

Mini-course Compositional Data Analysis

Network Analysis

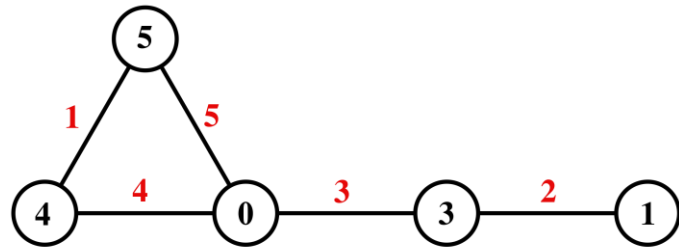
Agenda

1. Introduction to ecological / microbial networks
2. Association estimation
3. The SPIEC-EASI approach and R package
4. From associations to adjacencies
5. Network analysis

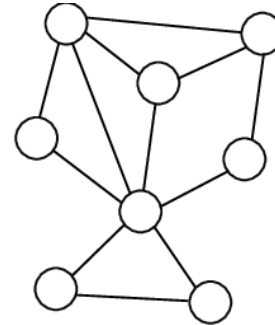
Introduction

Networks and graphs

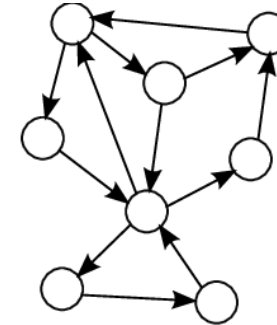
- The term **graph** is used in computer science and math



Graph with labeled **vertices** and **edges**

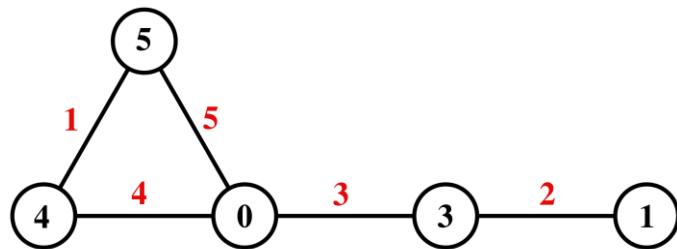


Directed graph



Undirected graph

- The term **network** is used in physics, biology, and social sciences



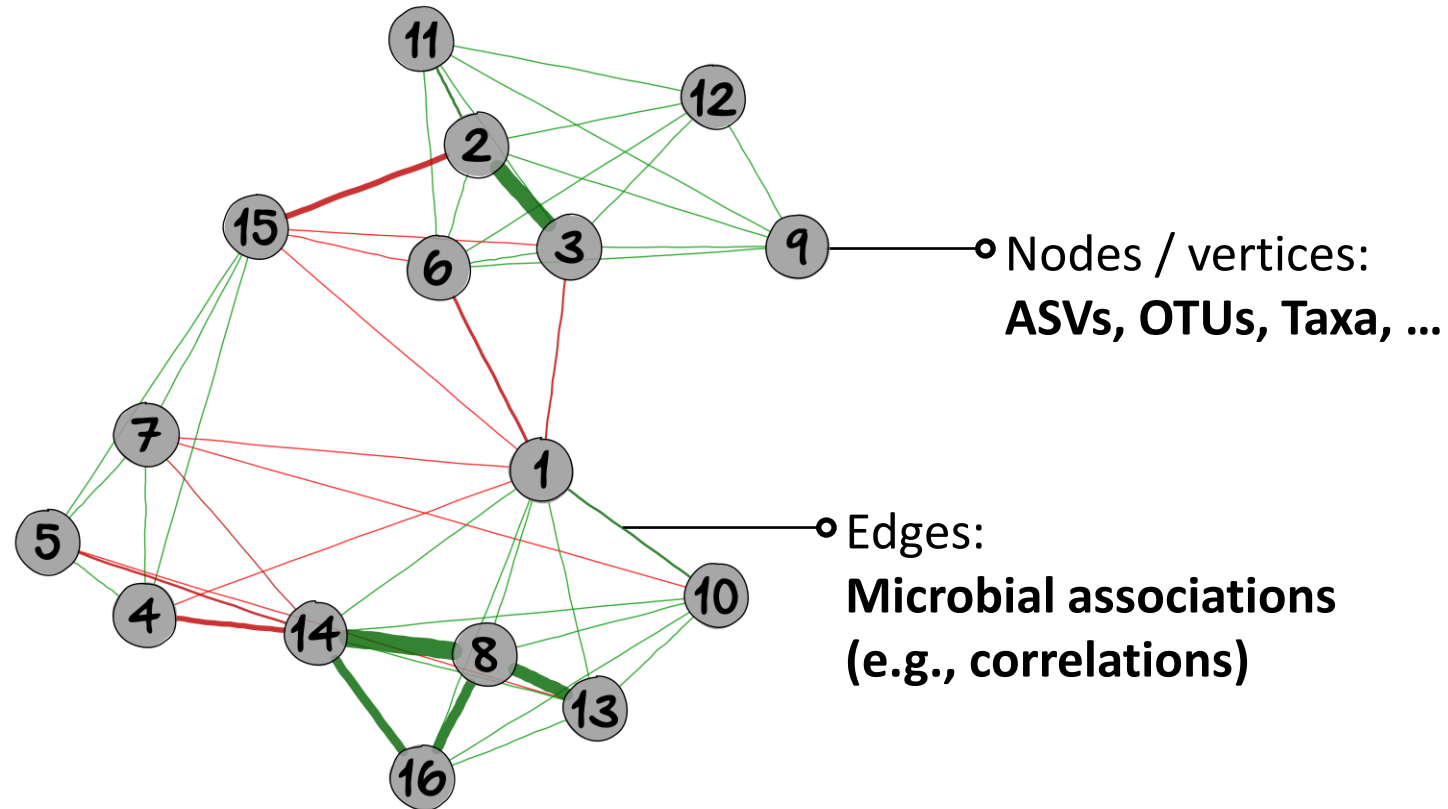
Network with labeled **nodes** and **links**

Networks in cell / molecular biology

- Protein-protein interaction (PPI) networks
- Gene regulatory (co-expression) networks
- Metabolic networks
- Signaling networks
- Neuronal networks
- Microbial (ecological) interaction networks

Microbiome networks

→ Insights into the organizational structure of a microbial community



**Association strength
and direction:**

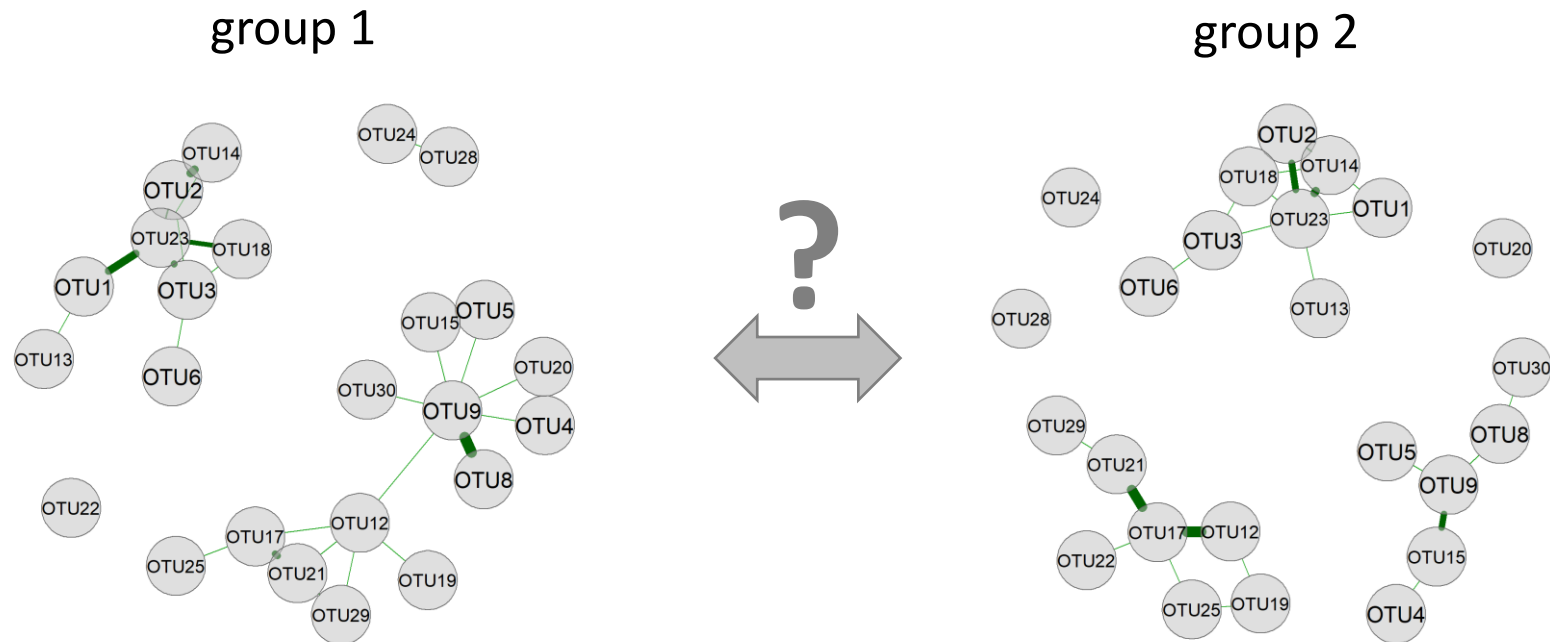
- Strongly positive
- Weakly positive
- Strongly negative
- Weakly negative

Layout:

Force-directed layout algorithm

Network comparison

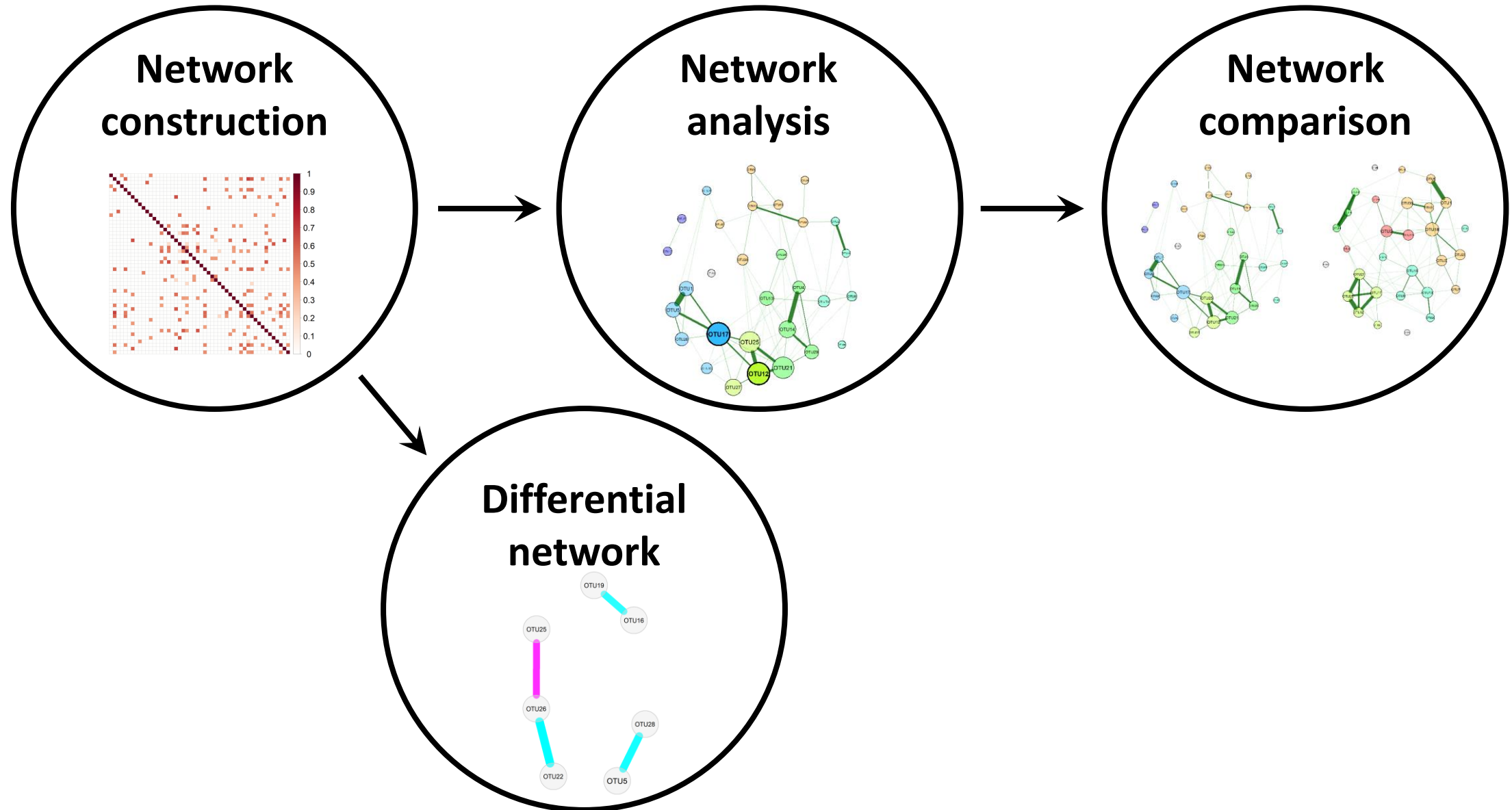
→ Does the microbial composition change across different conditions?



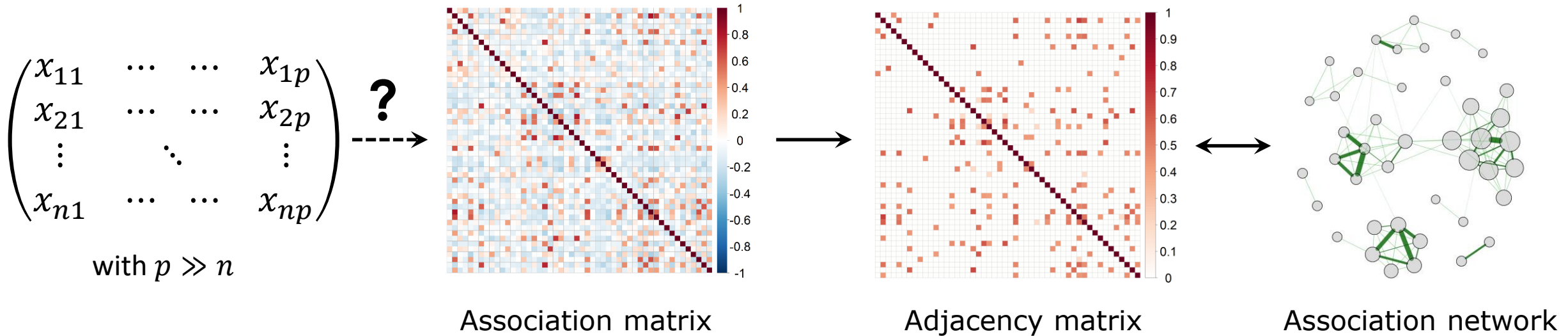
Possible groups:

- Cases and controls
- Two environmental states
- Two time points
- ...

Typical network analysis workflow



From sequencing data to networks

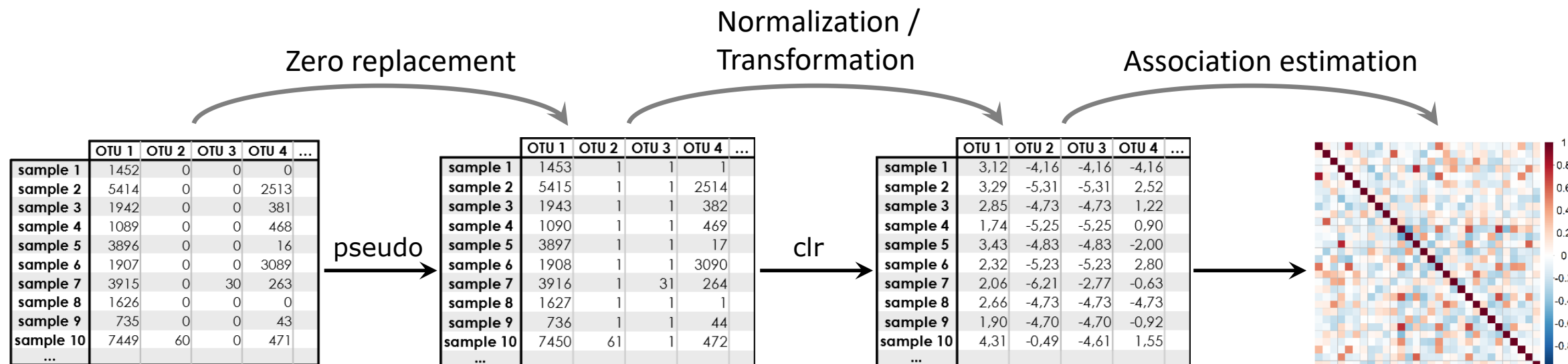


Data characteristics:

- We observe only a **sample** of the true microbial composition
- **Compositional** (only relative information)
- **Sequencing depth** (total number of reads) **varies** across samples
- **Zero-inflated**
- **High-dimensional** (number of taxa $p \gg$ sample size n)

Association estimation

Association estimation



- Pseudo counts
- Methods from **zCompositions** R package (Martín-Fernández et al., 2011)
- ...

- Centered log-ratio (**clr**) transformation
 - Variance stabilizing transformation (**VST**)
 - ...
- (Badri et al., 2020)

- Correlation
- Conditional dependence
- Proportionality

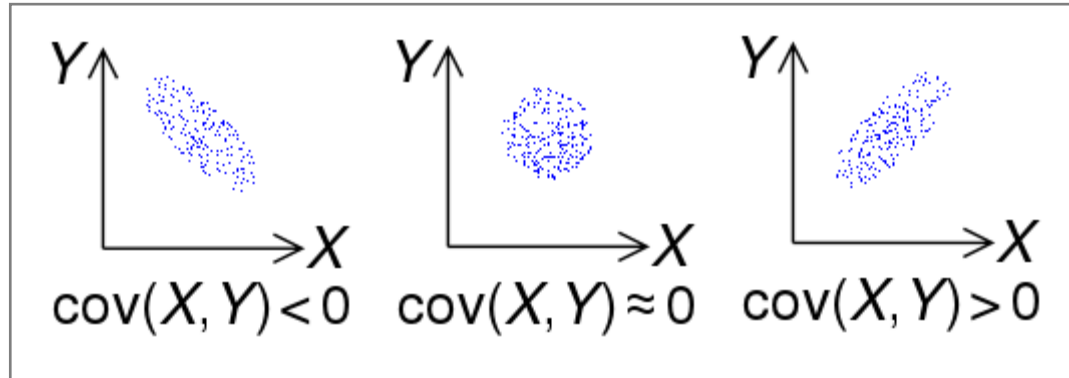
Packages often include the preprocessing steps, e.g.,

- SparCC
- SPRING
- SpiecEasi

References:

- Badri M, Kurtz ZD, Bonneau R, et al. Shrinkage improves estimation of microbial associations under different normalization methods. *NAR Genom Bioinform* 2020. doi: 10.1093/nargab/lqaa100.
- Martín-Fernández JA, Palarea-Albaladejo J, Olea RA. Dealing with zeros. *Compositional data analysis*, 2011, 43–58.

Correlation as association measure



Source: Wikipedia



Correlation:

$$\rho(X, Y) = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y}$$

How to estimate the covariance matrix?

Covariance estimation problem

Let x_1, \dots, x_n be independent observations of a p -dimensional random vector $X \in \mathbb{R}^{p \times 1}$, and $p \gg n$.

We want to find an estimate $\hat{\Sigma}$ of the **covariance matrix** $\Sigma = \mathbb{E}[(x - \mathbb{E}(x))(x - \mathbb{E}(x))^T] \in S^{p \times p}$

The **sample covariance matrix** S has entries

$$s_{ij} = \frac{1}{n-1} \sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)$$

where $\bar{x}_i = \frac{1}{n} \sum_{k=1}^n x_{ki}$ and x_{ki} is the k -th observation of the variable X_i .

Properties:

- S is an **unbiased** estimator of Σ
- In the Gaussian case, S is the **ML estimator** of the covariance matrix

$p \gg n$ case:

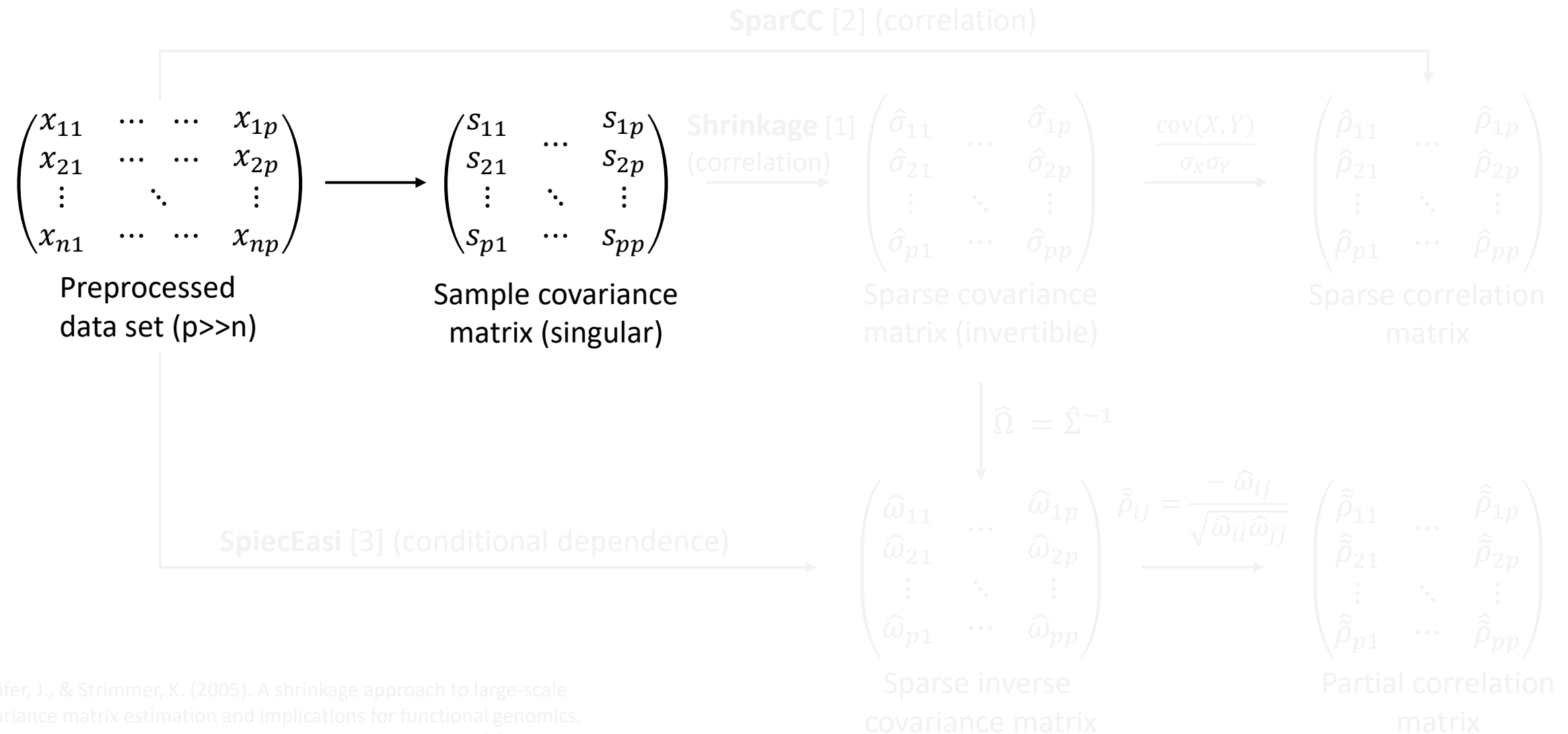
- S has no full rank
- S singular (non-invertible)
- S not positive definite

S is no good approximation if $p \gg n$!

$$X = \begin{pmatrix} x_{11} & \cdots & \cdots & x_{1p} \\ x_{21} & \cdots & \cdots & x_{2p} \\ \vdots & & \ddots & \vdots \\ x_{n1} & \cdots & \cdots & x_{np} \end{pmatrix}$$

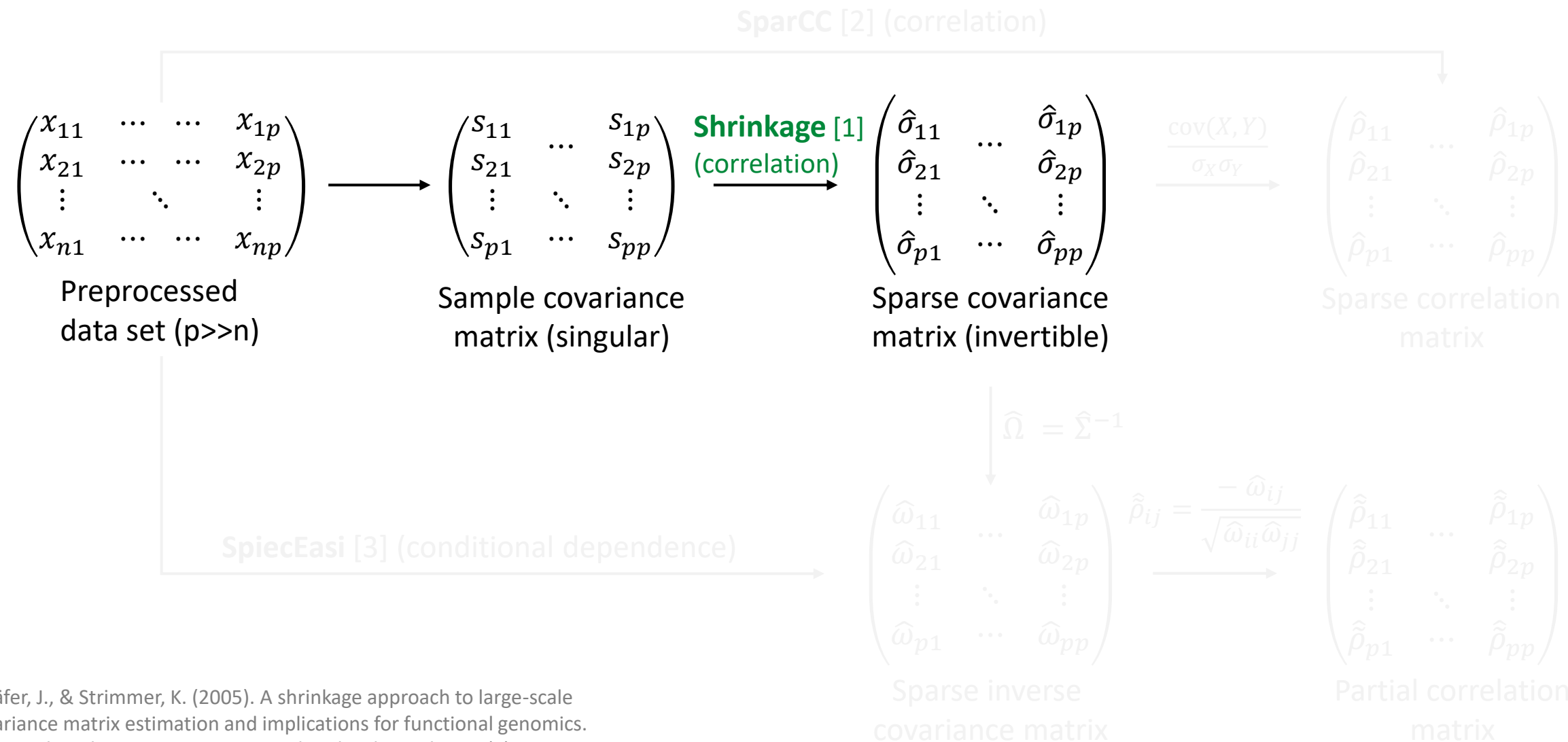
$$S = \begin{pmatrix} s_{11} & \cdots & s_{1p} \\ s_{21} & \cdots & s_{2p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{pmatrix}$$

Solutions for high-dimensional data (a few examples)



[1] Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 1–30.

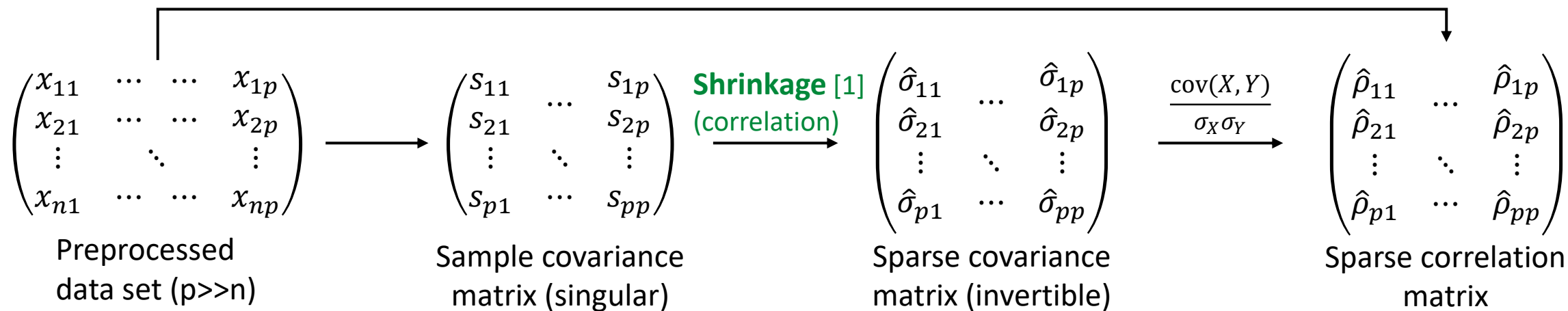
Solutions for high-dimensional data (a few examples)



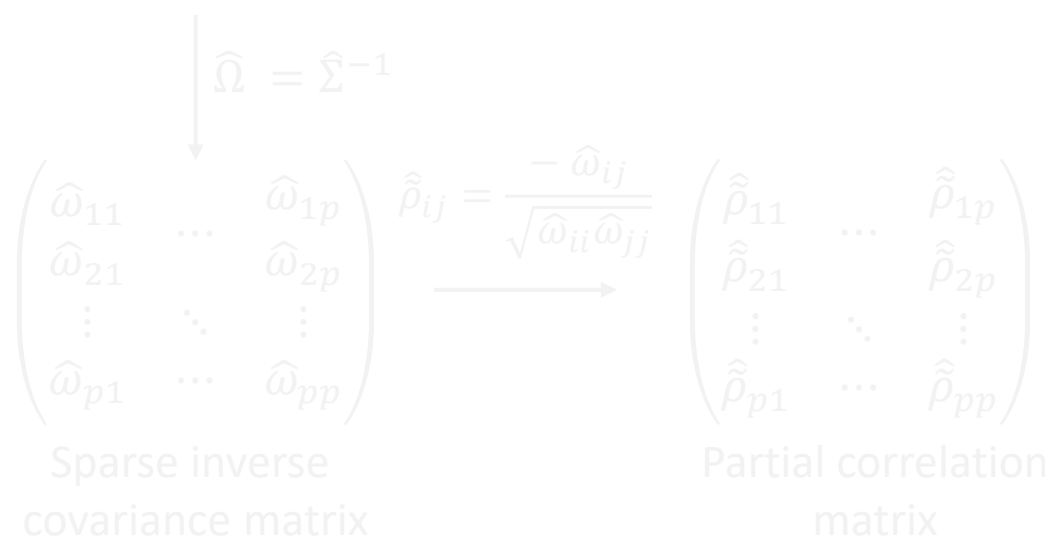
[1] Schäfer, J., & Strimmer, K. (2005). A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, 4(1), 1–30.

Solutions for high-dimensional data (a few examples)

SparCC [2] (correlation)



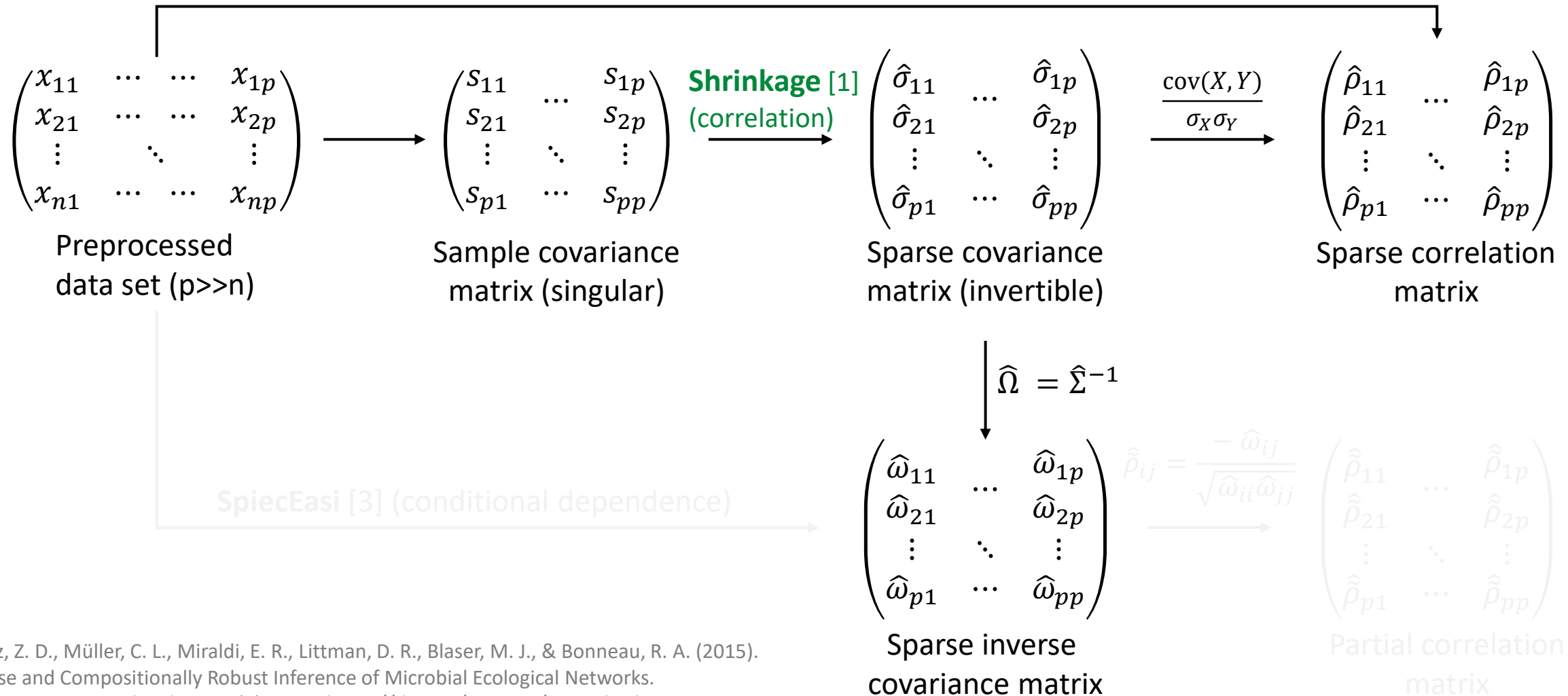
SpiecEasi [3] (conditional dependence)



[2] Friedman, J., & Alm, E. J. (2012). Inferring Correlation Networks from Genomic Survey Data. PLoS Computational Biology, 8(9), 1–11.

Solutions for high-dimensional data (a few examples)

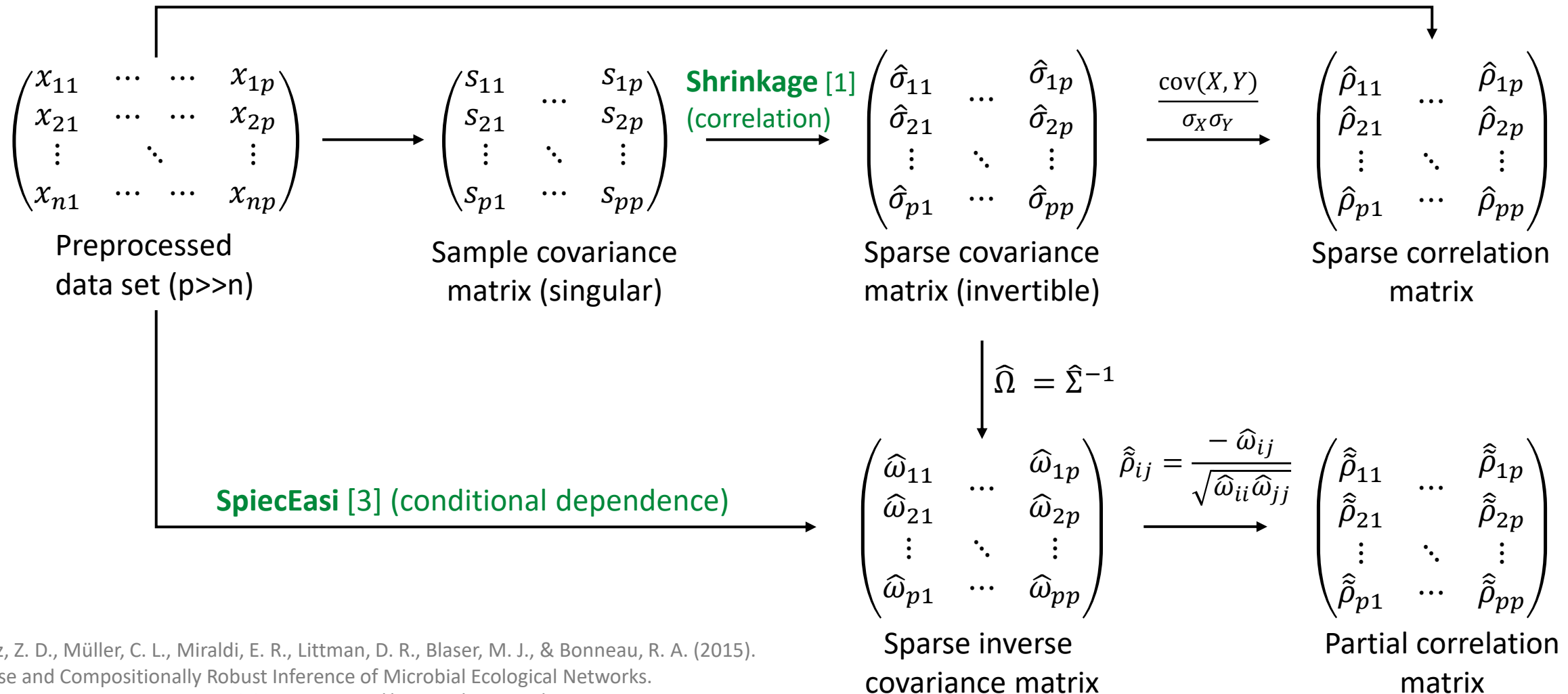
SparCC [2] (correlation)



[3] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25. <https://doi.org/10.1371/journal.pcbi.1004226>

Solutions for high-dimensional data (a few examples)

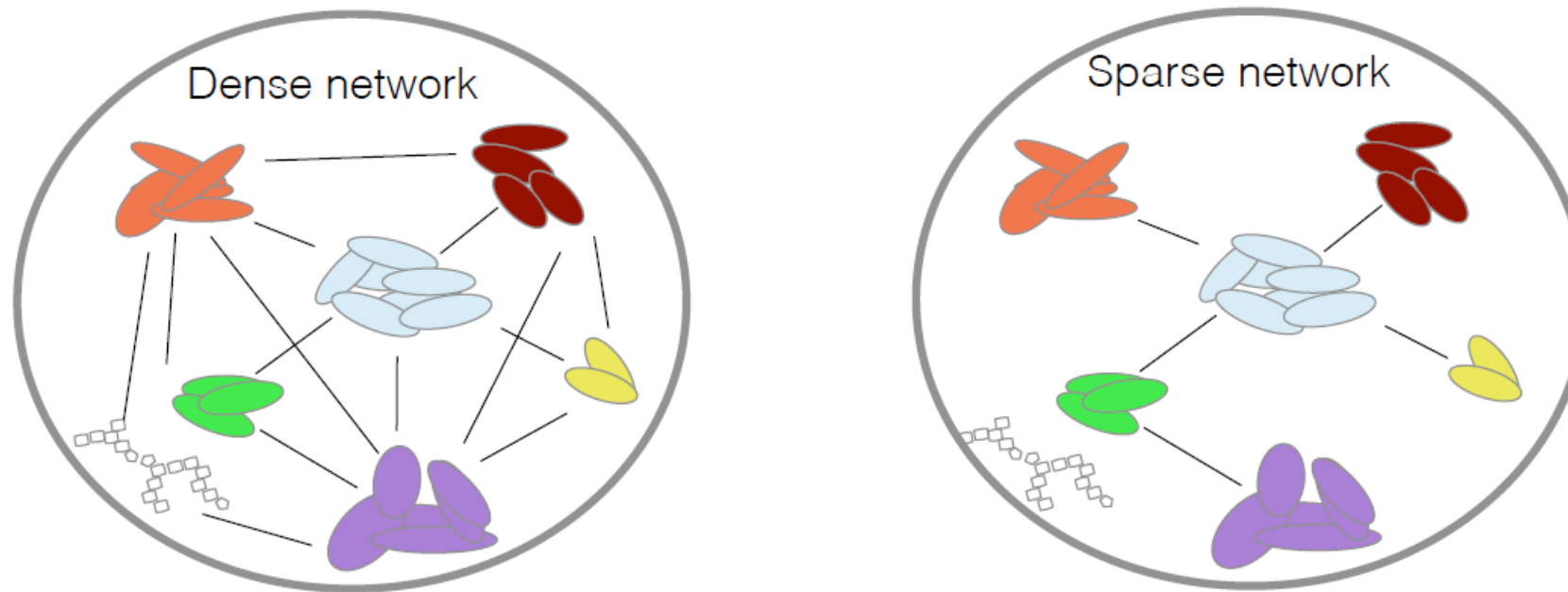
SparCC [2] (correlation)



[3] Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25. <https://doi.org/10.1371/journal.pcbi.1004226>

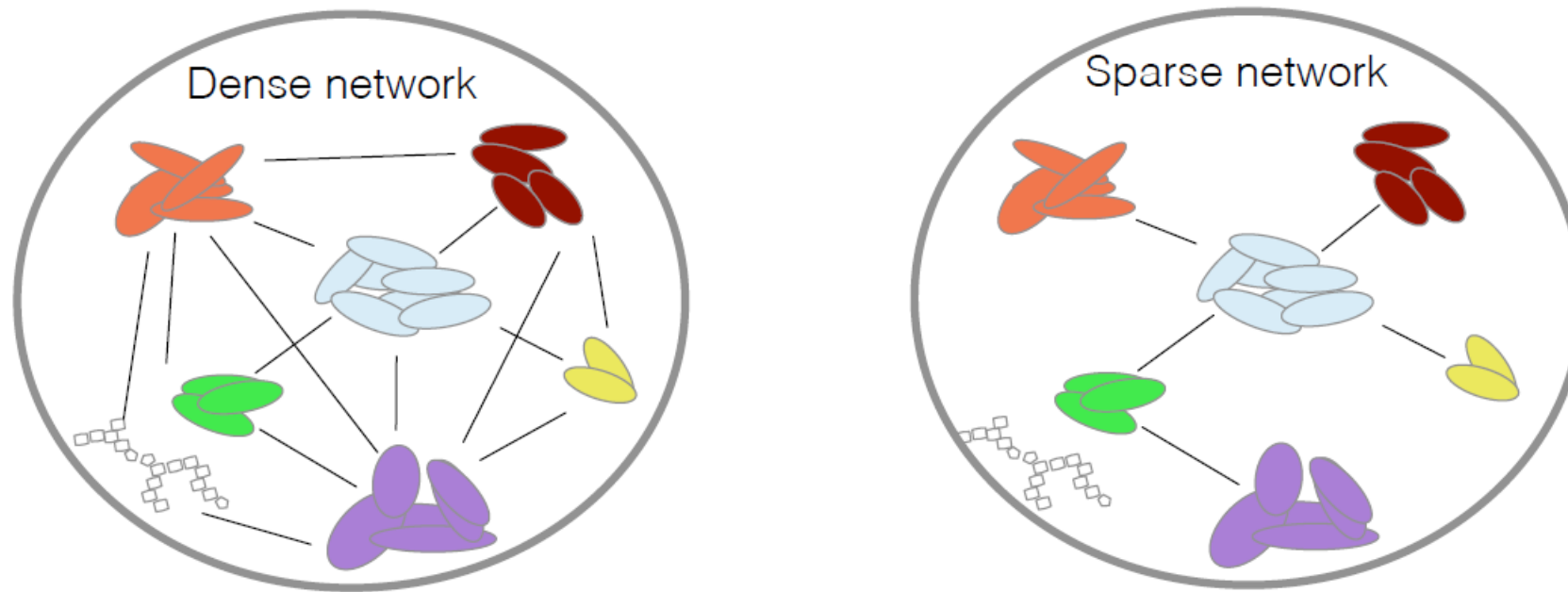
Key hypothesis for network inference

The inverse covariance matrix among transformed microbial relative abundances is **sparse**.



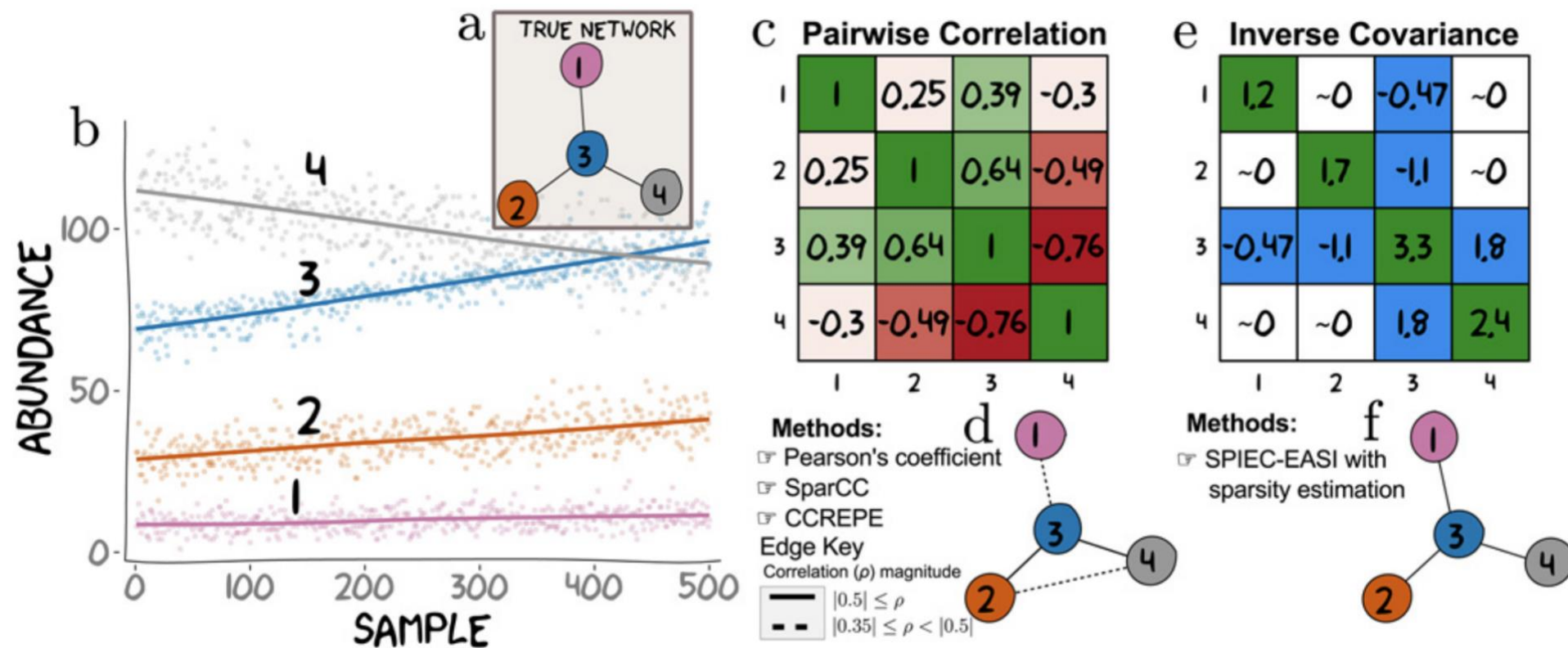
Key hypothesis for network inference

The network of interactions between the different microbes is **sparse**.



The SPIEC-EASI approach and R package

Conditional dependence vs. correlation



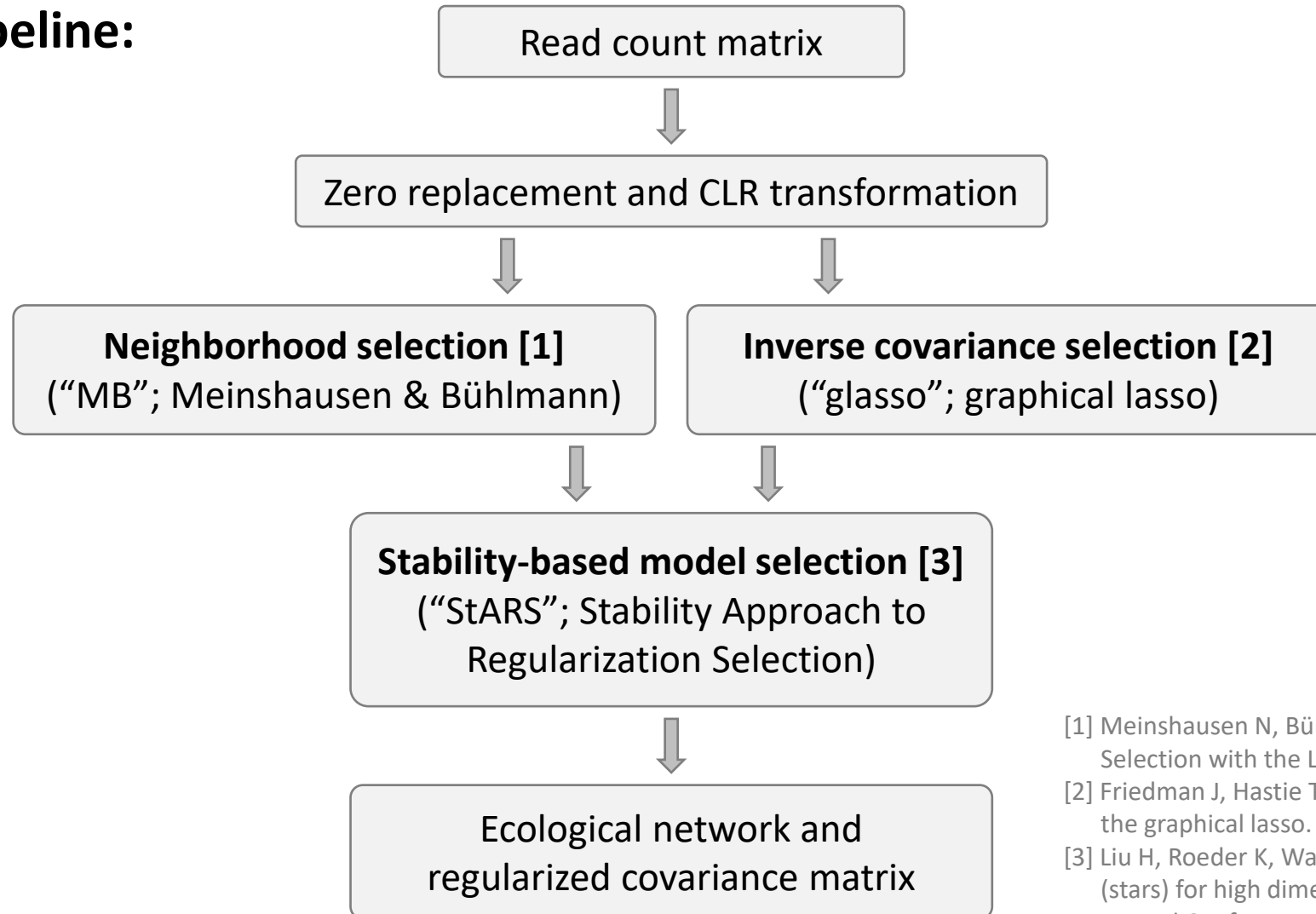
Conditional
dependence /
Partial correlation

Reference:
Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., & Bonneau, R. A. (2015). Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLoS Computational Biology, 11(5), 1–25.

SPIEC-EASI

(SParse InversE Covariance Estimation for Ecological Association Inference)

Pipeline:

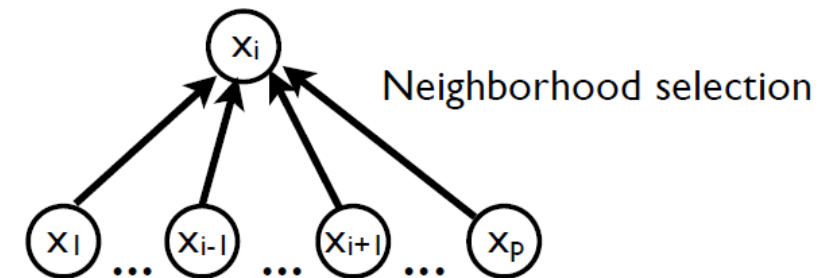


- [1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.
- [2] Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics (Oxford, England) 9: 432–441.
- [3] Liu H, Roeder K, Wasserman L (2010) Stability approach to regularization selection (stars) for high dimensional graphical models. Proceedings of the Twenty-Third Annual Conference on Neural Information Processing Systems (NIPS): 1–14.

Sparse neighborhood selection (MB)

- Uses sparse linear regression
- Proposed by Meinshausen and Bühlmann, 2006 (MB) [1]
- Idea: Find a **sparse weighted graph** by **node-wise linear regression**:
Use each column as response and solve the LASSO problem

$$\begin{array}{c} \text{OUT/gene } i \\ n \end{array} \left\{ \begin{array}{c} \text{all other OTUs/genes} \\ X^i \end{array} \right\} = \begin{array}{c} \text{all other OTUs/genes} \\ X^i \end{array} \times \begin{array}{c} \beta^* \\ p-1 \end{array} + \sigma \epsilon$$



[1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.

Sparse neighborhood selection (MB)

- Uses sparse linear regression
- Proposed by Meinshausen and Bühlmann, 2006 (MB) [1]
- Idea: Find a **sparse weighted graph** by **node-wise linear regression**:
Use each column as response and solve the LASSO problem

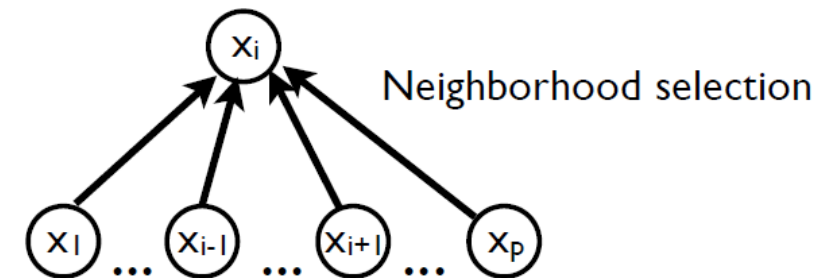
LASSO regression [2]:

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{\|Y - X\beta\|_2^2}{n} + \lambda \|\beta\|_1 \right\}$$

Likelihood term

Sparsity

$$n \begin{Bmatrix} \text{OUT/gene } i \\ X^i \end{Bmatrix} = \begin{Bmatrix} \text{all other OTUs/genes} \\ X^{-i} \end{Bmatrix} \times \begin{Bmatrix} \beta^* \end{Bmatrix}_{p-1} + \sigma \epsilon$$

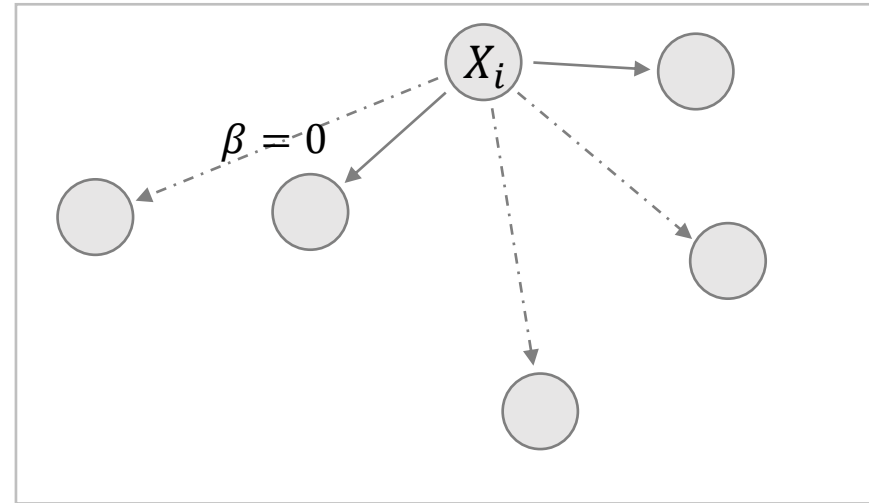


[1] Meinshausen N, Bühlmann P (2006) High Dimensional Graphs and Variable Selection with the Lasso. The Annals of Statistics 34: 1436–1462.

[2] Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Source: Journal of the Royal Statistical Society. Series B (Methodological), 58(1), 267–288.

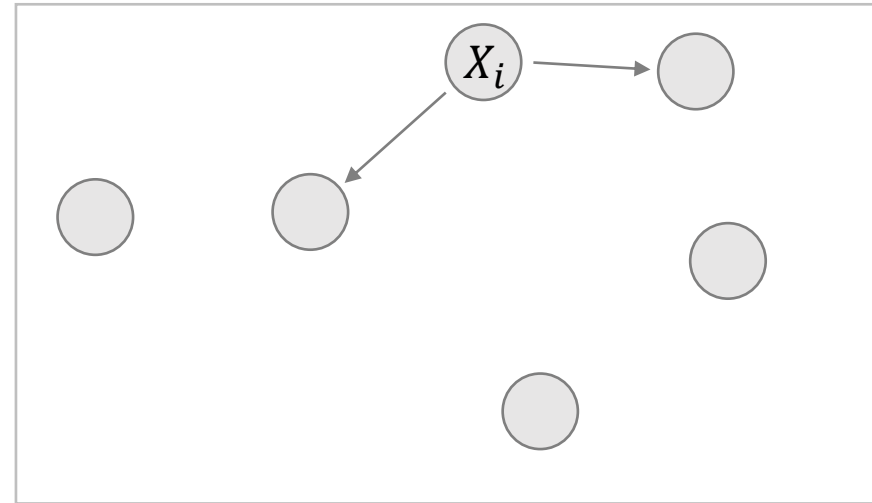
Sparse neighborhood selection (MB)

$$\begin{array}{c} \text{OUT/gene } i \\ n \end{array} \left\{ \begin{array}{c} \text{all other OTUs/genes} \\ X^{-i} \end{array} \right\} \times \left\{ \begin{array}{c} \beta^* \\ p-1 \end{array} \right\} + \sigma \epsilon$$



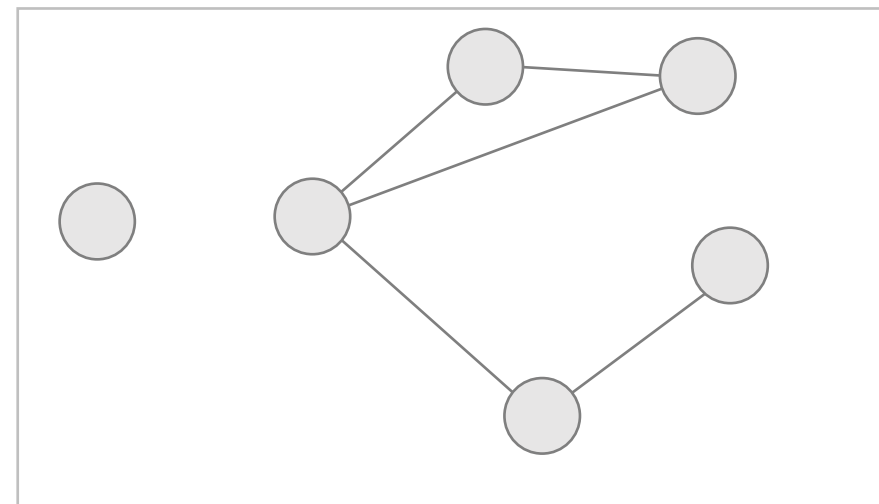
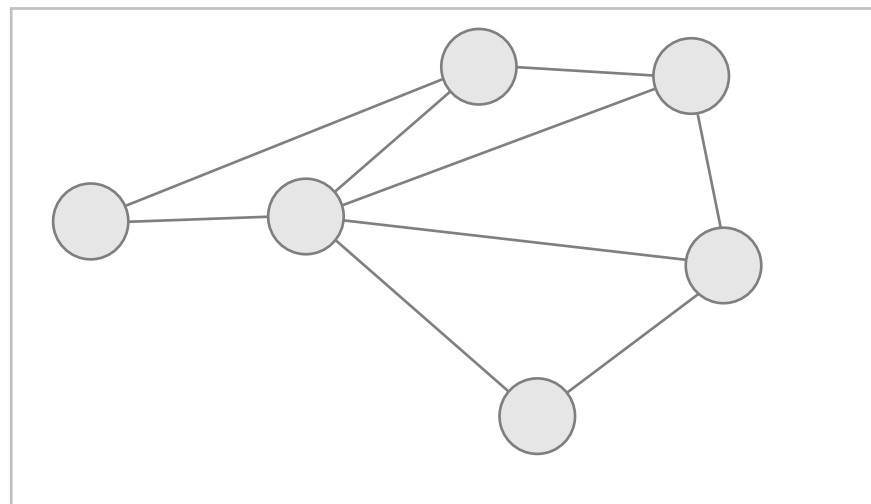
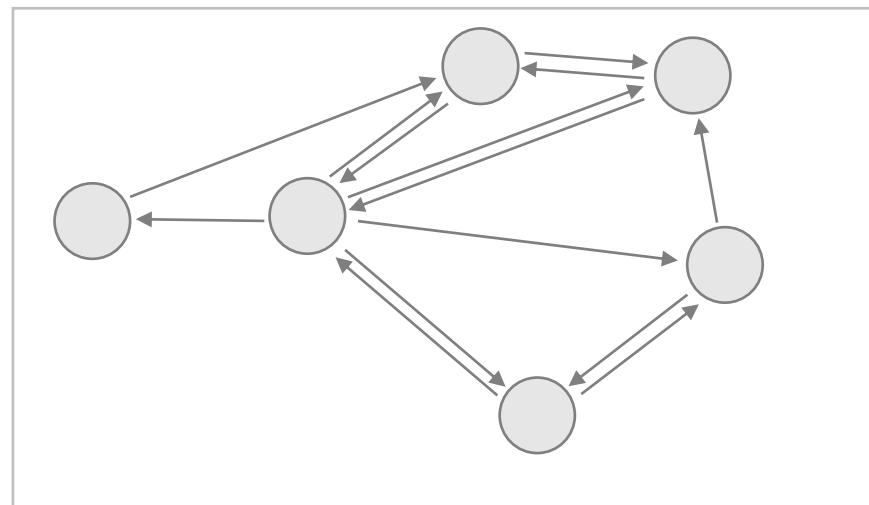
Sparse neighborhood selection (MB)

$$\begin{array}{c} \text{OUT/gene } i \\ n \end{array} \left\{ \begin{array}{c} \text{purple bar} \\ \mathbf{X}^i \end{array} \right\} = \begin{array}{c} \text{all other OTUs/genes} \\ \mathbf{X}^{-i} \end{array} \times \underbrace{\left\{ \begin{array}{c} \text{purple bar} \\ \beta^* \end{array} \right\}}_{p-1} + \sigma \epsilon$$



Sparse neighborhood selection (MB)

$$n \left\{ \begin{array}{c} \text{OUT/gene } i \\ \mathbf{X}^i \end{array} \right\} = \begin{array}{c} \text{all other OTUs/genes} \\ \mathbf{X}^{-i} \end{array} \times \underbrace{\begin{array}{c} \boxed{} \\ \beta^* \\ \boxed{} \end{array}}_{p-1} + \sigma \epsilon$$



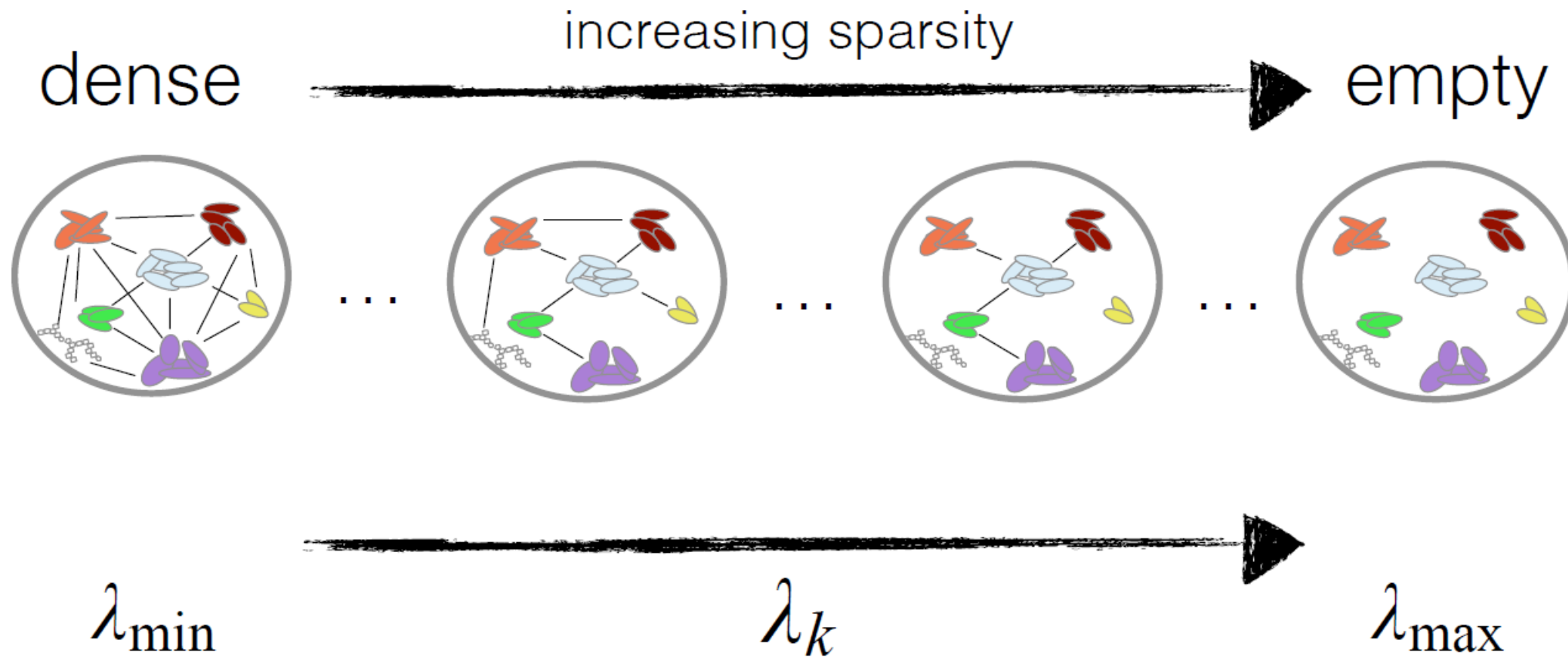
Sparse graphical model inference (GLASSO)

- Sparsity of the underlying network means that the **inverse** C^{-1} of the correlation (covariance) matrix C is **sparse**:
sparse Gaussian graphical model.
- Given: the sample correlation (covariance) matrix S
- Goal: Finding a sparse C^{-1} by convex optimization

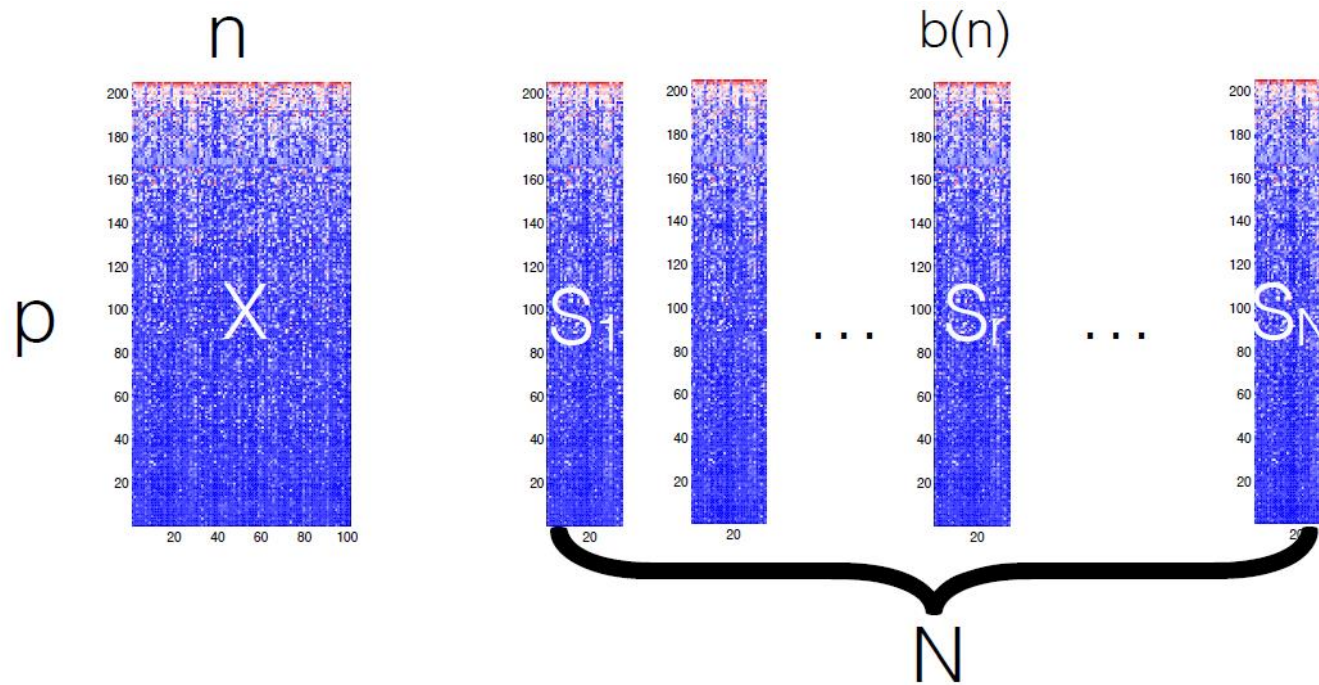
$$C^{-1} = \arg \min_{C^{-1} \in PD} -\log \det(C^{-1}) + \text{tr}(C^{-1} S) + \lambda \|C^{-1}\|_1$$

Likelihood term
Sparsity term

How to choose the tuning parameter λ ?



Stability-based model selection with StARS

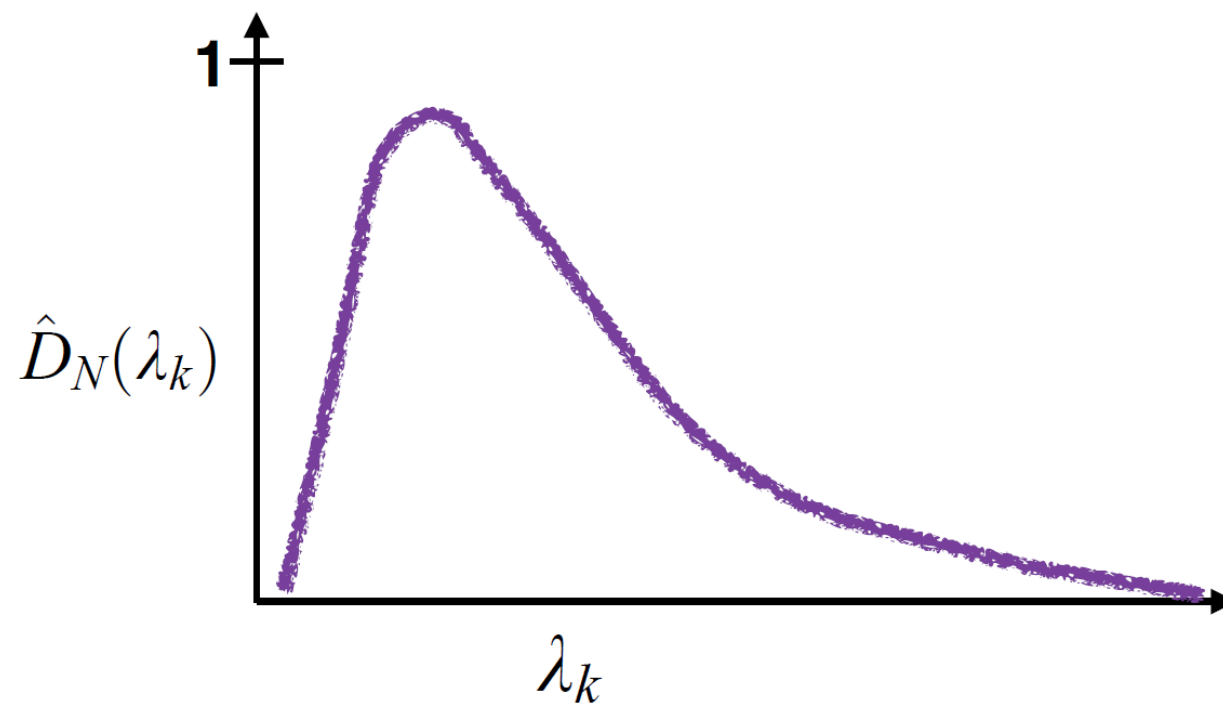
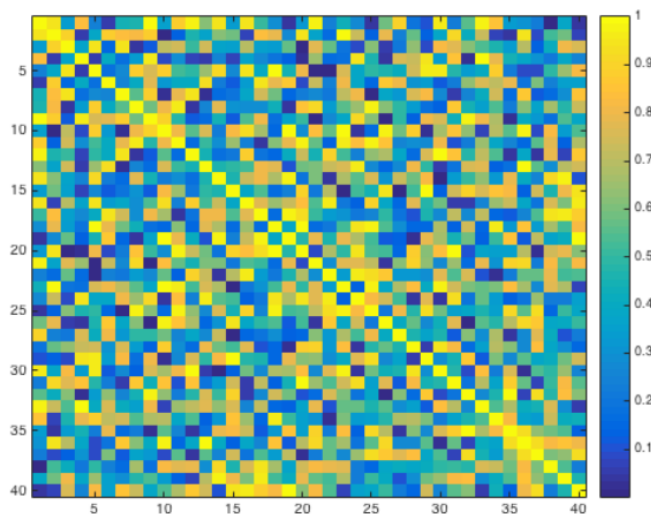


- Draw N random subsamples S_1, \dots, S_N
- For each subsample, estimate the graph using a sequence of λ values: $\{\lambda_1, \dots, \lambda_K\}$

Stability-based model selection with StARS

- For each lambda we estimate the matrix of edge probabilities
- For each lambda value sum up the **variances** of edge probabilities and call it $\hat{D}_N(\lambda_k)$
- Select the lambda where the sum of variances $\hat{D}_N(\lambda_k)$ is below a certain threshold β .

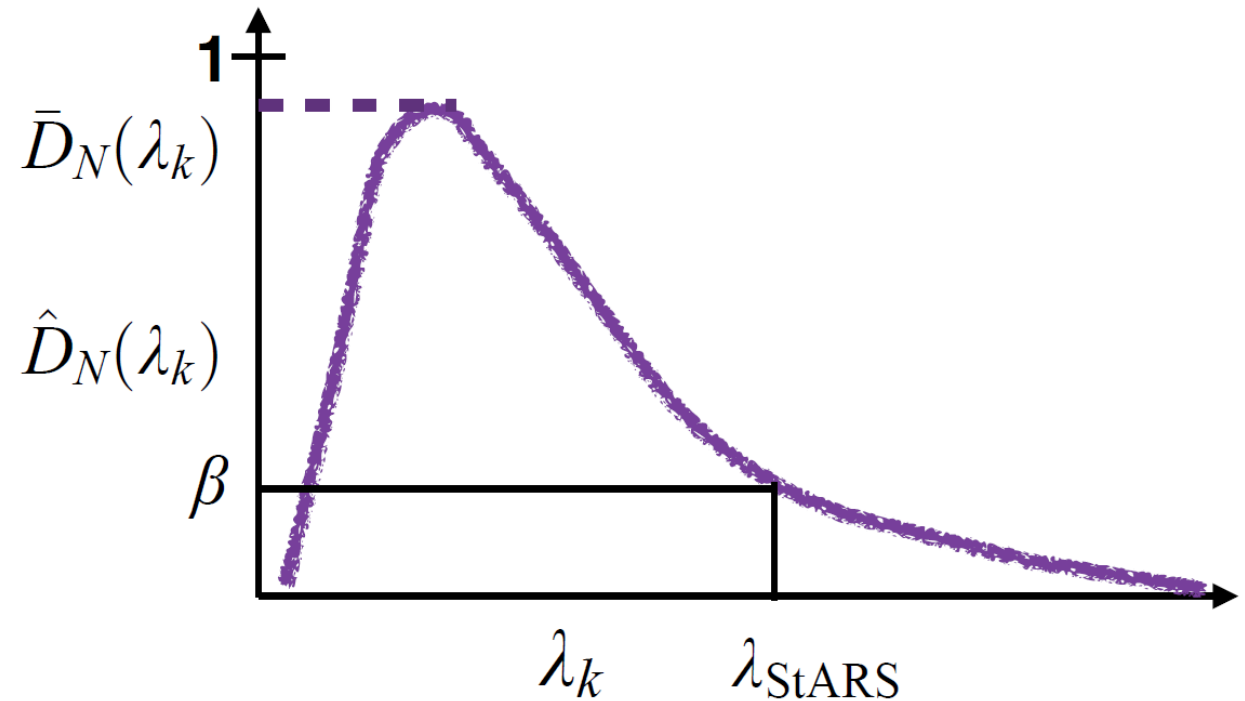
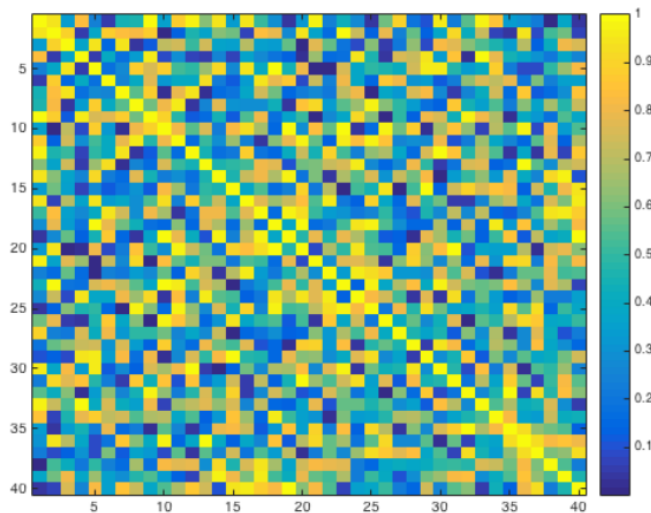
Matrix of edge probabilities



Stability-based model selection with StARS

- For each lambda we estimate the matrix of edge probabilities
- For each lambda value sum up the **variances** of edge probabilities and call it $\hat{D}_N(\lambda_k)$
- Select the lambda where the sum of variances $\hat{D}_N(\lambda_k)$ is below a certain threshold β .

Matrix of edge probabilities



Stability-based model selection with StARS

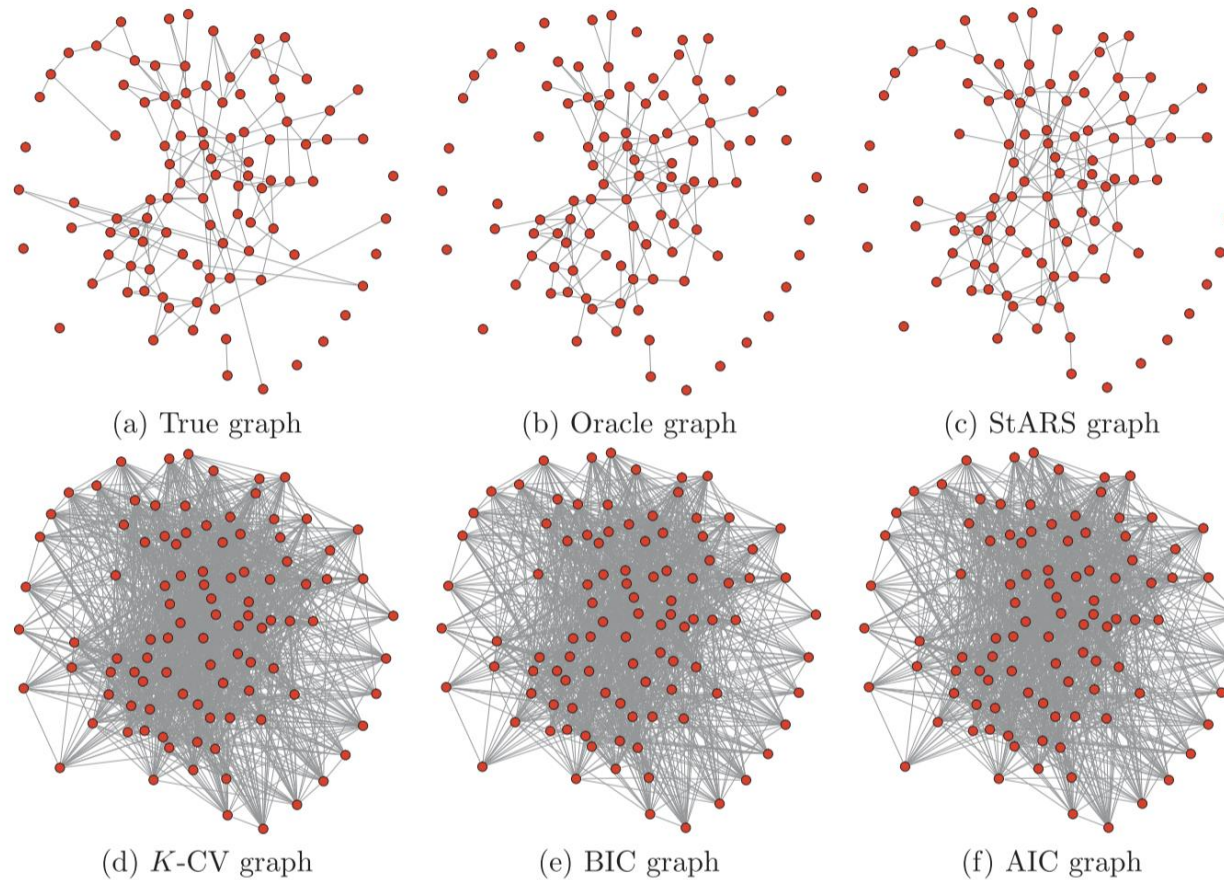
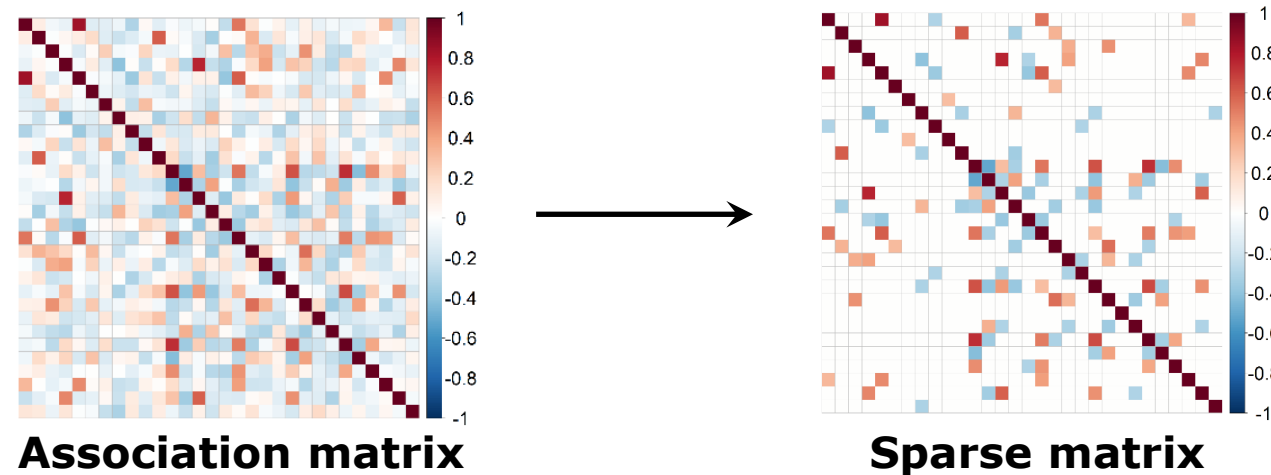


FIG 1. Comparison of different methods on the data from the neighborhood graphs ($n = 400, p = 100$).

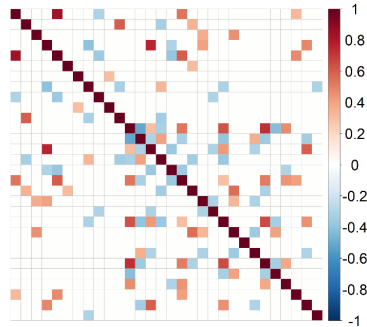
From associations to adjacencies

Further sparsification methods

- Thresholding
- Statistical tests
 - Applicable to other association measures, e.g., correlation
 - Student's t-test
 - Bootstrapping

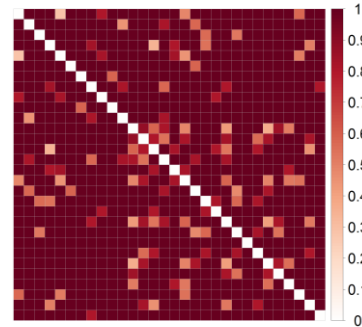


Inferring the adjacency matrix



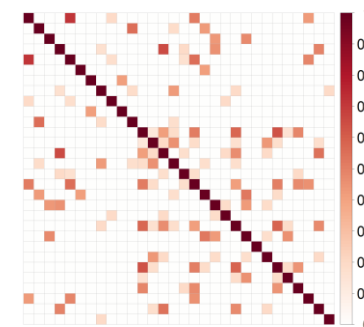
Sparse association matrix
with entries r_{ij} in $[-1,1]$

$$d_{ij} = ?$$



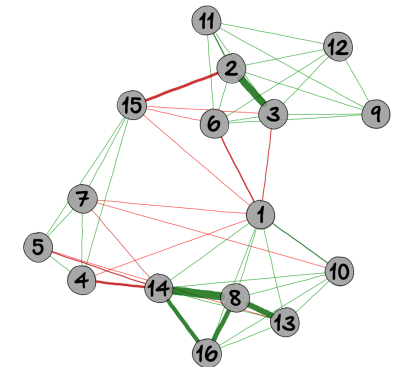
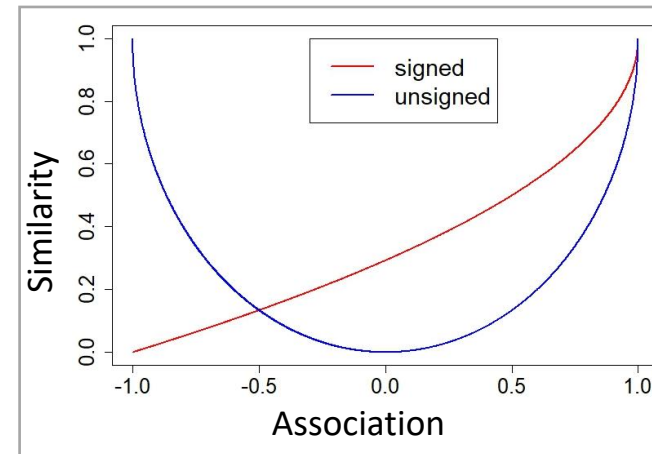
Dissimilarity matrix
(values in $[0,1]$)

$$s_{ij} = 1 - d_{ij}$$



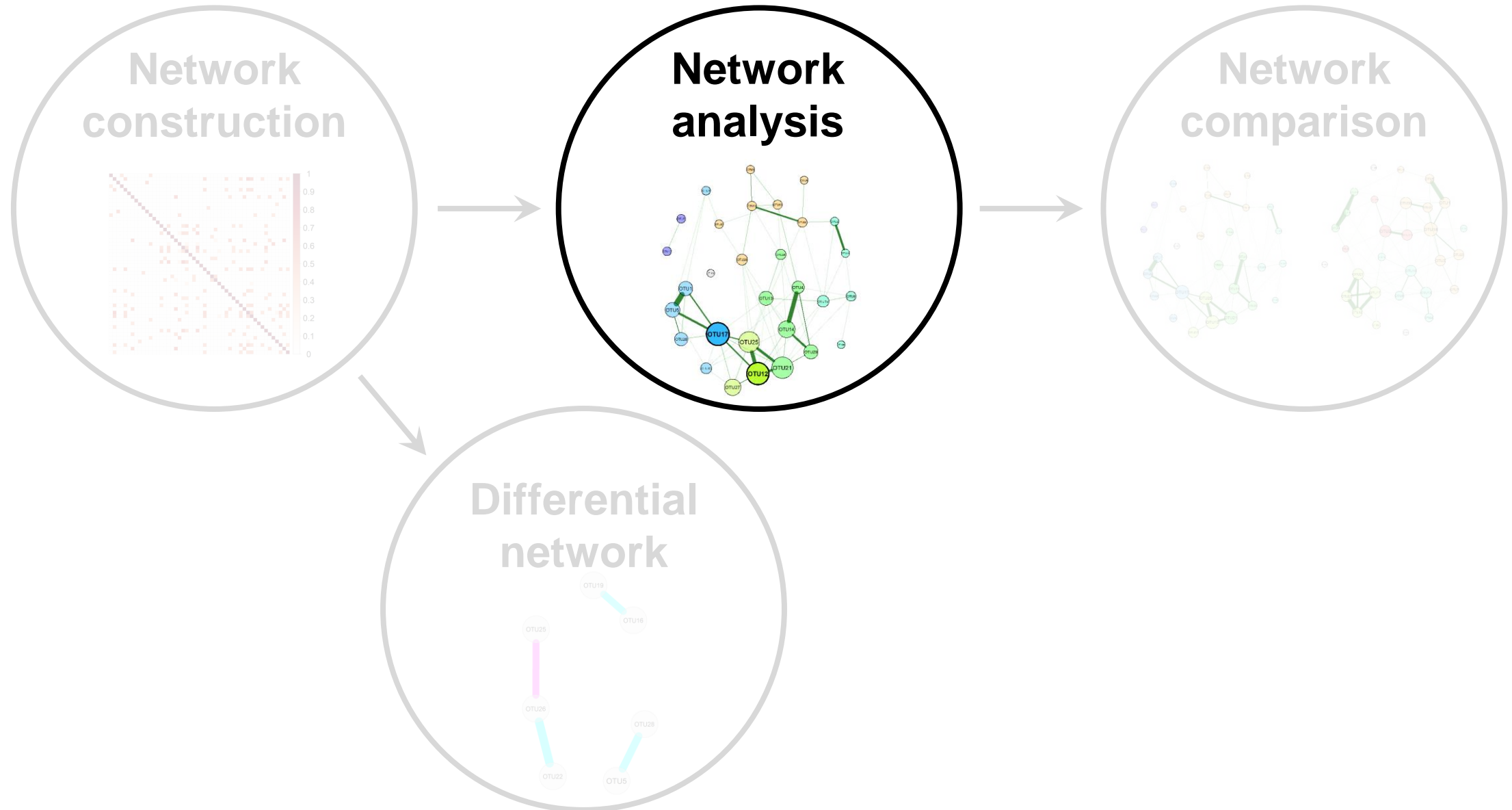
Similarity matrix
Adjacency matrix

- “unsigned”: $d_{ij} = \sqrt{1 - r_{ij}^2}$
→ low distance between strongly associated taxa
(positively as well as negatively)
- “signed”: $d_{ij} = \sqrt{0.5(1 - r_{ij})}$
→ distance is high for strongly negative associated taxa



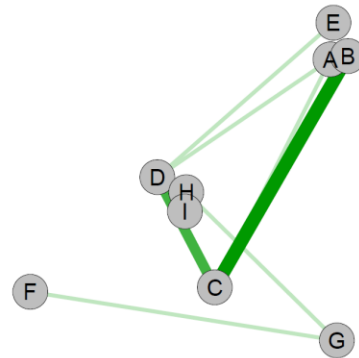
Network analysis

Typical network analysis workflow

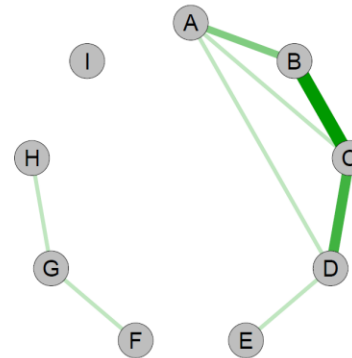


Layout

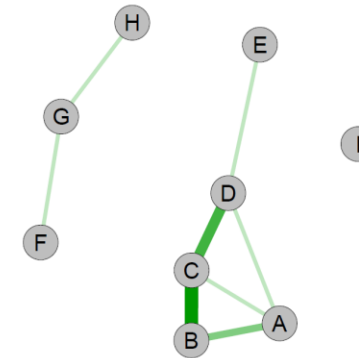
How to place nodes in the two-dimensional space?



random layout



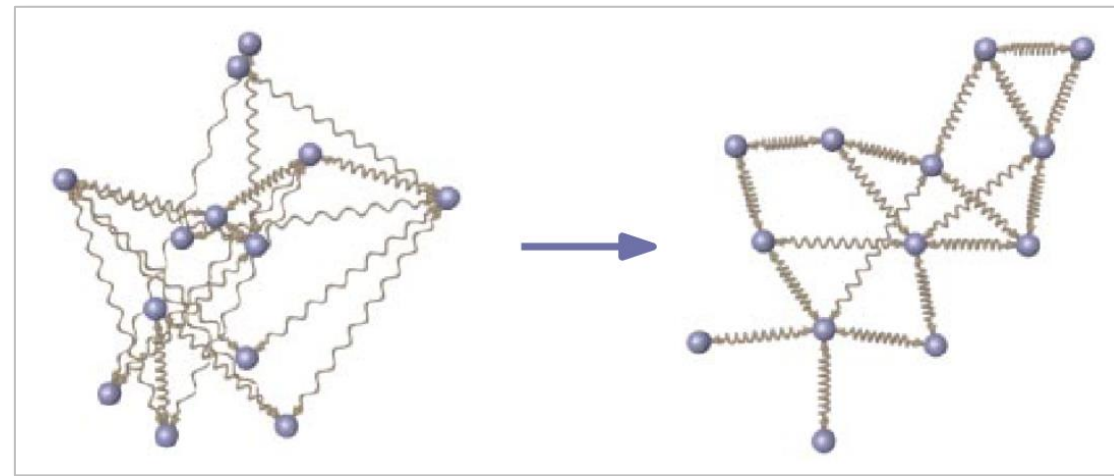
circular layout



force directed layout

Force-directed layout algorithm:

- Based on physical concepts: mechanical springs or electric repulsion
- E.g., Fruchterman-Reingold Layout



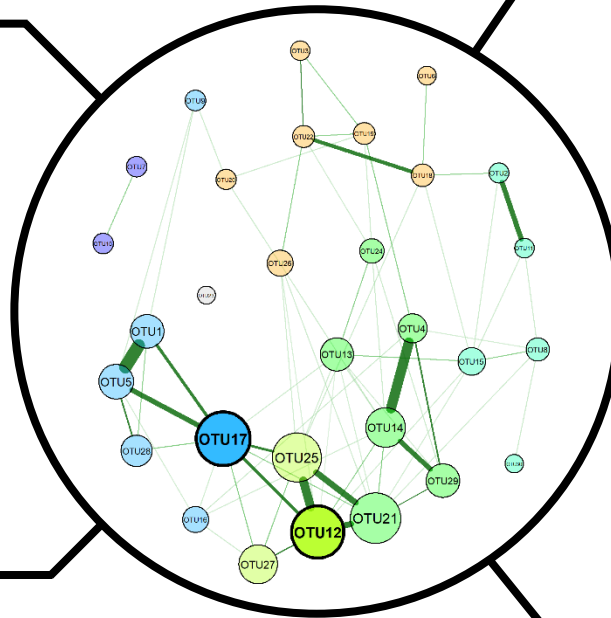
Network analysis

Centrality measures

Degree
Betweenness centrality
Closeness centrality
Eigenvector centrality

Clustering

Hierarchical clustering
Modularity Clustering
Fast greedy modularity optimization
Clustering based on edge betweenness



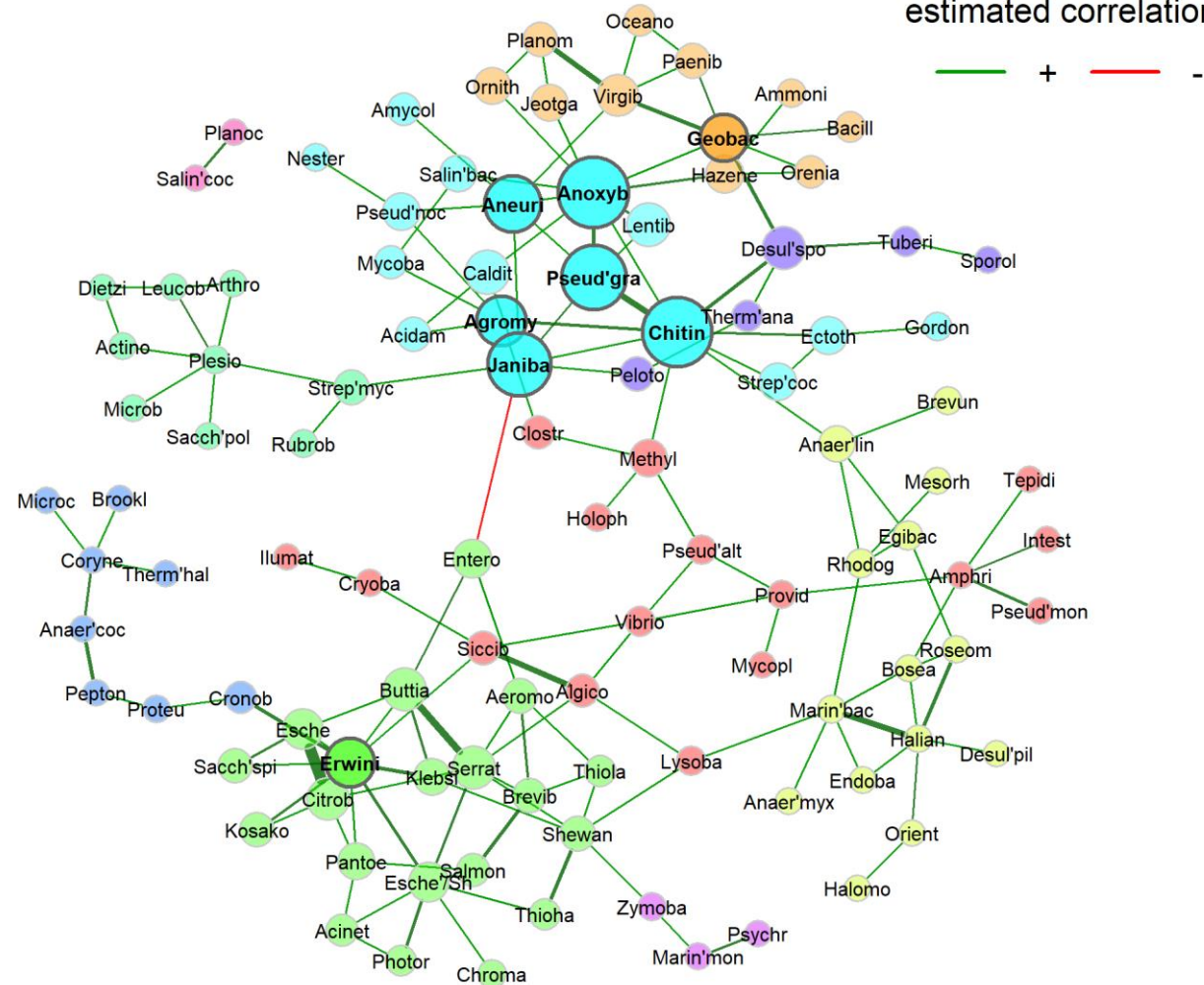
Global network properties

Modularity
Global clustering coefficient
Average path length
Average dissimilarity
Positive edge percentage
Edge density
Natural connectivity
Edge / vertex connectivity
Number of components

Hubs

Most central nodes
(regarding one or more
centrality values)

estimated correlation:



Visualization

Data features highlighted

Phylum:

- Actinobacteria
- Bacteroidetes
- Chloroflexi
- Fibrobacteres
- Firmicutes
- NA
- Proteobacteria
- Tenericutes

estimated correlation:

— + — -

