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**

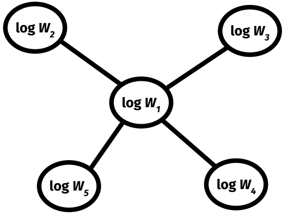
	OTU 1	OTU 2	...	OTU p
Sample 1	136	28	...	10
Sample 2	0	2	...	18
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
Sample n	54	25	...	5

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)

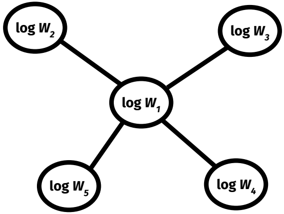
- Relationships among OTU abundances $W = (W_1, \ldots, W_p)$?
- Suppose $\log W \sim \text{Normal}(\,\cdot\,, \Omega)$
- **Non-zero entries of** $\Omega^{-1} \Leftrightarrow$ **conditional dependence, graphical model**



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** samples $\to \operatorname{clr} W$
- **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

$\Gamma \approx \Omega$

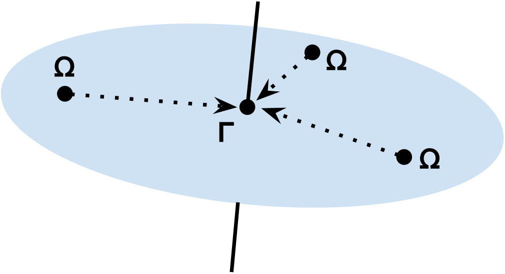
- Small or mostly negative correlations
- Approx. equal variances
- **Small "compositional effect"**

$\Gamma \not\approx \Omega$

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

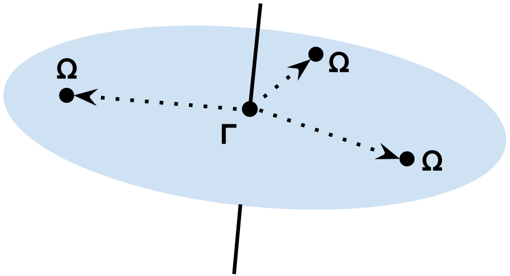
Covariance relationships and properties

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



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

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



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

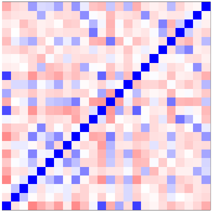
Covariance relationships and properties

Relationships can vary among potential Ωs

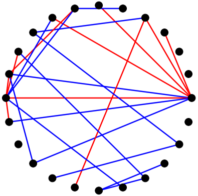
- Somewhat constrained, more so with larger p

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

$\widehat{\Gamma}$ (correlations)



Graph (24 edges)



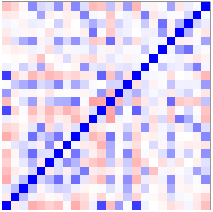
Covariance relationships and properties

Relationships can vary among potential Ωs

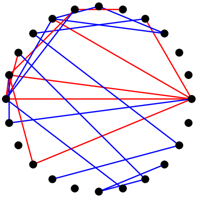
- Somewhat constrained, more so with larger p

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

Potential $\widehat{\Omega}$ #1



Graph (24 edges)



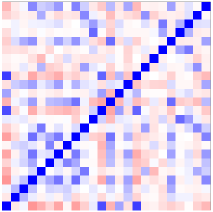
Covariance relationships and properties

Relationships can vary among potential Ωs

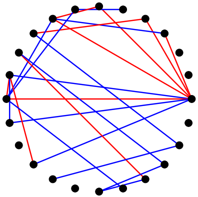
- Somewhat constrained, more so with larger p

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

Potential $\widehat{\Omega}$ #2



Graph (24 edges)

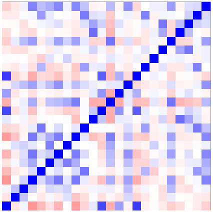


Relationships can vary among potential Ωs

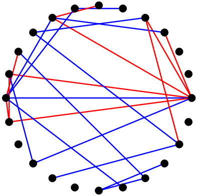
- Somewhat constrained, more so with larger p

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

Potential $\widehat{\Omega}$ #3



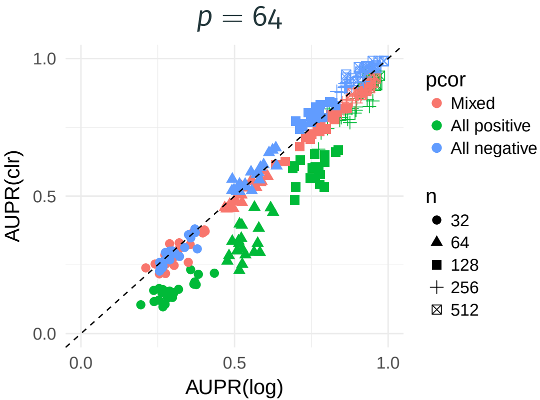
Graph (24 edges)



Graph selection performance

Performance with **small compositional effect**

- Comparable to graph selection from log data

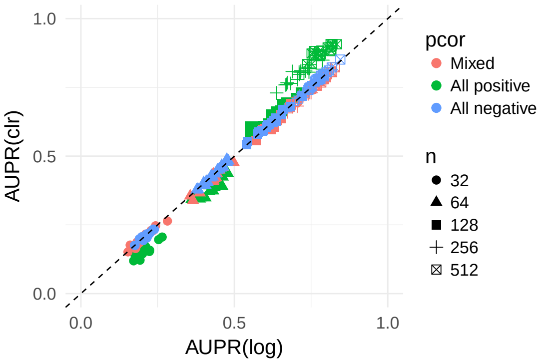


- Sparse graph
- Approx. equal variances
- Partial correlations (pcor) ± 0.25
- AUPR = area under precision-recall curve

Performance with **small compositional effect**

- Comparable to graph selection from log data

$p = 256$

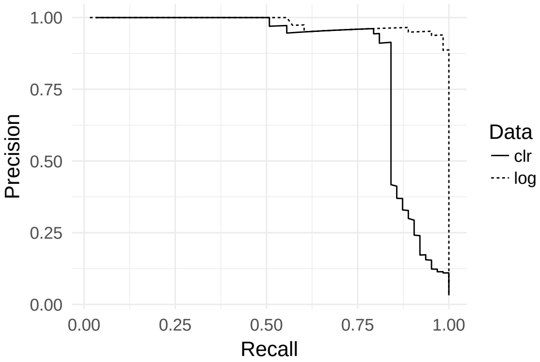


- Sparse graph
- Approx. equal variances
- Partial correlations (pcor) ± 0.25
- AUPR = area under precision-recall curve

Graph selection performance

Performance with **large compositional effect**

- Affected by distortion of covariances



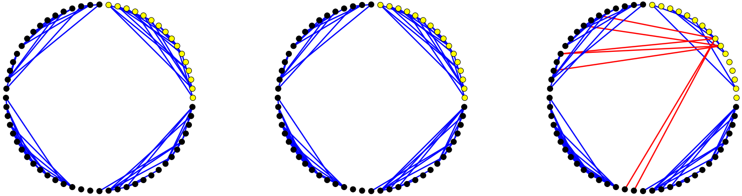
- Cluster graph, $p = 64$
- $25\times$ larger variances in one cluster
- $n = 1024$

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)