

Statistical Methods for Microbiome Data

Jing Ma

Public Health Sciences Division
Fred Hutch Cancer Research Center
jingma@fredhutch.org

Feb 8, 2018

Human microbiome

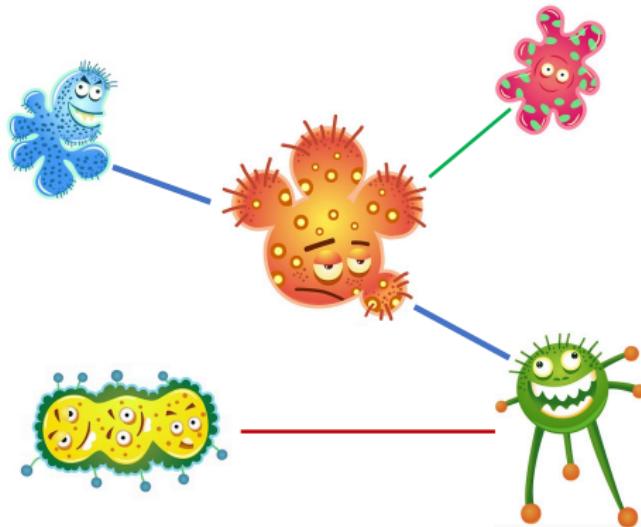


39 trillion bacterial cells > # of human cells

Part I: Testing Differences in Microbial Networks

Part II: Metagenomic Microbial Community Profiling

Networks

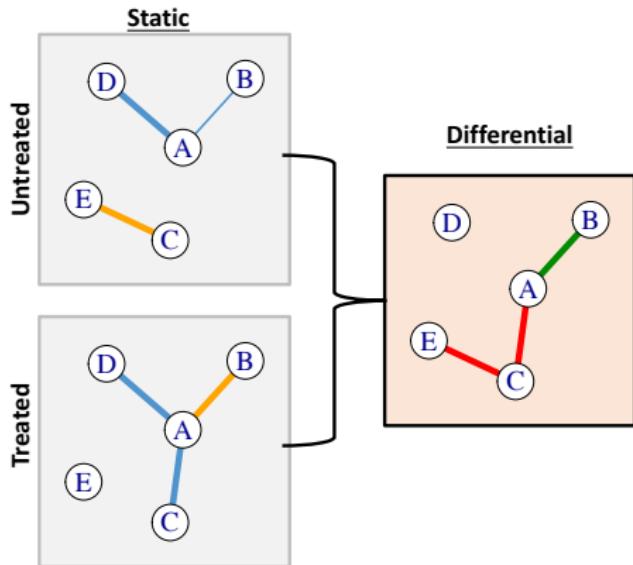


Networks

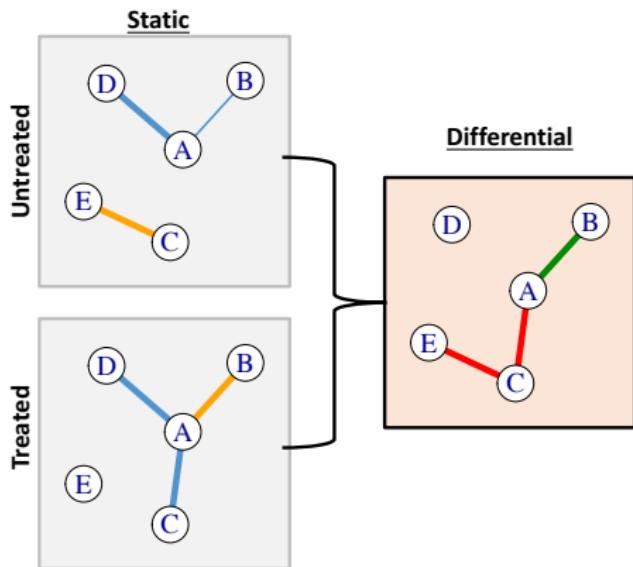


Question 1: What is a microbial network?

Differential network analysis



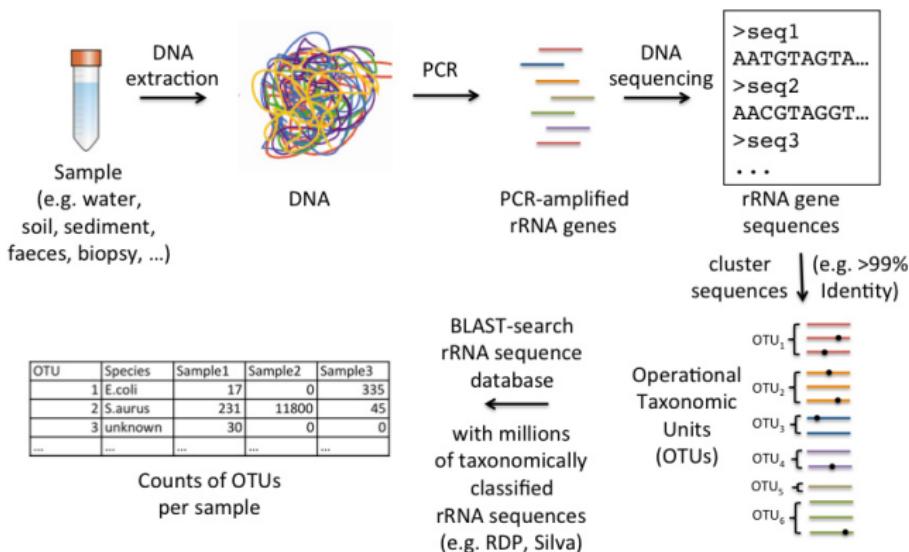
Differential network analysis



Question 2: How to test differences of microbial networks?

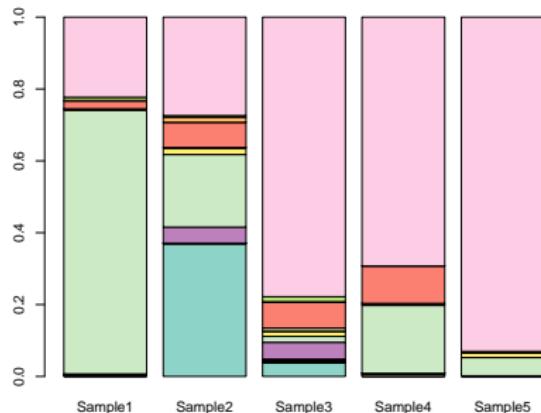
16S microbiome data

16S rRNA gene sequencing



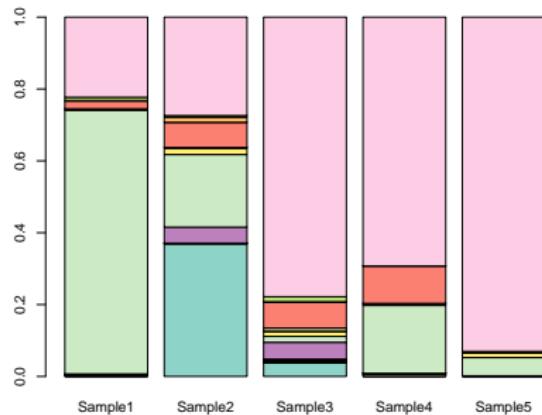
The microbiome data

- ▶ Microbiome data are **compositional**.



The microbiome data

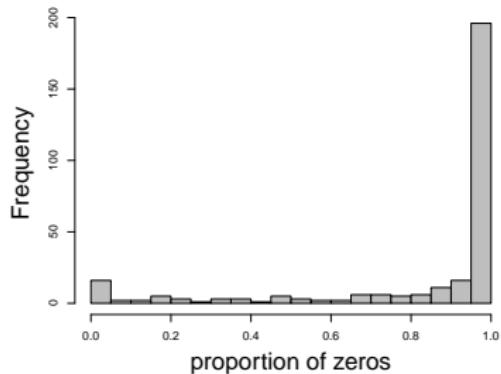
- ▶ Microbiome data are **compositional**.



- ▶ Methods that work well for normal random variables do not apply!

The microbiome data

- ▶ The compositional vector is very sparse.



Existing models for microbial relationships

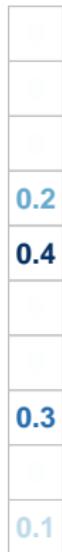
- ▶ Dissimilarity: [ReBoot](#) (Faust et al. 2012).
- ▶ Correlation: [SparCC](#) (Friedman and Alm 2012), [MENAP](#) (Deng et al. 2012), [CCLasso](#) (Fang et al. 2015), [REBACCA](#) (Ban et al. 2015).
- ▶ Probabilistic graphical models: [SPEIC-EASI](#) (Kurtz et al. 2015), [MINT](#) (Biswas et al. 2016).
- ▶ Limitations: marginal relationships, permutation-based significance test, zeros replaced with a pseudocount.

Microbial conditional dependency relationships

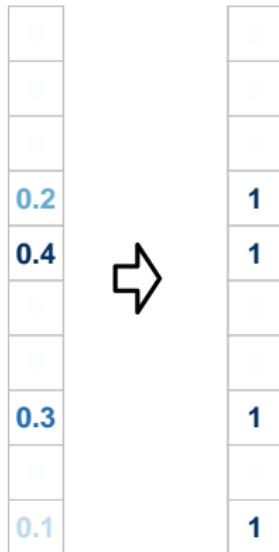
We want a model that

- ▶ captures the conditional dependency relationships among microbes,
- ▶ address the sparsity issue,
- ▶ infers differential network with **false discovery rate control**.

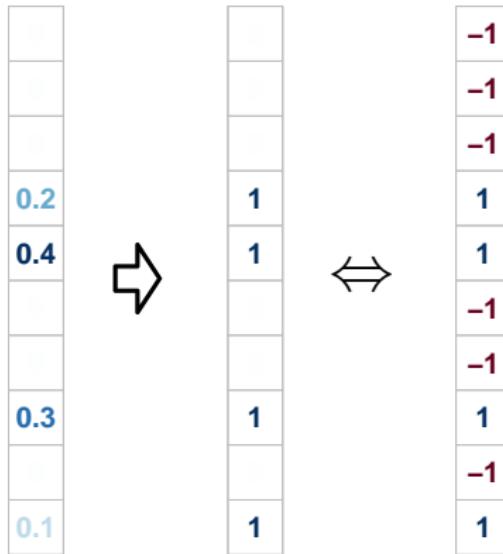
Our framework: binary Markov network (Ising model)



Our framework: binary Markov network (Ising model)



Our framework: binary Markov network (Ising model)

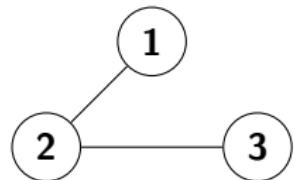


Our framework: binary Markov network (Ising model)

- ▶ Joint distribution $P_\Theta(X) \propto \exp \left\{ \sum_{1 \leq r < t \leq p} X_r X_t \theta_{rt} \right\}$.

Our framework: binary Markov network (Ising model)

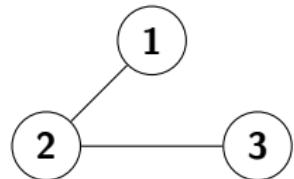
- ▶ Joint distribution $P_\Theta(X) \propto \exp \left\{ \sum_{1 \leq r < t \leq p} X_r X_t \theta_{rt} \right\}$.
- ▶ Conditional independence



$$\Theta = \begin{pmatrix} * & \theta_{12} & 0 \\ \theta_{21} & * & \theta_{23} \\ 0 & \theta_{32} & * \end{pmatrix}$$

Our framework: binary Markov network (Ising model)

- ▶ Joint distribution $P_\Theta(X) \propto \exp \left\{ \sum_{1 \leq r < t \leq p} X_r X_t \theta_{rt} \right\}$.
- ▶ Conditional independence



$$\Theta = \begin{pmatrix} * & \theta_{12} & 0 \\ \theta_{21} & * & \theta_{23} \\ 0 & \theta_{32} & * \end{pmatrix}$$

- ▶ Harris. Ecology (2016): small $p \leq 20$

Estimation of the Ising model

- ▶ Maximum likelihood: ok for small p , but intractable for large p :

$$P_{\Theta}(X) \propto \exp \left\{ \sum_{(r,t)} X_r X_t \theta_{rt} \right\}$$

¹Ravikumar et al. Ann. Stat. (2010)

Estimation of the Ising model

- ▶ Maximum likelihood: ok for small p , but intractable for large p :

$$P_{\Theta}(X) \propto \exp \left\{ \sum_{(r,t)} X_r X_t \theta_{rt} \right\}$$

- ▶ Nodewise (**penalized**) logistic regression¹ for large p :

$$P(X_r | X_{-r}) = \frac{\exp(X_r \sum_{j \neq r} X_j \theta_{rj})}{\exp(-X_r \sum_{j \neq r} X_j \theta_{rj}) + \exp(X_r \sum_{j \neq r} X_j \theta_{rj})}.$$

¹Ravikumar et al. Ann. Stat. (2010)

Inference beyond estimation

- ▶ Inference of a single network
 - ▶ done for Gaussian graphical model (GGM)², but not for Ising model!

²Liu. AOS (2013)

³Xia, Cai and Cai. Biometrika (2015)

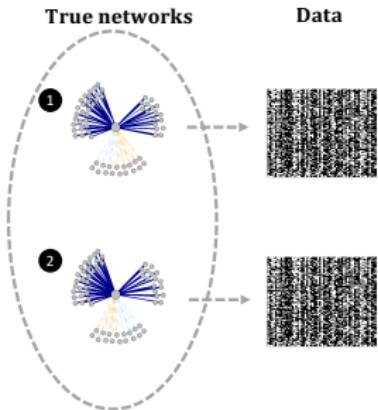
Inference beyond estimation

- ▶ Inference of a single network
 - ▶ done for Gaussian graphical model (GGM)², but not for Ising model!
- ▶ Two-sample (and multi-sample) inference
 - ▶ done for GGM³, but not for Ising model!

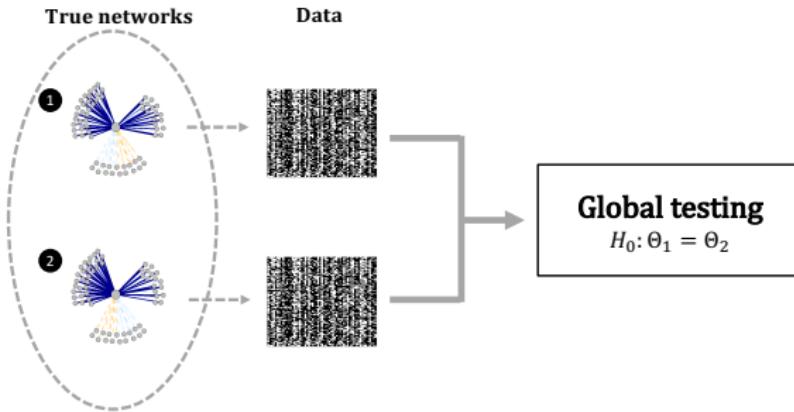
²Liu. AOS (2013)

³Xia, Cai and Cai. Biometrika (2015)

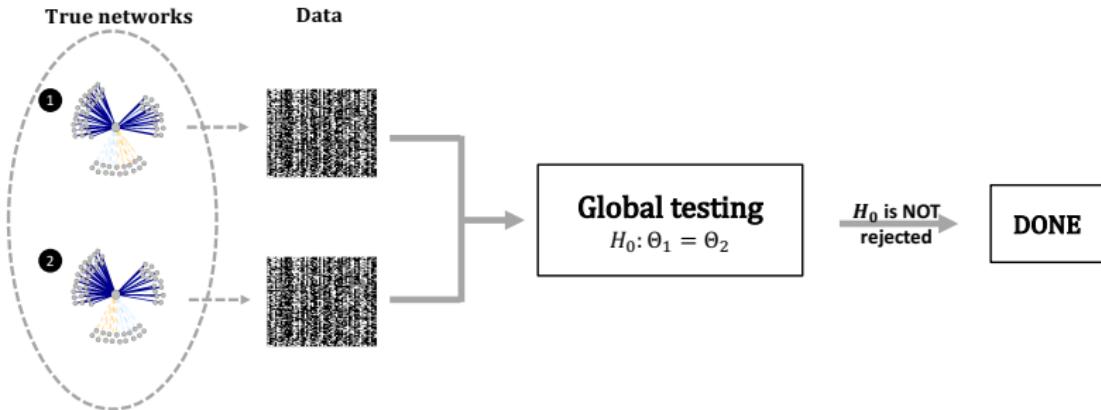
Inference of differential Markov network



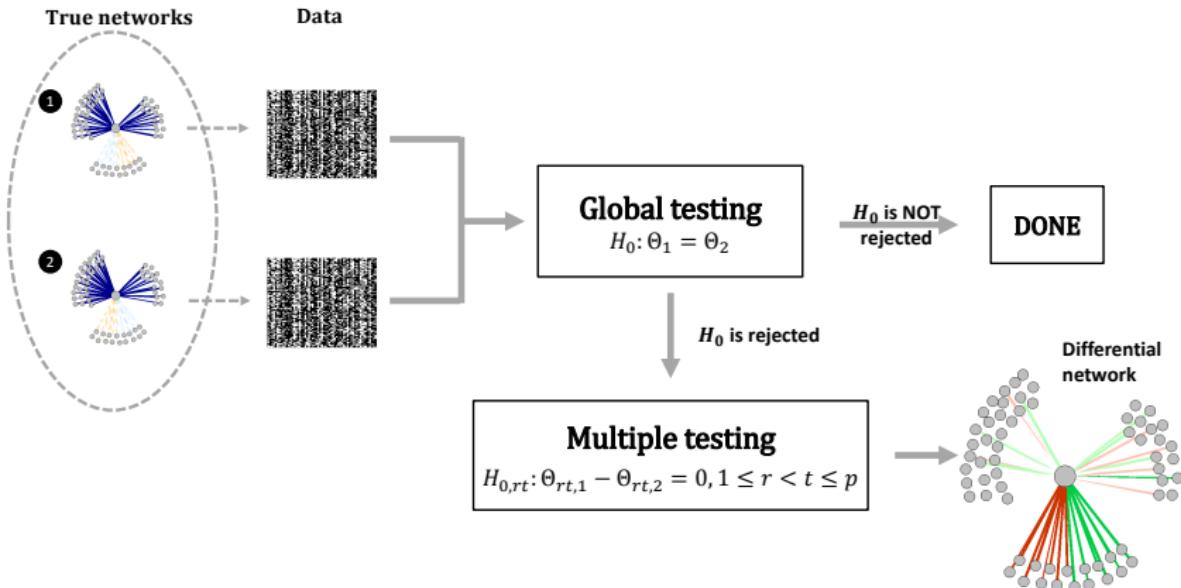
Inference of differential Markov network



Inference of differential Markov network



Inference of differential Markov network



Global testing $H_0 : \Theta_1 = \Theta_2$

$$H_0 : \max_{1 \leq r < t \leq p} |\theta_{rt,1} - \theta_{rt,2}| = 0.$$

Global testing $H_0 : \Theta_1 = \Theta_2$

$$H_0 : \max_{1 \leq r < t \leq p} |\theta_{rt,1} - \theta_{rt,2}| = 0.$$

Suppose we have **good** estimators $\check{\theta}_{rt,k}$ and their variances $\check{s}_{rt,k}$. Define

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

The test statistic for H_0 is

$$\max_{1 \leq r < t \leq p} W_{rt}^2.$$

Global testing $H_0 : \Theta_1 = \Theta_2$

$$H_0 : \max_{1 \leq r < t \leq p} |\theta_{rt,1} - \theta_{rt,2}| = 0.$$

Suppose we have **good** estimators $\check{\theta}_{rt,k}$ and their variances $\check{s}_{rt,k}$. Define

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

The test statistic for H_0 is

$$\max_{1 \leq r < t \leq p} W_{rt}^2.$$

Intuition:

- ▶ max statistic is most powerful against sparse alternatives.

What is a good estimator for Θ_k ?

- ▶ Nodewise penalized logistic regression for large p .

What is a good estimator for Θ_k ?

- ▶ Nodewise penalized logistic regression for large p .
 - ▶ Easy to implement, but the estimator is **biased**!

What is a good estimator for Θ_k ?

- ▶ Nodewise penalized logistic regression for large p .
 - ▶ Easy to implement, but the estimator is **biased**!
- ▶ Solution: debiasing via projection

Debiasing the Lasso via projection⁴

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.

⁴Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection⁴

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

⁴Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection⁴

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

- ▶ The debiased estimator (given $\hat{\beta}$) is

$$\check{\beta}_r = \beta_r^{lin} - \widehat{\text{bias}} = \hat{\beta}_r + \frac{v' (Y - \mathbf{Z}\hat{\beta})}{v' Z_r}.$$

⁴Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection⁴

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

- ▶ The debiased estimator (given $\hat{\beta}$) is

$$\check{\beta}_r = \beta_r^{lin} - \widehat{\text{bias}} = \hat{\beta}_r + \frac{v' (Y - \mathbf{Z}\hat{\beta})}{v' Z_r}.$$

- ▶ For an **optimal** direction v , $\check{\beta}_r \approx \beta_r + \text{variance}$.

⁴Zhang and Zhang, JRSSB (2014)

Solution for the Ising model

- ▶ Debiasing via projection and local Taylor expansion of

$$X_r = \dot{f}(X_{-r}\theta_r) + \varepsilon_r,$$

where $f(u) = \log(e^u + e^{-u})$ and ε_r is sub-Gaussian.

Solution for the Ising model

- ▶ Debiasing via projection and local Taylor expansion of

$$X_r = \dot{f}(X_{-r} \theta_r) + \varepsilon_r,$$

where $f(u) = \log(e^u + e^{-u})$ and ε_r is sub-Gaussian.

- ▶ The debiased estimator $\check{\theta}_{rt}$ satisfies

$$\sqrt{n}(\check{\theta}_{rt} - \theta_{rt}^*) \rightarrow \mathcal{N}(0, s_{rt}).$$

Global testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Global testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Global testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Step 3 Form the test statistic

$$M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2.$$

Global testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Step 3 Form the test statistic

$$M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2.$$

Step 4 Reject H_0 if $M_{n,p}$ is large.

Theory: global testing

Test statistic

$$M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2.$$

Theorem (M, Xia, Cai and Li)

Under the null and some regularity conditions, for any $z \in \mathbb{R}$,

$M_{n,p} - 4 \log p + \log \log p \rightarrow$ Type I extreme value distribution,

as $n_1, n_2, p \rightarrow \infty$.

Theory: global testing

Test statistic

$$M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2.$$

Theorem (M, Xia, Cai and Li)

Under the null and some regularity conditions, for any $z \in \mathbb{R}$,

$M_{n,p} - 4 \log p + \log \log p \rightarrow$ Type I extreme value distribution,

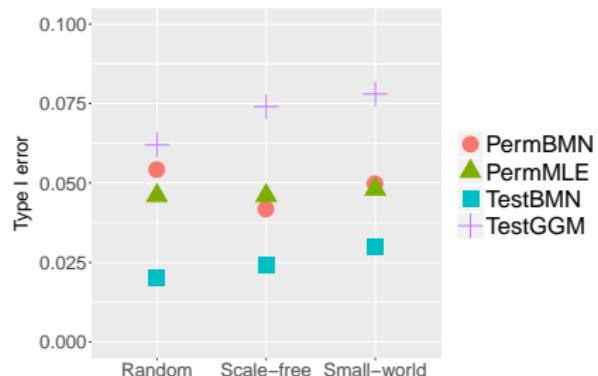
as $n_1, n_2, p \rightarrow \infty$.

Intuition:

- ▶ $W_{rt} \rightarrow \mathcal{N}(0, 1)$ under the null.
- ▶ W_{rt} 's are weakly dependent under mild assumptions.

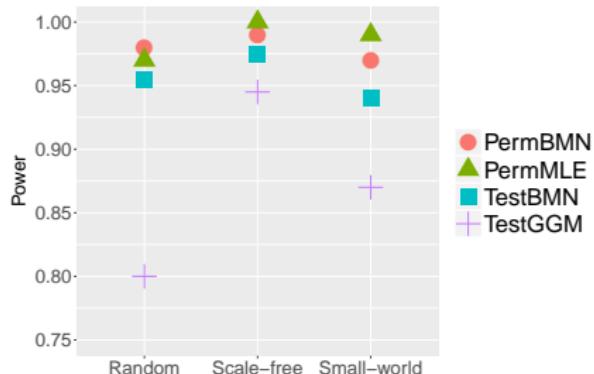
Simulation results: type I error

- ▶ $p = 100$, $\Theta_1 = \Theta_2 = \Theta_0$.
- ▶ Generate data $\{X^{(i)}\}_{i=1}^n \sim P_{\Theta_1}$ and $\{Y^{(i)}\}_{i=1}^n \sim P_{\Theta_2}$ by Gibbs sampling.
- ▶ Run global testing with $\alpha = 5\%$.

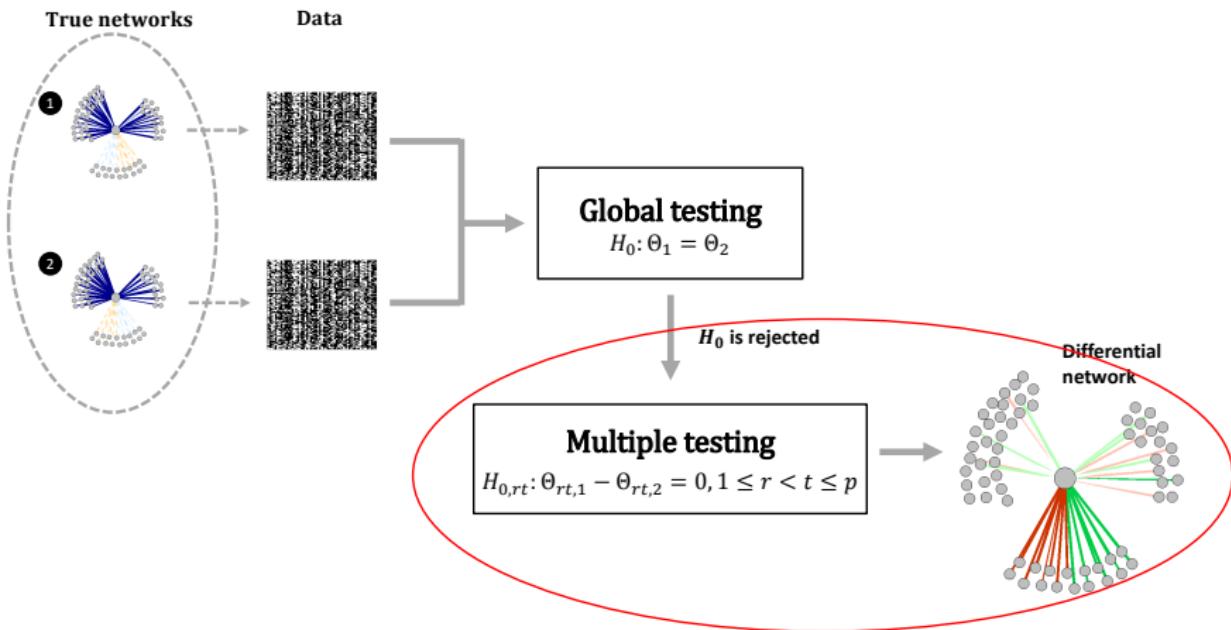


Simulation results: power

- ▶ $p = 100, \Theta_1 = \Theta_0 - \Delta, \Theta_2 = \Theta_0 + \Delta$ where $\|\Delta\|_0 = 10$.
- ▶ Generate data $\{X^{(i)}\}_{i=1}^n \sim P_{\Theta_1}$ and $\{Y^{(i)}\}_{i=1}^n \sim P_{\Theta_2}$ by Gibbs sampling.
- ▶ Run global testing with $\alpha = 5\%$.



Inference of differential Markov network



Multiple testing

$$H_{0,rt} : \theta_{rt,1} - \theta_{rt,2} = 0, \quad 1 \leq r < t \leq p.$$

Multiple testing

$$H_{0,rt} : \theta_{rt,1} - \theta_{rt,2} = 0, \quad 1 \leq r < t \leq p.$$

- ▶ Test statistic for individual hypothesis:

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}} \rightarrow_{\text{null}} \mathcal{N}(0, 1).$$

- ▶ Reject $H_{0,rt}$ if $|W_{rt}| \geq \tau$.

Multiple testing

$$H_{0,rt} : \theta_{rt,1} - \theta_{rt,2} = 0, \quad 1 \leq r < t \leq p.$$

- ▶ Test statistic for individual hypothesis:

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}} \xrightarrow{\text{null}} \mathcal{N}(0, 1).$$

- ▶ Reject $H_{0,rt}$ if $|W_{rt}| \geq \tau$.

Q: how to choose τ to ensure false discovery rate control?

What is the FDR?

Challenge: for any given τ

- ▶ Number of rejections: $R(\tau) = \sum_{1 \leq r < t \leq p} I(|W_{rt}| \geq \tau)$.

What is the FDR?

Challenge: for any given τ

- ▶ Number of rejections: $R(\tau) = \sum_{1 \leq r < t \leq p} I(|W_{rt}| \geq \tau)$.
- ▶ Number of **false** rejections: $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{rt}| \geq \tau)$.

What is the FDR?

Challenge: for any given τ

- ▶ Number of rejections: $R(\tau) = \sum_{1 \leq r < t \leq p} I(|W_{rt}| \geq \tau)$.
- ▶ Number of **false** rejections: $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{rt}| \geq \tau)$.
- ▶ Need to control

$$\text{FDR}(\tau) := E[\text{FDP}(\tau)], \quad \text{FDP}(\tau) := \frac{R_0(\tau)}{R(\tau) \vee 1}.$$

What is the FDR?

Solution:

- ▶ W_{rt} 's are only weakly dependent under some mild assumptions.

What is the FDR?

Solution:

- ▶ W_{rt} 's are only weakly dependent under some mild assumptions.
- ▶ $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{rt}| \geq \tau) \approx$ sum of i.i.d. random variables

$$\frac{R_0(\tau)}{|\mathcal{H}_0|} \approx 2\{1 - \Phi(\tau)\}, \quad \text{where } \Phi(\cdot) \text{ is c.d.f. of } \mathcal{N}(0, 1).$$

What is the FDR?

Solution:

- ▶ W_{rt} 's are only weakly dependent under some mild assumptions.
- ▶ $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{rt}| \geq \tau) \approx$ sum of i.i.d. random variables

$$\frac{R_0(\tau)}{|\mathcal{H}_0|} \approx 2\{1 - \Phi(\tau)\}, \quad \text{where } \Phi(\cdot) \text{ is c.d.f. of } \mathcal{N}(0, 1).$$

- ▶ Assuming sparsity, number of true nulls $|\mathcal{H}_0| \approx (p^2 - p)/2$.

What is the FDR?

Solution:

- ▶ W_{rt} 's are only weakly dependent under some mild assumptions.
- ▶ $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{rt}| \geq \tau) \approx$ sum of i.i.d. random variables

$$\frac{R_0(\tau)}{|\mathcal{H}_0|} \approx 2\{1 - \Phi(\tau)\}, \quad \text{where } \Phi(\cdot) \text{ is c.d.f. of } \mathcal{N}(0, 1).$$

- ▶ Assuming sparsity, number of true nulls $|\mathcal{H}_0| \approx (p^2 - p)/2$.
- ▶ We thus have

$$\widehat{\text{FDP}}(\tau) = \frac{2\{1 - \Phi(\tau)\}(p^2 - p)/2}{R(\tau) \vee 1}.$$

Multiple testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Multiple testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Step 3 Find $\hat{\tau}$

$$\hat{\tau} = \inf\{0 \leq \tau \leq \sqrt{4 \log p - 2 \log \log p} : \widehat{\text{FDP}}(\tau) \leq \alpha\}.$$

If the above $\hat{\tau}$ does not exist, $\hat{\tau} = \sqrt{4 \log p}$.

Multiple testing procedure

Step 1 Given presence/absence data X and Y , obtain debiased $\check{\theta}_{rt,k}$ (and $\check{s}_{rt,k}$) for $1 \leq r < t \leq p$ and $k = 1, 2$.

Step 2 For $1 \leq r < t \leq p$, calculate W_{rt}

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}.$$

Step 3 Find $\hat{\tau}$

$$\hat{\tau} = \inf\{0 \leq \tau \leq \sqrt{4 \log p - 2 \log \log p} : \widehat{\text{FDP}}(\tau) \leq \alpha\}.$$

If the above $\hat{\tau}$ does not exist, $\hat{\tau} = \sqrt{4 \log p}$.

Step 4 Reject $H_{0,rt}$ if $W_{rt} \geq \hat{\tau}$ for $1 \leq r < t \leq p$.

Theory: multiple testing

Theorem (M, Xia, Cai and Li)

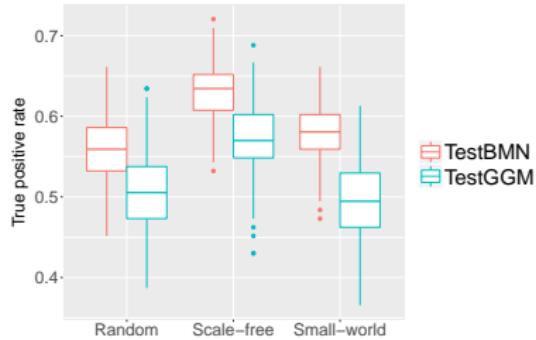
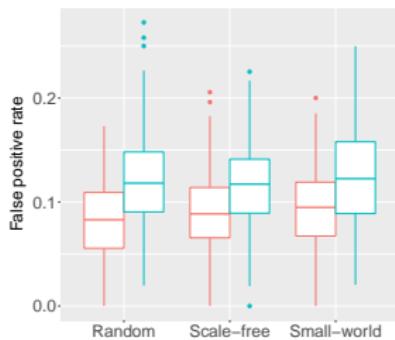
Let $q_0 = |\mathcal{H}_0|$ and $q = (p^2 - p)/2$. Under some regularity conditions, our multiple testing procedure asymptotically controls the false discovery rate, i.e.

$$\frac{\text{FDR}(\hat{\tau})}{\alpha q_0/q} \rightarrow 1, \quad \frac{\text{FDP}(\hat{\tau})}{\alpha q_0/q} \rightarrow 1,$$

as $n_1, n_2, p \rightarrow \infty$.

Simulation results

- ▶ $p = 100, \Theta_1 = \Theta_0 - \Delta, \Theta_2 = \Theta_0 + \Delta$ where $\|\Delta\|_0 = 0.04 \cdot \binom{p}{2}$.
- ▶ Generate data $\{X^{(i)}\}_{i=1}^n \sim P_{\Theta_1}$ and $\{Y^{(i)}\}_{i=1}^n \sim P_{\Theta_2}$ by Gibbs sampling.
- ▶ Run multiple testing with $\alpha = 10\%$.



Gut microbiome in UK twins



Fig: Goodrich et al. Cell Host & Microbe.
(2016)

Data

- ▶ 16S rRNA sequencing of the gut microbiome.
- ▶ Very rare bacterial genera⁵ were removed, leaving $p = 59$.
- ▶ Only one member from each family was used.
- ▶ Young: $18 \leq \text{age} \leq 43$, $n_1 = 171$.
- ▶ Elderly: $74 \leq \text{age} \leq 89$, $n_2 = 180$.

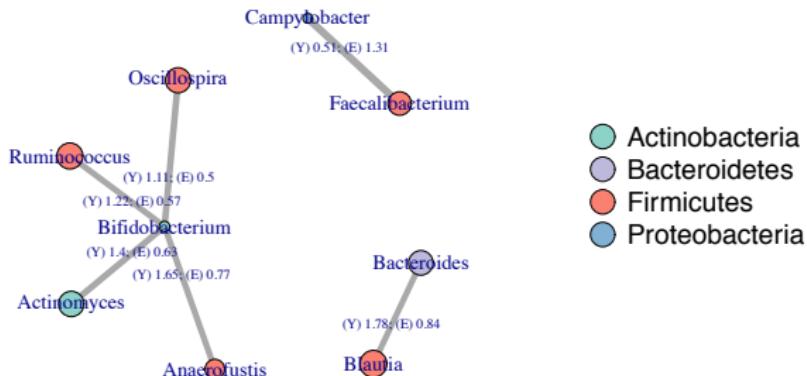
⁵ Taxonomic rank: Species → Genus → Family → Order → Class → Phylum → Kingdom

Results: differential network

- ▶ Global testing p -value = 0.009.

Results: differential network

- ▶ Global testing p -value = 0.009.
- ▶ Differential network obtained via multiple testing with FDR = 15% (Edge: differential interactions; Edge label: *odds ratio*).



Implications of differential network

Campylobacter – Faecalibacterium:

- ▶ Young OR = 0.51: presence in **Faecalibacterium** is associated with lower odds of presence in **Campylobacter**, a **competitive** relationship.
- ▶ Elderly OR = 1.31: presence in **Faecalibacterium** is associated with higher odds of presence in **Campylobacter**, a **collaborative** relationship.

Implications of differential network

Aging is characterized by chronic low-grade inflammation (inflammaging).

- ▶ Abundance of **Faecalibacterium** negatively associated with age (Franceschi et al. Trends Endocrinol Metab. 2017).
- ▶ **Ruminococcus** enriched in immune-mediated inflammatory diseases (Forbes et al. Front Microbiol. 2016).
- ▶ Abundance of **Oscillospira** enriched in inflammatory diseases (Konikoff and Gophna. Trends Microbiol. 2016).

Summary

- ▶ Learn conditional dependency relationships among microbes using Markov networks.



Summary



- ▶ Learn conditional dependency relationships among microbes using Markov networks.
- ▶ Differential network analysis identifies systematic changes in microbial interactions associated with age.

Part I: Testing Differences in Microbial Networks

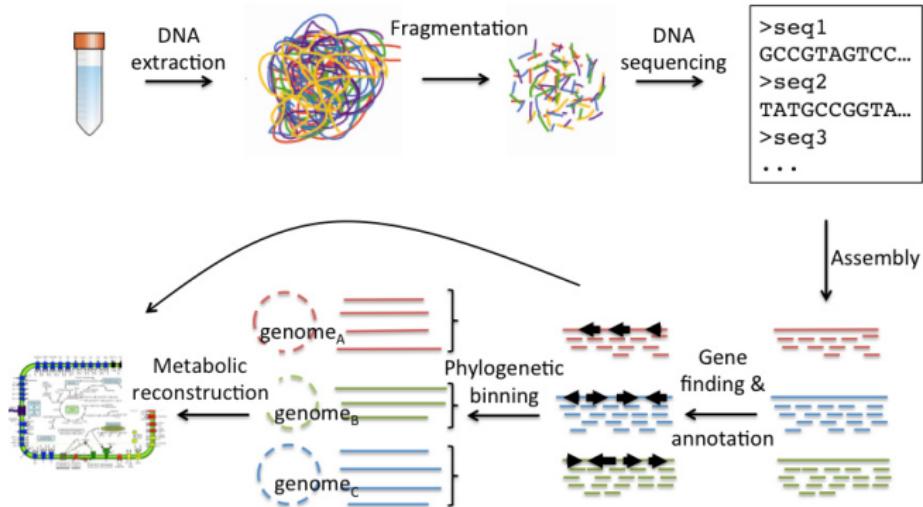
Part II: Metagenomic Microbial Community Profiling

Penn PLEASE study (pediatric Crohn's disease)

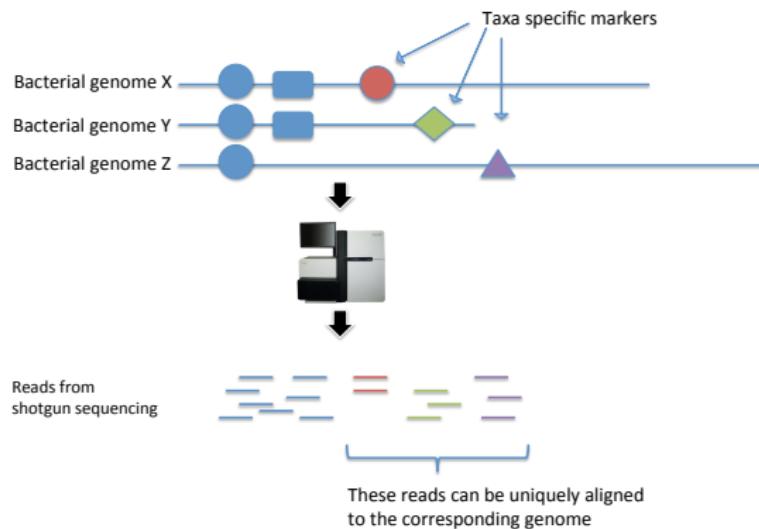
Data

- ▶ 90 subjects × 4 times points disease samples
- ▶ 26 healthy controls
- ▶ Shotgun metagenomics: ave 11×10^6 paired-end reads per sample

Shotgun metagenomics



Shotgun metagenomics - align reads to taxa-specific markers⁶



⁶Segata et al. Nature Methods (2012)

Abundance quantification based on taxa-specific markers

- ▶ Observe read counts per marker (x_{ijk}), total counts per sample (N_i), marker length (ℓ_{jk})
- ▶ Estimate the relative abundance of each species (θ_{ij})

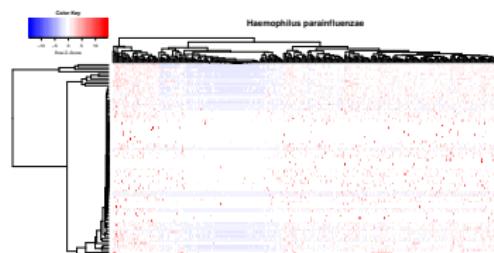
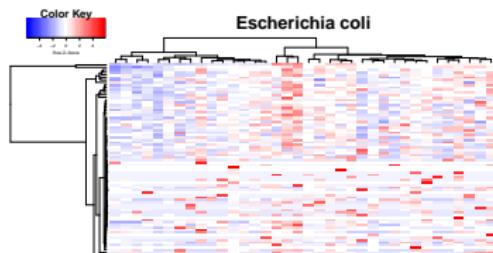
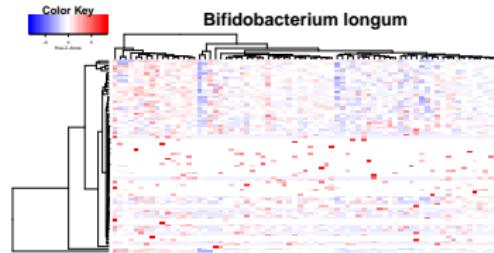
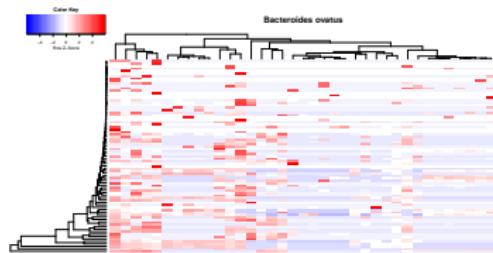
Sample	Species 1			Species j			Species p			Total reads		
	1	\dots	K_1	\dots	1	\dots	K_j	\dots	1	\dots	K_p	
	ℓ_{11}	\dots	ℓ_{1,K_1}		ℓ_{j1}	\dots	ℓ_{j,K_2}		ℓ_{p1}	\dots	ℓ_{p,K_p}	
1												N_1
\vdots						\vdots						\vdots
i				\dots	x_{ijk}							N_i
\vdots						\vdots		$i = 1, \dots, n$				\vdots
.						.		$j = 1, \dots, p$.
n							$k = 1, \dots, K_j$					N_n

Challenges

- ▶ Reads are from thousands of different microbial genomes with different abundances, but there are many unknown species
- ▶ Data are very sparse with excess of zeros
- ▶ Outliers

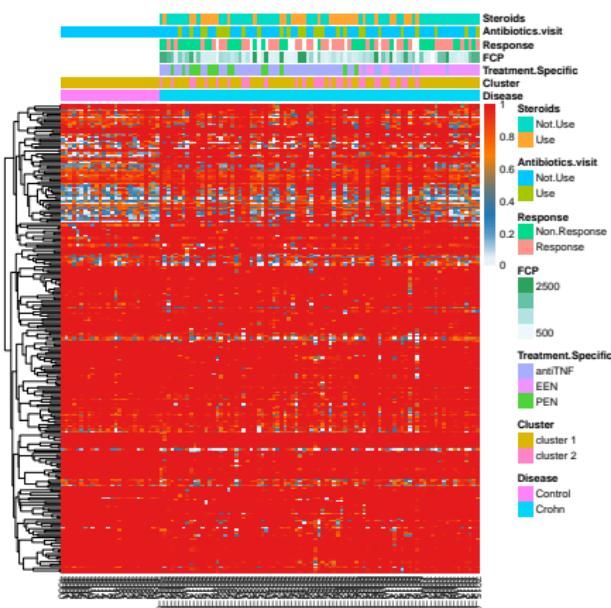
Penn PLEASE samples and marker-specific effects

- rows: samples; columns: markers



Proportion of observed zeros

- rows: species; columns: samples



The Zero-inflated Poisson (ZIP) model

For a given species j with K_j markers,

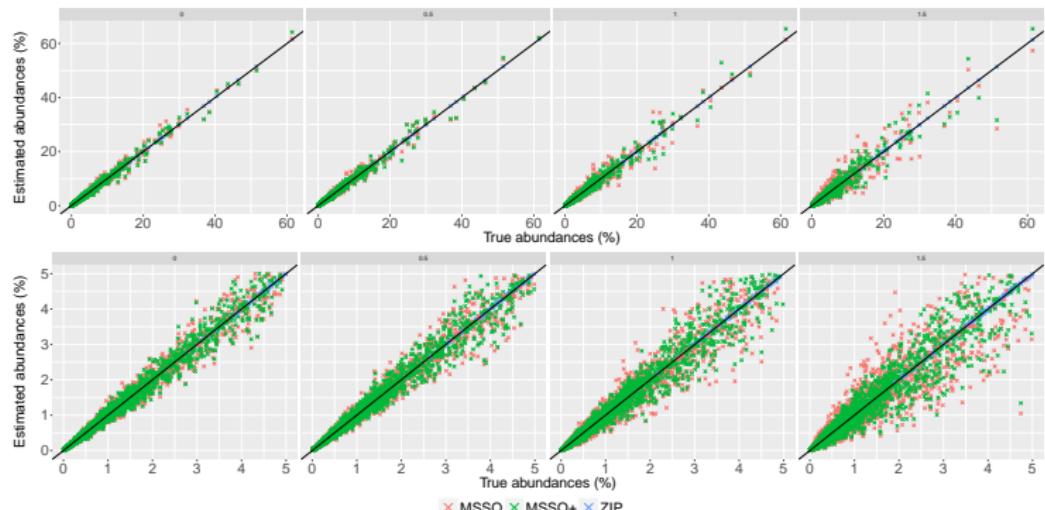
$$x_{ijk} = \begin{cases} 0 & \pi_{ij}, \\ \text{Poisson}(\alpha_i N_i \ell_{jk} \phi_{jk} \theta_{ij}) & 1 - \pi_{ij}, \end{cases} \quad i = 1, \dots, n; k = 1, \dots, K_j$$

- ▶ $\pi_{ij} > 0$: probability of the zero state
- ▶ $\alpha_i > 0$: normalizing constant for sample i
- ▶ N_i : total number of reads for sample i
- ▶ ℓ_{jk} : marker length
- ▶ $\theta_{ij} \geq 0$: relative abundance; $\sum_{j=1}^p \theta_{ij} = 1$
- ▶ $\phi_{jk} \geq 0$: marker-specific effect; $\sum_{k=1}^{K_j} \phi_{jk} = K_j$

If $\pi_{ij} = 0$ and $\alpha_i = 1$, it reduces to the Poisson model (Chen et al. 2016).

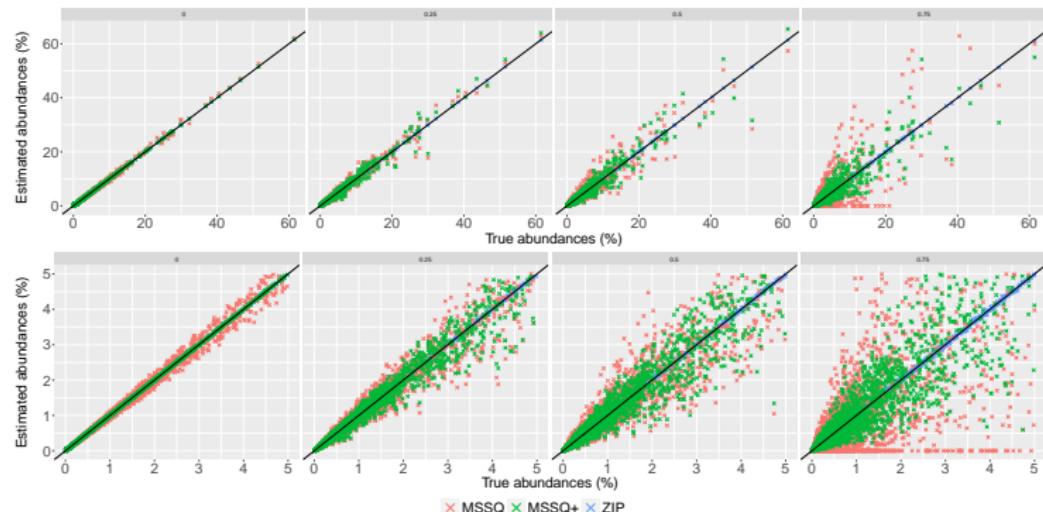
Simulation: abundance estimation with varying marker effects

- ▶ 50 samples, 50 species, $\pi_{ij} \sim \text{Unif}[0.45, 0.55]$

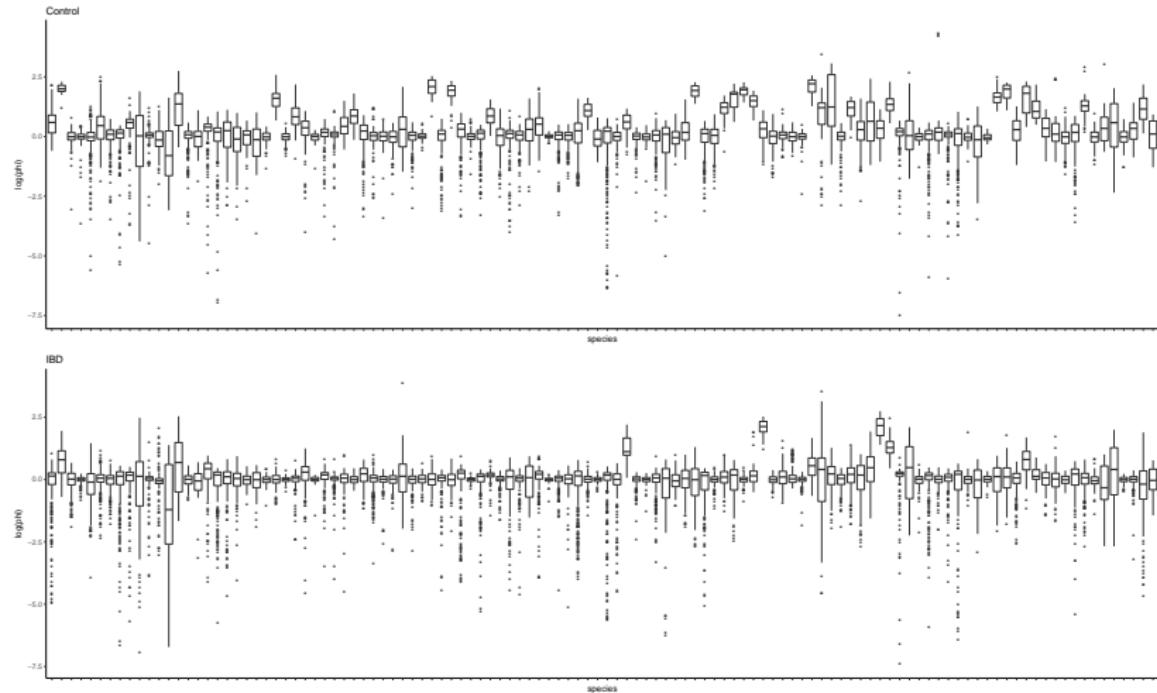


Simulation: abundance estimation with varying zero inflation

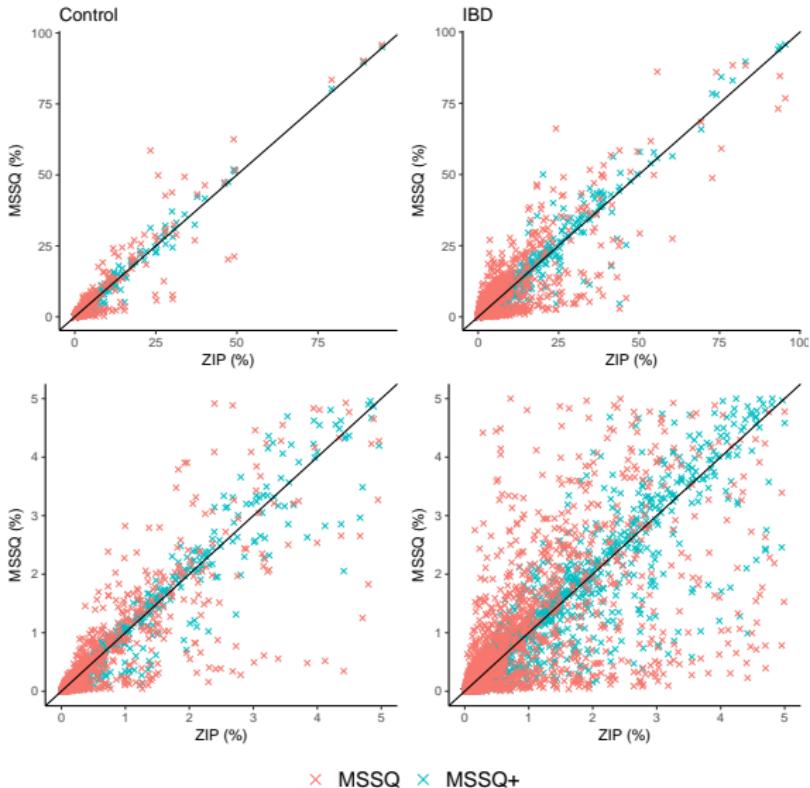
- ▶ 50 samples, 50 species, $\phi_{jk} \sim \log \mathcal{N}(0, 1.5^2)$



PLEASE data - marker specific effects



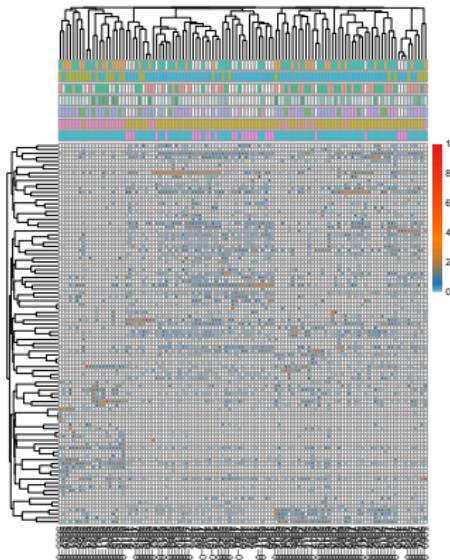
PLEASE data - abundance estimation



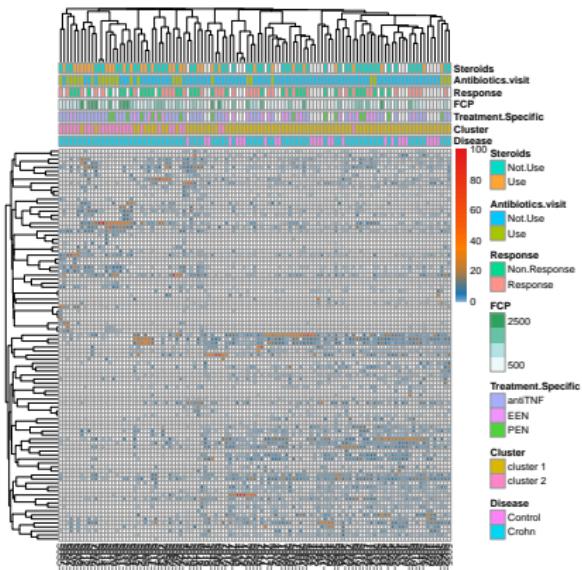
PLEASE data - clustering

114 species visualized after removing low abundant ones (< 1%)

MSSQ



ZIP



Open questions

- ▶ Differential abundance analysis: variability of species abundance

Acknowledgement



Hongzhe Li



Tony Cai



Yin Xia

Manuscript is available upon request.

Code is available at <https://github.com/drjingma/TestBMN>.

Supplementary Slides

Debiasing for linear regression

- $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.

Debiasing for linear regression

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

Debiasing for linear regression

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

- ▶ The debiased estimator (given $\hat{\beta}$) is

$$\check{\beta}_r = \beta_r^{lin} - \widehat{\text{bias}} = \hat{\beta}_r + \frac{v'(Y - \mathbf{Z}\hat{\beta})}{v' Z_r}.$$

Debiasing for linear regression

- ▶ $Y = \mathbf{Z}\beta + \varepsilon, \mathbf{Z} \in \mathbb{R}^{n \times p}$.
- ▶ Projecting Y onto $v \in \mathbb{R}^n$ yields

$$\beta_r^{lin} = \frac{v' Y}{v' Z_r} = \beta_r + \underbrace{\frac{v' \varepsilon}{v' Z_r}}_{\text{variance}} + \underbrace{\sum_{j:j \neq r} \frac{v' Z_j \beta_j}{v' Z_r}}_{\text{bias}}.$$

- ▶ The debiased estimator (given $\hat{\beta}$) is

$$\check{\beta}_r = \beta_r^{lin} - \widehat{\text{bias}} = \hat{\beta}_r + \frac{v'(Y - \mathbf{Z}\hat{\beta})}{v' Z_r}.$$

- ▶ For an **optimal** direction v , $\check{\beta}_r \approx \beta_r + \text{variance}$.

Debiasing for logistic regression

Back to our case:

$$X_r = \dot{f}(X_{-r}, \theta_r) + \varepsilon_r,$$

where $f(u) = \log(e^u + e^{-u})$ and ε_r is sub-Gaussian.

Debiasing for logistic regression

Back to our case:

$$X_r = \dot{f}(X_{-r} \theta_r) + \varepsilon_r,$$

where $f(u) = \log(e^u + e^{-u})$ and ε_r is sub-Gaussian.

- Linear approximation: local Taylor expansion ($\hat{u}_r = X_{-r} \hat{\theta}_r$) yields

$$\underbrace{X_r - \dot{f}(\hat{u}_r) + \ddot{f}(\hat{u}_r) X_{-r} \hat{\theta}_r}_Y = \underbrace{\ddot{f}(\hat{u}_r) X_{-r}}_Z \theta_r + (Re + \varepsilon_r).$$

Debiasing for logistic regression

Back to our case:

$$X_r = \dot{f}(X_{-r} \theta_r) + \varepsilon_r,$$

where $f(u) = \log(e^u + e^{-u})$ and ε_r is sub-Gaussian.

- Linear approximation: local Taylor expansion ($\hat{u}_r = X_{-r} \hat{\theta}_r$) yields

$$\underbrace{X_r - \dot{f}(\hat{u}_r)}_Y + \underbrace{\ddot{f}(\hat{u}_r) X_{-r} \hat{\theta}_r}_Z = \underbrace{\ddot{f}(\hat{u}_r) X_{-r}}_Z \theta_r + (Re + \varepsilon_r).$$

- Given an initial estimator $\hat{\theta}_r$ and score vector $v_{rt}^{(i)}$, the debiased estimator is

$$\check{\theta}_{rt} = \hat{\theta}_{rt} + \frac{\sum_{i=1}^{n_1} v_{rt}^{(i)} \{ X_r^{(i)} - \dot{f}(\hat{u}_r^{(i)}) \}}{\sum_{i=1}^{n_1} v_{rt}^{(i)} \ddot{f}(\hat{u}_r^{(i)}) X_t^{(i)}} \approx \hat{\theta}_{rt} + \underbrace{\frac{\sum_{i=1}^{n_1} v_{rt}^{(i)} \varepsilon_r^{(i)}}{\sum_{i=1}^{n_1} v_{rt}^{(i)} \ddot{f}(\hat{u}_r^{(i)}) X_t^{(i)}}}_{\text{variance}}.$$

Score vectors for debiasing

How to choose $v_{rt}^{(i)}$?

$$\check{\theta}_{rt} = \theta_{rt} + \underbrace{\frac{\sum_{i=1}^{n_1} v_{rt}^{(i)} \varepsilon_r^{(i)}}{\sum_{i=1}^{n_1} v_{rt}^{(i)} \check{f}(\hat{u}_r^{(i)}) X_t^{(i)}}}_{\text{noise}} + \text{bias} + REM,$$

where REM is small given good $\hat{\theta}_r$. Principles for picking V are

- ▶ $E[V \varepsilon_r] = 0$,
- ▶ $\min \langle V, V \rangle$ subject to $\langle V, X_t \rangle = 1$,
- ▶ $\langle V, h(X_{-\{r,t\}}) \rangle = 0$ for any measurable function h .

Thus we can pick $v_{rt}^{(i)}$ as the residual

$$v_{rt}^{(i)} = (X_t^{(i)} + 1)/2 - g(X_{-\{r,t\}}^{(i)}, \hat{\theta}_r, \hat{\theta}_t), \quad i = 1, \dots, n.$$

Variance of $\check{\theta}_{rt}$

- The de-biased estimator is

$$\check{\theta}_{rt} \approx \theta_{rt} + \frac{n^{-1} \sum_{i=1}^n v_{rt}^{(i)} \varepsilon_r^{(i)}}{n_1^{-1} \sum_{i=1}^{n_1} v_{rt}^{(i)} \ddot{f}(\hat{u}_r^{(i)}) X_t^{(i)}}.$$

- Let v_{rt}^o be the oracle score vector and $F_{rt} = 4E_{\Theta_1}[(v_{rt}^o)^2 \ddot{f}(u_r)]$.
- Define

$$\tilde{\theta}_{rt} := \theta_{rt} + \frac{n^{-1} \sum_{i=1}^n v_{rt}^{o,(i)} \varepsilon_r^{(i)}}{F_{rt}/2} \approx \check{\theta}_{rt}.$$

- The variance is

$$\text{Var}(\check{\theta}_{rt}) \approx \text{Var}(\tilde{\theta}_{rt}) = \frac{1}{F_{rt}} \approx \left\{ 4n^{-1} \sum_{i=1}^n (v_{rt}^{(i)})^2 \ddot{f}(X_{-r}^{(i)} \hat{\theta}_r) \right\}^{-1} := \check{s}_{rt}.$$

Assumptions

- ▶ $\log p = o(n_k^{1/3})$ and $n_1 \asymp n_2$
- ▶ $\max_{1 \leq r \leq p} \|\hat{\theta}_{r,k} - \theta_{r,k}\|_1 = o_p(\{\log p\}^{-1})$.
- ▶ $\max_{1 \leq r \leq p} \|\hat{\theta}_{r,k} - \theta_{r,k}\|_2 = o_p(\{n_k \log p\}^{-1/4})$.
- ▶ $\max_{1 \leq t \leq p} |\mathcal{A}_t(\xi)| = o(p^\gamma)$ for $0 < \gamma < 1$, where for $\xi > 0$
$$\mathcal{A}_t(\xi) = \{r : |\sinh(2\theta_{rt,1})| \geq (\log p)^{-2-\xi} \text{ or } |\sinh(2\theta_{rt,2})| \geq (\log p)^{-2-\xi}\}.$$

Theory: global testing

Challenge:

- ▶ $M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2.$

Theory: global testing

Challenge:

- ▶ $M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2$.
- ▶ Entry-wise statistics W_{rt} and $W_{r't'}$ are dependent!

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}$$

Theory: global testing

Challenge:

- ▶ $M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2$.
- ▶ Entry-wise statistics W_{rt} and $W_{r't'}$ are dependent!

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}$$

Solution:

- ▶ Under some mild assumptions:
 - ▶ Θ_1 and Θ_2 are sparse: robustness of microbial communities.
 - ▶ The number of *large* coefficients ($\theta_{rt,k}$) is small: a few high activity microbial interactions.

Theory: global testing

Challenge:

- ▶ $M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2$.
- ▶ Entry-wise statistics W_{rt} and $W_{r't'}$ are dependent!

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}$$

Solution:

- ▶ Under some mild assumptions:
 - ▶ Θ_1 and Θ_2 are sparse: robustness of microbial communities.
 - ▶ The number of large coefficients ($\theta_{rt,k}$) is small: a few high activity microbial interactions.
- ▶ W_{rt} and $W_{r't'}$ are only weakly dependent!

Theory: global testing

Challenge:

- ▶ $M_{n,p} = \max_{1 \leq r < t \leq p} W_{rt}^2$.
- ▶ Entry-wise statistics W_{rt} and $W_{r't'}$ are dependent!

$$W_{rt} = \frac{\check{\theta}_{rt,1} - \check{\theta}_{rt,2}}{\sqrt{\check{s}_{rt,1}/n_1 + \check{s}_{rt,2}/n_2}}$$

Solution:

- ▶ Under some mild assumptions:
 - ▶ Θ_1 and Θ_2 are sparse: robustness of microbial communities.
 - ▶ The number of large coefficients ($\theta_{rt,k}$) is small: a few high activity microbial interactions.
- ▶ W_{rt} and $W_{r't'}$ are only weakly dependent!
- ▶ $M_{n,p} - 4 \log p + \log \log p \rightarrow \exp\{-(8\pi)^{-1/2} e^{-z/2}\}!$

Theory: multiple testing

- ▶ Given $\alpha > 0$, want

$$\tau^* = \inf \left\{ 0 \leq \tau \leq 2\sqrt{\log p} : \frac{R_0(\tau)}{R(\tau) \vee 1} \leq \alpha \right\}.$$

Theory: multiple testing

- ▶ Given $\alpha > 0$, want

$$\tau^* = \inf \left\{ 0 \leq \tau \leq 2\sqrt{\log p} : \frac{R_0(\tau)}{R(\tau) \vee 1} \leq \alpha \right\}.$$

- ▶ $\tau \leq 2\sqrt{\log p}$ because $W_{rt} \rightarrow \mathcal{N}(0, 1)$ under $H_{0,rt}$. Thus

$$\max_{(r,t) \in \mathcal{H}_0} |W_{rt}| \leq 2\sqrt{\log p}, \quad a.s.$$

Theory: multiple testing

- ▶ Given $\alpha > 0$, want

$$\tau^* = \inf \left\{ 0 \leq \tau \leq 2\sqrt{\log p} : \frac{R_0(\tau)}{R(\tau) \vee 1} \leq \alpha \right\}.$$

- ▶ $\tau \leq 2\sqrt{\log p}$ because $W_{rt} \rightarrow \mathcal{N}(0, 1)$ under $H_{0,rt}$. Thus

$$\max_{(r,t) \in \mathcal{H}_0} |W_{rt}| \leq 2\sqrt{\log p}, \quad a.s.$$

- ▶ Under weak dependence of W_{rt} 's, for
 $0 \leq \tau \leq b_p = \sqrt{4 \log p - 2 \log(\log p)}$,

$$\frac{R_0(\tau)}{|\mathcal{H}_0|} \approx 2\{1 - \Phi(\tau)\}, \quad |\mathcal{H}_0| \approx (p^2 - p)/2,$$

where $\Phi(\cdot)$ is the c.d.f. of $\mathcal{N}(0, 1)$.