Introduction to Sparse Gaussian graphical models for biological network inference

Cartable

Julien Chiquet

MIA Paris

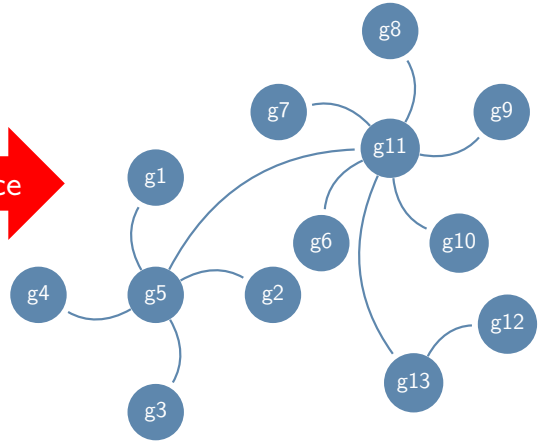
October the 13, 2016

A challenging problem

≈ 10s/1,000s assays

≈ 1,000s/1,000,000s features

Inference



1. Nodes are fixed
   - restricted to a set of interest
2. Edges (interactions) are inferred
   - based upon statistical concepts

Main statistical challenges

1. (Ultra) High dimensionality ($n < p$, $n \lll p$)
2. Heterogeneity/structure of the data

Exploratory research

By pointing important actors (genes, OTU), it may assist the biologist in

1. formulating a hypothesis for further experiments,
2. unraveling main tendencies at play in complex systems.

Outline

Canonical framework: sparse GGM

Accounting for some biological features

Network inference for enhancing other methods

Outline

Canonical framework: sparse GGM

Accounting for some biological features

Network inference for enhancing other methods

Gaussian Graphical Model: canonical settings

Biological experiments in comparable Gaussian conditions

Profiles of a set $\mathcal{P} = \{1, \ldots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

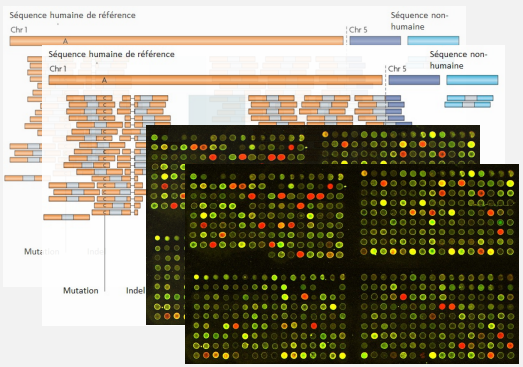
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \ldots, X^n)$ of exp. stacked in an $n \times p$ data matrix $\mathbf{X}$.

Conditional independence structure

$$(i,j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\setminus\{i,j\}} \Leftrightarrow \rho_{ij|\setminus\{i,j\}} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} = 0.$$

The data

Stacking $(X^1, \ldots, X^n)$, we met the usual individual/variable table $\mathbf{X}$



stacked in

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \ldots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \ldots & x_n^p \end{pmatrix}$$

Gaussian Graphical Model: canonical settings

Biological experiments in comparable Gaussian conditions

Profiles of a set $\mathcal{P} = \{1, \ldots, p\}$ of genes is described by $X \in \mathbb{R}^p$ such as

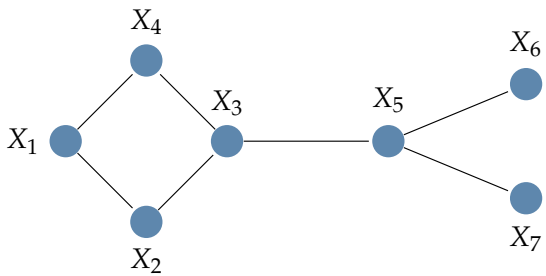
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \ldots, X^n)$ of exp. stacked in an $n \times p$ data matrix $\mathbf{X}$.

Conditional independence structure

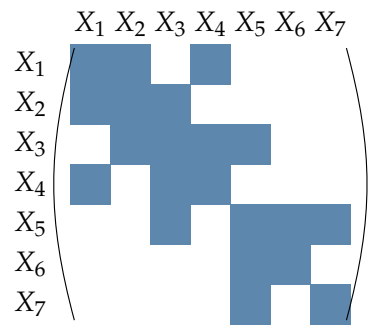
$$(i,j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\setminus\{i,j\}} \Leftrightarrow \rho_{ij|\setminus\{i,j\}} = -\frac{\Theta_{ij}}{\sqrt{\Theta_{ii}\Theta_{jj}}} = 0.$$

Graphical interpretation

$$\mathcal{G} = (\mathcal{P}, \mathcal{E})$$



$\boldsymbol{\Theta}$



$\rightsquigarrow$ "Covariance" selection

Existing inference approach I

Limited-order partial correlations

Partial order correlation

For some sets $\mathcal{U}$ with $|\mathcal{U}| \leq q$ and $\mathcal{V} = \mathcal{U} \cup \{i, j\}$, the q-order partial correlation, for $q \in \{0, \ldots, p-2\}$, is

$$\rho_{ij|\mathcal{U}} = -\frac{\Theta^{\mathcal{V}}_{ij}}{\sqrt{\Theta^{\mathcal{V}}_{ii}\Theta^{\mathcal{V}}_{jj}}} \quad \text{where } \boldsymbol{\Theta}^{\mathcal{V}} = (\boldsymbol{\Sigma}_{\mathcal{V}\mathcal{V}})^{-1}.$$

Basic procedure

- test the hypotheses $\rho_{ij|\mathcal{U}} = 0$ for *every* $\mathcal{U}$ such that $|\mathcal{U}| = q$,
- $i \leftrightarrow j \in \mathcal{G}$ iff all hypotheses are rejected.

Developments: Wille and Buhlmann (2006); Castelo and Roverato (2006); Verzelen, Villers (2008) ...

Existing inference approach I

Limited-order partial correlations

Partial order correlation

For some sets $\mathcal{U}$ with $|\mathcal{U}| \leq q$ and $\mathcal{V} = \mathcal{U} \cup \{i, j\}$, the q-order partial correlation, for $q \in \{0, \ldots, p-2\}$, is

limitations

- Computationally expansive (C_q^{p-2} tests + mat. inversion).
- Remains an approximation of the true graph
- Need mutiple-test correction
- Not adapted to high-dimensional data

Basic procedure

- test the hypotheses $\rho_{ij|\mathcal{U}} = 0$ for *every* $\mathcal{U}$ such that $|\mathcal{U}| = q$,
- $i \leftrightarrow j \in \mathcal{G}$ iff all hypotheses are rejected.

Developments: Wille and Buhlmann (2006); Castelo and Roverato (2006); Verzelen, Villers (2008) ...

Existing inference approach II

Bayesian GGM

For $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ be the conditional graph associated to $\mathbf{X} \sim \mathcal{N}(0, \mathbf{\Theta}^{-1})$,

$$\mathbb{P}(\mathcal{G}, \mathbf{\Theta}|\mathbf{X}) \propto \mathbb{P}(\mathbf{X}|\mathcal{G}, \mathbf{\Theta})\mathbb{P}(\mathbf{\Theta}|\mathcal{G})\mathbb{P}(\mathcal{G})$$

with $\mathbb{P}(\mathbf{X}|\mathcal{G}, \mathbf{\Theta})$ the Gaussian multivariate likelihood.

Priors

- Uniform distribution over a set $\mathcal{G}_S$ or truncated Poisson

$$\mathbb{P}(\mathcal{G}) = \frac{1}{|\mathcal{G}_S|}, \quad \mathbb{P}(\mathcal{G}) \propto \frac{\gamma^{|\mathcal{E}|}}{|\mathcal{E}|!}.$$

- $\mathcal{G}$-Wishart over the space $\mathbb{P}_\mathcal{G}$ of p.d matrices with same support as $\mathcal{G}$

$$\mathbb{P}(\mathbf{\Theta}|\mathcal{G}) = \frac{1}{Z(\mathbf{T})}|\mathbf{\Theta}|^{(d-2)/2} \exp\left\{\frac{1}{2}\mathrm{tr}(\mathbf{T}\mathbf{\Theta})\right\}.$$

  Z is computed by MCMC schemes.

$\rightsquigarrow$ See Loïc's work tomorrow

Existing inference approach III

Regularization/penalized likelihood approach

Let $\mathbf{\Theta}$ be the model parameter to infer (related to the edges).

Constraint optimization approach

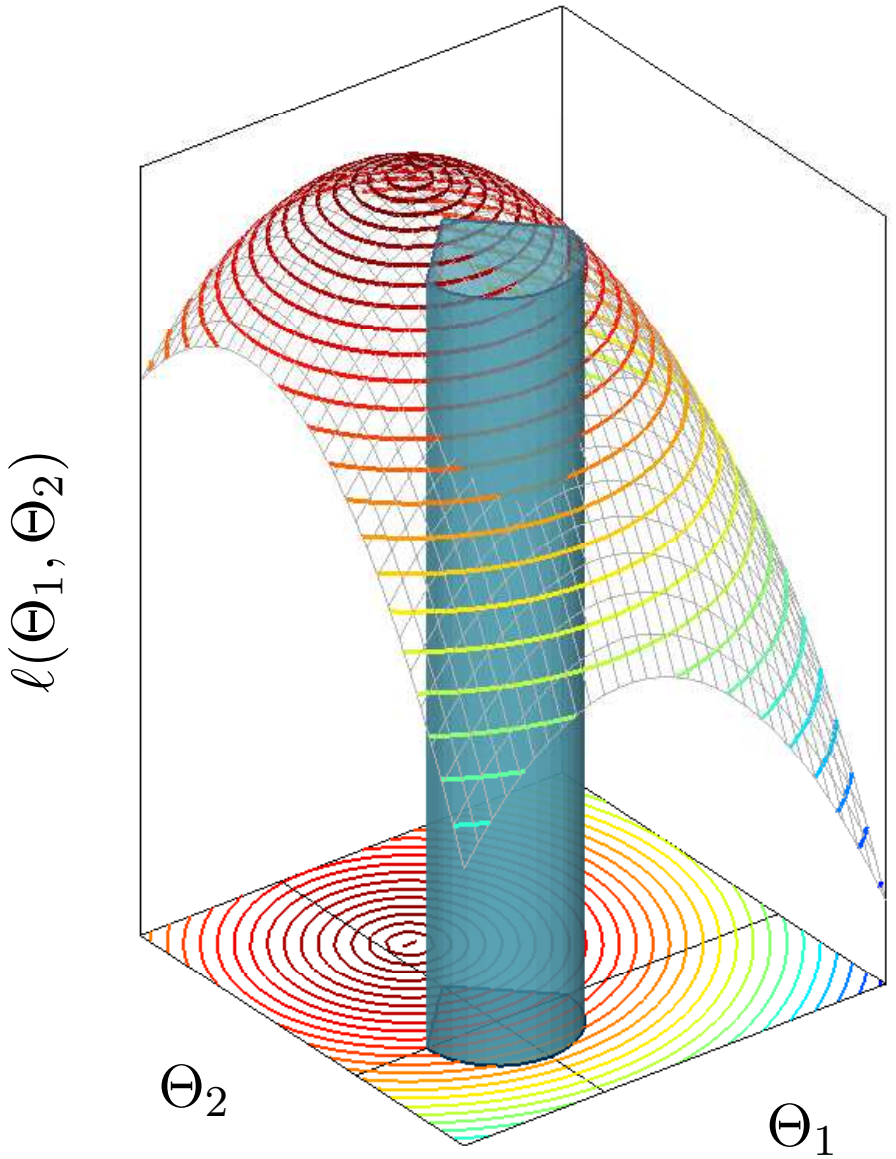
$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}} \log \ell(\mathbf{\Theta}; \mathbf{X}) \quad \text{s.t.} \quad \Omega(\mathbf{\Theta}) \leq c$$

Convex optimization approach

$$\hat{\mathbf{\Theta}}_\lambda = \arg\min_{\mathbf{\Theta}} - \log \ell(\mathbf{\Theta}; \mathbf{X}) \; + \; \lambda \; \mathrm{pen}_{\ell_1}(\mathbf{\Theta}),$$

- $\log \ell$ is the model log-likelihood,
- Ω and c define a feasible set.
- pen is a penalty function controlled by λ.

A geometric view of sparsity



$$\begin{cases} \underset{\Theta_1,\Theta_2}{\text{maximize}} & \ell(\Theta_1, \Theta_2) \\ \text{s.t.} & \Omega(\Theta_1, \Theta_2) \leq c \end{cases}$$

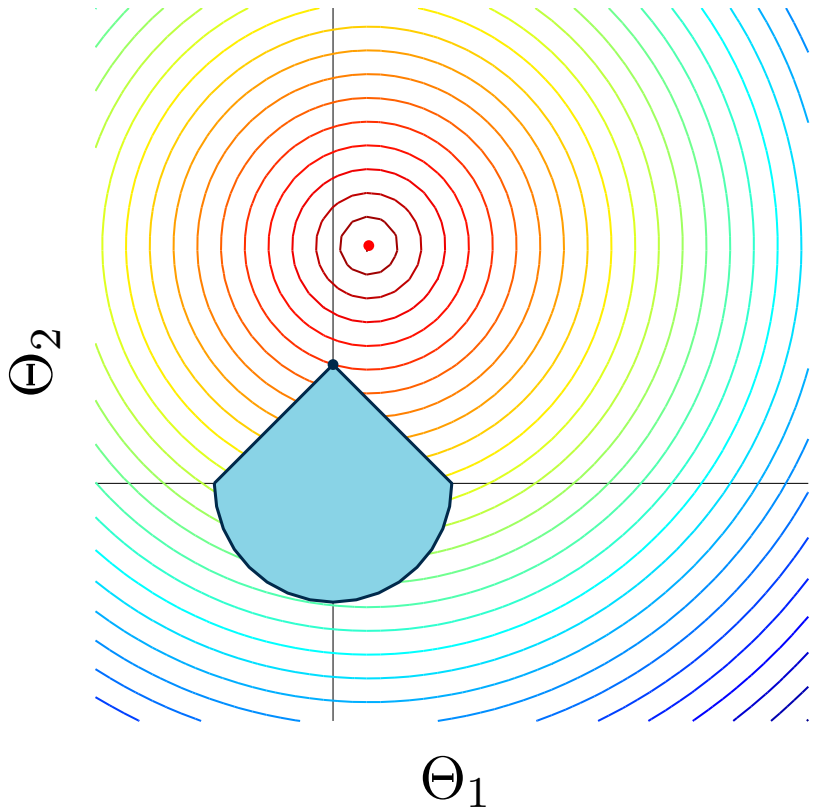
$$\Updownarrow$$

$$\underset{\Theta_1,\Theta_2}{\text{minimize}} -\ell(\Theta_1, \Theta_2) + \lambda\Omega(\Theta_1, \Theta_2)$$

$\Omega \equiv \text{pen}_{\ell_1}$ is a penalty tuned by $\lambda > 0$.
It performs

1. *regularization* ($n \ll p$),
2. *selection* (induced by ℓ_1),
3. *can be seen* as a log-prior on $\mathbf{\Theta}$.

A geometric view of sparsity



$$\begin{cases} \underset{\Theta_1,\Theta_2}{\text{maximize}} & \ell(\Theta_1,\Theta_2) \\ \text{s.t.} & \Omega(\Theta_1,\Theta_2) \leq c \end{cases}$$

$$\Updownarrow$$

$$\underset{\Theta_1,\Theta_2}{\text{minimize}} -\ell(\Theta_1,\Theta_2) + \lambda\Omega(\Theta_1,\Theta_2)$$

$\Omega \equiv \text{pen}_{\ell_1}$ is a penalty tuned by $\lambda > 0$.
It performs

1. *regularization* ($n \ll p$),
2. *selection* (induced by ℓ_1),
3. *can be seen* as a log-prior on $\boldsymbol{\Theta}$.

Gold standard penalized approach

Use ℓ_1 for both regularizing and promoting *sparsity*

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta} \in \mathbb{S}_+} \ell(\mathbf{\Theta}; \mathbf{X}) - \lambda \|\mathbf{\Theta}\|_1$$

- symmetric, positive-definite
- solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).
- `R` packages `huge`, `QUIC`, `fastclime`, `flare`, ...

Extensions to non-Gaussian case

- Simple transformation: often surprisingly efficient
  $\rightsquigarrow \log(1 + \mathbf{X}), \sqrt{\mathbf{X}}$, compute Spearman's correlation
- Non-paranormal transformation (Liu et al 2009)
  $\rightsquigarrow$ copula
- Poisson models (Allen et al, Gallopin et al.)

Properties

Theoretical results

- Selection consistency (Ravikumar et al. 2009-'12). For an "appropriate" λ,

$$n \approx \mathcal{O}(d^2 \log(p)) \text{ with } d = \max_{j \in \mathcal{P}} (\text{degree}_j)$$

- Ultra high-dimension phenomenon (Verzelen, 2011). Occur when

$$\frac{d \log(p/d)}{n} \geq 1/2, \qquad (\text{e.g., } n = 50, p = 200, d \geq 8).$$

Computational capability ('14 NIPS submissions)

- Solve GLASSO/CLIME for $p = 10^6$ (on 400 cores).
- based on alternating direction method of multipliers (ADMM)
- $+$ *many* tricks

Model selection: what λ?

Cross-validation

Optimal in terms of prediction, not in terms of selection

Information based criteria

- GGMSelect (Girault *et al*, '12) selects among a family of candidates.
- Adapt IC to sparse high dimensional problems, e.g.

$$\text{EBIC}_\gamma(\widehat{\mathbf{\Theta}}_\lambda) = -2\text{loglik}(\widehat{\mathbf{\Theta}}_\lambda; \mathbf{X}) + |\mathcal{E}_\lambda|(\log(n) + 4\gamma \log(p)),$$

Resampling/subsampling

Keep edges frequently selected on an range of λ after sub-samplings

- Stability Selection (Meinshausen and Bühlman, 2010, Bach 2008)
- Stability approach to Regularization Selection (StaRS) (Liu, 2010).

Limitations towards biological network inference

- Sparse GGM
  - \+ very solid statistical and computational framework
  - \+ extend to non strictly normal distribution (NGS)

- Guillem's talk + DREAM challenge
  - \+ competitive to other inference methods
  - − performances remain questionable on real data, as for other methods

Ideas

Strengthen the inference by accounting for biological features

1. structure of the network (organization of biological mechanisms)
2. sample heterogeneity (structure of the population)
3. horizontal integration (use multiple data and platforms)

Outline

Canonical framework: sparse GGM

Accounting for some biological features

- Accounting for latent organisation of the network
- Accouting for sample heterogeneity
- Accouting for multiscale data with multiattribute models

Network inference for enhancing other methods

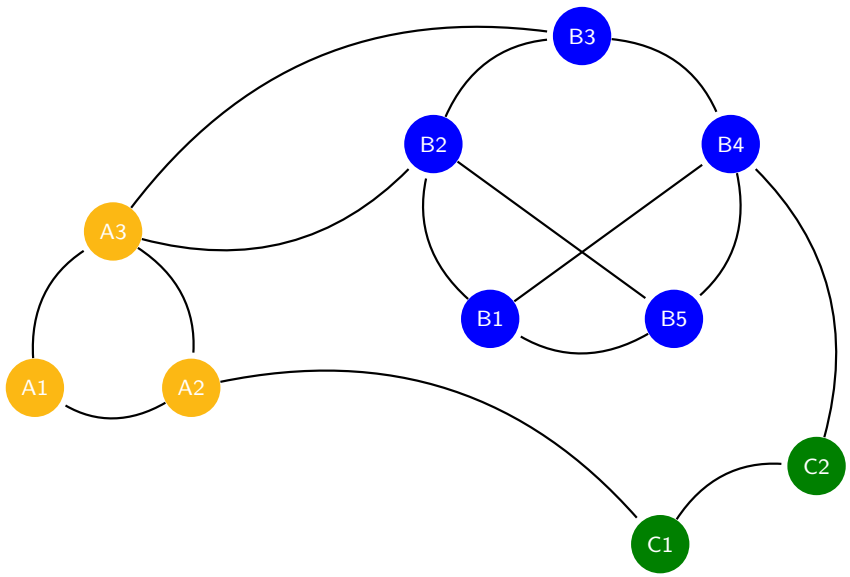
Outline

Handling with the data structure and scarcity

By introducing some prior

Priors should be biologically grounded

1. not too many genes effectively interact: sparsity,
2. networks are organized: latent clustering.



Structured regularization

SIMoNe: Statistical Inference for MOdular NEtworks

$$\arg\max_{\mathbf{\Theta},\mathbf{Z}} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda \, \|\mathbf{P}_{\mathbf{Z}} \star \mathbf{\Theta}\|_{\ell_1},$$

where $\mathbf{P}_{\mathbf{Z}}$ is a matrix of weights depending on a underlying latent structure $\mathbf{Z}$ (depicted through a stochastic block model).

$\rightsquigarrow$ Cluster-driven inference via an EM-like strategy.


Ambroise, Chiquet, Matias. Inferring sparse GGM with latent structure, EJS, 2009.


Marlin, Schmidt, Murphy: similar Bayesian work UCI 2010.


Wong et al., close update: *Adaptive Graphical Lasso*, 2014.


Chiquet et al., SIMoNe R-package *(needs updates...)*, Note Bioinformatics, 2009.

How to come up with a latent clustering?

Inference: Stochastic Bloc Model (SBM) cf. Timothée's talk

- Spread the nodes into $\mathcal{Q}$ classes with $\mathbb{P}(i \in q) = \alpha_q$;
- Connexion probabilities depend upon node classes:

$$\mathbb{P}(i \leftrightarrow j | i \in \text{class } q, j \in \text{class } \ell) = \pi_{q\ell}.$$

EM-strategy - conditional expectation to maximize

$$\begin{aligned} Q\left(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(m)}\right) &= \mathbb{E}\left\{\log \ell(\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z})|\mathbf{X}; \boldsymbol{\Theta}^{(m)}; \boldsymbol{\pi}, \boldsymbol{\alpha}\right\} \\ &= \sum_{\mathbf{Z} \in \mathcal{Z}} \mathbb{P}\left(\mathbf{Z}|\boldsymbol{\Theta}^{(m)}\right) \log \ell(\mathbf{X}, \boldsymbol{\Theta}, \mathbf{Z}). \end{aligned}$$

- The E step requires a variational estimation ($\hat{\mathbf{Z}}$)
- The M step is a weighted graphical-Lasso problem ($\hat{\boldsymbol{\Theta}}$)
- The weights are such that $P_{\mathbf{Z}} \propto 1 - \hat{\pi}_{q\ell}$.

Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.* Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

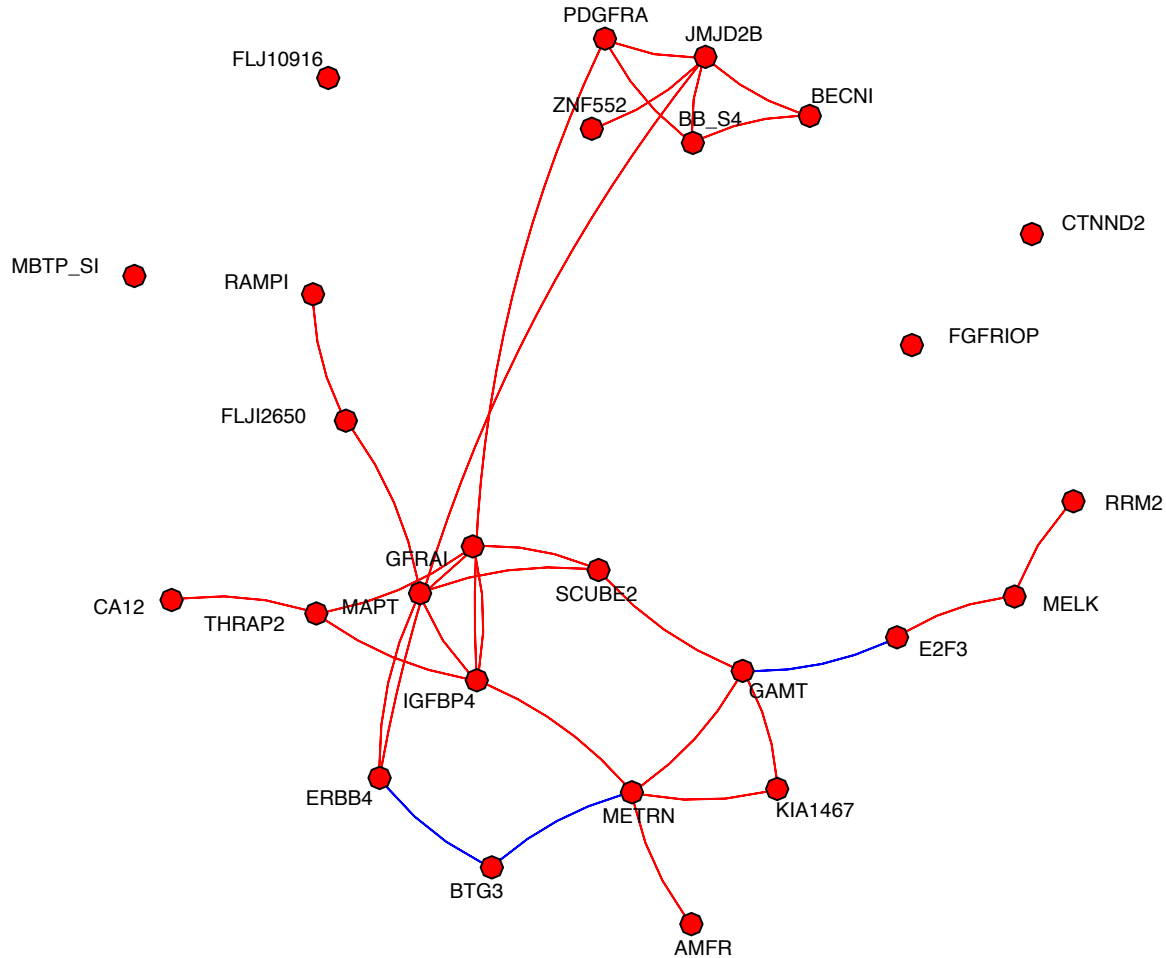


Figure: Pooling the data, Neighborhood Selection

Illustration on breast Cancer

Prediction of the outcome of preoperative chemotherapy

Hess *et al.*
Journal. of Clinical Oncology, 2006.

Data set

133 patients classified as

1. pathologic complete response,
2. residual disease,

according to a signature of 26 genes (small network).

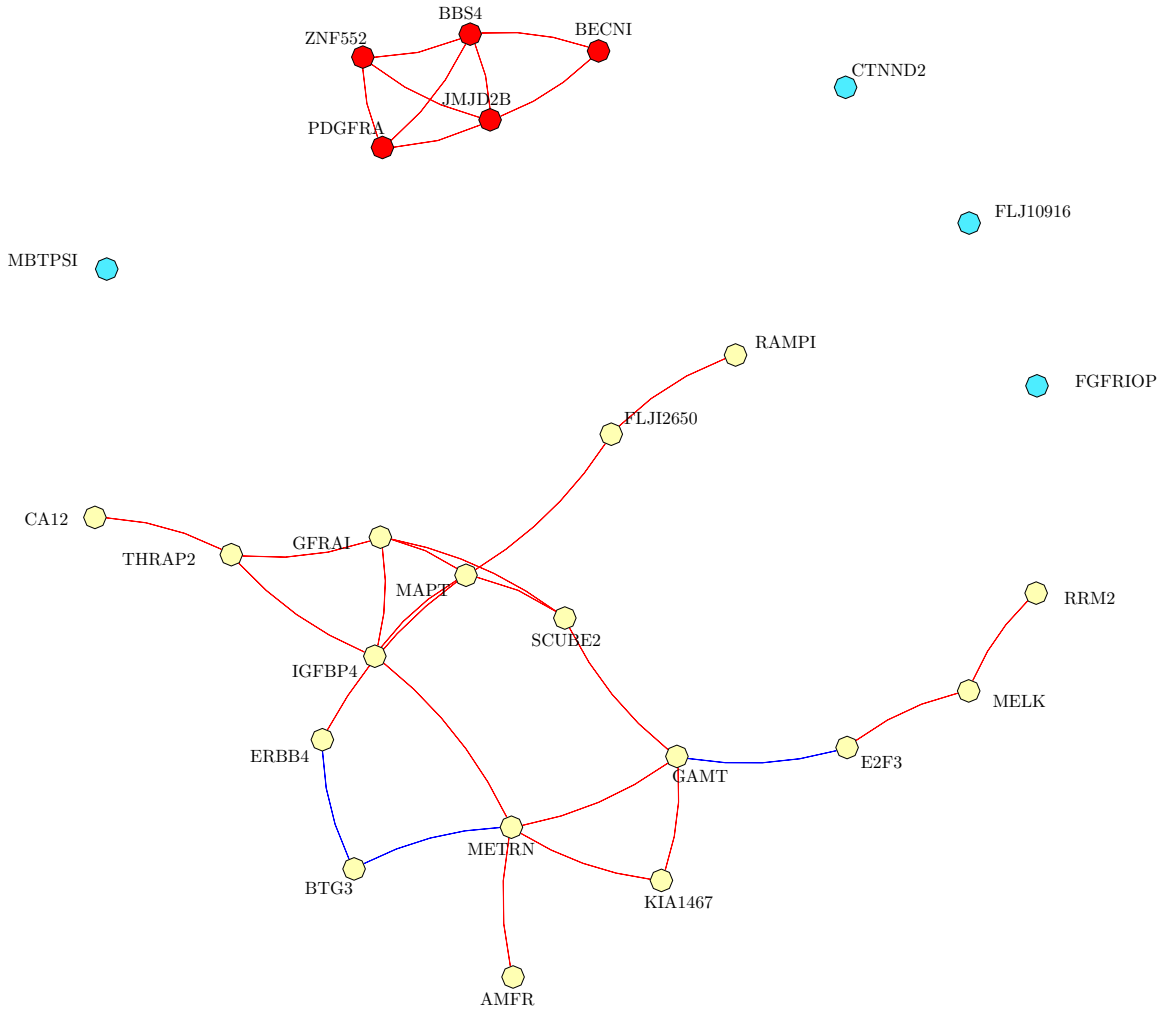


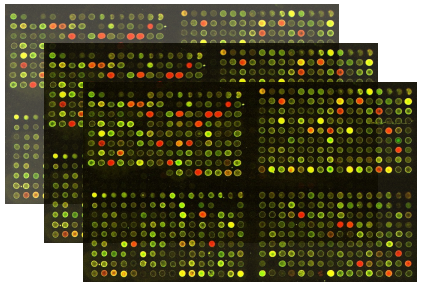
Figure: Pooling the data, SIMoNE with clustering

Outline

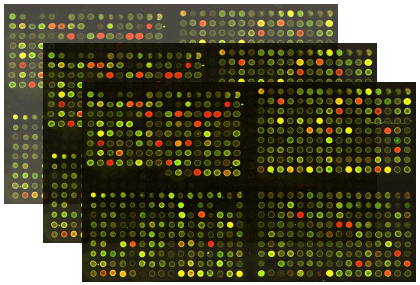
Handling scarcity and heterogeneity of data

Merge several experimental conditions

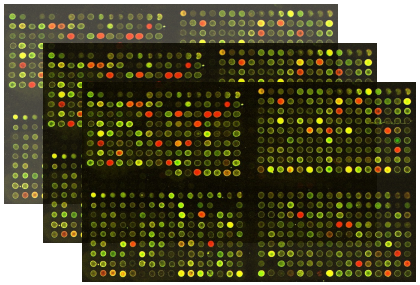
condition 1



condition 2



condition 3



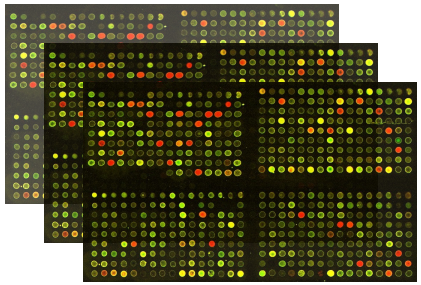
Multiple inference of GGM

$$\arg\max_{\mathbf{\Theta}^{(c)},c=1\ldots,C} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)};\mathbf{S}^{(c)}) - \lambda\, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

Handling scarcity and heterogeneity of data

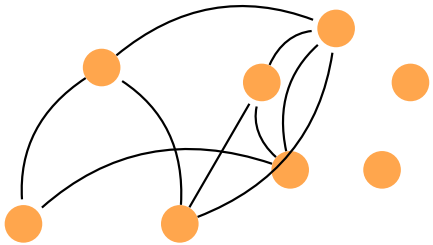
Inferring each graph independently does not help

condition 1

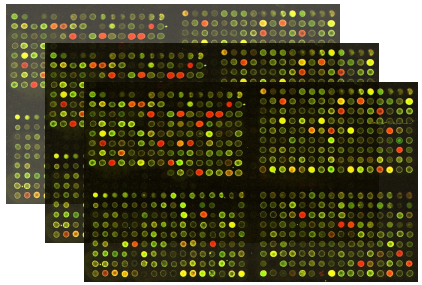


$(X_1^{(1)}, \ldots, X_{n_1}^{(1)})$

inference

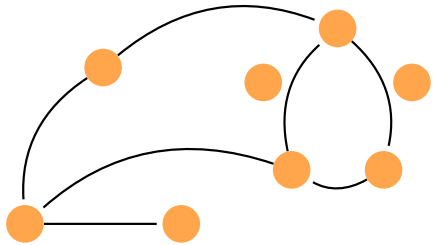


condition 2

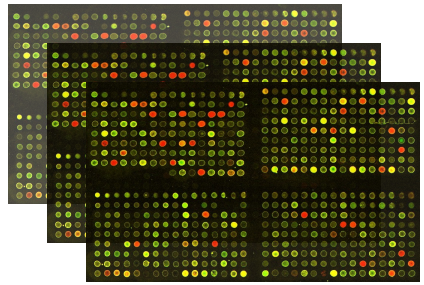


$(X_1^{(2)}, \ldots, X_{n_2}^{(2)})$

inference

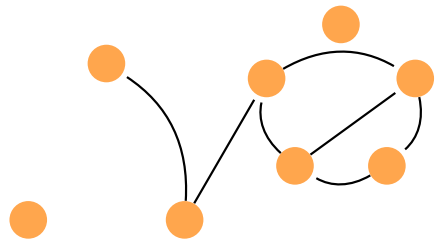


condition 3



$(X_1^{(3)}, \ldots, X_{n_3}^{(3)})$

inference

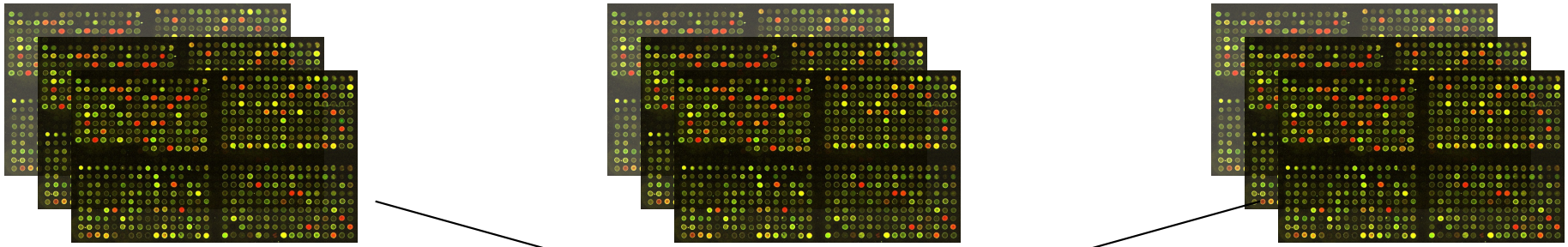


Multiple inference of GGM

Handling scarcity and heterogeneity of data

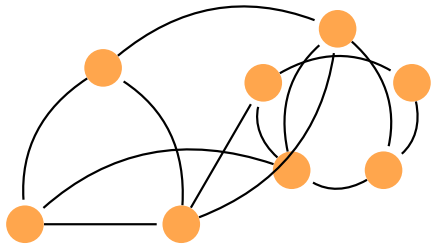
By pooling all the available data (like we just have with Hess' data set)

condition 1 condition 2 condition 3



$$(X_1, \ldots, X_n),\ n = n_1 + n_2 + n_3.$$

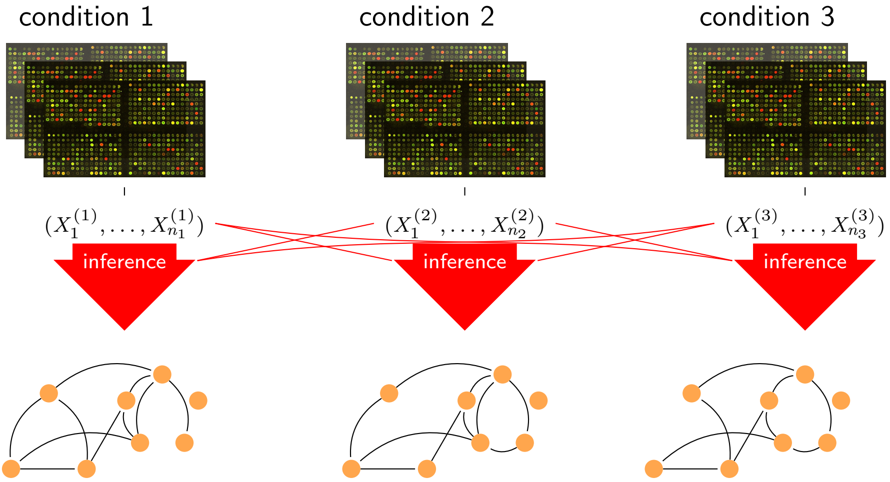
inference



Multiple inference of GGM

Handling scarcity and heterogeneity of data

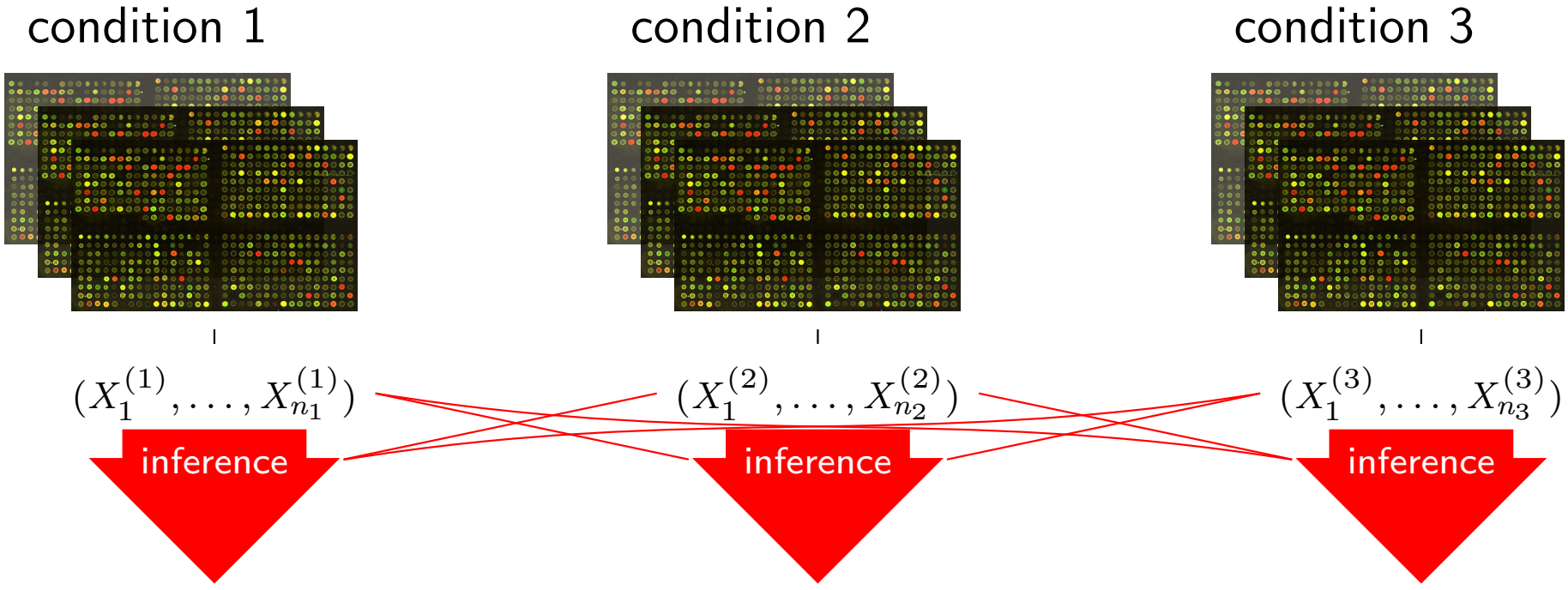
By breaking the separability



Multiple inference of GGM

Handling scarcity and heterogeneity of data

By breaking the separability



Multiple inference of GGM

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \ell(\mathbf{\Theta}^{(c)}; \mathbf{S}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\mathbf{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\mathbf{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\mathbf{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Intertwined-Lasso

- $\overline{\mathbf{S}} = \frac{1}{n} \sum_{t=1}^{T} n_t \mathbf{S}^{(t)}$ is the "pooled-tasks" covariance matrix.
- $\widetilde{\mathbf{S}}^{(t)} = \alpha \mathbf{S}^{(t)} + (1-\alpha)\overline{\mathbf{S}}$ is a mixture between specific and pooled covariance matrices.

A multitask approach

Chiquet, Grandvalet, Ambroise, Statistics and Computing 2010/11

Break the separability

Joint the optimization problem by either modifying

$$\underset{\boldsymbol{\Theta}^{(c)}, c=1\ldots,C}{\arg\max} \sum_{c=1}^{C} \tilde{\ell}(\boldsymbol{\Theta}^{(c)}; \tilde{\mathbf{S}}^{(c)}) - \lambda \, \mathrm{pen}_{\ell_1}(\boldsymbol{\Theta}^{(c)}).$$

1. the fitting term
2. the regularization term

Sparsity with grouping effect

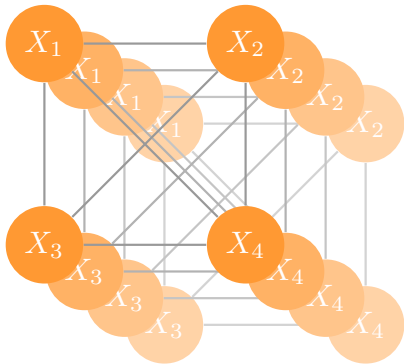
- Group-Lasso (Yuan and Lin 2006, Grandvalet and Canu, 1998),
- Cooperative-Lasso (Chiquet et al, AoAS, 2012),

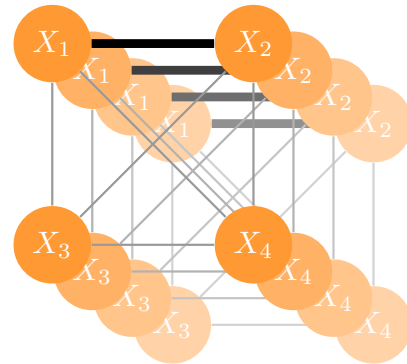
Grouping effects induced

Potential groups

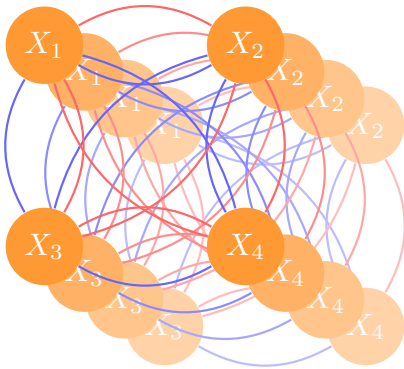
Group(s) induced by edges $(1, 2)$

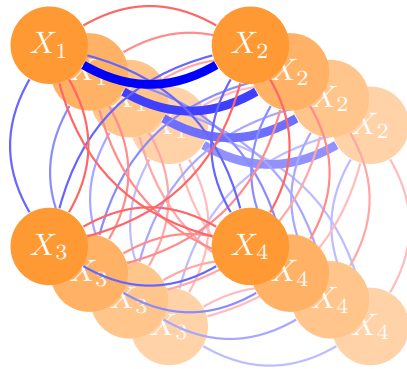
Group-Lasso

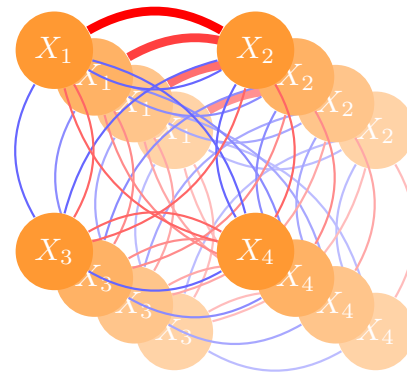




Cooperative-Lasso







Revisiting the Hess *et al.* data set

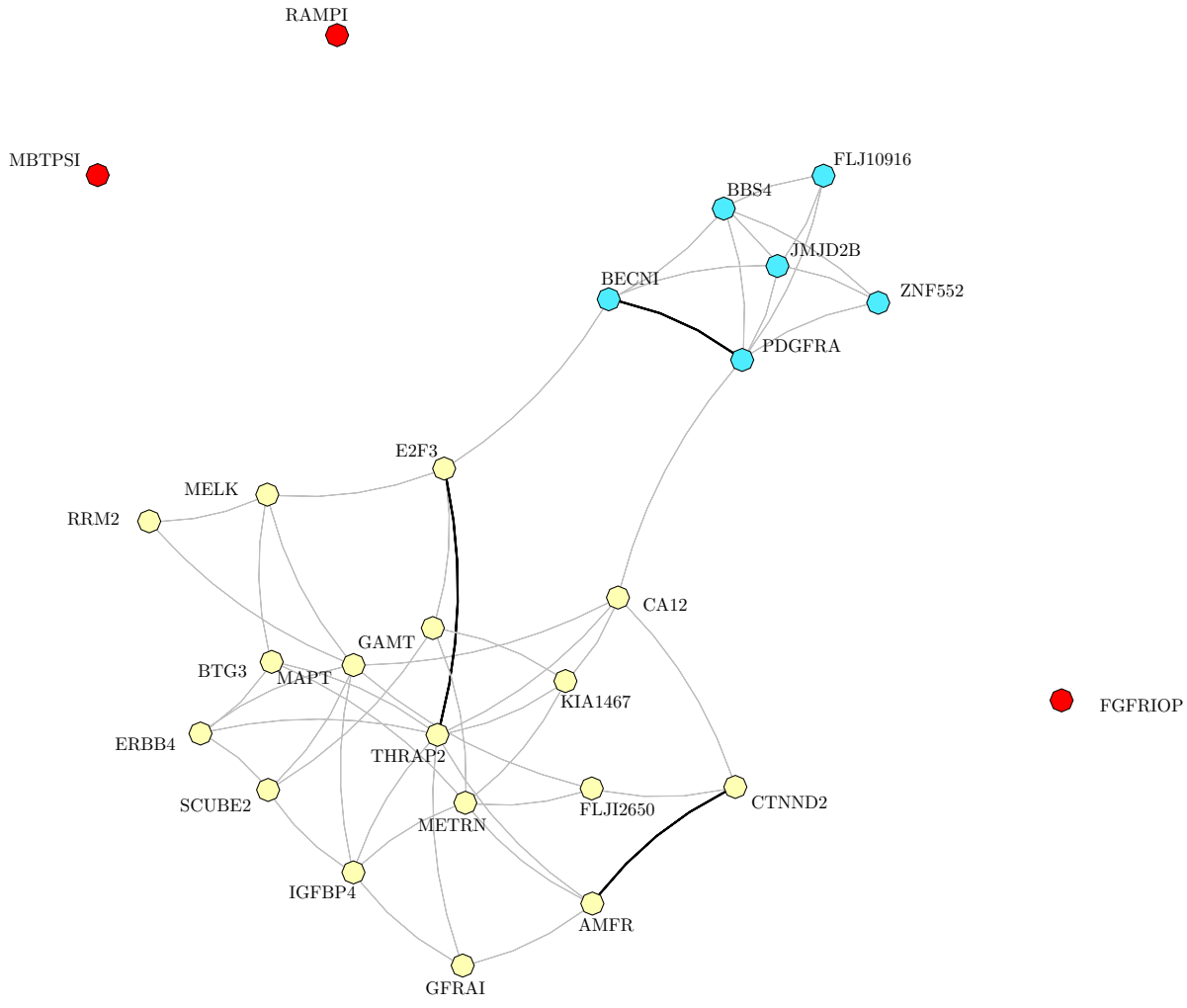


Figure: Cooperative-Lasso applied on the two sets of patients (PCR/noPCR). Bold edges are different in the finally selection graph.

Application: ER status in Breast cancer

Dataset: 466 patients with breast cancer

provided by Guedj *et al.*,
A refined molecular taxonomy of breast cancer, *Oncogene*, 2011.

Objective: identify changes in regulatory mechanisms

- ER^+/ER^-: breast cancer growth stimulated by estrogen hormones,
- ER^+ tackled with anti-hormonal therapies,
- ER^- found clinically more aggressive.

Jeanmougin, Charbonnier, Guedj and Chiquet, Network inference in breast cancer with Gaussian graphical models and extensions.

Probabilistic graphical models for genetics, Oxford University Press, 2014.

Application: ER status in Breast cancer

Network inference with cooperative-Lasso on 200 candidate genes (partial view)

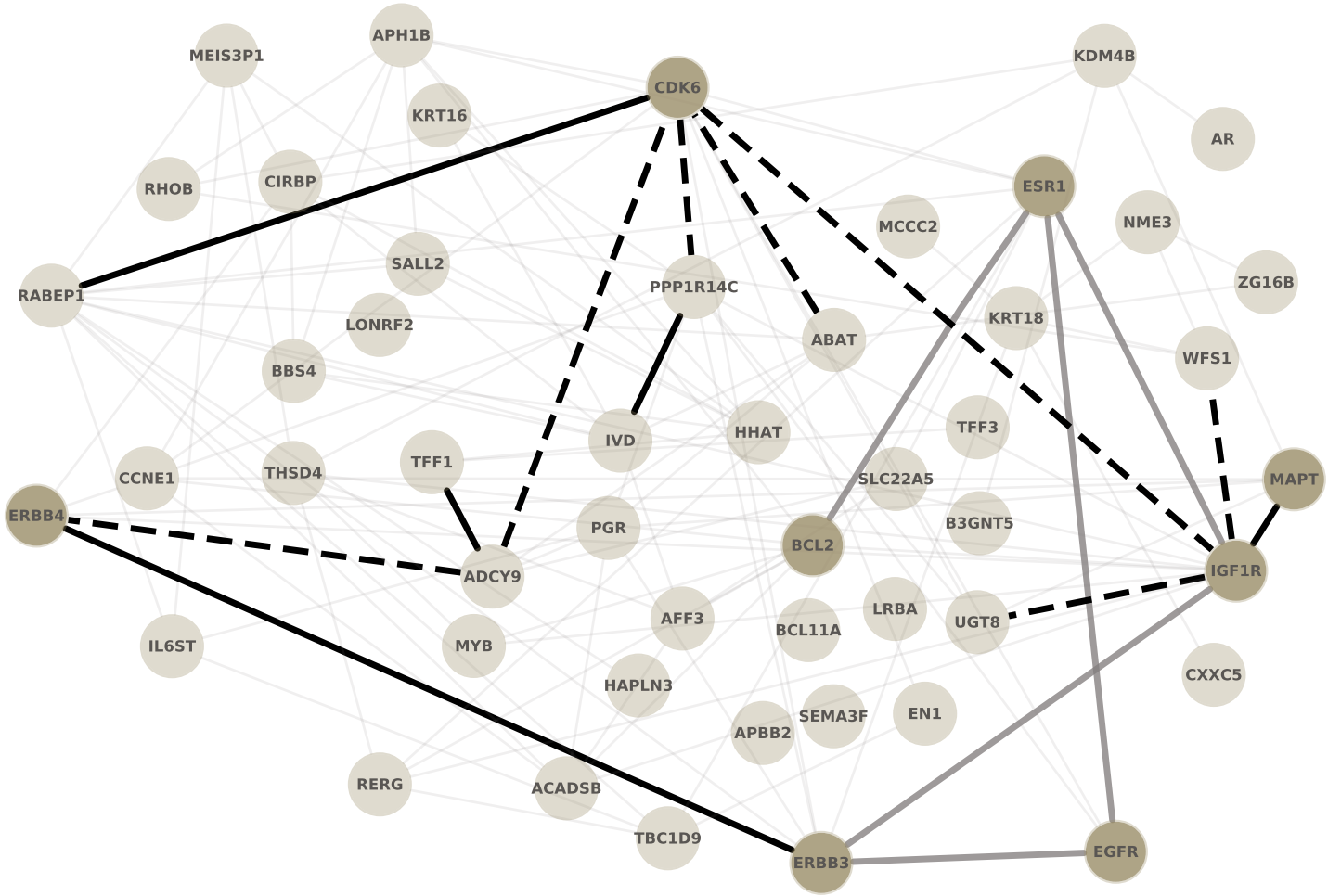


Figure: The dashed black edges are inferred only under the ER- condition and the solid black edges are only predicted under the ER+ condition. Gray are common to both conditions

Application: ER status in Breast cancer

Network inference with the cooperative-Lasso fits known anti-apoptotic mechanisms

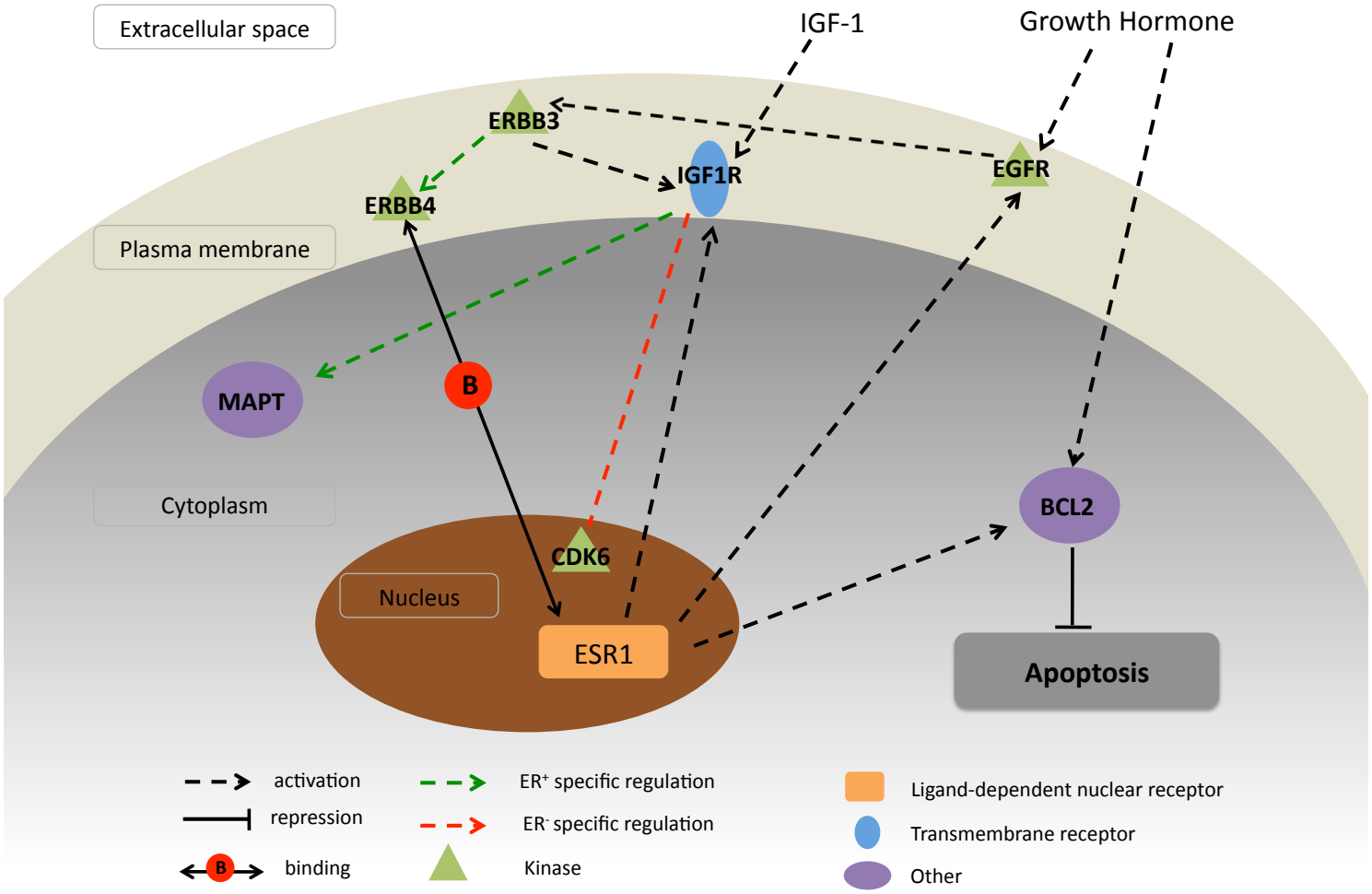
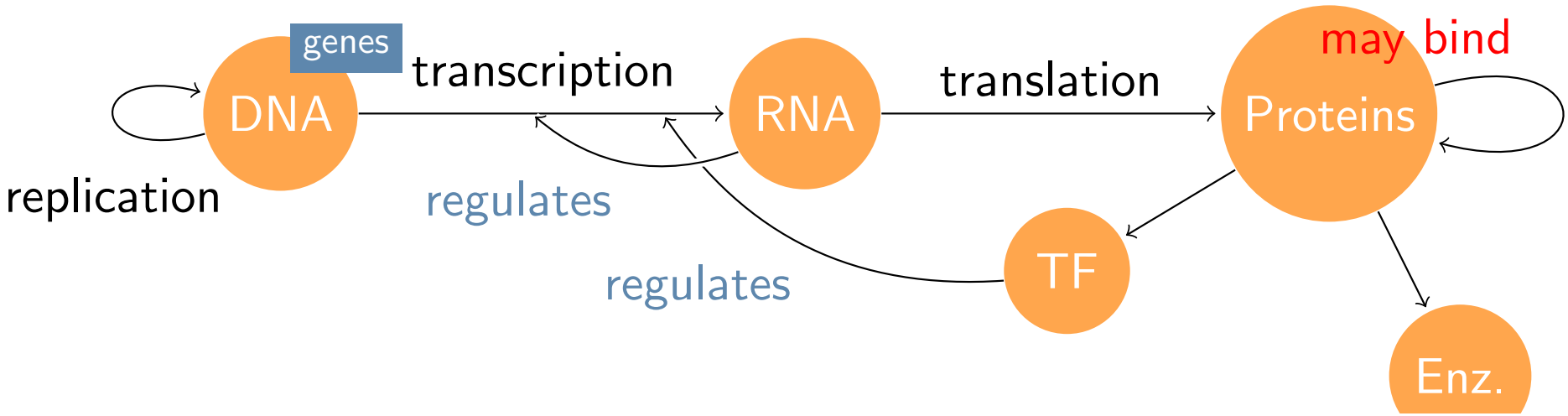


Figure: Most edges are supported by the literature (except two)

Outline

Why Multi-attribute Networks?

Joint work with E. Kolaczyk (Boston) and C. Ambroise (Évry)



Data integration

- Omic technologies can profile cells at different levels: DNA, RNA, protein, chromosomal, and functional.
- multiple molecular profiles combined on the same set of biological samples can be *synergistic*.

Multiattribute GGM

Consider e.g. some p genes of interest and the $K = 2$ omic experiments

1. X_{i1} is the expression profile of gene i (transcriptomic data),
2. X_{i2} is the corresponding protein concentration (proteomic data).

Define a block-wise precision matrix

- $X = (X_1, \ldots, X_p)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ in $\mathbb{R}^{pK}$,
- $X_i = (X_{i1}, \ldots, X_{iK})^\intercal \in \mathbb{R}^K$.

$$\mathbf{\Theta} = \mathbf{\Sigma}^{-1} = \begin{bmatrix} \mathbf{\Theta}_{11} & & \mathbf{\Theta}_{1p} \\ & \ddots & \\ \mathbf{\Theta}_{p1} & & \mathbf{\Theta}_{pp} \end{bmatrix}, \qquad \mathbf{\Theta}_{ij} \in \mathcal{M}_{K,K}, \ \forall (i,j) \in \mathcal{P}^2.$$

Graphical Interpretation

Define $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ as the multivariate analogue of the *conditional graph*:

$$(i,j) \in \mathcal{E} \Leftrightarrow \mathbf{\Theta}_{ij} \neq \mathbf{0}_{KK}.$$

Multivariate Neighborhood selection

The penalized multivariate regression approach

For each node /gene, recover its neighborhood by solving

$$\arg \min_{\mathbf{B}_i \in \mathcal{M}_{(p-1)K,K}} \frac{1}{2N} \left\| \mathbf{X}_i - \mathbf{X}_{\setminus i} \mathbf{B}_i \right\|_F^2 + \lambda \Omega(\mathbf{B}_i),$$

Choice of Penalty

Group-based penalty to activate the set of attributes simultaneously on a given link:

$$\Omega(\mathbf{B}_i) = \sum_{j \in \mathcal{P} \setminus i} \|\mathbf{B}_{ij}\| \quad , \quad \mathbf{B}_{ij} \in \mathcal{M}_{KK}$$

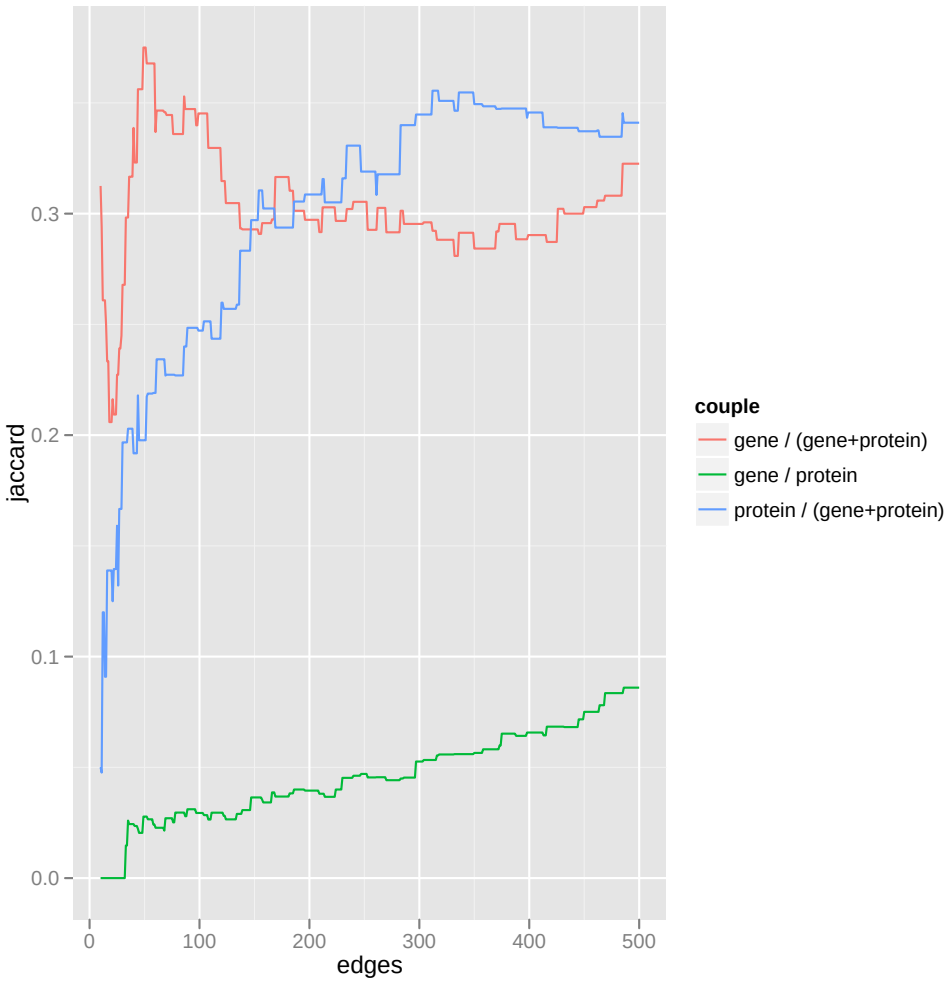
- $\|M\| = \|M\|_F = \left(\sum_{i,j} M_{ij}^2 \right)^{1/2}$, the Frobenius norm,
- $\|M\| = \|M\|_\infty = \max_{i,j} |M_{ij}|$, the sup norm (shared magnitude),
- $\|M\| = \|M\|_\star = \sum \text{eig}(M)$, the nuclear norm (rank penalty).

Illustration on the NCI-60 data set

Molecular profile data on a panel of 60 diverse human cancer cell lines

1. Protein: reverse-phase lysate arrays (RPLA) for 92 antibodies;
2. Gene : Human Genome U95 affymetrix ($\sim$ 9,000 genes).

$\rightsquigarrow$ consensus set with 91 protein and corresponding gene profiles.



Jaccard's similarity index

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

$\rightsquigarrow$ multiattribute network shares a high Jaccard index with both uni attribute networks.

Outline

Canonical framework: sparse GGM

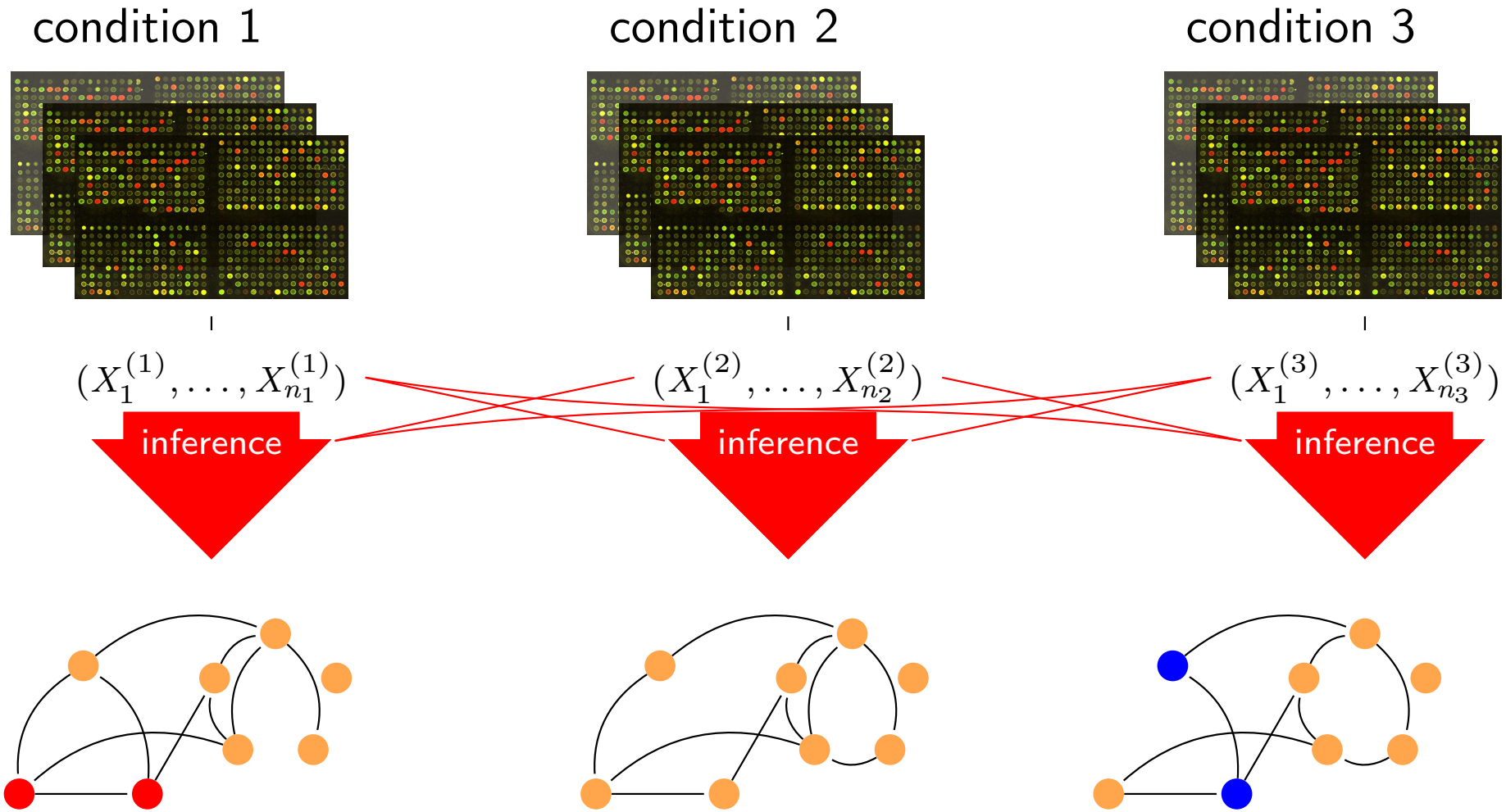
Accounting for some biological features

Network inference for enhancing other methods

- Differential analysis
- Multivariate regression

Outline

Context: multitask framework (Trung Ha's thesis)



Genes expressions might be shifted by 2 nonindependent phenomenons:

1. Its average expression level of genes.
2. Its relations with others genes.

Model setup for differential analysis

Idea

Share information for the multiple learning task corresponding to both

1. interactions (networks/precision matrix) and,
2. average expression levels (means).

to perform differential analysis.

Hypothesis testing for differential analysis

Assume $X^k = (X_1^k, \dots X_p^k) \sim \mathcal{N}\left(\boldsymbol{\mu}^k, \boldsymbol{\Sigma} = \boldsymbol{\Theta}^{-1}\right)$. For each gene j, we test

$$\begin{cases} H_0: & \mu_j^k = \mu_j^k, \quad \forall (k, k') \\ H_1: & \exists (k, k') : \mu_j^k \neq \mu_j^{k'} \end{cases},$$

where the genes are related by $\boldsymbol{\Sigma}$.

Coupling the two problems

GGM: linear regression point of view

Expression of gene j in task k for the ith replicate is linearly explained by the other genes

$$X_{ij}^k = \mu_j^k + \sum_{j' \neq j} \beta_{jj'} (X_{j'}^k - \mu_{j'}^k) + \varepsilon_j^k, \quad \varepsilon_j \sim \mathcal{N}(0, 1/\Theta_{jj}).$$

where $\beta_{jj'} = -\Theta_{jj'}/\Theta_{j'j'}$ explains the relation between genes j and j'.

Strategy: fused the vector of means corrected by the covariance

$$\underset{\boldsymbol{\mu}^k, \Theta^k}{\text{minimize}} \sum_{k=1}^{K} \text{RSS}(\boldsymbol{\mu}^k, \Theta) + \lambda_1 \|\Theta)\|_1 + \lambda_2 \sum_{k<k'} \omega_{kk'} \left\| \boldsymbol{\mu}^k - \boldsymbol{\mu}^{k'} \right\|_1$$

where

- the red penalty regularizes the network component
- the blue penalty favors fusion of means across tasks.

Coupling the two problems

GGM: linear regression point of view

Expression of gene j in task k for the ith replicate is linearly explained by the other genes

$$X_{ij}^k = \mu_j^k + \sum_{j' \neq j} \beta_{jj'}(X_{j'}^k - \mu_{j'}^k) + \varepsilon_j^k, \quad \varepsilon_j \sim \mathcal{N}(0, 1/\Theta_{jj}).$$

where $\beta_{jj'} = -\Theta_{jj'}/\Theta_{j'j'}$ explains the relation between genes j and j'.

Strategy: fused the vector of means corrected by the covariance

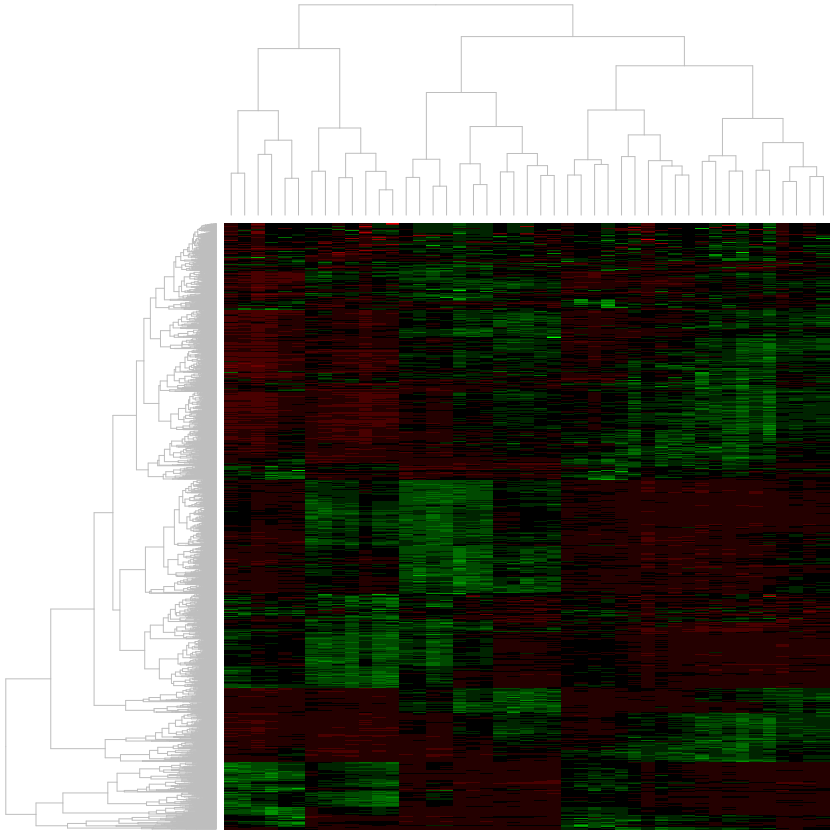
$$\underset{\boldsymbol{\mu}^k, \boldsymbol{\Theta}^k}{\text{minimize}} \sum_{k=1}^{K} \text{RSS}(\boldsymbol{\mu}^k, \boldsymbol{\Theta}) \; + \; \lambda_1 \|\boldsymbol{\Theta})\|_1 \; + \; \lambda_2 \sum_{k<k'} \omega_{kk'} \left\| \boldsymbol{\mu}^k - \boldsymbol{\mu}^{k'} \right\|_1$$

where

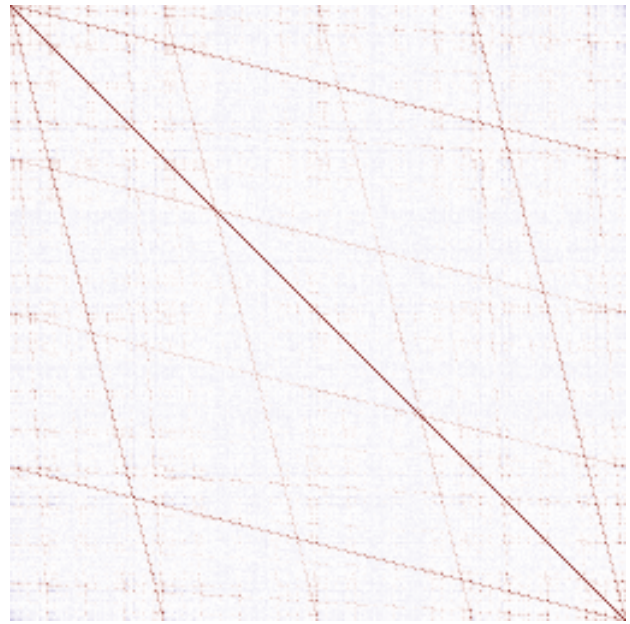
- the red penalty regularizes the network component
- the blue penalty favors fusion of means across tasks.

Outline

Example: regulatory motif in *P. falciparum* (malaria)



Clustering of gene expression
(row: genes; columns: conditions)



Correlation pattern between 4-size
counts in gene promotor regions

Goal

Task: selecting regulatory motifs

Inference in multivariate linear regression

Consider n samples and let for individual i

- $\mathbf{y}_i$ be the q-dimensional vector of responses,
- $\mathbf{x}_i$ be the p-dimensional vector of predictors,
- $\mathbf{B}$ be the $p \times q$ matrix of regression coefficients
- $\boldsymbol{\varepsilon}_i$ be a noise term with a q-dimensional covariance matrix $\boldsymbol{\Sigma}$.

$$\mathbf{y}_i = \mathbf{B}^T \mathbf{x}_i + \boldsymbol{\varepsilon}_i, \quad \boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}), \quad \forall i = 1, \ldots, n,$$

Matrix notation

Let $\mathbf{Y}(n \times q)$ and $\mathbf{X}(n \times p)$ be the data matrices, then

$$\mathbf{Y} = \mathbf{X}\mathbf{B} + \boldsymbol{\varepsilon}, \quad \mathrm{vec}(\boldsymbol{\varepsilon}) \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n \otimes \boldsymbol{\Sigma}).$$

remark

If $\mathbf{X}$ is a design matrix, this is called the "General Linear Model" (GLM), but Mathematics are the same.

Regularized MLR

Double penalization

Consider regularizing both $\mathbf{B}$ and the inverse covariance $\mathbf{\Sigma}^{-1}$

$$(\hat{\mathbf{B}}, \hat{\mathbf{\Sigma}}) = \arg\max_{\mathbf{B},\mathbf{\Sigma}} \left\{ \log \ell(\mathbf{Y}, \mathbf{X}; \mathbf{\Sigma}, \mathbf{B}) + \lambda \text{pen}_{\ell_1}(\mathbf{B}) + \lambda_2 \text{pen}(\mathbf{\Sigma}^{-1}) \right\}$$

$\rightsquigarrow$ $\mathbf{\Sigma}^{-1}$ can be seen as a network between responses

Goal: enhancing the selection of relevant variables in $\mathbf{B}$

- by carrying the general trend carried by $\mathbf{\Sigma}$
- exact recovery of edges in $\hat{\mathbf{\Sigma}}^{-1}$ is no more needed

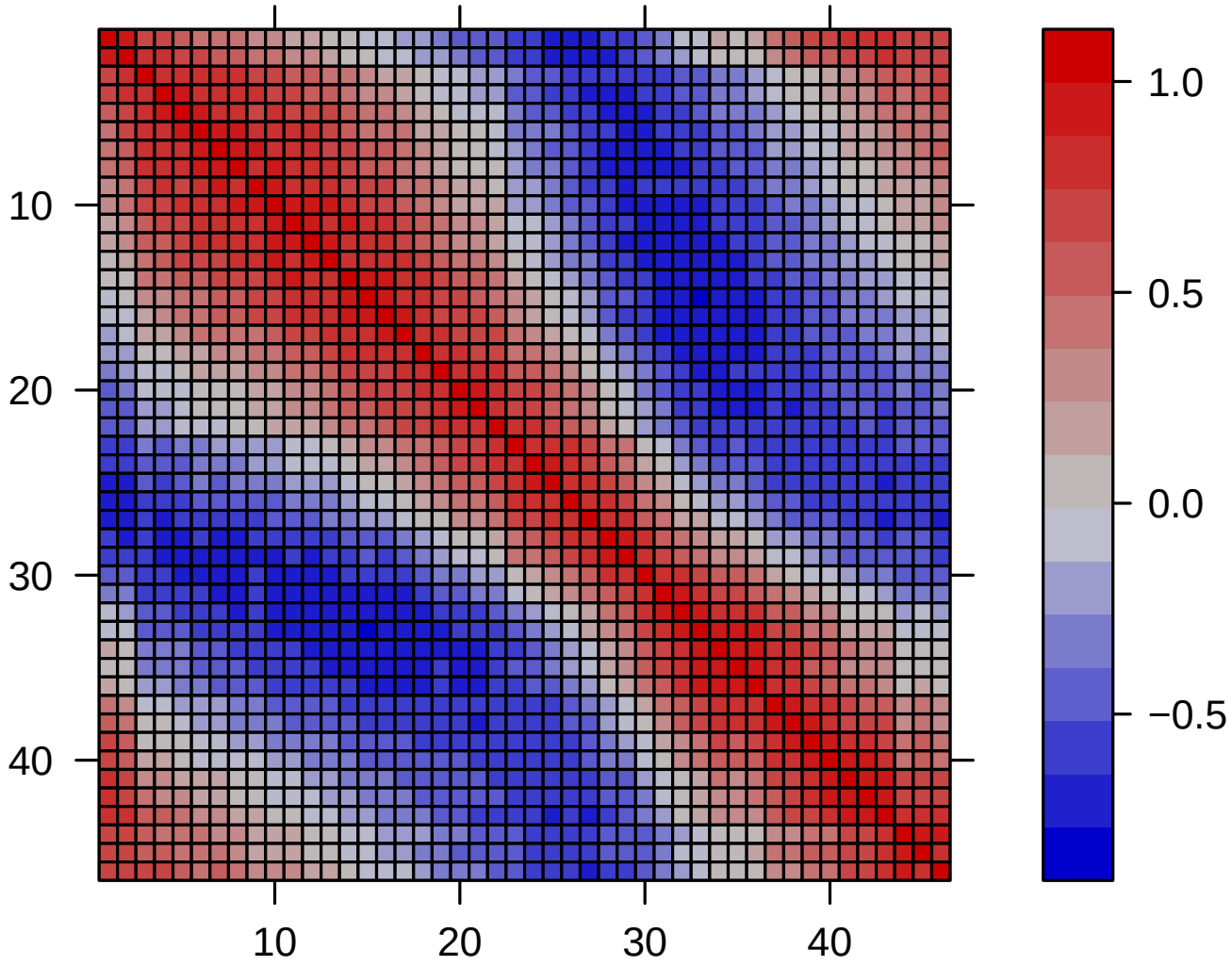
Applications

- with T. Mary-Huard, S. Robin, Multi-trait genomic selection
- with M. Brégère (MSc.), M. Perrot (PhD) C. Lévy-Leduc, omics.

Example: multiple assays in transcriptomics

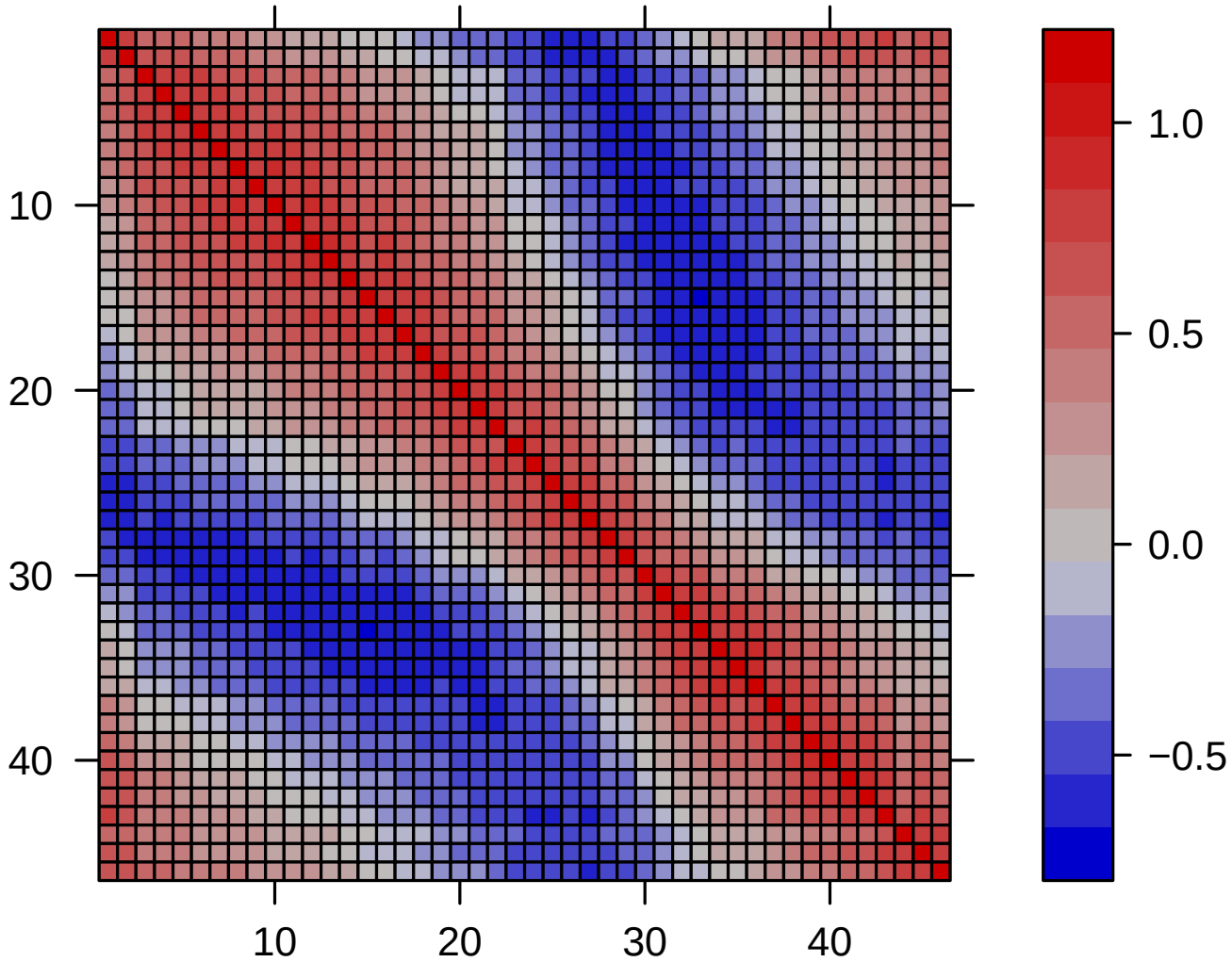
Stress condition in Plasmodium

```
load("plasmodium_expression.Rdata")
image(Matrix(cor(Y)), xlab="", ylab="", sub="")
```



Estimating the covariance between the assays

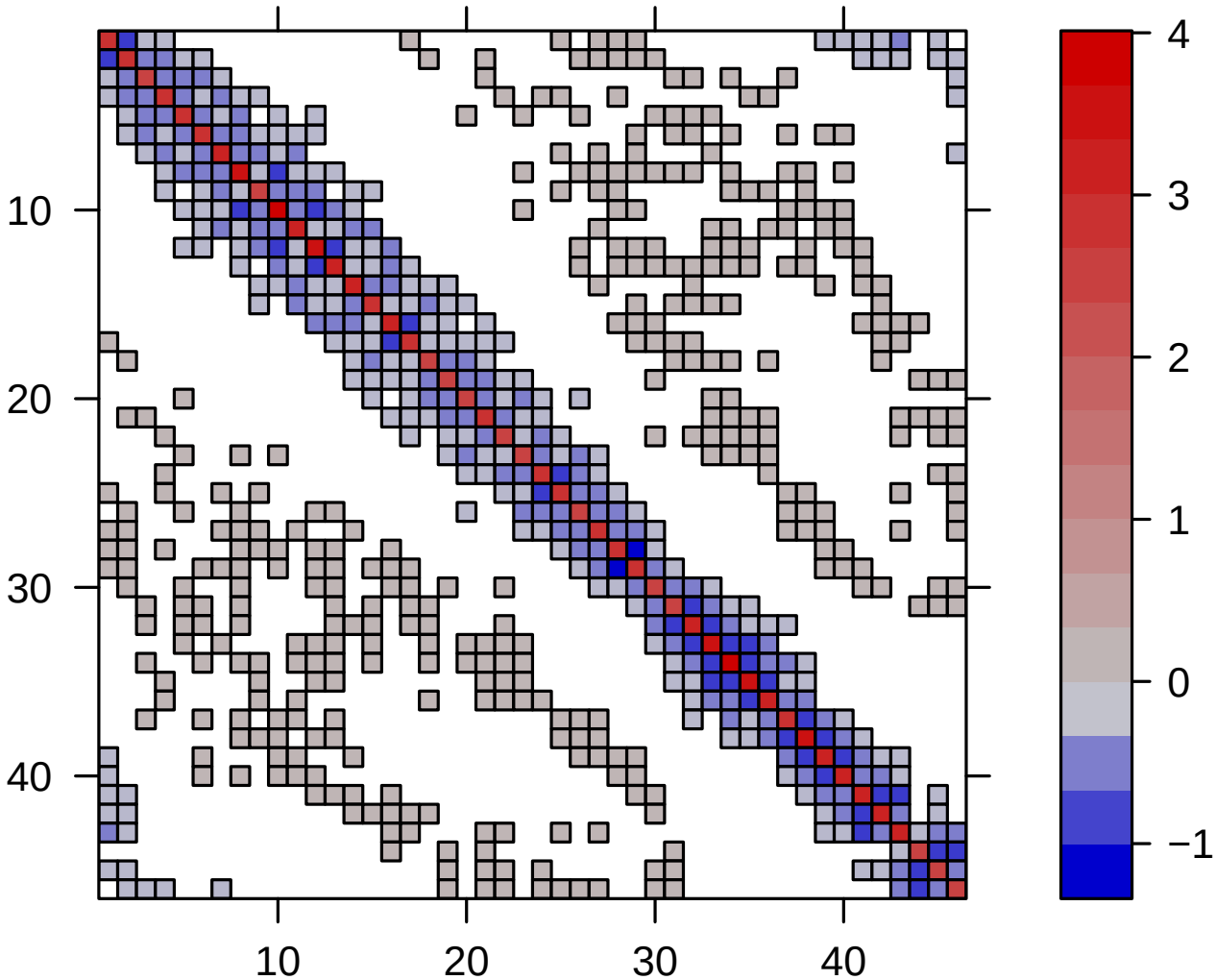
```
library(huge)
huge.out <- huge(as.matrix(Y), method="glasso", cov.output=TRUE, verbose=FALSE)
out  <- huge.select(huge.out, verbose=FALSE)
```



Sparse inverse covariance allows compressing information

```
cat(sum(abs(out$opt.icov) != 0)/2, "param. among", p*2/2, "potential param.")

## 380 param. among 46 potential param.
```



Conclusion

Sparse Gaussian Graphical Model

Well established framework with a vast, growing literature

1. Nice modeling tool (conditional dependencies),
2. Good theoretical framework (which I have not much talked about),
3. Powerful algorithms
   - that scale the dimension (large p large n)
   - that allow resampling/parallelization (for robustness)

$\rightsquigarrow$ Great tool for covariance estimation/selection in a reasonably high dimensional settings.

Still...

- an interaction is not even well defined
- $\rightsquigarrow$ carefull with interpretation of the networks
- metagenomics data do have some specificities
- $\rightsquigarrow$ adaptation needed

Thanks

To my coauthors and to you for your patience and for listening...