Covariance Properties and Graph Selection for High-Dimensional Compositional Data

Camden Lopez

WNAR 2017, Student Paper Session 3

Outline

- Compositional microbiome data
- Graphical model selection using SPIEC-EASI
- Covariance relationships and properties
- Graph selection performance

Compositional microbiome data

16S amplicon sequencing

- Sample $\rightarrow$ DNA $\rightarrow$ 16S sequences $\rightarrow$ OTU counts
- **Relative abundances only**

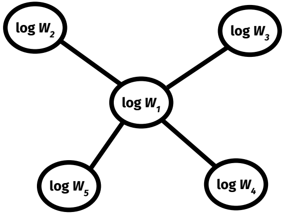
	OTU 1	OTU 2	...	OTU p
Sample 1	136	28	...	10
Sample 2	0	2	...	18
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Sample n	54	25	...	5

OTU = operational taxonomic unit

Graphical model inference using SPIEC-EASI

SPIEC-EASI: **SP**arse **I**nvers**E** **C**ovariance Estimation for **E**cological **AS**sociation **I**nference (Kurtz et al. 2015)

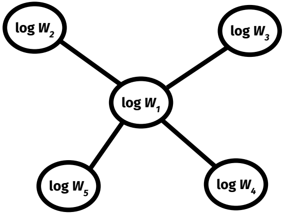
- Relationships among OTU abundances $W = (W_1, \ldots, W_p)$?
- Suppose $\log W \sim \mathrm{Normal}(\,\cdot\,, \Omega)$
- **Non-zero entries of** $\Omega^{-1} \Leftrightarrow$ **conditional dependence, graphical model**



Graphical model inference using SPIEC-EASI

SPIEC-EASI: **SP**arse **I**nvers**E** **C**ovariance Estimation for **E**cological **AS**sociation **I**nference (Kurtz et al. 2015)

- **Observe** $W \times ? \to \log W + ?$
- **Centering** $\log W + ? \to \operatorname{clr} W$
- **Assumption**: $\operatorname{cov}(\operatorname{clr} W) = \Gamma \approx \Omega = \operatorname{cov}(\log W)$
- **Graphical model inference**: $\widehat{\Gamma} \to \widehat{\Omega^{-1}}$ (graphical lasso, e.g.)



Covariance relationships and properties

Properties of Γ

- $\gamma_{ij} = \omega_{ij} - \overline{\omega}_{i\cdot} - \overline{\omega}_{\cdot j} + \overline{\omega}_{\cdot\cdot}$
- Rows/col's sum to zero
- p fewer free parameters than Ω

Covariance relationships and properties

Properties of Γ

- $\gamma_{ij} = \omega_{ij} - \overline{\omega}_{i\cdot} - \overline{\omega}_{\cdot j} + \overline{\omega}_{\cdot\cdot}$
- Rows/col's sum to zero
- p fewer free parameters than Ω

$\Gamma \approx \Omega$

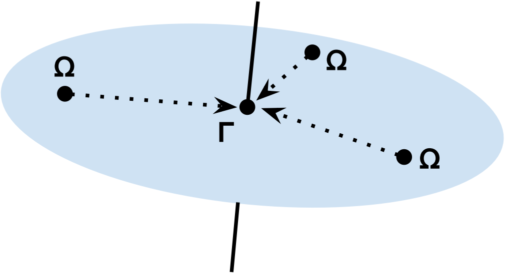
- Small or mostly negative correlations
- Approx. equal variances
- **Small "compositional effect"**

$\Gamma \not\approx \Omega$

- Mostly positive correlations
- Unequal variances
- **Large "compositional effect"**

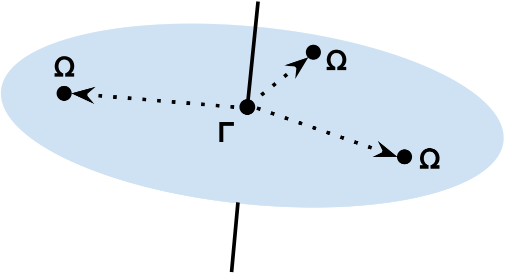
Covariance relationships and properties

One $\Gamma \leftrightarrow$ **many** Ω



- For each Γ, p-dimensional space of **potential** Ω**s**

One $\Gamma \leftrightarrow$ many Ω



- Can solve for **potential Ωs** (must check $\Omega \succ 0$)

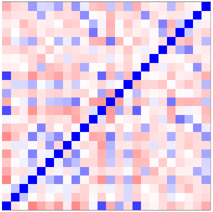
Covariance relationships and properties

Relationships can vary among potential Ωs

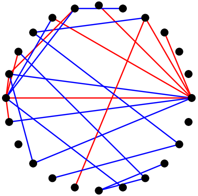
- Somewhat constrained, but not entirely

Example ($p = 24$, red = negative, blue = positive)

$\widehat{\Gamma}$ (correlations)



Graph (24 edges)



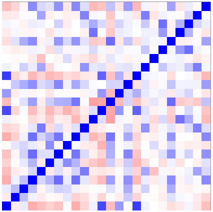
Covariance relationships and properties

Relationships can vary among potential Ωs

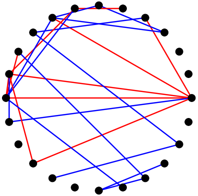
- Somewhat constrained, but not entirely

Example ($p = 24$, red = negative, blue = positive)

Potential $\widehat{\Omega}$ #1



Graph (24 edges)



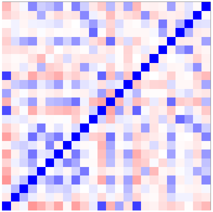
Covariance relationships and properties

Relationships can vary among potential Ωs

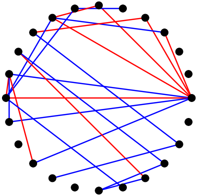
- Somewhat constrained, but not entirely

Example ($p = 24$, red = negative, blue = positive)

Potential $\widehat{\Omega}$ #2



Graph (24 edges)



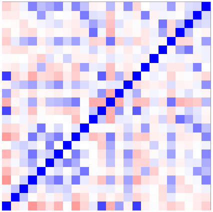
Covariance relationships and properties

Relationships can vary among potential Ωs

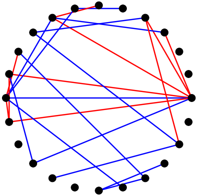
- Somewhat constrained, but not entirely

Example ($p = 24$, red = negative, blue = positive)

Potential $\widehat{\Omega}$ #3



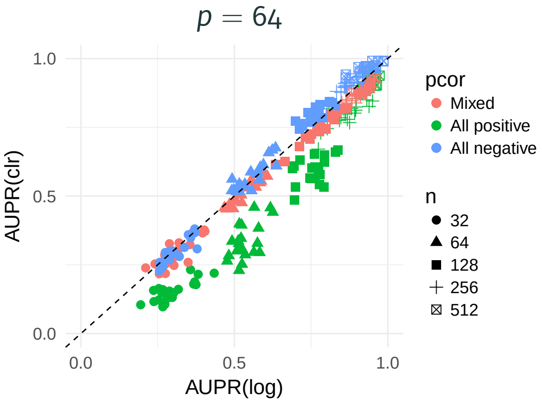
Graph (24 edges)



Graph selection performance

Performance with **small compositional effect**

- Comparable to graph selection from log data



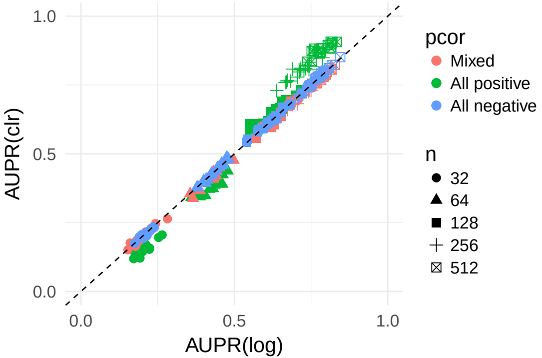
- Sparse graph
- Approx. equal variances
- Partial correlations (pcor) ± 0.25
- AUPR = area under precision-recall curve

Graph selection performance

Performance with **small compositional effect**

- Comparable to graph selection from log data

$p = 256$

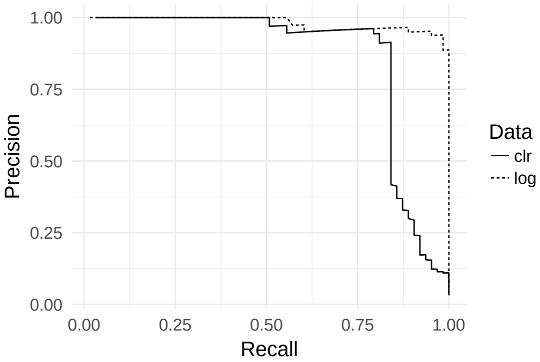


- Sparse graph
- Approx. equal variances
- Partial correlations (pcor) ± 0.25
- AUPR = area under precision-recall curve

Graph selection performance

Performance with **large compositional effect**

- Affected by distortion of covariances



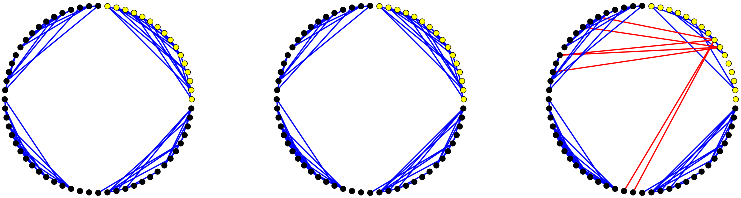
- Cluster graph, $p = 64$
- $25\times$ larger variances in one cluster
- $n = 1024$

Graph selection performance

Performance with **large compositional effect**

- Affected by distortion of covariances

Graph | Data = log | Data = clr



Summary

- **Limitation of compositional data**: One $\Gamma \leftrightarrow$ many Ω, uncertainty about $\log W$ relationships
- **SPIEC-EASI graph selection** for $\log W$ based on $\operatorname{clr} W$ data **performs well** … provided the compositional effect is not too large
- **Large compositional effect** distorts covariances and causes erroneous edges in graph

References & Acknowledgements

SPIEC-EASI paper:

- Kurtz., Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology* 11, e1004226.

Thank you Oregon State University faculty:

- Yuan Jiang, Duo Jiang, Sarah Emerson (Statistics)
- Thomas Sharpton (Microbiology and Statistics)

Centered log-ratio transformation

Given sample $x = (x_1, \dots, x_p)$ with geometric mean $g(x) = (\prod_{i=1}^{p} x_i)^{\frac{1}{p}}$,

$$\begin{aligned} \operatorname{clr}(x) &= \left(\log \frac{x_1}{g(x)}, \dots, \log \frac{x_p}{g(x)}\right) \\ &= \left(\log x_1 - \frac{1}{p}\sum_{i=1}^{p} \log x_i, \dots, \log x_p - \frac{1}{p}\sum_{i=1}^{p} \log x_i\right) \end{aligned}$$

and for $c > 0$,

$$\operatorname{clr}(cx) = \operatorname{clr}(x)$$

Covariance relationships

Matrix form:

$$
\begin{aligned}
\operatorname{clr} W &= \mathrm{G} \log W \\
\mathrm{G} &= \begin{pmatrix} 1 - \frac{1}{p} & \cdots & -\frac{1}{p} \\ \vdots & \ddots & \vdots \\ -\frac{1}{p} & \cdots & 1 - \frac{1}{p} \end{pmatrix} \\
\Rightarrow \operatorname{cov}(\operatorname{clr} W) &= \mathrm{G} \operatorname{cov}(\log W) \mathrm{G} \\
\Rightarrow \Gamma &= \mathrm{G} \Omega \mathrm{G}
\end{aligned}
$$

Graphical lasso estimation of Ω^{-1}

$$\widehat{\Omega^{-1}}_{\text{glasso}} = \underset{\Omega^{-1} \succeq 0}{\arg\max} \left[\log\det(\Omega^{-1}) - \operatorname{tr}(\widehat{\Omega}\Omega^{-1}) - \lambda \|\Omega^{-1}\|_1 \right]$$

$$\widehat{\Omega^{-1}}_{\text{SPIEC-EASI}} = \underset{\Omega^{-1} \succeq 0}{\arg\max} \left[\log\det(\Omega^{-1}) - \operatorname{tr}(\widehat{\Gamma}\Omega^{-1}) - \lambda \|\Omega^{-1}\|_1 \right]$$

Solving for potential Ω

Option 1: Choose $\overline{\omega}_{1\cdot}, \ldots, \overline{\omega}_{p\cdot}$

$$\gamma_{ij} = \omega_{ij} - \overline{\omega}_{i\cdot} - \overline{\omega}_{\cdot j} + \overline{\omega}_{\cdot\cdot}$$

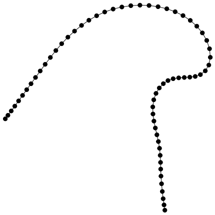
$$\Rightarrow \omega_{ij} = \gamma_{ij} + \overline{\omega}_{i\cdot} + \overline{\omega}_{\cdot j} - \overline{\omega}_{\cdot\cdot}$$

Option 2: Choose $\omega_{11}, \ldots, \omega_{pp}$

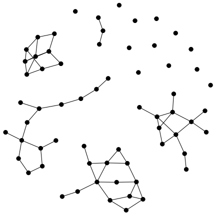
$$\omega_{ij} = \gamma_{ij} + \frac{1}{2}\left(\omega_{ii} - \gamma_{ii} + \omega_{jj} - \gamma_{jj}\right)$$

Graphs used in simulations

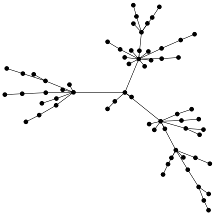
$p = 64$, $e = p - 1$



Band



Cluster



Scale-free