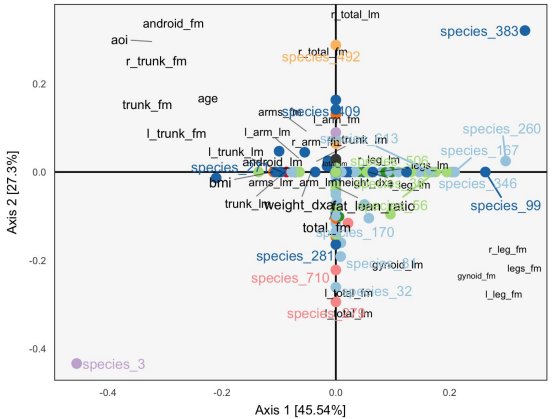


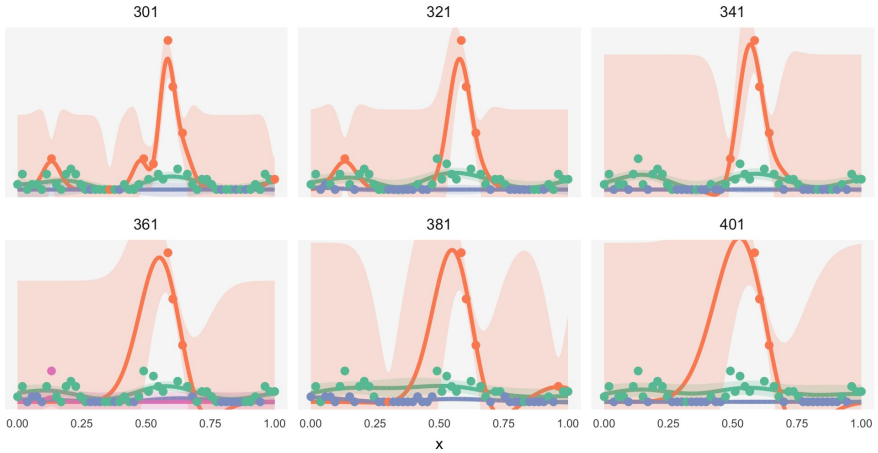


Other [Workflow Guides]

- Distill, critique, and compare existing approaches
- Guide for turning raw reads into abundance matrices
- Methods for segmenting dynamic regimes
- A survey of multitable modeling techniques





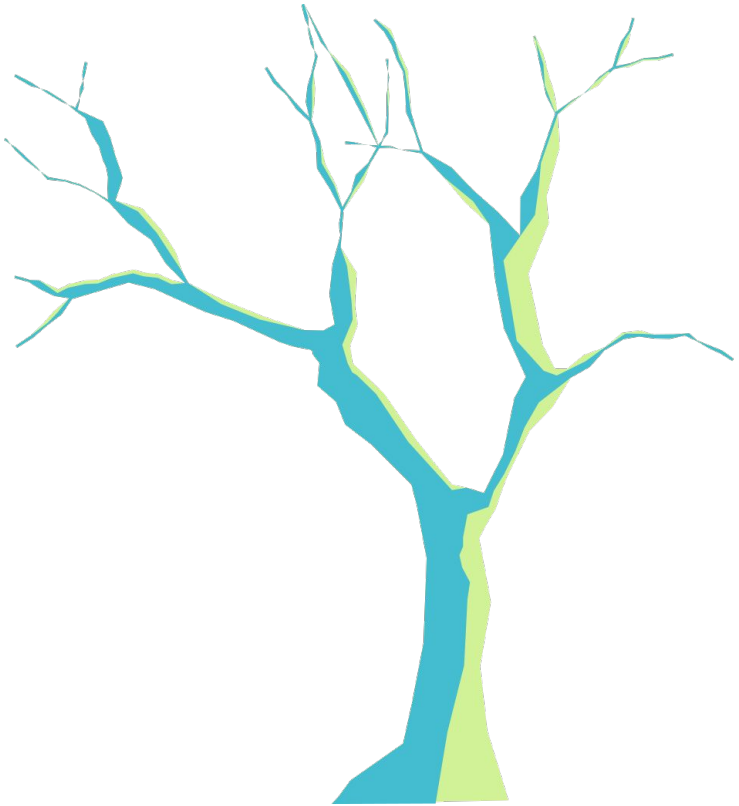


treelapse = trees + timelapse

Visualizing Hierarchies of Time Series

Tree-Structured Statistical Problems

- Differential Abundance
  - Compare bacterial abundances across conditions, for different species
  - Identify the taxonomic subtrees whose bacteria are differentially abundant

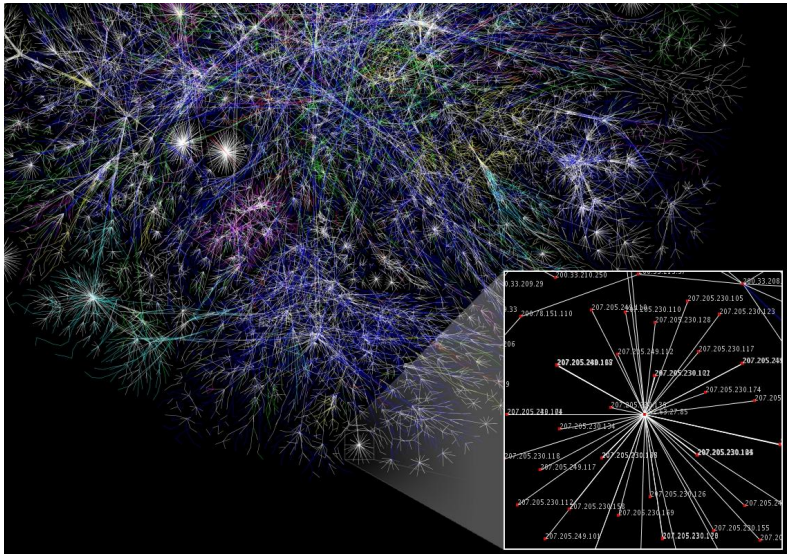


Tree-Structured Statistical Problems

- Differential Abundance
  - Compare bacterial abundances across conditions, for different species
  - Identify the taxonomic subtrees whose bacteria are differentially abundant
- Differential Dynamics
  - Describe changes in bacterial abundances, at the largest subtree where the pattern appears
  - Has an ecological flavor, with emphases on role of niches and environmental changes

Visualization Principles: Focus + Context

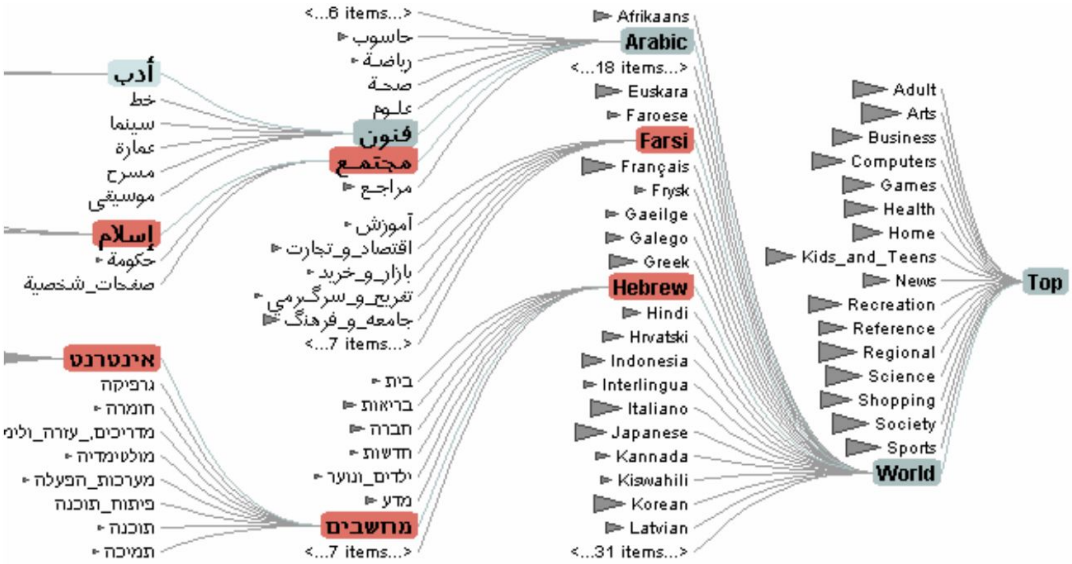
- Study a dataset across scales
  - Full network vs. neighborhood of a node
  - Full time series vs. short time window
- Interactivity allows transitions across scales
- *Focusing* on elements of interest while retaining *context*



A partial map of the internet, from The Opte Project, distributed under a Creative Commons license.

Degree-of-Interest (DOI) Trees

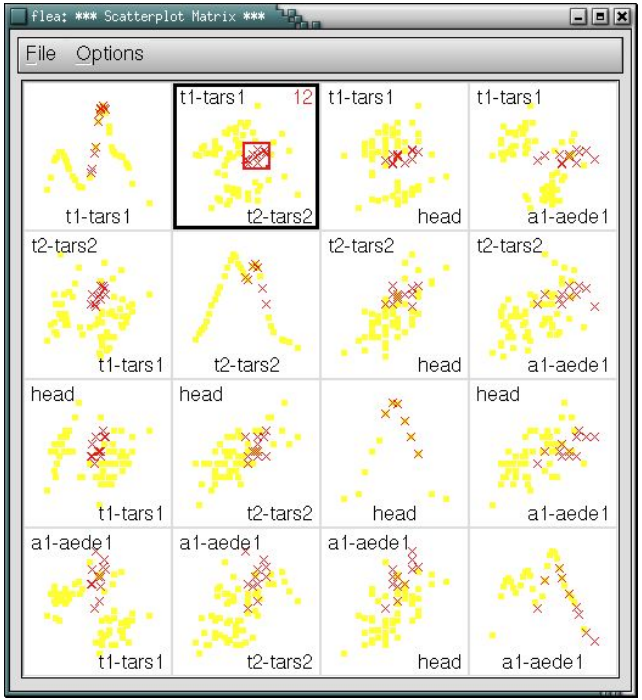
- An application of the focus + context idea to tree structured data
- Easily navigate the tree across different scales (e.g., species vs. phylum)



Screenshot of a DOI tree, from [Heer and Card 2004].

Visualization Principles: Linking

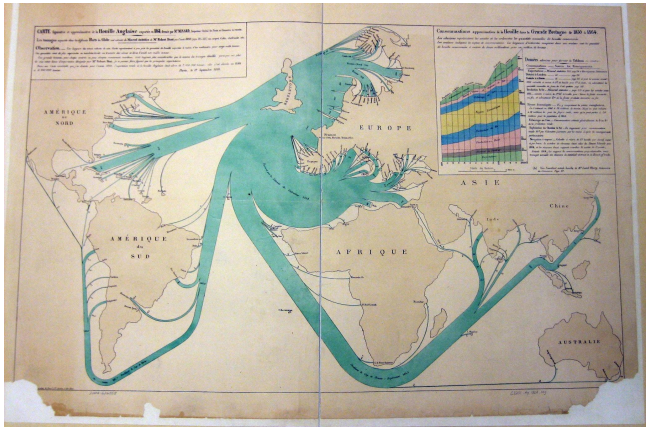
- Alternative representations reveal covariation [Becker and Cleveland 1987, Buja et. al. 1996]
- Can be useful in high-dimensional settings
- Conditional probability and database query interpretations
    - What are values other variables, conditional on constraints for some of them?



Linked scatterplot brushing, as implemented in GGobi [Voigt 2002].

Visualization Proposals

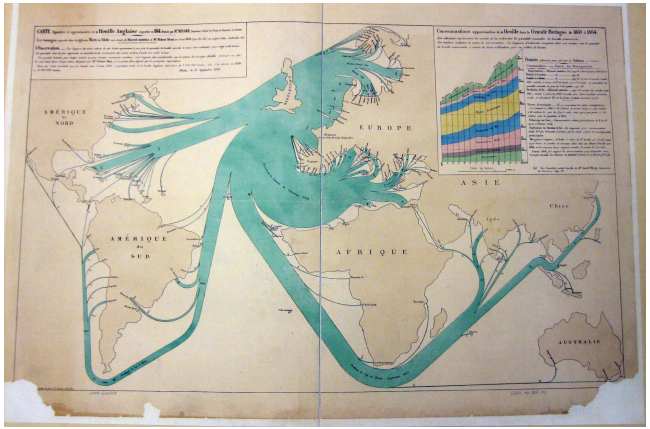
- Differential Abundance
    - DOI Tree: Usual DOI, but node sizes reflect species abundances
    - DOI Sankey: Extension to multiple groups, distinguished by color
    - Instead of separate trees, split branches into several groups



A sankey diagram of British coal exports, made in 1868 by Charles Joseph Minard.

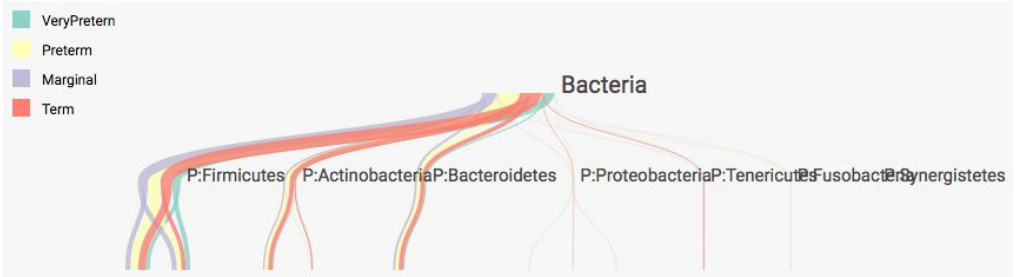
Visualization Proposals

- Differential Abundance
  - DOI Tree: Usual DOI, but node sizes reflect species abundances
  - DOI Sankey: Extension to multiple groups, distinguished by color
  - Instead of separate trees, split branches into several groups



A sankey diagram of British coal exports, made in 1868 by Charles Joseph Minard.

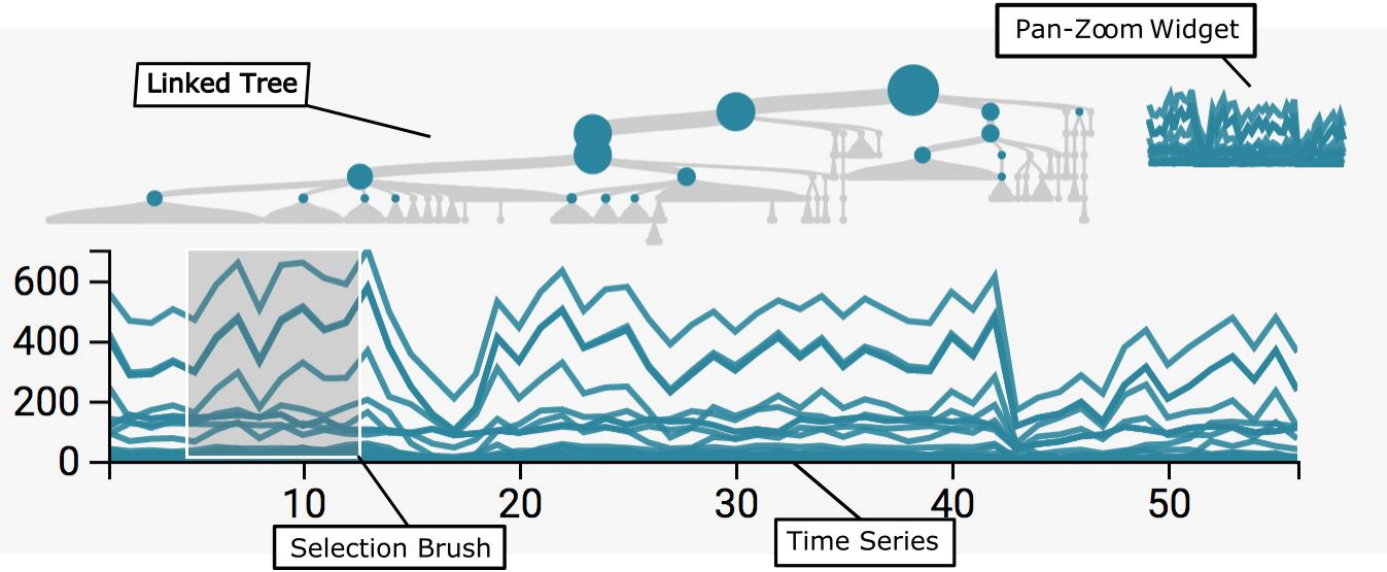
[Demo]



An example DOI Sankey, applied to a microbiome taxonomy.

Visualization Proposals

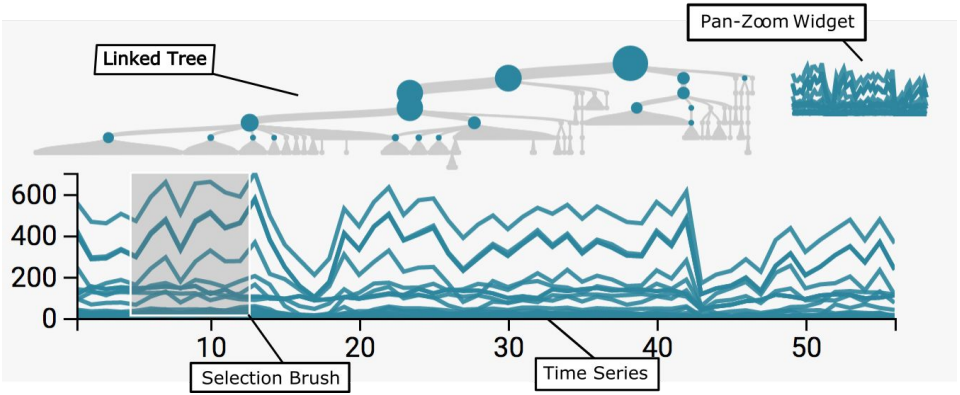
- Differential Dynamics
  - One-to-one mapping between nodes and time series
  - Timebox Trees: Link timeboxes and the tree, with selections over the time series
  - Treeboxes: Link timeboxes and the tree, with selections over the tree



Antibiotics revisited

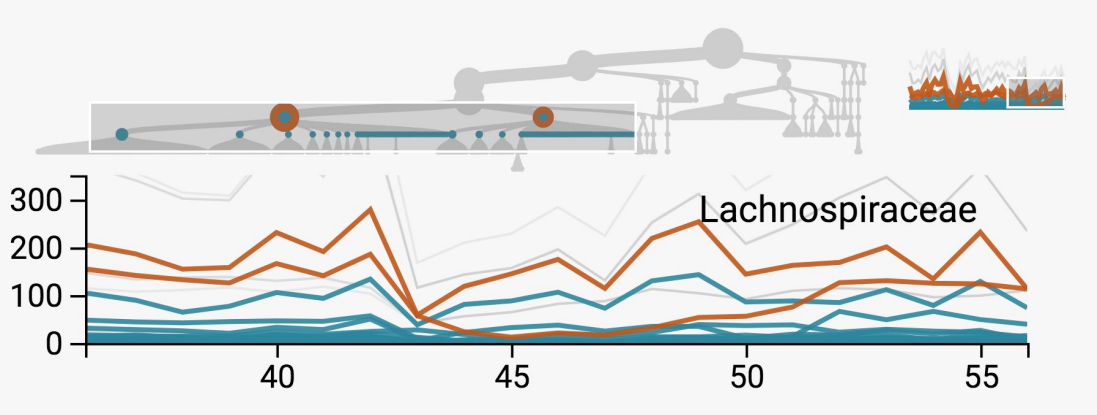
- We apply these methods to the data of [Dethlefsen and Relman 2011], which we analyzed earlier using latent variable models
- In contrast to that work, our goal here is interactive visualization of species abundance trajectories

Demo



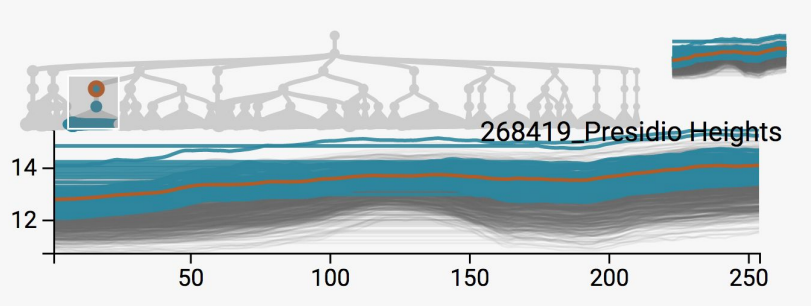
Differential Response among Firmicutes

- Drawing boxes over the tree and searching across suborders of Firmicutes suggests a differential responses between Lachnospiraceae and Ruminococcaceae
- Note that we are focusing on the time window surrounding the second antibiotic time course

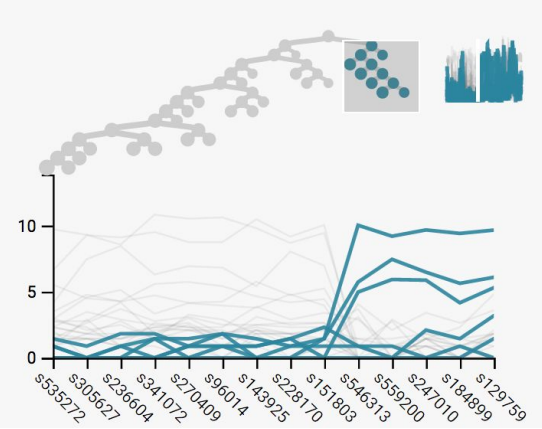


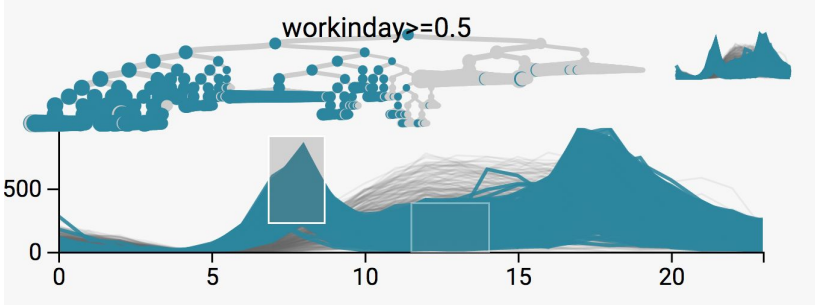
Beyond Phylogenies

Hierarchical structure is present in many contexts



Spatial Hierarchies: Nodes / series are geographic regions





Regression Trees: Nodes / series are averages of samples on one side of a split

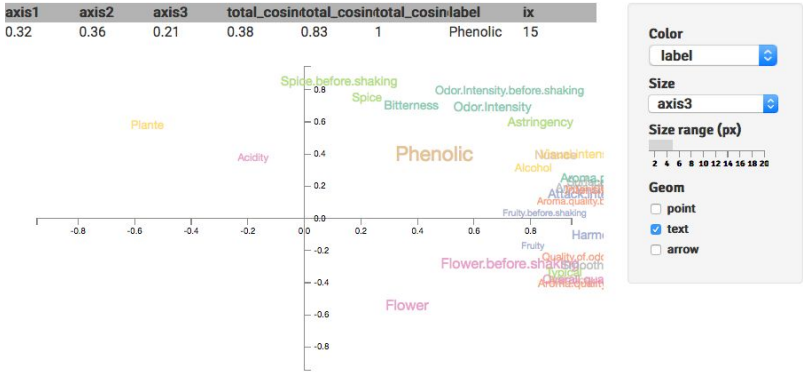
Hierarchical Clustering: Nodes / series are centroids within clusters

Other Visualization Projects

- **centroidview**: Combine differential dynamics with differential abundance
  - Demo
  - Package: github.com/krisrs1128/centroidview

Other Visualization Projects

- **centroidview**: Combine differential dynamics with differential abundance
  - Demo
  - Package: github.com/krisrs1128/centroidview
- **mvarVis**: Streamline interpretation of multivariate analysis results
  - Demo
  - Package: github.com/krisrs1128/mvarVis



Conclusion

- Microbiome studies are a source of richly structured, high-dimensional data
- Interactive data visualization and probabilistic modeling can guide the discovery and representation of latent structure

treelapse
- https://krisrs1128.github.io/treelapse/
- Interactive Visualization of Hierarchically Structured Data (Sankaran and Holmes 2017)

Latent variable modeling:
- https://github.com/krisrs1128/microbiome_plvm
- Latent Variable Modeling for the Microbiome (Sankaran and Holmes 2017 (arXiv))

Support
- NSF TR01 AI112401
- NIH T32 5T32GM096982-04
- Stanford Weiland Graduate Fellowship

References

Becker, Richard A., and William S. Cleveland. "Brushing scatterplots." *Technometrics* 29, no. 2 (1987): 127-142.

Bogomolov, Marina, Christine B. Peterson, Yoav Benjamini, and Chiara Sabatti. "Testing hypotheses on a tree: new error rates and controlling strategies." arXiv preprint arXiv:1705.07529 (2017).

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. "Latent dirichlet allocation." *Journal of machine Learning research* 3, no. Jan (2003): 993-1022.

Blei, David M., and John D. Lafferty. "Dynamic topic models." In *Proceedings of the 23rd international conference on Machine learning*, pp. 113-120. ACM, 2006.

Buja, Andreas, Dianne Cook, and Deborah F. Swayne. "Interactive high-dimensional data visualization." *Journal of computational and graphical statistics* 5, no. 1 (1996): 78-99.

Callahan, Ben J., Kris Sankaran, Julia A. Fukuyama, Paul J. McMurdie, and Susan P. Holmes. "Bioconductor workflow for microbiome data analysis: from raw reads to community analyses." *F1000Research* 5 (2016).

Canny, John. "GaP: a factor model for discrete data." In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 122-129. ACM, 2004.

Dethlefsen, Les, and David A. Relman. "Incomplete recovery and individualized responses of the human distal gut microbiota to repeated antibiotic perturbation." *Proceedings of the National Academy of Sciences* 108, no. Supplement 1 (2011): 4554-4561.

Earle, Kristen A., Gabriel Billings, Michael Sigal, Joshua S. Lichtman, Gunnar C. Hansson, Joshua E. Elias, Manuel R. Amieva, Kerwyn Casey Huang, and Justin L. Sonnenburg. "Quantitative imaging of gut microbiota spatial organization." *Cell host & microbe* 18, no. 4 (2015): 478-488.

Fukuyama, Julia, Laurie Rumker, Kris Sankaran, Pratheepa Jeganathan, Les Dethlefsen, David A. Relman, and Susan P. Holmes. "Multidomain analyses of a longitudinal human microbiome intestinal cleanout perturbation experiment." *PLoS Computational Biology* 13, no. 8 (2017): e1005706.

Heer, Jeffrey, and Stuart K. Card. "DOITrees revisited: scalable, space-constrained visualization of hierarchical data." In *Proceedings of the working conference on Advanced visual interfaces*, pp. 421-424. ACM, 2004.

Pritchard, Jonathan K., Matthew Stephens, and Peter Donnelly. "Inference of population structure using multilocus genotype data." *Genetics* 155, no. 2 (2000): 945-959.

Schloss, Patrick D., and Jo Handelsman. "The last word: books as a statistical metaphor for microbial communities." *Annu. Rev. Microbiol.* 61 (2007): 23-34.

Voigt, Robert. "An extended scatterplot matrix and case studies in information visualization. published as diplomarbeit." (2002).

Acknowledgements

Thank you Professor Holmes for all the valuable guidance, and for helping me discover what it means to be a scholar.

Thank you Professors Switzer, Baiocchi, Efron, and Relman for serving on the committee and offering useful feedback on my research.

Thank you to Lester and Persi for shaping some of my basic views of statistics and its role in the world.

Thank you to the Holmes lab -- Claire, Lan, Julia, Christof, Pratheepa, Nikos, Ben, Sergio, Joey M. -- for creating a welcoming and intellectually stimulating environment.

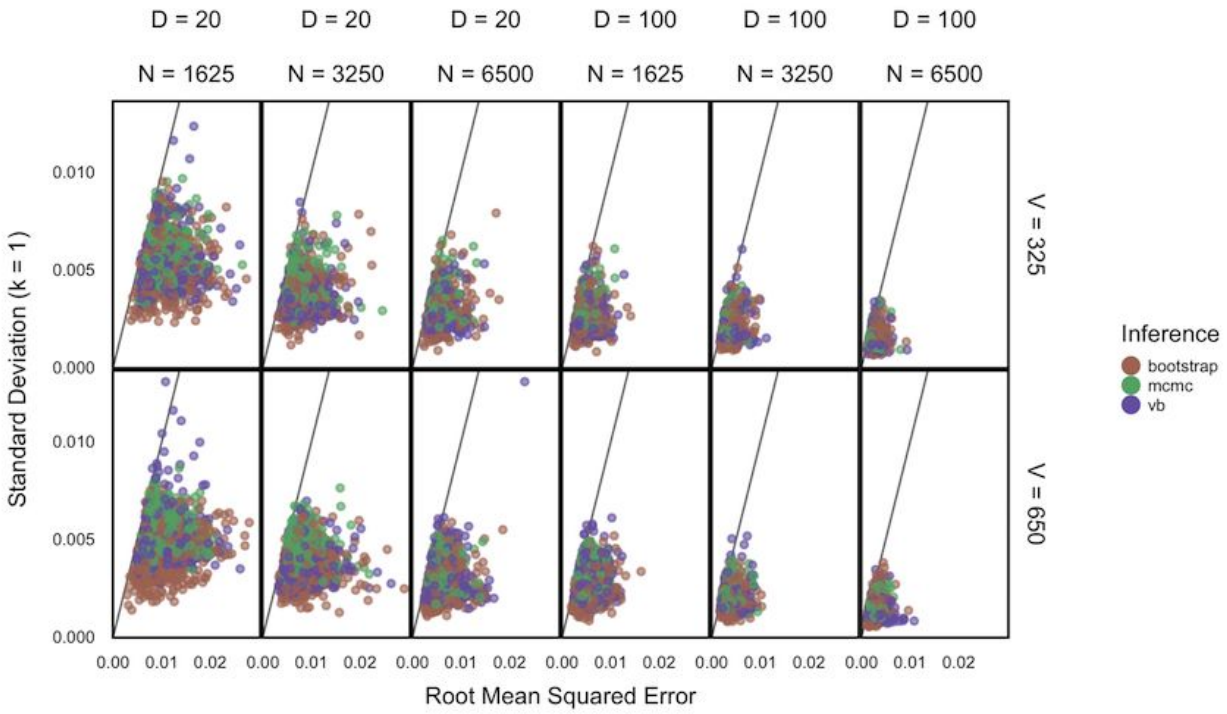
Thank you project collaborators -- Les, Lina, Suzanne, Shuo, Yan -- for the creative ideas and fun discussions.

Thank you friends -- Joey A., Jessy, Subhabrata, Zhou, Mona, Stephen, Keli, Pragya, Gene, Robin, Wanning, Chris, Ling, Prof. Zerlang, ... -- for making life here so interesting.

And finally thank you to my family for teaching me everything that was important, besides just statistics.

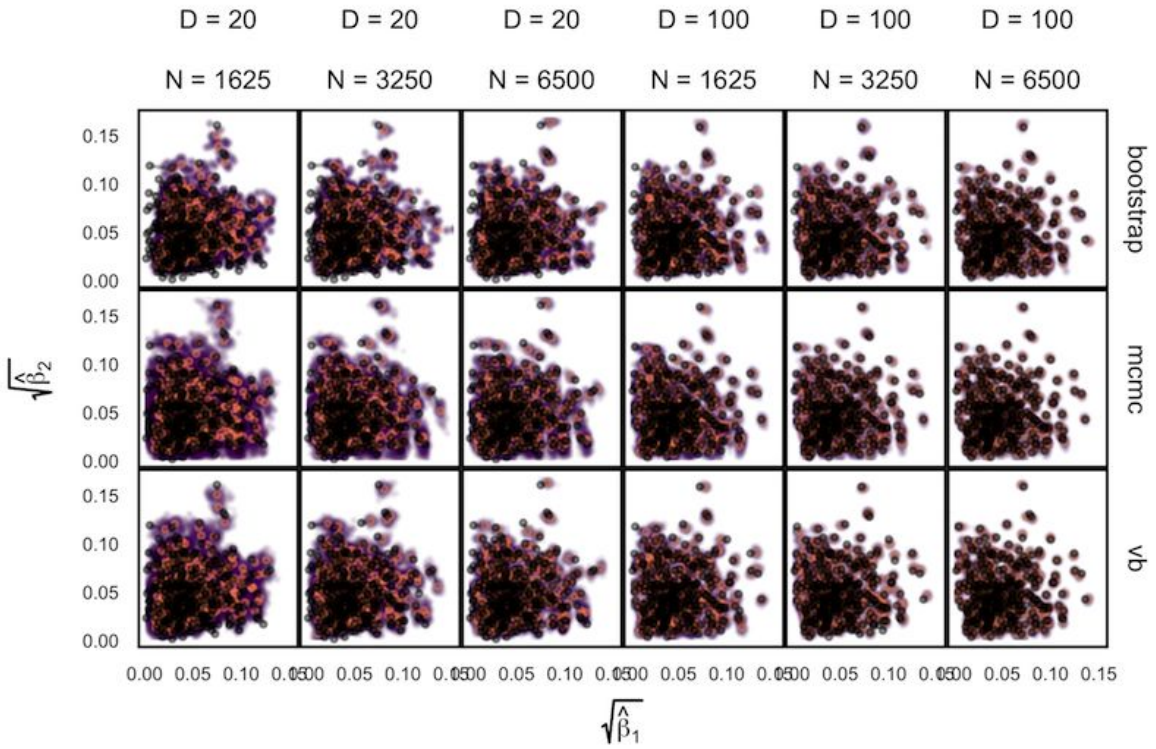
Simulation Study

- We perform a simulation study to better understand algorithms
  - D: Number of samples
  - V: Number of species
  - N: Sequencing depth
- Colors are inference techniques
- Variational Bayes and the Bootstrap generally underestimate posterior uncertainty, but all become precise with large N



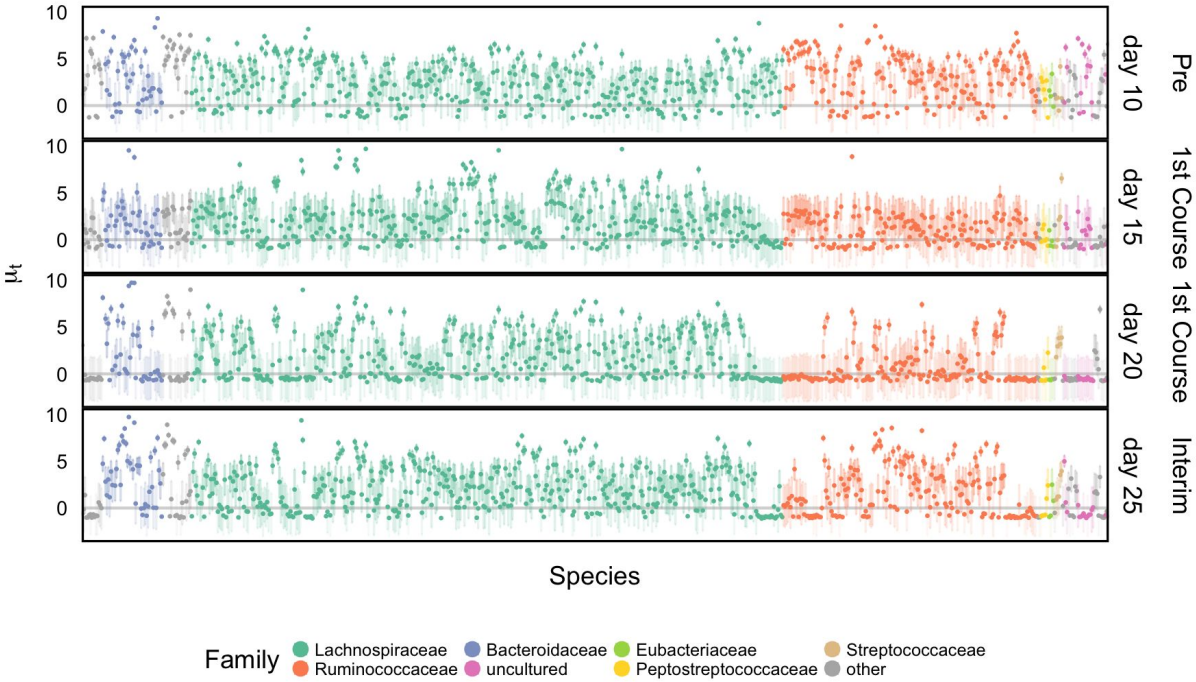
Simulation Study

- We perform a simulation study to better understand algorithms
  - D: Number of samples
  - V: Number of species
  - N: Sequencing depth
- Colors are inference techniques
- Variational Bayes and the Bootstrap generally underestimate posterior uncertainty, but all become precise with large N



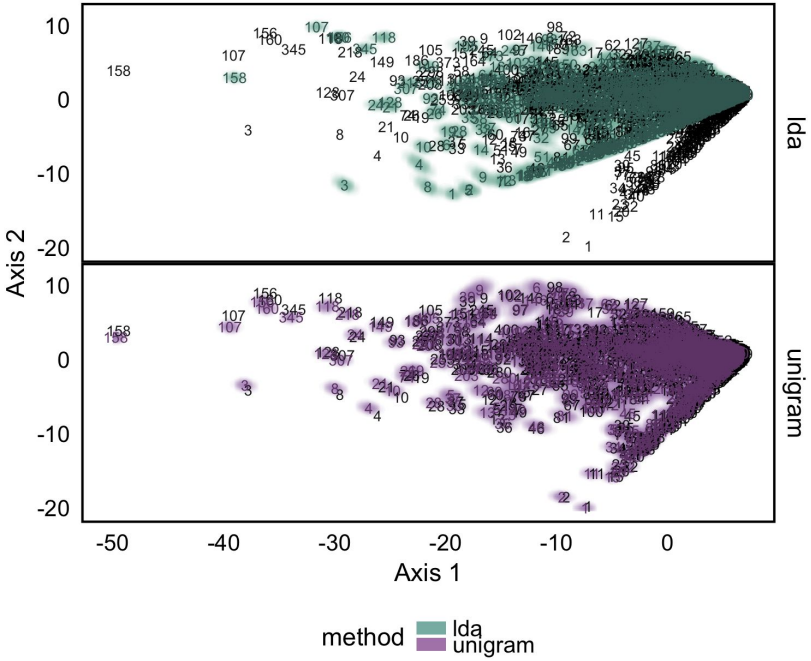
Unigram Smoothed Means: $\mu_{t(d)v}$

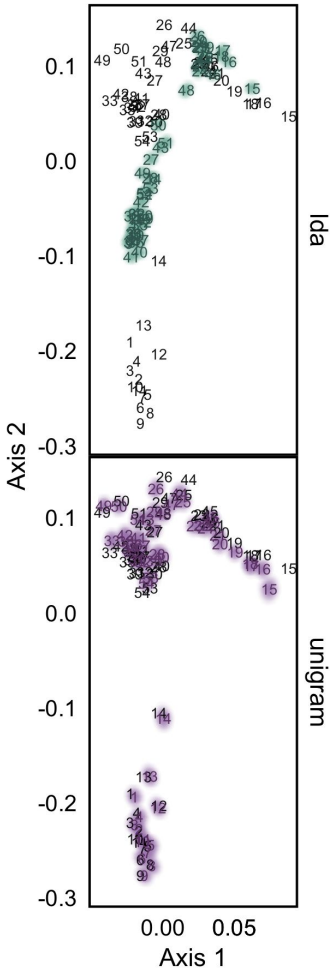
- Intervals represent posteriors for the smoothed means
- Each row is a timepoint, each interval is a species, colors are taxonomic groups
- The third panel down is the first one where antibiotics has been applied, note the drop in diversity



Model Criticism

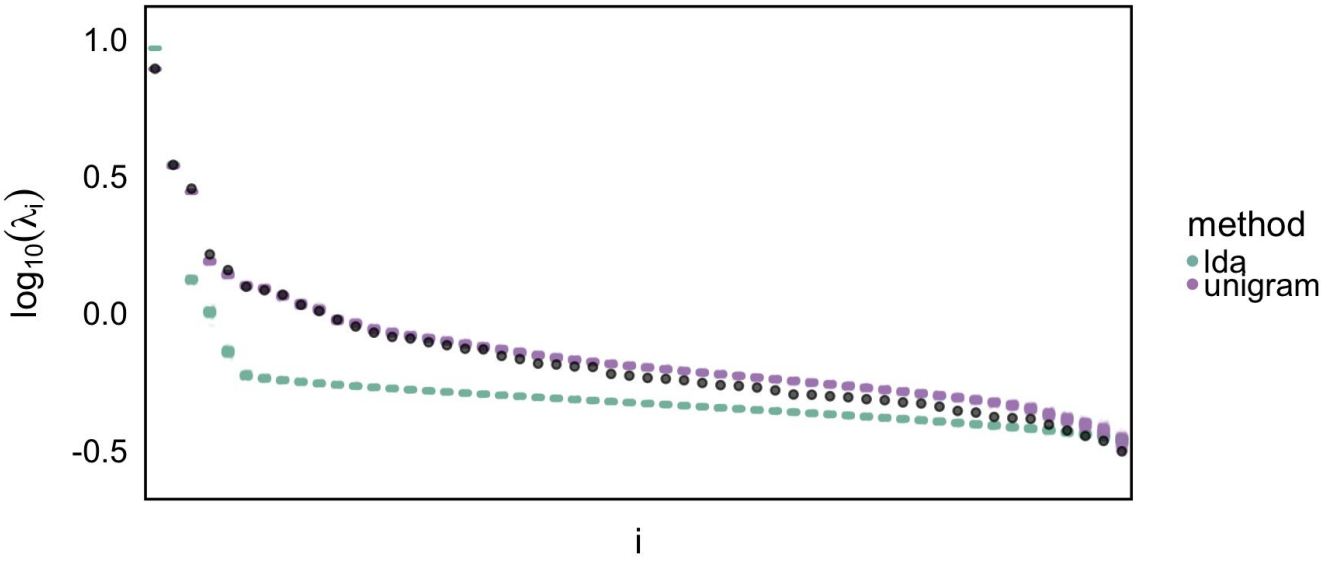
- We can perform PCA on the posterior predictive samples
- Loadings (species) are given on the left, while scores (samples) are given on the right
- Black text are observed scores / loadings, blue and purple clouds are posterior predictive samples
- Scores and loadings have been aligned using a procrustes rotation





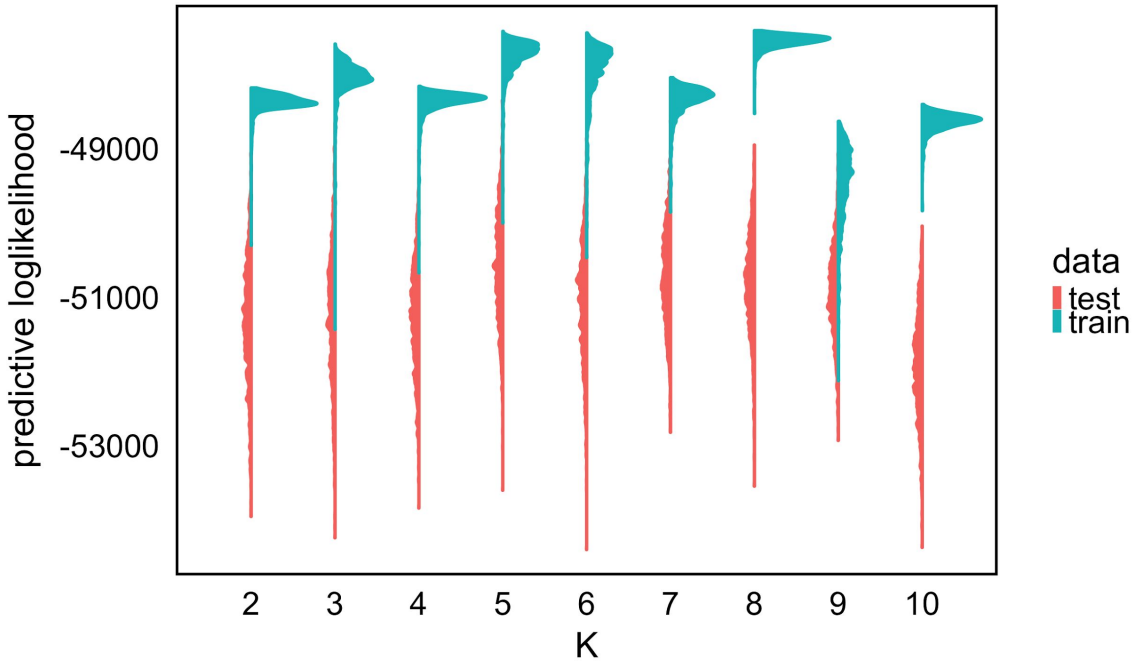
Model Criticism

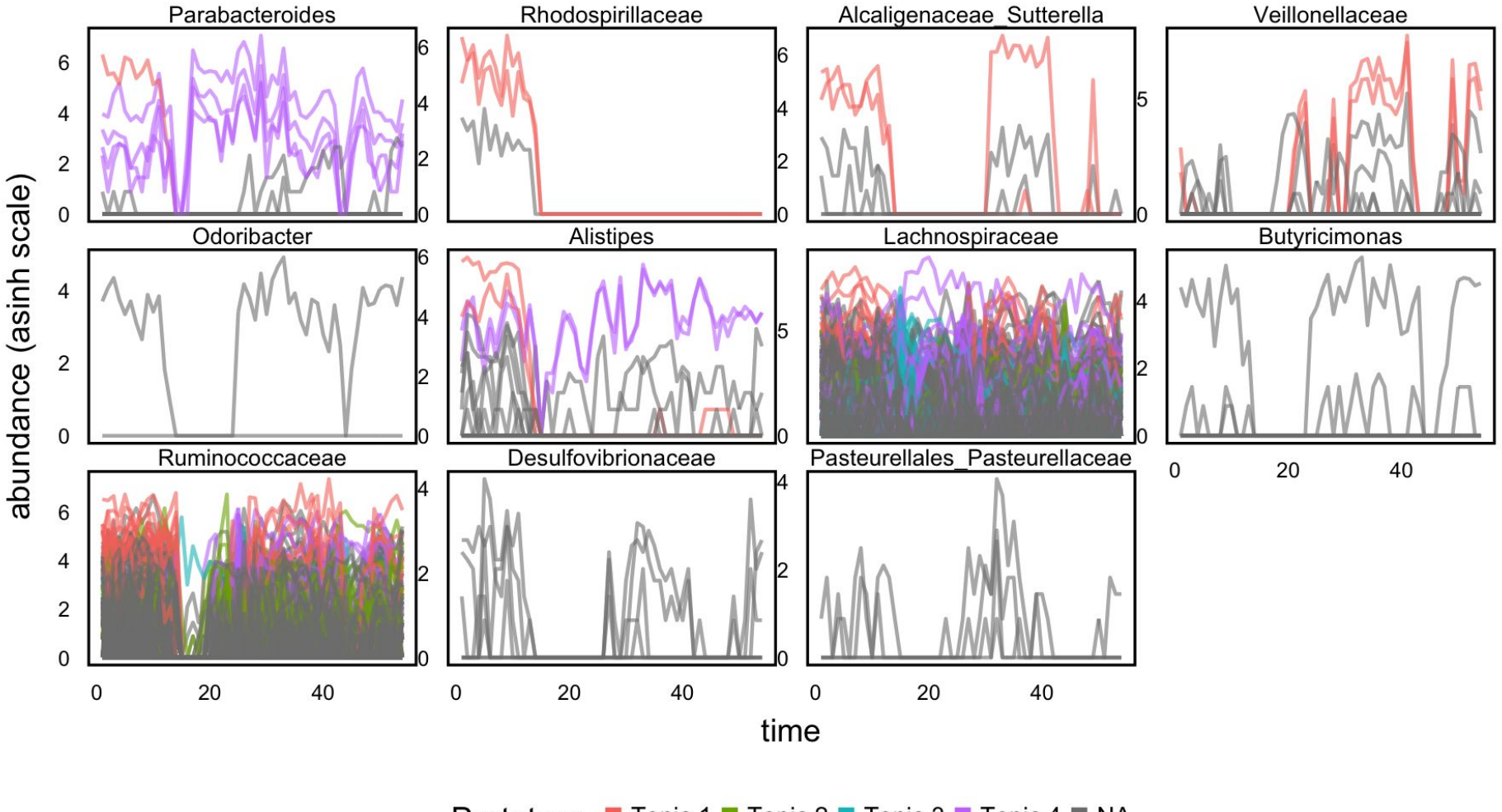
- Eigenvalues from real and posterior predictive samples
- Can be used to guide choice of tuning parameters



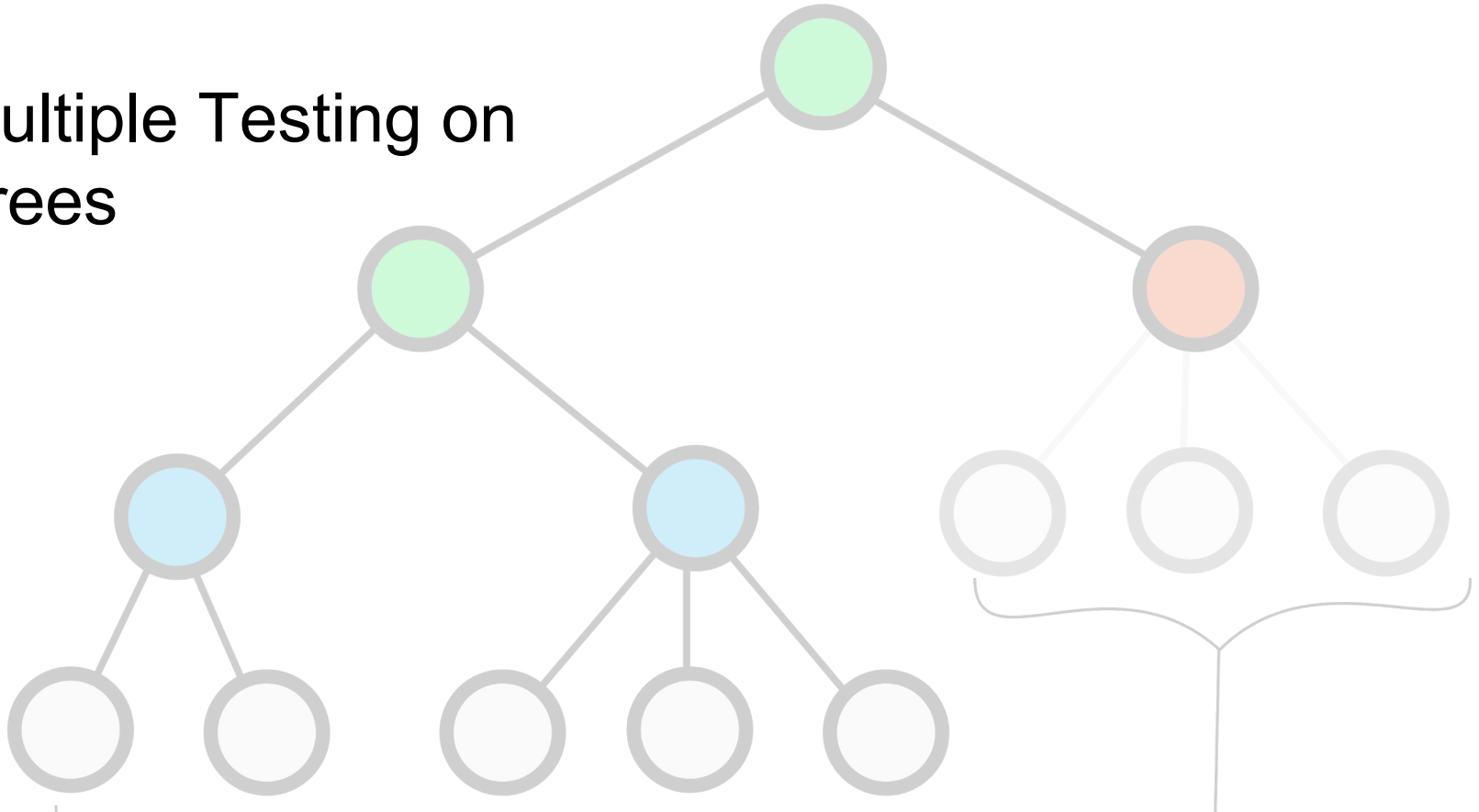
Train / Test log-likelihood

- We can choose k by evaluating the log-likelihood on test data
- Approximate predictive log-likelihoods by sampling the posterior
- Occam's razor effect: larger K don't necessarily increase training log-likelihoods




Parabacteroides
Rhodospirillaceae
Alcaligenaceae_Sutterella
Veillonellaceae
Odoribacter
Alistipes
Lachnospiraceae
Butyricimonas
Ruminococcaceae
Desulfovibrionaceae
Pasteurellales_Pasteurellaceae
abundance (asinh scale)
time
Prototype
Topic 1
Topic 2
Topic 3
Topic 4
NA

Multiple Testing on Trees

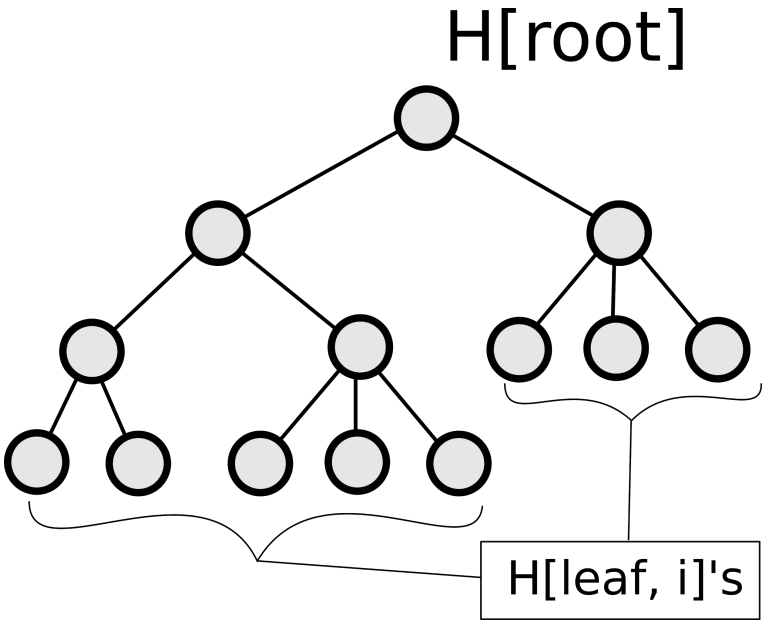


Problem and Approaches

- What is an inferential approach to tree-structured differential abundance?
  - FDR control only applies at the level of individual hypothesis
  - How can we leverage indirect evidence across the tree?
- This question is related to a few recent research programs
  - Testing groups of hypothesis: Group BH, p-filter, Multilayer Knockoffs
  - Hierarchical search: TreeQTL, DAGGER
  - Isolated proposals from bioinformatics: HALLA, OiMAT, StructFDR

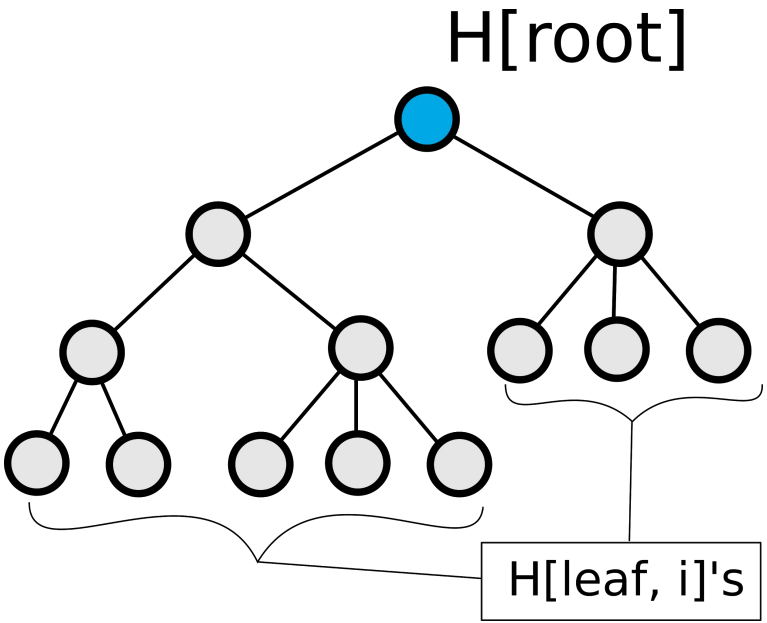
A simple algorithm and its properties

A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



A simple algorithm and its properties

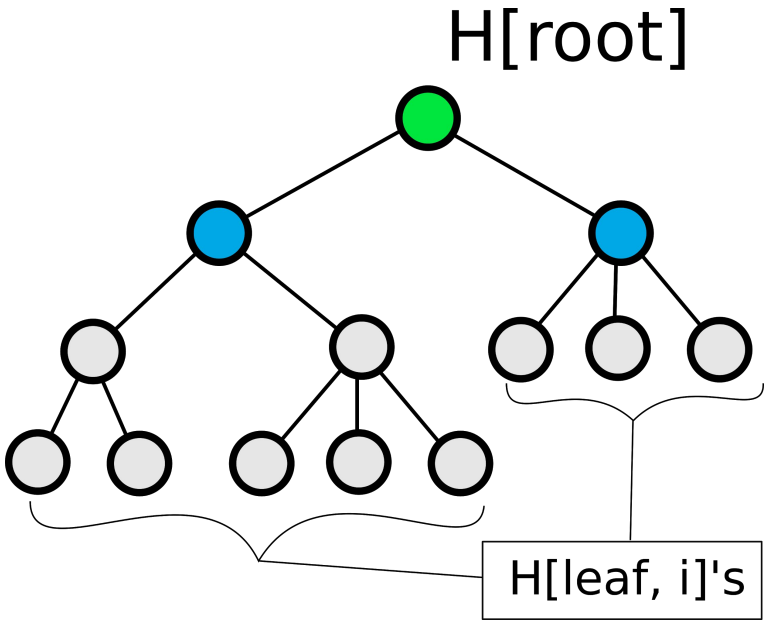
A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



- Test the root. If this fails to reject, report no discoveries.

A simple algorithm and its properties

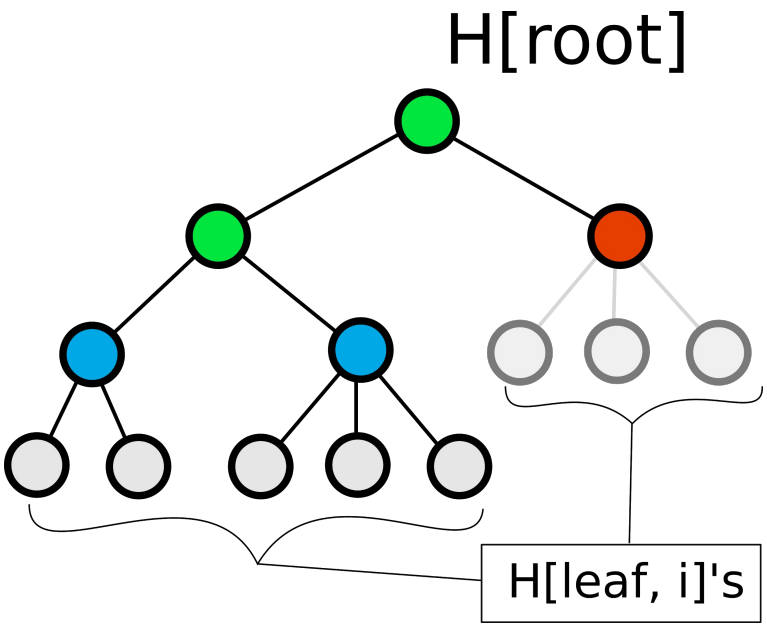
A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



- Test the root. If this fails to reject, report no discoveries.
- If it is rejected, test the children hypothesis with BH.

A simple algorithm and its properties

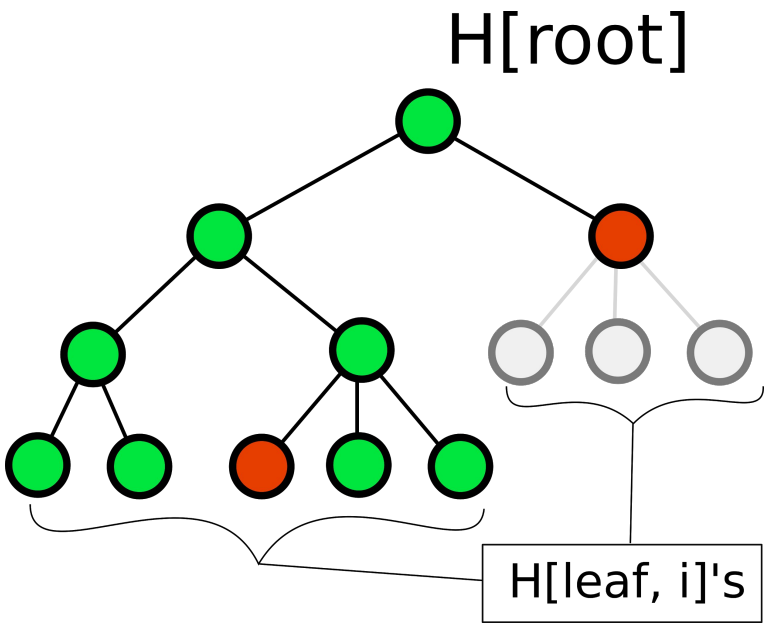
A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



- Test the root. If this fails to reject, report no discoveries.
- If it is rejected, test the children hypothesis with BH.
- If it fails to reject, don’t test any of the descendant hypothesis

A simple algorithm and its properties

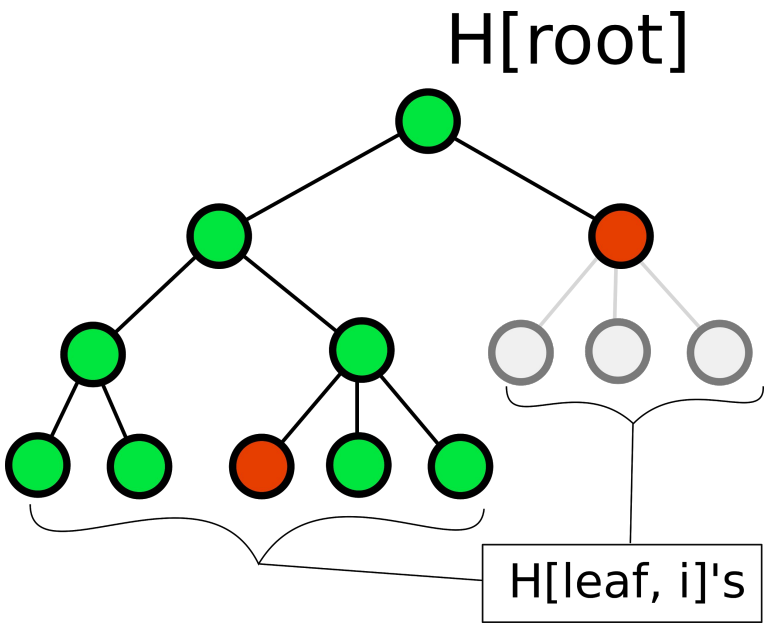
A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



- Test the root. If this fails to reject, report no discoveries.
- If it is rejected, test the children hypothesis with BH.
- If it fails to reject, don't test any of the descendant hypothesis
- Continue testing in this fashion down to the leaves

A simple algorithm and its properties

A proposal of Yekutieli [2008] is to proceed top-down through a hierarchy, using an FDR controlling procedure for each collection of sibling nodes.



- This procedure was shown to guarantee different types of false discovery rates
- Tree FDR $\leq 2\delta^* q$
- Level-L FDR $\leq 2\delta^* Lq$

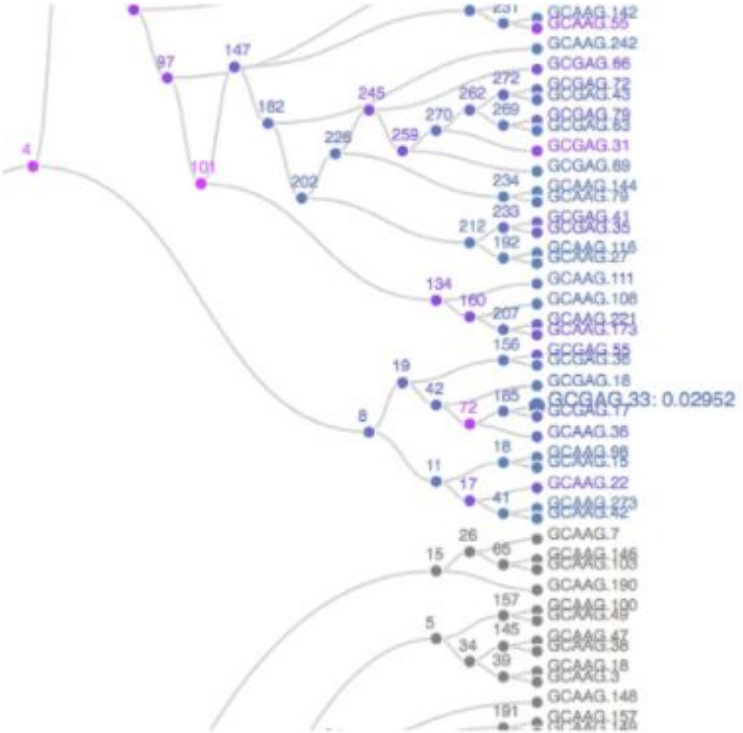
q Sibling-wise BH level

L Depth in the tree

δ^* Universal constant (< 1.44)

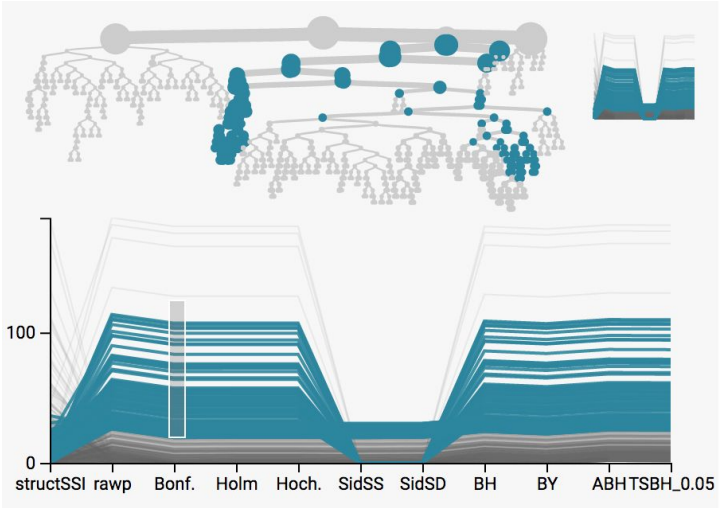
Implementation

- We designed and implemented the structSSI (“structured simultaneous and selective inference”) R package [Sankaran and Holmes 2014], implementing this method and a group testing procedure
- Includes utilities for
  - Computing statistics along the tree
  - Interactively inspecting results of confirmatory analysis is better than printing tables of p-values



Implementation

- We designed and implemented the structSSI (“structured simultaneous and selective inference”) R package [Sankaran and Holmes 2014], implementing this method and a group testing procedure
- Includes utilities for
  - Computing statistics along the tree
  - Interactively inspecting results of confirmatory analysis is better than printing tables of p-values



Developments

- Since our work on structSSI, alternative approaches for testing along trees have been proposed
- Petersen et. al. [2016] (TreeQTL), Bogomolov et. al [2017], and Ramdas et. al [2017] (p-filter) describe variations of Yekutieli [2008] that control a selective version of FDR
- Barber and Ramdas [2017] (DAGGER) and Katsevich and Sabatti [2017] (Multilayer Knockoff Filter) describe multilayer approaches to FDR control in the multiple testing and knockoff settings

References

Barber, Rina Foygel, and Aaditya Ramdas. "The p-filter: multi-layer FDR control for grouped hypotheses." *arXiv preprint arXiv:1512.03397* (2015).

Bogomolov, Marina, Christine B. Peterson, Yoav Benjamini, and Chiara Sabatti. "Testing hypotheses on a tree: new error rates and controlling strategies." arXiv preprint arXiv:1705.07529 (2017).

Hu, James X., Hongyu Zhao, and Harrison H. Zhou. "False discovery rate control with groups." *Journal of the American Statistical Association* 105, no. 491 (2010): 1215-1227.

Peterson, Christine B., Marina Bogomolov, Yoav Benjamini, and Chiara Sabatti. "TreeQTL: hierarchical error control for eQTL findings." *Bioinformatics* 32, no. 16 (2016): 2556-2558.

Ramdas, Aaditya, Jianbo Chen, Martin J. Wainwright, and Michael I. Jordan. "DAGGER: A sequential algorithm for FDR control on DAGs." *arXiv preprint arXiv:1709.10250* (2017).

Sankaran, Kris, and Susan Holmes. "structSSI: Simultaneous and Selective Inference for Grouped or Hierarchically Structured Data." *Journal of statistical software* 59, no. 13 (2014): 1.

Yekutieli, Daniel. "Hierarchical false discovery rate–controlling methodology." *Journal of the American Statistical Association.* 103, no. 481 (2008): 309-316.