Differential Network Biology: Testing Differences in Microbial Networks

Jing Ma

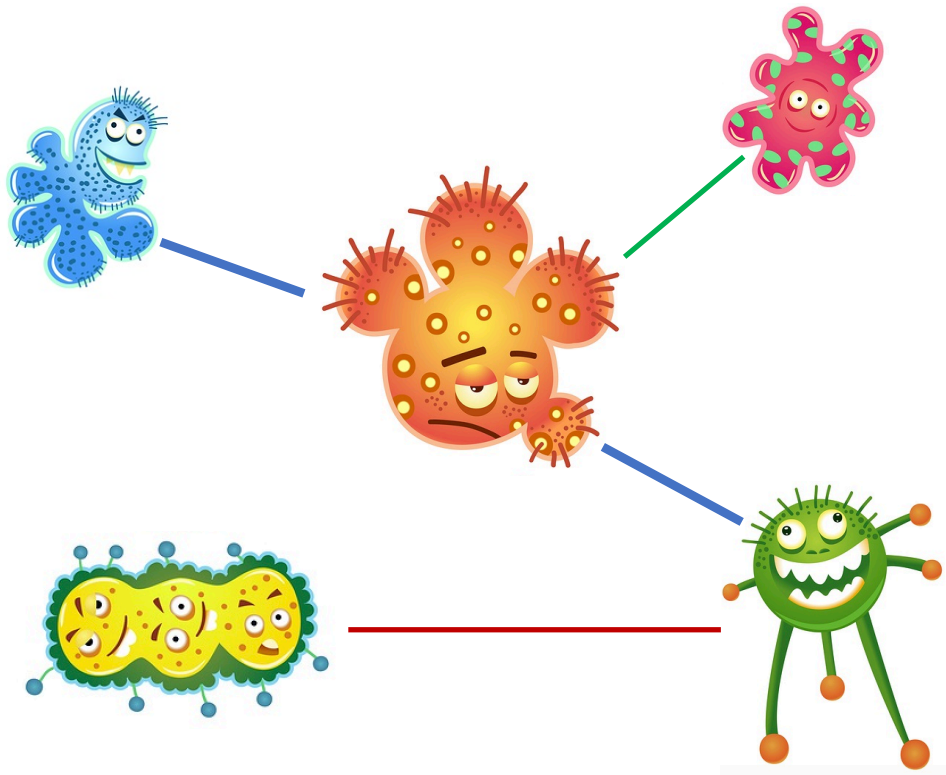
Public Health Sciences Division
Fred Hutch Cancer Research Center
jingma@fredhutch.org

January 4, 2018

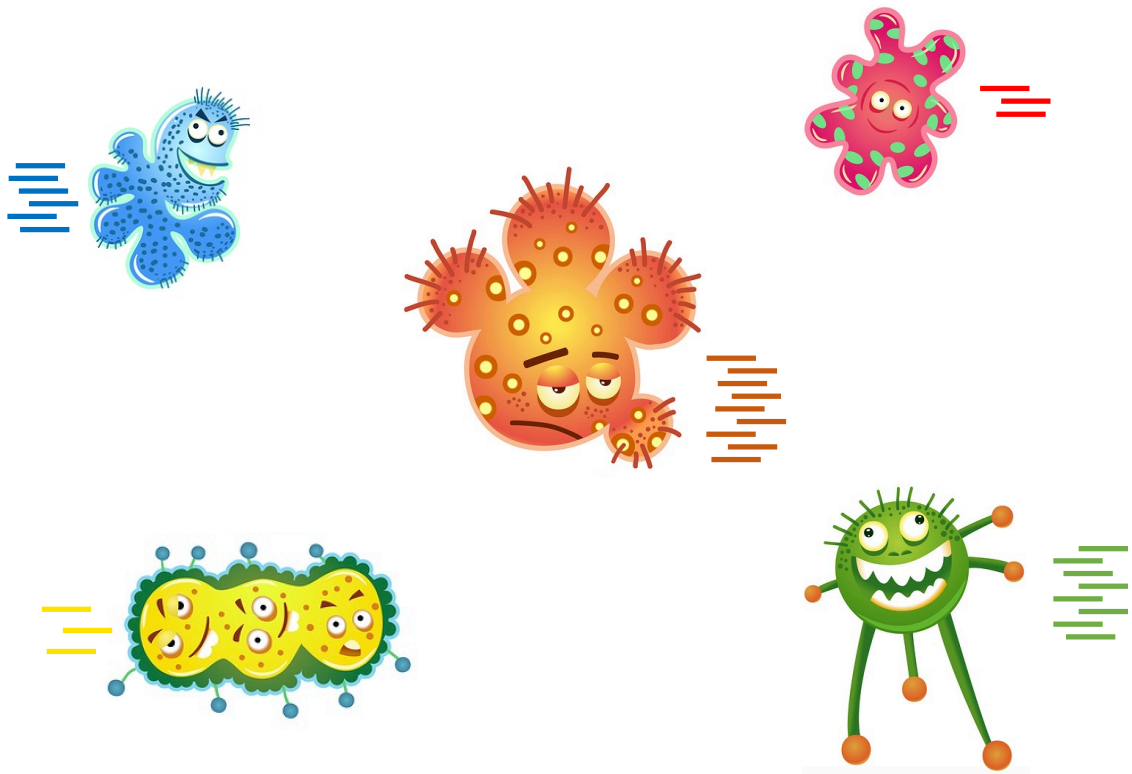
Human microbiome

39 trillion bacterial cells > # of human cells

Networks

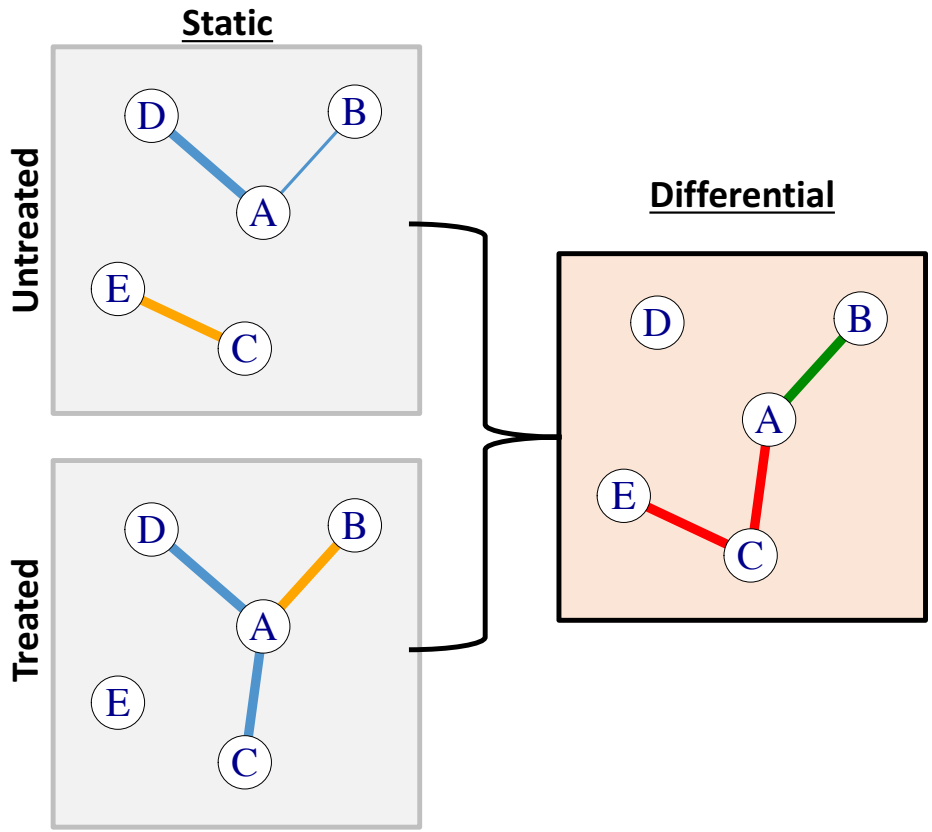


Networks

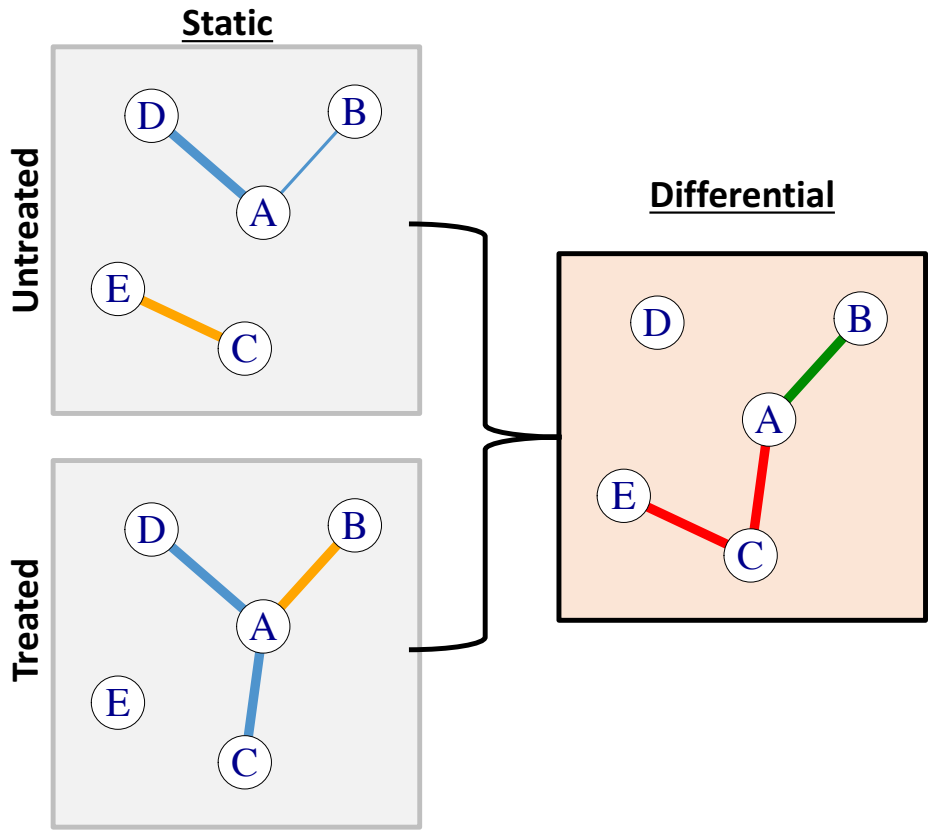


Question 1: What is a microbial network?

Differential network analysis



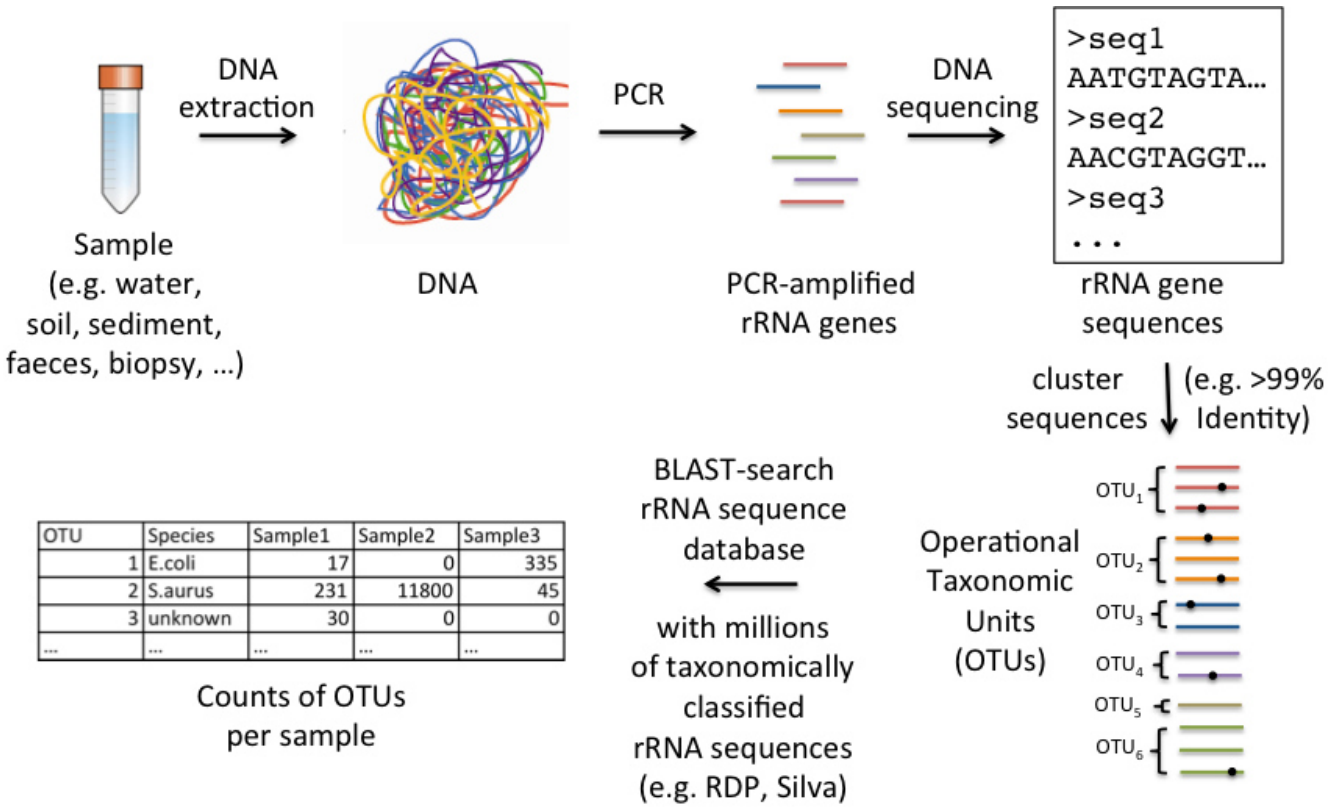
Differential network analysis



Question 2: How to test differences of microbial networks?

The microbiome data

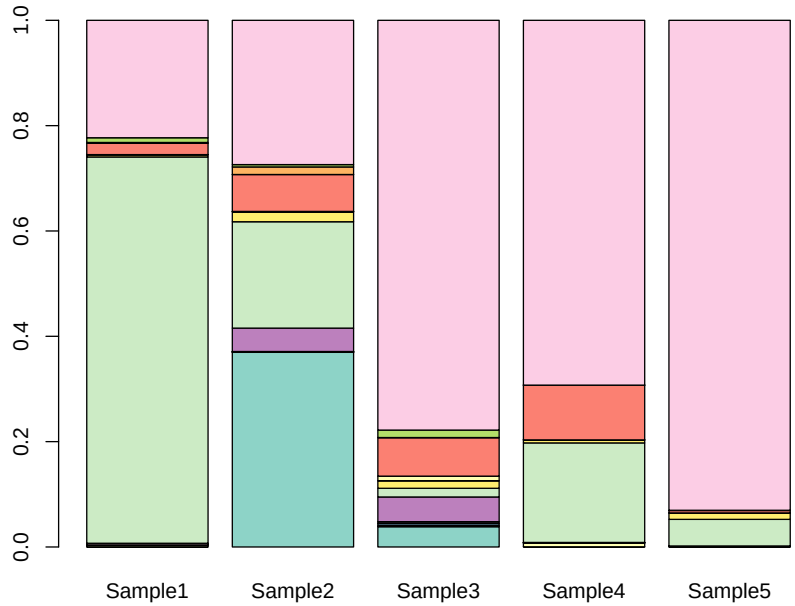
16S rRNA gene sequencing



OTU	Species	Sample1	Sample2	Sample3
1	E.coli	17	0	335
2	S.aurus	231	11800	45
3	unknown	30	0	0
...	...	...	...	...

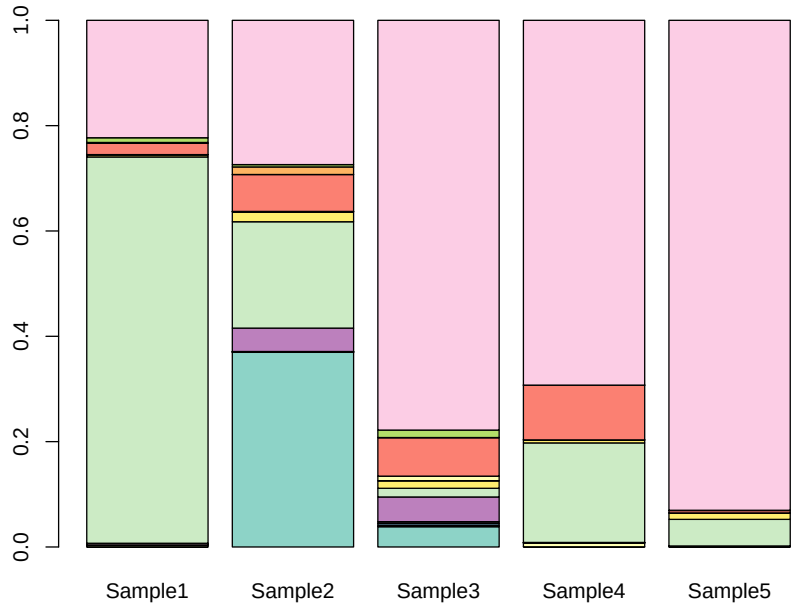
The microbiome data

- Microbiome data are compositional.



The microbiome data

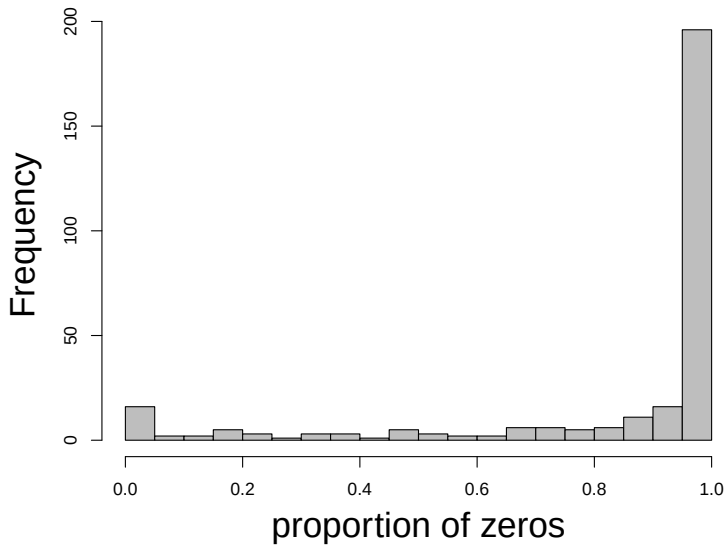
- Microbiome data are compositional.



- 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: SPIEC-EASI (Kurtz et al. 2015), MINT (Biswas et al. 2016).
- Limitations: marginal relationships, permutation-based significance test, zeros replaced with a pseudocount.

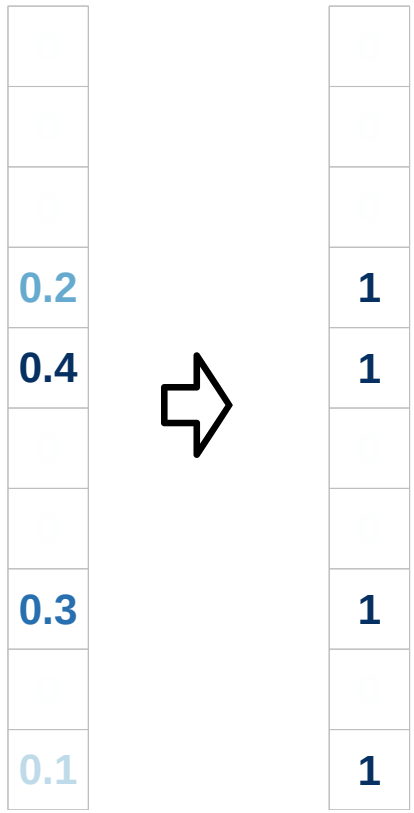
Microbial conditional dependency relationships

We want a model that

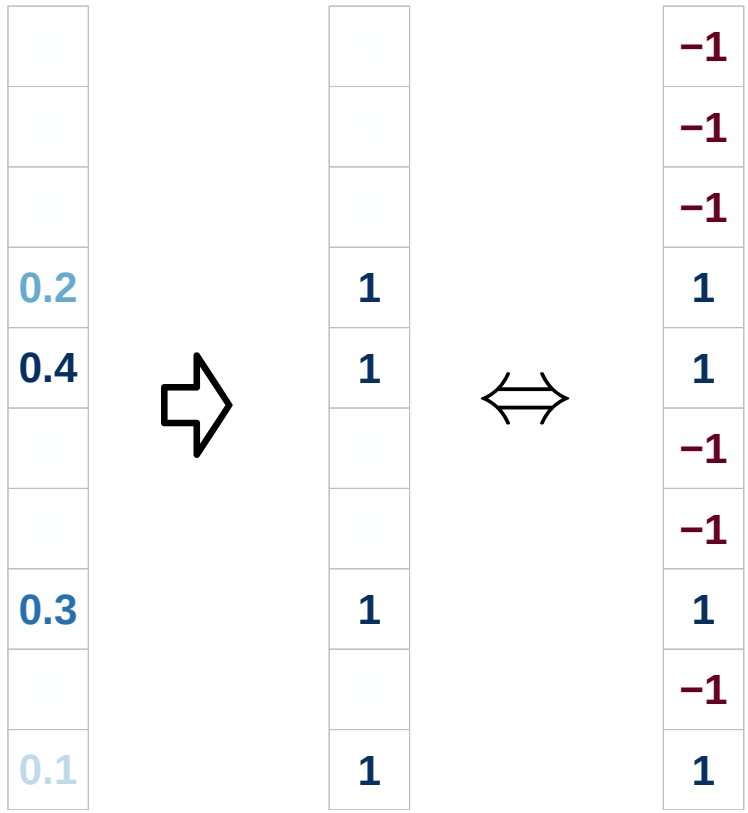
- captures the conditional dependency relationships among microbes,
- address the sparsity issue,
- infers differential network with false discovery rate control.

Our framework: binary Markov network (Ising model)

Our framework: binary Markov network (Ising model)



Our framework: binary Markov network (Ising model)

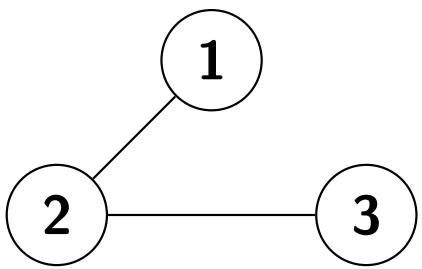


Our framework: binary Markov network (Ising model)

- Joint distribution $\mathrm{P}_\Theta(X) \propto \exp\left\{ \sum_{1 \le r < t \le p} X_r X_t \theta_{rt} \right\}$.

Our framework: binary Markov network (Ising model)

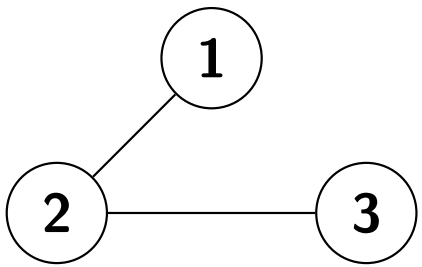
- Joint distribution $\mathrm{P}_\Theta(X) \propto \exp\left\{ \sum_{1 \leq r < t \leq p} X_r X_t \theta_{rt} \right\}$.
- Conditional independence



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

Our framework: binary Markov network (Ising model)

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



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

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

Estimation of the Ising model

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

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

[1] Ravikumar et al. Ann. Stat. (2010)

Estimation of the Ising model

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

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

- Nodewise (penalized) logistic regression[1] for large p:

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

[1] Ravikumar 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!

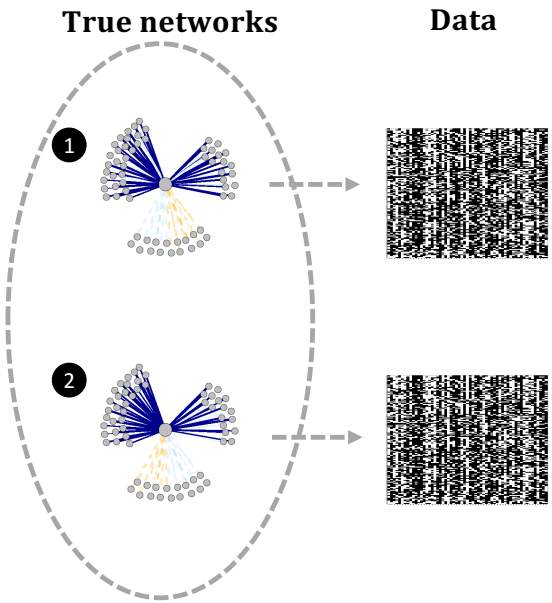
[2] Liu. AOS (2013)
[3] Xia, Cai and Cai. Biometrika (2015)

Inference beyond estimation

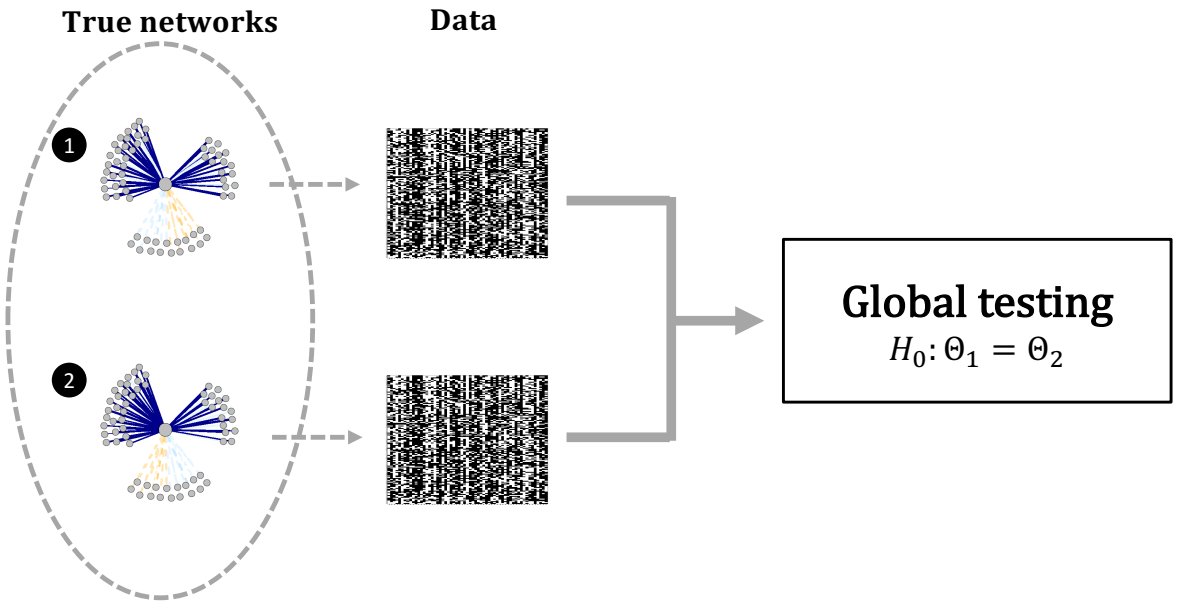
- Inference of a single network
  - done for Gaussian graphical model (GGM)[2], but not for Ising model!
- Two-sample (and multi-sample) inference
  - done for GGM[3], but not for Ising model!

[2] Liu. AOS (2013)
[3] Xia, Cai and Cai. Biometrika (2015)

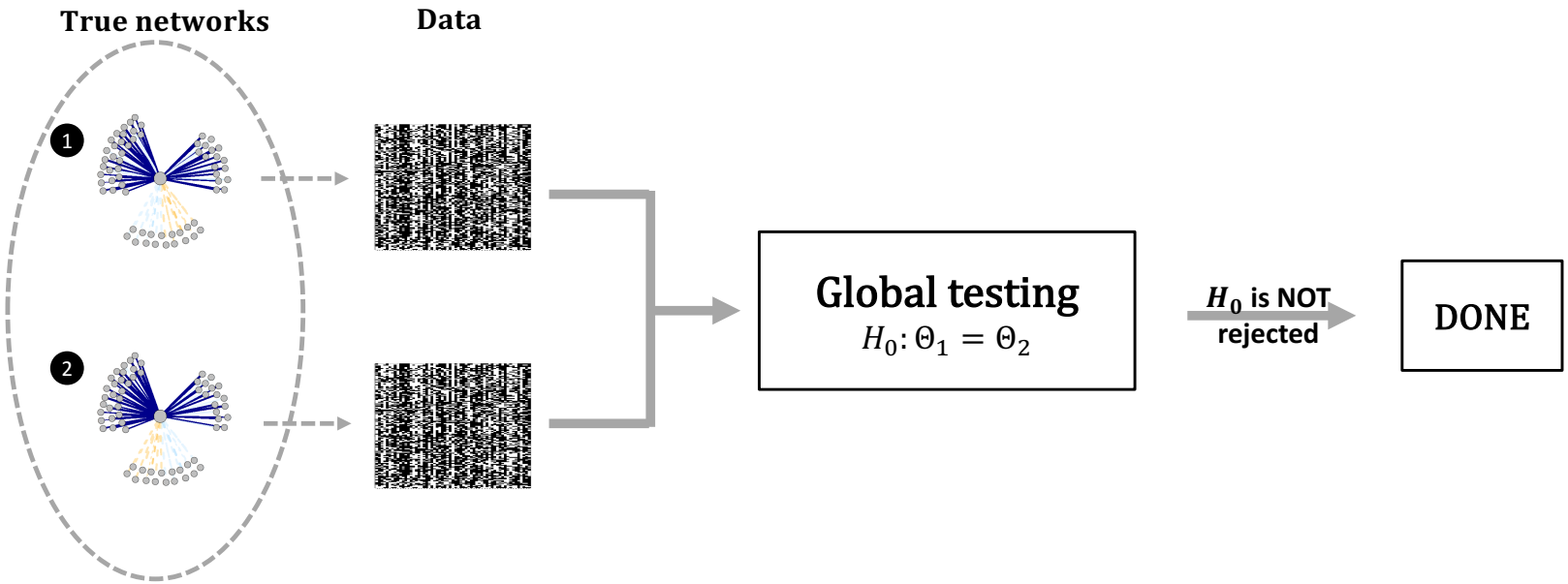
Inference of differential Markov network



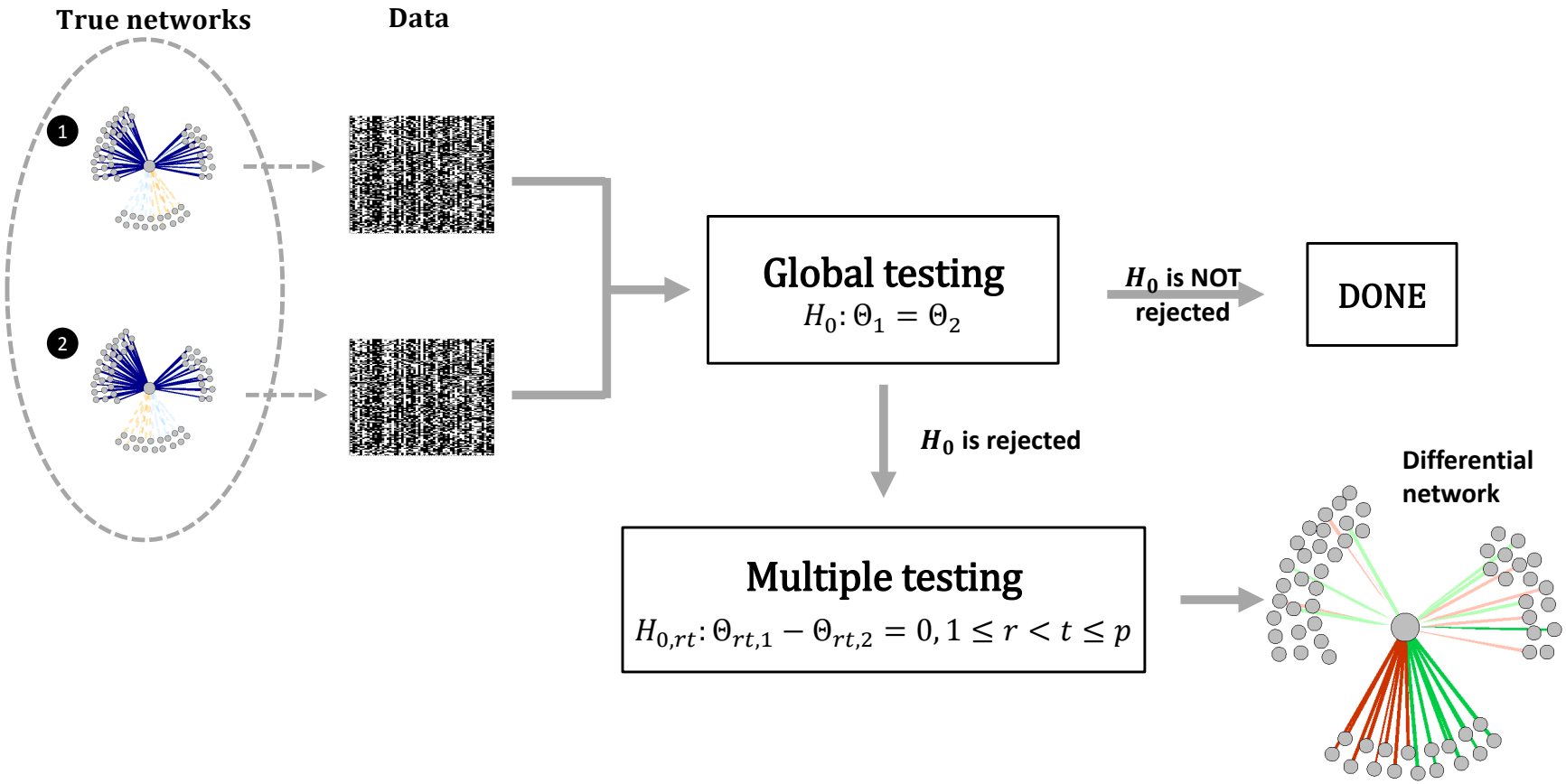
Inference of differential Markov network



Inference of differential Markov network



Inference of differential Markov network



Global testing $H_0 : \Theta_1 = \Theta_2$

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

Global testing $H_0 : \Theta_1 = \Theta_2$

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

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

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

Global testing $H_0 : \Theta_1 = \Theta_2$

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

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

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

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 projection[4]

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

[4] Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection[4]

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

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

[4] Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection[4]

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

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

[4] Zhang and Zhang, JRSSB (2014)

Debiasing the Lasso via projection[4]

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

[4] Zhang 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

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

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

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

Theorem (M, Xia, Cai and Li)

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

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

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

Theory: global testing

Test statistic

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

Theorem (M, Xia, Cai and Li)

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

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

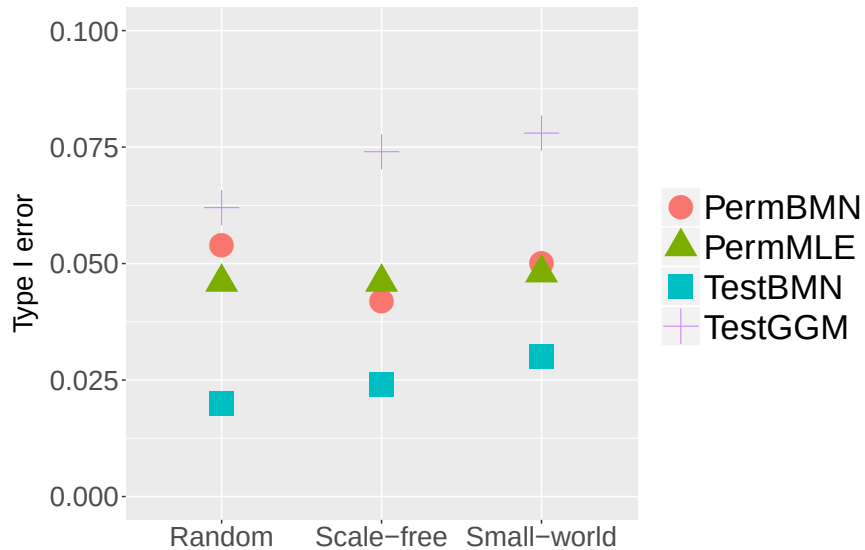
as $n_1, n_2, p \to \infty$.

Intuition:

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

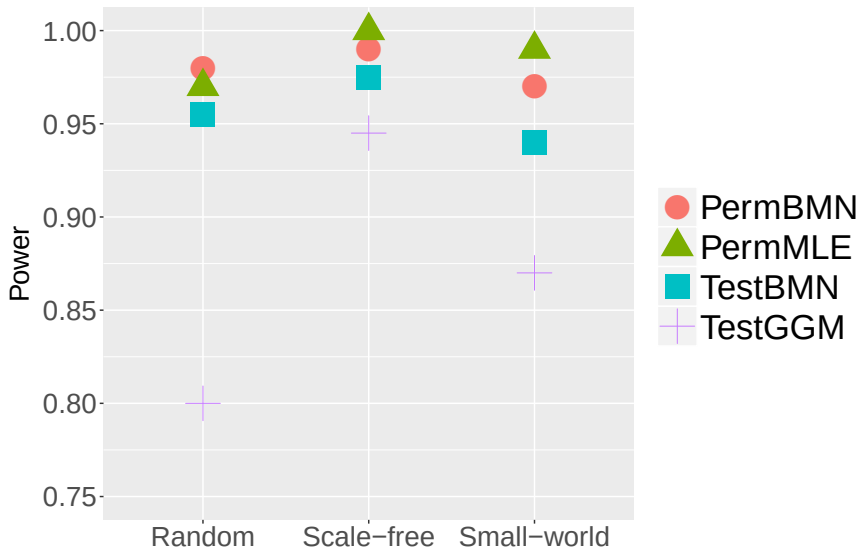
Simulation results: type I error

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

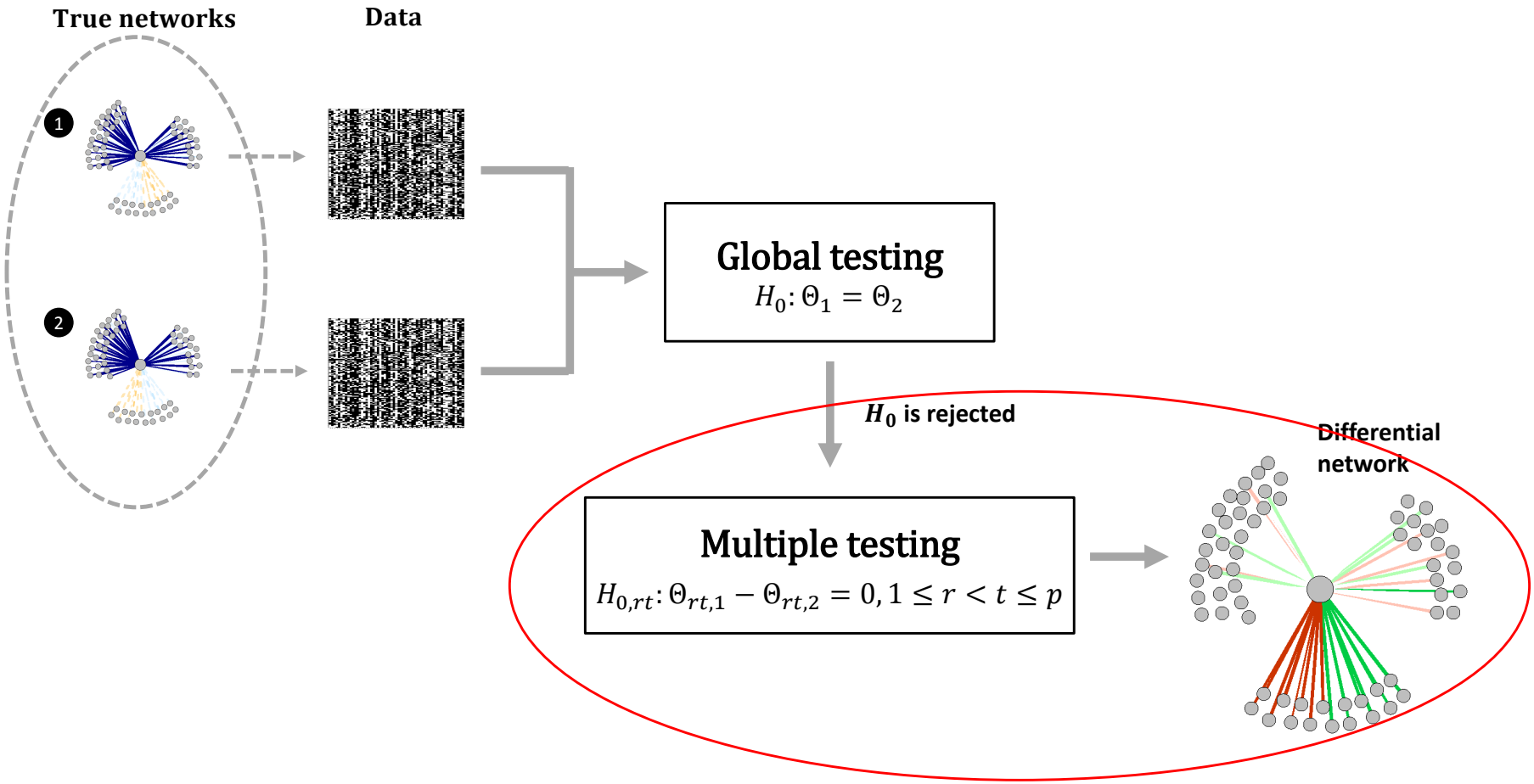


Simulation results: power

- $p = 100, \Theta_1 = \Theta_0 - \Delta, \Theta_2 = \Theta_0 + \Delta$ where $\|\Delta\|_0 = 10$.
- Generate data $\{X^{(i)}\}_{i=1}^n \sim \mathrm{P}_{\Theta_1}$ and $\{Y^{(i)}\}_{i=1}^n \sim \mathrm{P}_{\Theta_2}$ 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:

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

- 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: $R(\tau) = \sum_{1 \le r < t \le p} I(|W_{rt}| \ge \tau)$.

What is the FDR?

Challenge: for any given τ

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

What is the FDR?

Challenge: for any given τ

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

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

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

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

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 $\hat{\tau}$

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

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 $\hat{\tau}$

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

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)

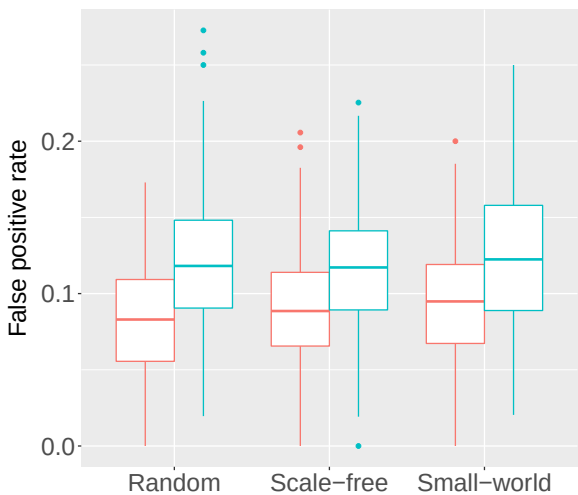
Let $q_0 = |\mathcal{H}_0|$ and $q = (p^2 - p)/2$. Under some regularity conditions, our multiple testing procedure asymptotically controls the false discovery rate, i.e.

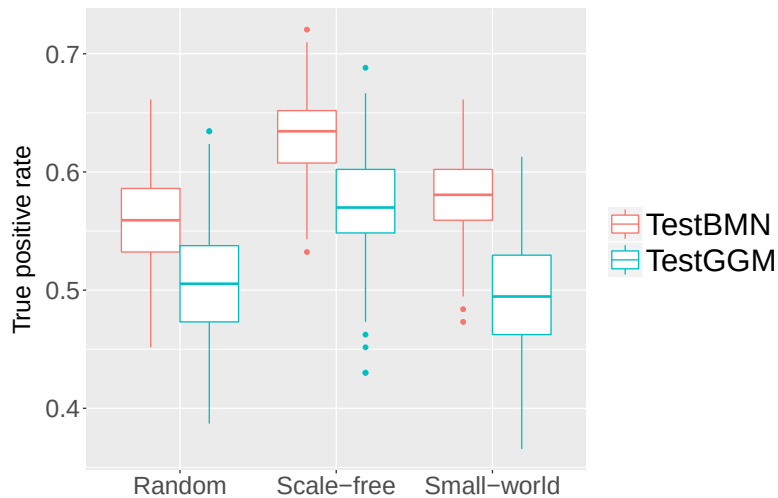
$$\frac{\mathrm{FDR}(\hat{\tau})}{\alpha q_0/q} \to 1, \quad \frac{\mathrm{FDP}(\hat{\tau})}{\alpha q_0/q} \to 1,$$

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

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 $\{X^{(i)}\}_{i=1}^n \sim \mathrm{P}_{\Theta_1}$ and $\{Y^{(i)}\}_{i=1}^n \sim \mathrm{P}_{\Theta_2}$ 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 genera[5] were removed, leaving $p = 59$.
- Only one member from each family was used.
- Young: $18 \leq \text{age} \leq 43$, $n_1 = 171$.
- Elderly: $74 \leq \text{age} \leq 89$, $n_2 = 180$.

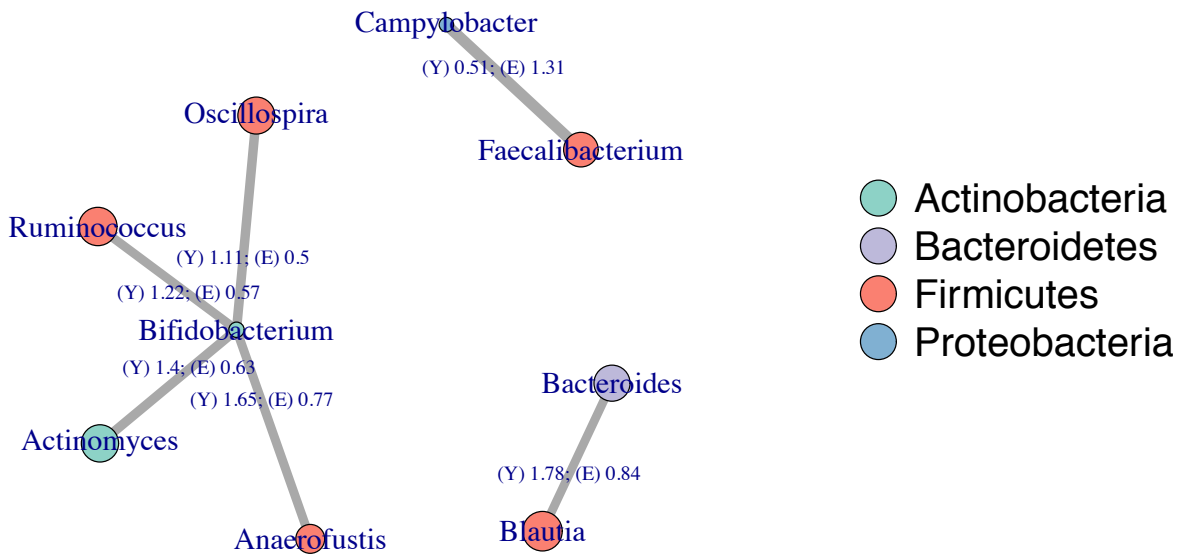
[5] Taxonomic rank: Species → Genus → Family → Order → Class → Phylum → Kingdom

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:

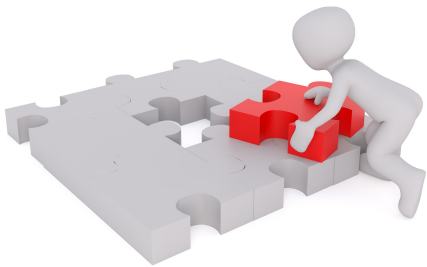
- Young OR $= 0.51$: presence in Faecalibacterium is associated with lower odds of presence in Campylobacter, a competitive relationship.
- Elderly OR $= 1.31$: presence in Faecalibacterium is associated with higher odds of presence in Campylobacter, a collaborative relationship.

Implications of differential network

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

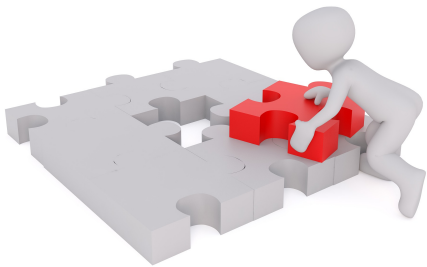
- Abundance of Faecalibacterium negatively associated with age (Franceschi et al. Trends Endocrinol Metab. 2017).
- Ruminococcus enriched in immune-mediated inflammatory diseases (Forbes et al. Front Microbiol. 2016).
- Abundance of Oscillospira enriched in inflammatory diseases (Konikoff and Gophna.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

Summary



- Learn conditional dependency relationships among microbes using Markov networks.

Summary

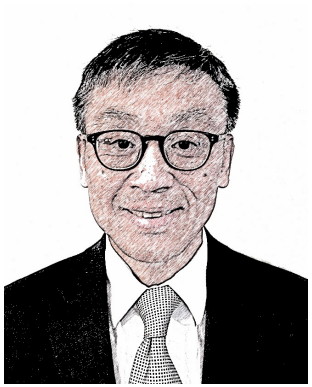


- Learn conditional dependency relationships among microbes using Markov networks.
- Differential network analysis identifies systematic changes in microbial interactions associated with age.

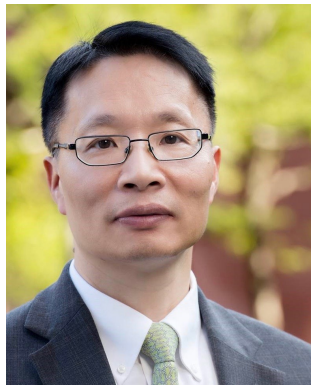
What's next?

- Presence/absence loses information – higher resolution?
- Multi-omics: microbiome, metabolomics, ...

Acknowledgement



Hongzhe Li



Tony Cai

Yin Xia

Manuscript is available upon request.
Code is available at `https://github.com/drjingma/TestBMN`.

Symposium

Microbiome

Data to Knowledge

Friday, March 16, 2018

Fred Hutchinson Cancer Research Center Campus

Symposium Agenda (subject to change)

Making Sense of the Human Microbiome

Speakers: Meredith Hullar, Elhanan Borenstein, David Fredricks

Microbiome and Infectious Diseases

Speakers: William DePaolo, Alison Roxby, Nina Salama

Bioinformatics for the Microbiome

Speakers: Ben Callahan, Daniel McDonald, Noah Hoffman

Microbiome Data Analysis

Speakers: Hongzhe Li, Shyamal Peddada, Jing Ma

Sponsored by Fred Hutchinson Cancer Research Center Public Health Sciences Division
Organized by Jing Ma, Michael Wu and Ruth Etzioni

For more information, visit fredhutch.org/microbiome2018

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, $\check{\beta}_r \approx \beta_r$+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

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

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

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

Score vectors for debiasing

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

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

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

- $\mathrm{E}[V\varepsilon_r] = 0$,
- $\min\langle V, V\rangle$ 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

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

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 $F_{rt} = 4\mathrm{E}_{\Theta_1}[(v_{rt}^o)^2 \ddot{f}(u_r)]$.
- 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

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

Assumptions

- $\log p = o(n_k^{1/3})$ and $n_1 \asymp n_2$
- $\max_{1 \le r \le p} \|\hat{\theta}_{r,k} - \theta_{r,k}\|_1 = o_{\mathrm{p}}\left(\{\log p\}^{-1}\right)$.
- $\max_{1 \le r \le p} \|\hat{\theta}_{r,k} - \theta_{r,k}\|_2 = o_{\mathrm{p}}\left(\{n_k \log p\}^{-1/4}\right)$.
- $\max_{1 \le t \le p} |\mathcal{A}_t(\xi)| = o(p^{\gamma})$ for $0 < \gamma < 1$, where for $\xi > 0$

$$\mathcal{A}_t(\xi) = \{r : |\sinh(2\theta_{rt,1})| \ge (\log p)^{-2-\xi} \text{ or } |\sinh(2\theta_{rt,2})| \ge (\log p)^{-2-\xi}\}.$$

Theory: global testing

Challenge:

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

Theory: global testing

Challenge:

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

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

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

- $M_{n,p} = \max_{1 \le r < t \le p} W_{rt}^2$.
- 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!
- $M_{n,p} - 4\log p + \log\log p \to \exp\{-(8\pi)^{-1/2} e^{-z/2}\}$!

Theory: multiple testing

- Given $\alpha > 0$, want

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

Theory: multiple testing

- Given $\alpha > 0$, want

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

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

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

Theory: multiple testing

- Given $\alpha > 0$, want

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

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

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

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

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

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