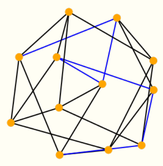
Reconstruction quality of a biological network when its constituting elements are partially observed

Victor Picheny, Matthieu Vignes, Nathalie Villa-Vialaneix
INRA, UR875 MIA-T, France & Massey University, IFS, NZ -
firstname.lastname@toulouse.inra.fr

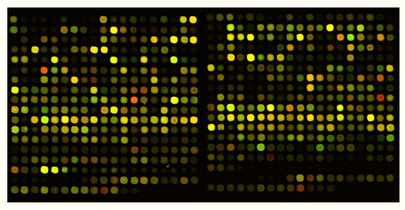
Application framework: impact of selecting genes on network inference

Ideal network inference



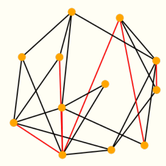
true biological network
(unknown)

$\longrightarrow$ observations



transcriptomic data

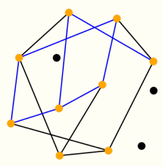
$\longrightarrow$ estimation



predicted network
(from full data)

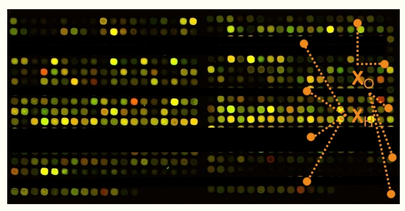
red edges:
false positive

Inference with missing nodes



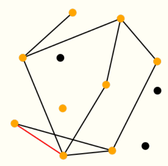
projected biological network
with gene selection

$\longrightarrow$ observations
(incomplete)



transcriptomic data

$\longrightarrow$ estimation



??? predicted network ???
(from partial data)

blue edges:
false negative

Research questions: Evaluate the impact of gene sampling on the estimation of network

Theoretical framework: Gaussian Graphical Model framework

Gene expression: $X \sim \mathcal{N}(0, \Sigma)$, sample size: n, number of genes: p and $X = (X_O, X_H)$, with X_H **not observed**.

non-zero entries of $S = \Sigma^{-1} \Leftrightarrow$ edges of full graph

Questions:

Influence of:

- ratio of missing variables r
- missing node context: random or peculiar nodes (*e.g.* big/small degree or large/small betweenness)

How to estimate the errors?

- compare to a graph whose links reflect path existence in the "true" graph (*induced*)
- compare to a graph inherited from edges of the "true" graph only (*projected*)
- compare to a graph learnt from complete data

Method 1 (naive approach)

graphical Lasso [1] on observed data

$$\Sigma_{OO}^{-1} = \underbrace{S_{OO}}_{\text{to be estimated}} - \underbrace{S_{OH}(S_{HH})^{-1}S_{HO}}_{\text{biais}}$$

Method 2: CPW-S+L [2]

Question of *identifiability* of the 2 components of Σ_{OO}^{-1}:

- **sparse** S_{OO} and
- **low-rank** $S_{OO}(S_{HH})^{-1}S_{HO}$

$\rightarrow$ *via* an algebraic study of sparse and low-rank matrix varieties.
More specifically: **transversality** of tangent spaces $T_*(S_{OO})$ and $T_\star(S_{OH}(S_{HH})^{-1}S_{HO}) \leftrightarrows$ statistical identifiability.
Assumptions:

- sparsity = few non-zeros per column/row $\leftrightarrows$ no dense subgraph.
- $S_{OH}(S_{HH})^{-1}S_{HO}$ has row/column spaces not too aligned with coordinate axes $\leftrightarrows$ marginalisation effect over X_H's is "spread out" over many X_O's.

penalised likelihood method leads to consistent estimate *via*:

$$(S_{OO}, \widehat{S_{OH}(S_{HH})^{-1}S_{HO}}) = \underset{(S,L), S-L\succ 0, L\succeq 0}{\operatorname{argmin}} -l(S-L, \Sigma_{OO}) + \lambda\left[\gamma\|S\|_{l_1} + tr(L)\right]$$

with $l(S, \Sigma) = \log\det(S) - \mathrm{tr}(S\Sigma) + c$, the GGM log-likelihood.

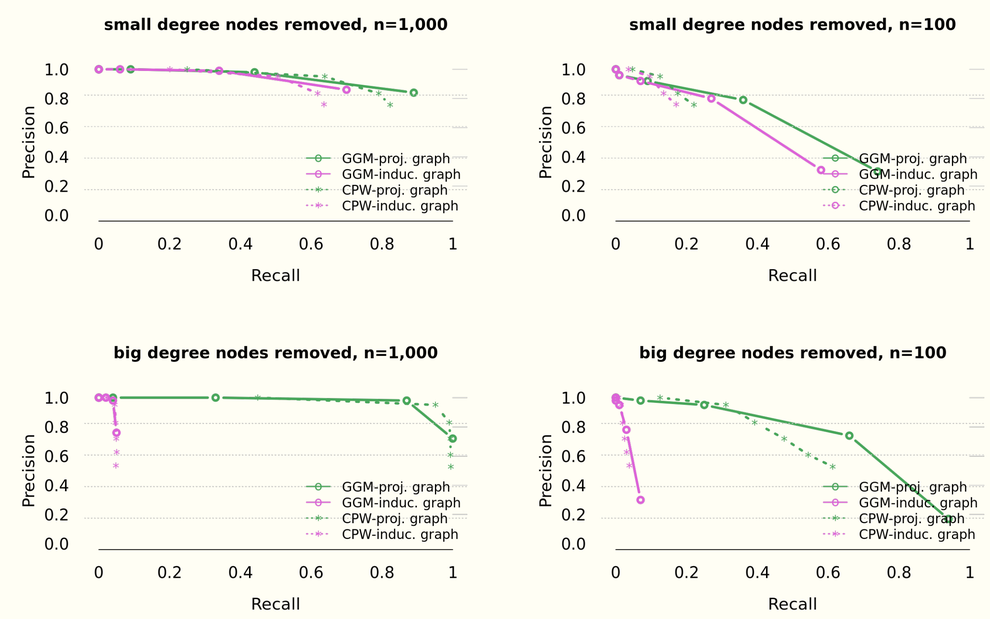
Experimental setup

Tests on simulated data sets:

- data simulated according to a **GGM** with $\mathbf{p = 100}$ genes
- sample size: $\mathbf{n = 100}$ and $\mathbf{1,000}$
- ratio of missing variables: $r = 0$ (full graph), 5%, 10%, $\mathbf{20}\%$ and 30%
- missing node "context": with **large/low degree**, high/low or at random.

(10 replicate networks)

Selected results: precision vs. recall curves



(precision $= \frac{TP}{TP+FP}$, recall $= \frac{TP}{TP+FN}$)

References

[1] J. Friedman, T. Hastie, and R. Tibshirani (2007) Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432-441.
[2] V. Chandrasekaran, P.A. Parillo, and A.S. Willsky (2012). Latent variable graphical model selection via convex optimization. *Annals of Statistics* **40**:1935–1967.