Covariance Properties and Graph Selection for High-Dimensional Compositional Data

Camden Lopez

June 6, 2017

Outline

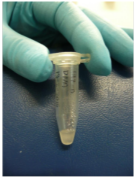
- Compositional microbiome data
- Graphical model and graphical lasso
- Centered log-ratio transformation and SPIEC-EASI
- Covariance properties and alternative covariances
- Graph selection performance

Compositional microbiome data

16S amplicon sequencing:

1. Sample from environment
2. Extract DNA
3. Isolate and amplify 16S genes
4. Classify 16S genes by operational taxonomic unit (OTU)


1


2


3


4

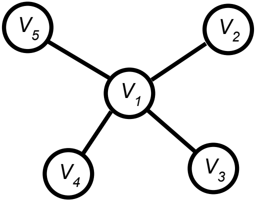
Compositional microbiome data

Sample	OTU 1	OTU 2	$\cdots$	OTU p
1	y_{11}	y_{12}	$\cdots$	y_{1p}
2	y_{21}	y_{22}	$\cdots$	y_{2p}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	y_{n1}	y_{n2}	$\cdots$	y_{np}

- y_{ij} = # sequences mapped to OTU j in sample i
- **Only relative proportions/ratios informative**

Graphical model

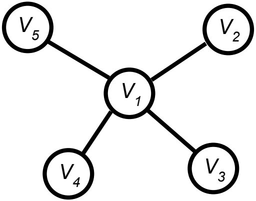
Graphical model representation of $V = (V_1, \ldots, V_p)$



- **Nodes = variables, edges = conditional dependence**

Graphical model

Graphical model representation of $V = (V_1, \ldots, V_p)$



- **Nodes = variables, edges = conditional dependence**
- Assuming $V \sim \mathrm{N}(\mu, \Sigma)$, V_i and V_j conditionally dependent if and only if $(\Sigma^{-1})_{ij} \neq 0$

Goal of inference

Notation:

$$W = (W_1, \ldots, W_p) = \text{absolute abundances of OTUs}$$
$$\Omega = \text{cov}(\log W)$$

Goal of inference

Notation:

$$W = (W_1, \ldots, W_p) = \text{absolute abundances of OTUs}$$
$$\Omega = \mathrm{cov}(\log W)$$

Goal:

- Infer conditional dependence relationships among $\log W_1, \ldots, \log W_p$
- Assuming $\log W \sim N(\,\cdot\,, \Omega)$, infer non-zero entries of Ω^{-1}

Centered log-ratio transformation

Centered log-ratio (clr) transformation:

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

Centered log-ratio transformation

Centered log-ratio (clr) transformation:

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

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

Centered log-ratio transformation

Centered log-ratio (clr) transformation:

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

$$z = \mathrm{clr}(x) = \left(\log \frac{x_1}{g(x)}, \ldots, \log \frac{x_p}{g(x)}\right)$$
$$= \left(\log x_1 - \frac{1}{p}\sum_{i=1}^{p} \log x_i, \ldots, \log x_p - \frac{1}{p}\sum_{i=1}^{p} \log x_i\right)$$

Notation:

- $\Gamma = \mathrm{cov}(\mathrm{clr}\, W)$

SPIEC-EASI

SParse **I**nvers**E** **C**ovariance estimation for **E**cological **AS**sociation **I**nference (SPIEC-EASI)

- Estimate $\Gamma = \mathrm{cov}(\mathrm{clr}\, W)$ from compositional data
- Infer non-zero entries of Ω^{-1} using $\hat{\Gamma}$, assuming $\Gamma \approx \Omega$

SPIEC-EASI

SParse **I**nvers**E** **C**ovariance estimation for **E**cological **AS**sociation **I**nference (SPIEC-EASI)

- Estimate $\Gamma = \text{cov}(\text{clr}\, W)$ from compositional data
- Infer non-zero entries of Ω^{-1} using $\hat{\Gamma}$, assuming $\Gamma \approx \Omega$

Graphical lasso estimate:

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

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

Covariance properties

Matrix form:

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

Covariance properties

Matrix form:

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

Entry form:

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

- Rows (and columns) of Γ sum to zero

Covariance properties

$\Gamma \approx \Omega$ when $-\overline{\omega}_{i\cdot} - \overline{\omega}_{j\cdot} + \overline{\omega}_{\cdot\cdot} \approx 0$

- Ω row averages all small ("sparse" covariances)
- Larger p helps if Ω row sums increase slower than p
- **Small "compositional effect"**

Covariance properties

$\Gamma \approx \Omega$ when $-\overline{\omega}_{i\cdot} - \overline{\omega}_{j\cdot} + \overline{\omega}_{\cdot\cdot} \approx 0$

- Ω row averages all small ("sparse" covariances)
- Larger p helps if Ω row sums increase slower than p
- **Small "compositional effect"**

$\Gamma \not\approx \Omega$

- Some or all Ω row averages large
- Many positive correlations, or some extremely large variances
- **Large "compositional effect"**

Alternative covariances

Given Γ, what could Ω be?

- Γ has p fewer free parameters than Ω:
  $\frac{1}{2}p(p-1)$ vs. $\frac{1}{2}p(p+1)$
- Each Γ associated with a p dimensional space of possible Ωs

Alternative covariances

Given Γ, what could Ω be?

- Γ has p fewer free parameters than Ω: $\frac{1}{2}p(p-1)$ vs. $\frac{1}{2}p(p+1)$
- Each Γ associated with a p dimensional space of possible Ωs

Solving for Ω, given Γ and choosing $\omega_{11}, \ldots, \omega_{pp}$:

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

Alternative covariances

Given Γ, what could Ω be?

- Γ has p fewer free parameters than Ω: $\frac{1}{2}p(p-1)$ vs. $\frac{1}{2}p(p+1)$
- Each Γ associated with a p dimensional space of possible Ωs

Solving for Ω, given Γ and choosing $\omega_{11}, \ldots, \omega_{pp}$:

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

- Check that Ω is **positive definite** (valid covariance)

Alternative covariances: $p = 2$

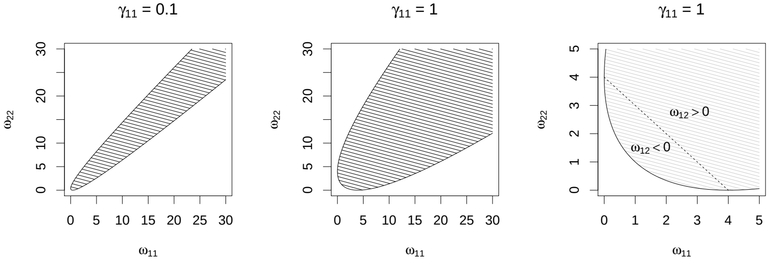
Can analyze and visualize $p = 2$ case:

- Ω positive definite iff $\det(\Omega) > 0$
- Can find where $\omega_{12} < 0$, $\omega_{12} = 0$, or $\omega_{12} > 0$

Alternative covariances: $p = 2$

Can analyze and visualize $p = 2$ case:

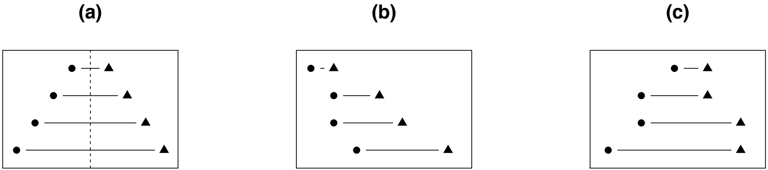
- Ω positive definite iff $\det(\Omega) > 0$
- Can find where $\omega_{12} < 0$, $\omega_{12} = 0$, or $\omega_{12} > 0$



Alternative covariances: $p = 2$

Example:

- (a) clr abundances
- (b) log abundances with $\omega_{11} = 0.7$, $\omega_{22} = 4.3$, $\omega_{12} = 1.7$
- (c) log abundances with $\omega_{11} = 0.7$, $\omega_{22} = 0.3$, $\omega_{12} = -0.3$



Alternative covariances: larger p

More complicated to analyze...

- Investigated several examples by trial-and-error

Alternative covariances: larger p

More complicated to analyze...

- Investigated several examples by trial-and-error

Example: Γ corresponding to Ω used in $p = 64$ simulations

- Similar **limits on potential variances** $\omega_{11}, \ldots, \omega_{pp}$

	ω_{11}	ω_{22}	ω_{33}	ω_{44}	ω_{55}	...
min	0.22	0.30	0.18	0.22	0.18	...
max	0.39	0.47	0.36	0.40	0.36	...

- Unconditional relationships (Ω entries) and conditional relationships (Ω^{-1}) can **vary widely**

Alternative covariances: larger p

Partial correlations

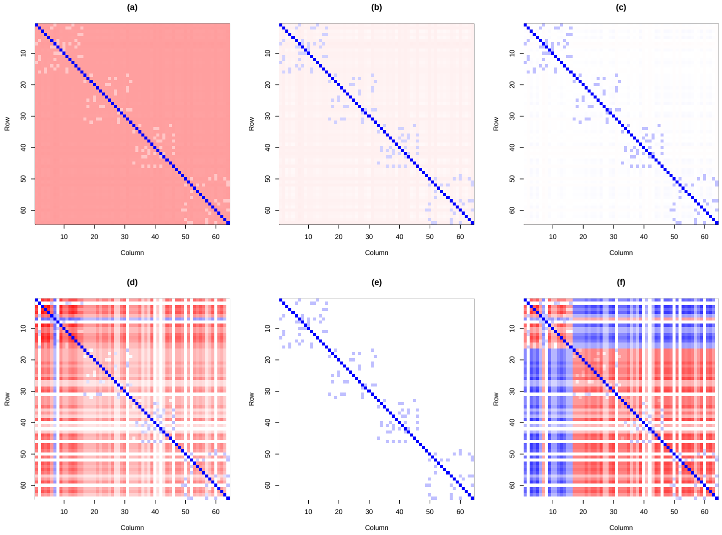
- Indicate signs and strengths of conditional relationships
- Blue = positive, red = negative partial correlation

Alternative covariances: larger p

Partial correlations

- Indicate signs and strengths of conditional relationships
- Blue = positive, red = negative partial correlation
- (a) $\omega_{ii} = \gamma_{ii} + 10^{-4}$ for all i
- (b) $\omega_{ii} = \gamma_{ii} + 10^{-3}$ for all i
- (c) $\omega_{ii} = \gamma_{ii} + 10^{-2}$ for all i
- (d) Small ω_{ii} in first cluster, others equal to values in initial Ω
- (e) All ω_{ii} equal to values in initial Ω
- (d) Large ω_{ii} in first cluster, others equal to values in initial Ω

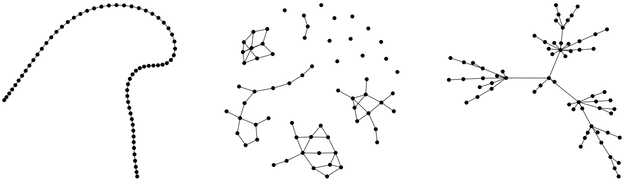
Alternative covariances: larger p



Performance: simulation setup

How well does it work in ideal settings?

- Three sparse graph structures ($p - 1$ edges in each)
- Graph $\rightarrow \Omega^{-1}$ (p. cor. ± 0.25) $\rightarrow \Omega$
- $\log W \sim \mathrm{N}(0, \Omega)$



Performance: simulation setup

Performance metric:

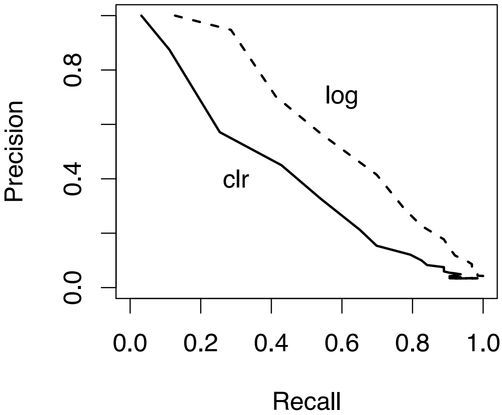
- Area under precision-recall curve (AUPR)
- Points on curve $\leftrightarrow$ graphical lasso solutions for $\lambda_{min} < \cdots < \lambda_{max}$
- Using both log and clr data for comparison

$$\text{Recall} = \frac{\#\ \text{correctly selected edges}}{\#\ \text{edges in true graph}}$$

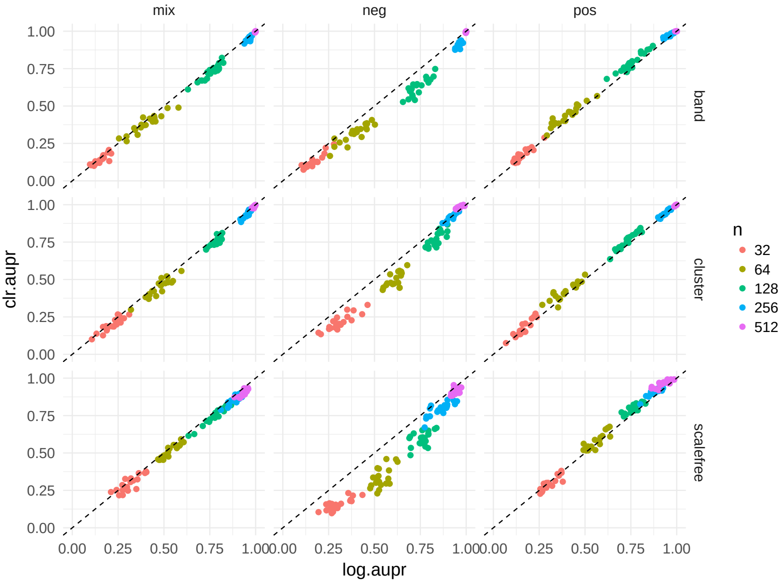
$$\text{Precision} = \frac{\#\ \text{correctly selected edges}}{\#\ \text{edges selected}}$$

Performance: simulation setup

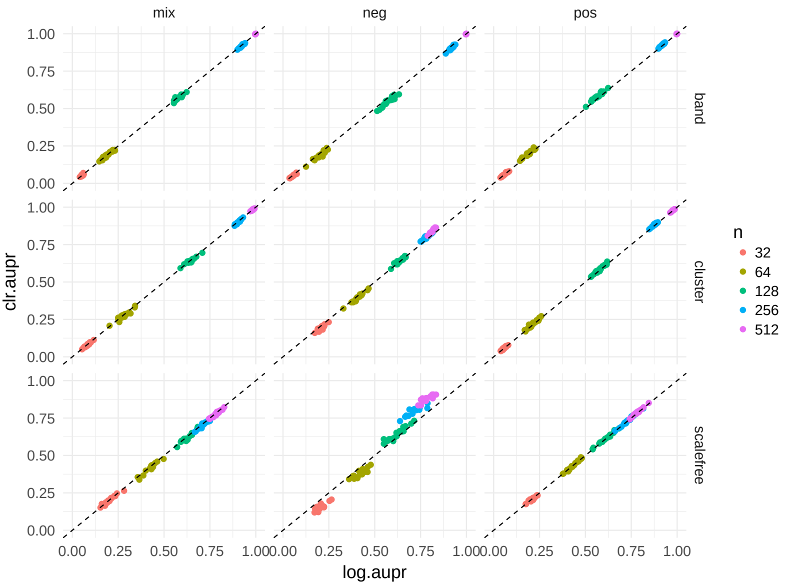
Precision-recall curves (example):



Performance: $p = 64$



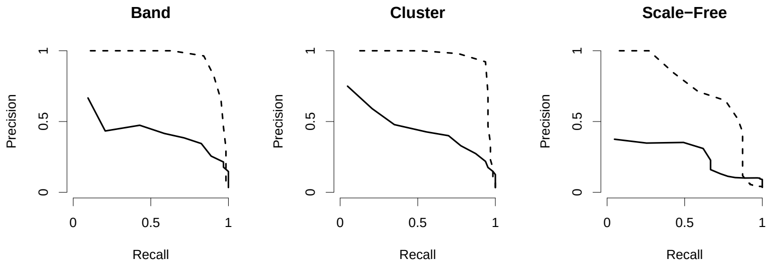
Performance: $p = 256$



Performance: large compositional effect

What if there's a large compositional effect?

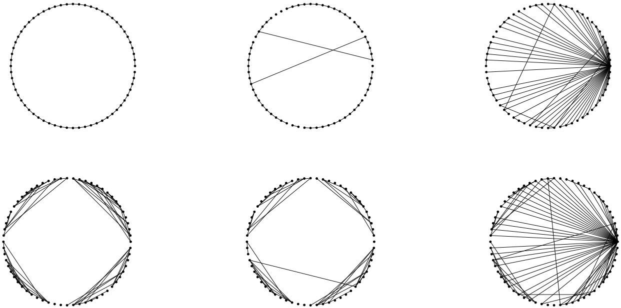
- Same setting as $p = 64$ simulations, but ω_{11} 400 times larger
- (—) = clr, (- - -) = log



Performance: large compositional effect

Spurious edges due to compositional effect:

- Truth (left) vs. log data (center) vs. clr data (right)



Conclusion

Graph selection:

- Performs well when there isn't a large compositional effect
- Otherwise can select spurious edges

Limitation of compositional data:

- Covariances of clr data could correspond to a variety of log abundance covariances
- Variety of possible conditional and unconditional relationships

References

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Chapman and Hall, London, UK.

Friedman, J., Hastie, T., and Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics* 9, 432–441.

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.