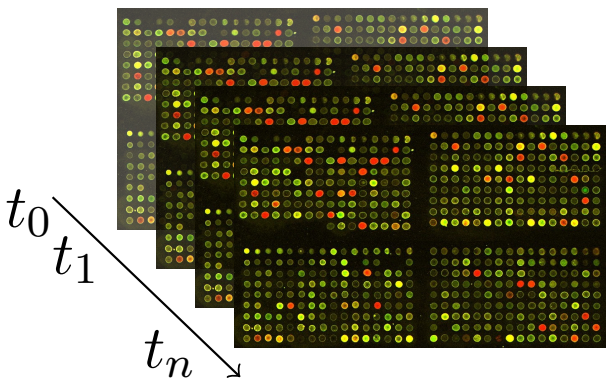
Inferring Gene Regulatory Networks from Time-Course Gene Expression Data

Camille Charbonnier *and* Julien Chiquet *and* Christophe Ambroise

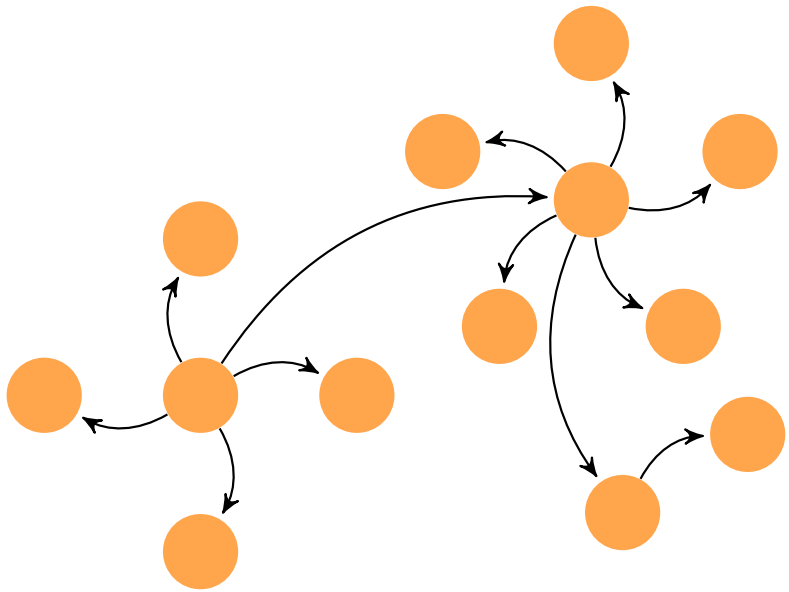
Laboratoire Statistique et Génome,
La génopole - Université d'Évry

JOBIM – 7-9 septembre 2010

Problem



Inference



≈ 10s microarrays over time

≈ 1000s probes ("genes")

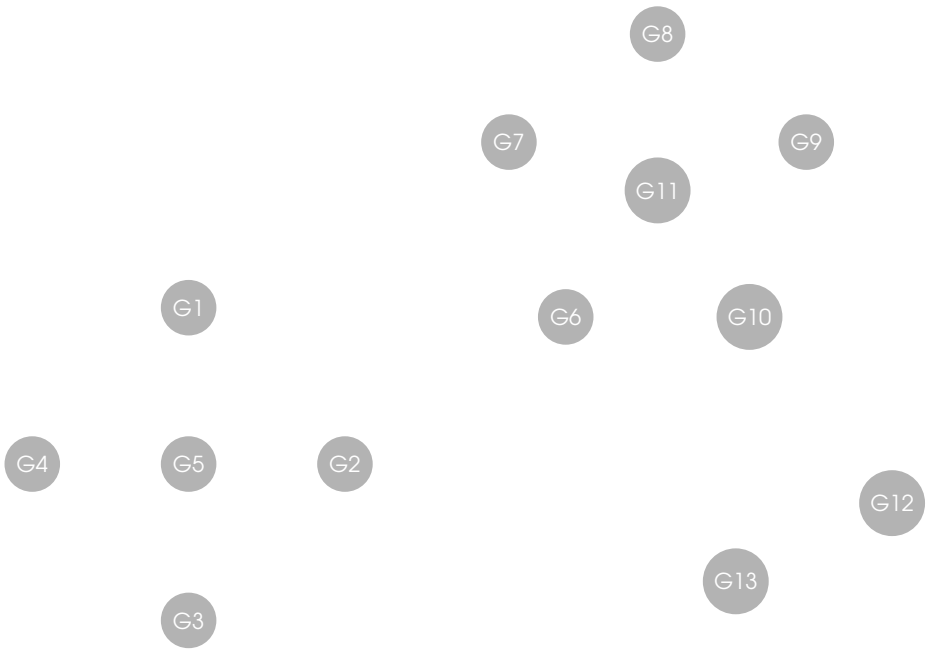
Which interactions?

The main statistical issue is the high dimensional setting.

Handling the scarcity of the data
By introducing some prior

Priors should be biologically grounded

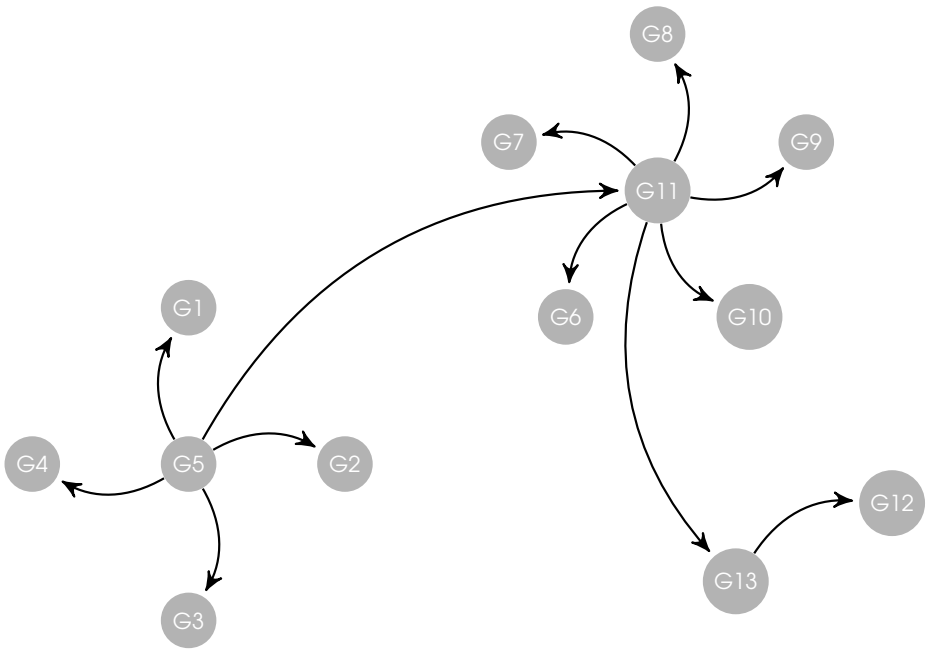
1. few genes effectively interact (sparsity),
2. networks are organized (latent clustering),



Handling the scarcity of the data
By introducing some prior

Priors should be biologically grounded

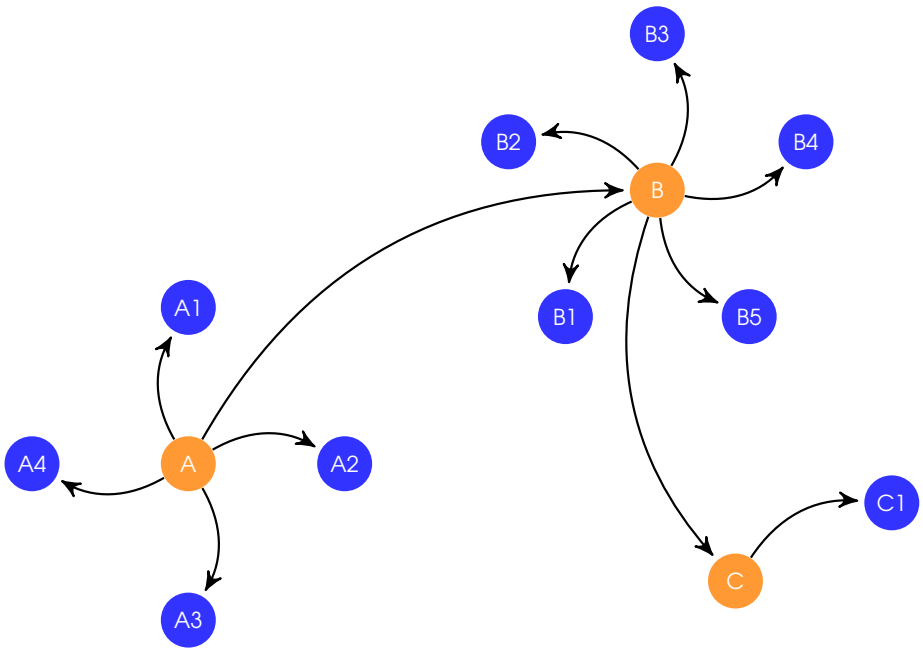
1. few genes effectively interact (sparsity),
2. networks are organized (latent clustering),



Handling the scarcity of the data
By introducing some prior

Priors should be biologically grounded

1. few genes effectively interact (sparsity),
2. networks are organized (latent clustering),



Outline

Statistical models

- Gaussian Graphical Model for Time-course data
- Structured Regularization

Algorithms and methods

- Overall view
- Model selection

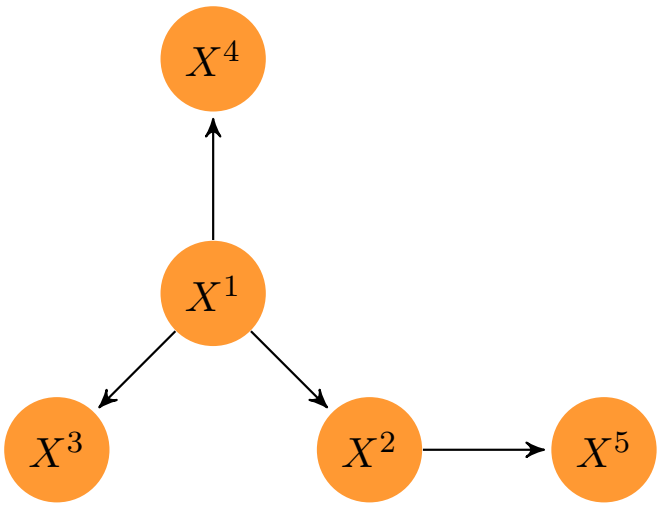
Numerical experiments

- Inference methods
- Performance on simulated data
- E. coli S.O.S DNA repair network

Gaussian Graphical Model for Time-course data

Collecting gene expression

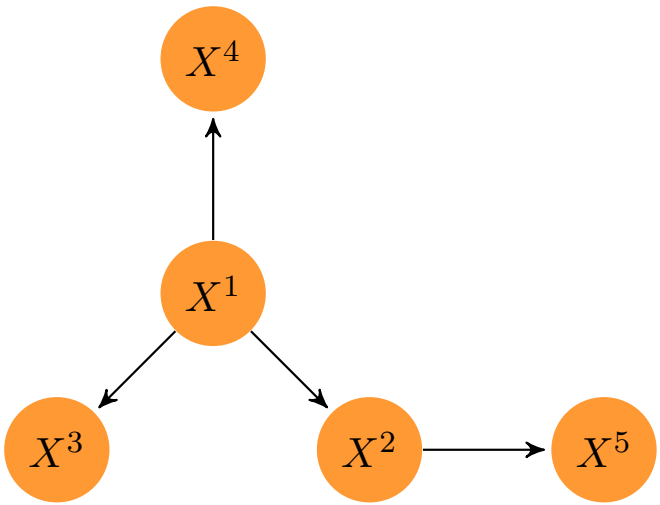
1. Follow-up of one single experiment/individual;
2. Close enough time-points to ensure
   - dependency between consecutive measurements;
   - homogeneity of the Markov process.



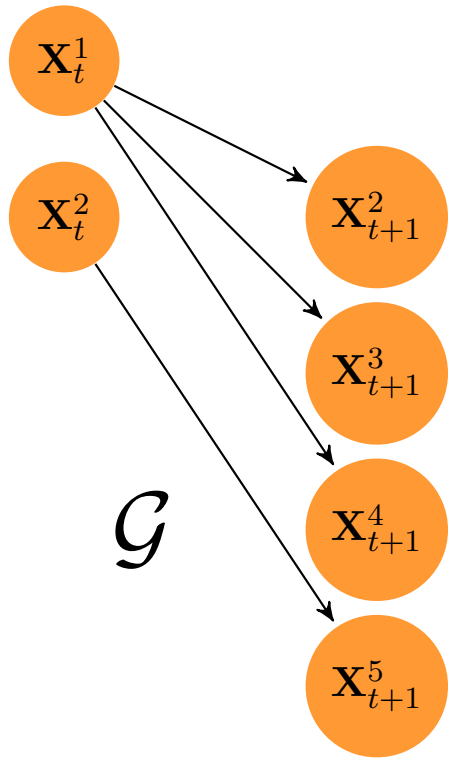
Gaussian Graphical Model for Time-course data

Collecting gene expression

1. Follow-up of one single experiment/individual;
2. Close enough time-points to ensure
   - dependency between consecutive measurements;
   - homogeneity of the Markov process.



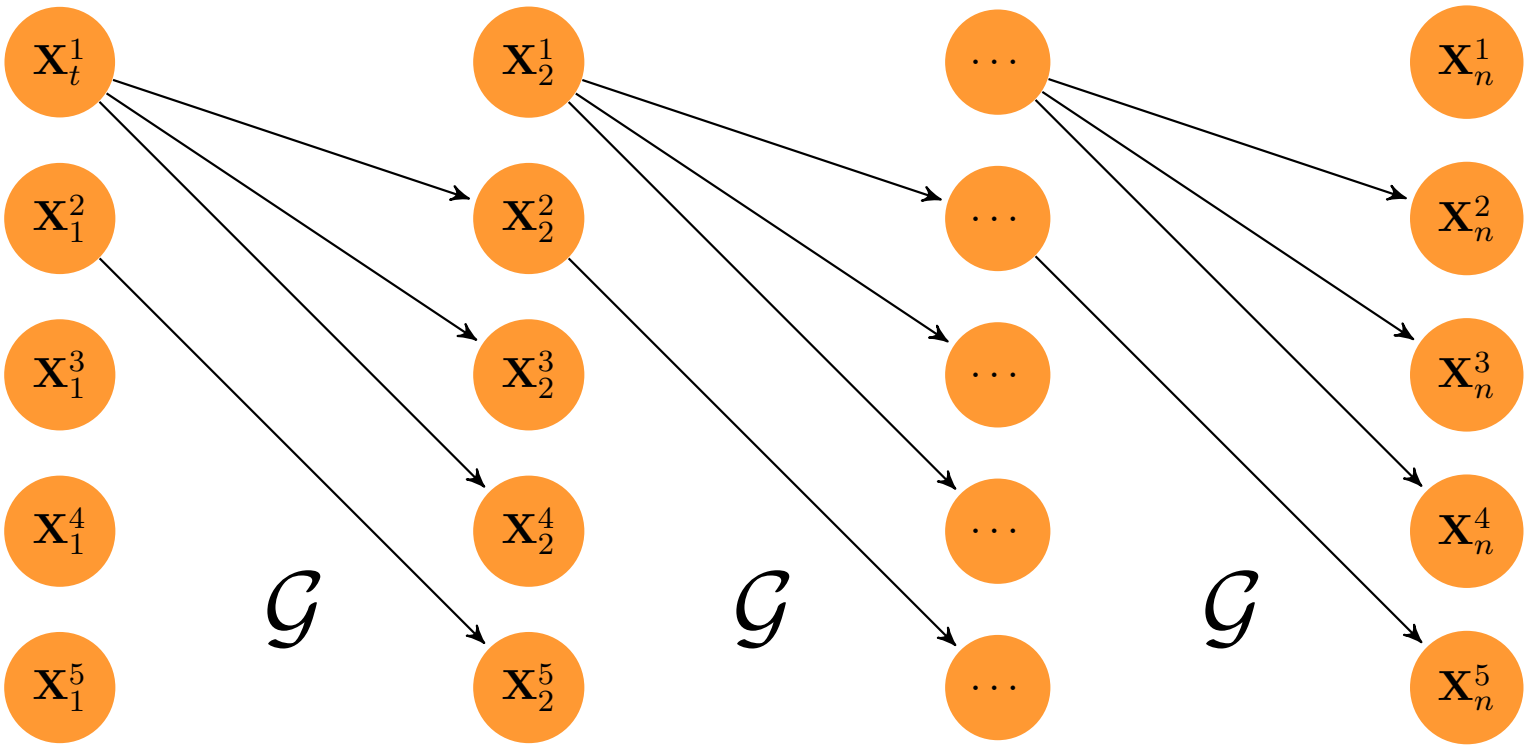
stands for



Gaussian Graphical Model for Time-course data

Collecting gene expression

1. Follow-up of one single experiment/individual;
2. Close enough time-points to ensure
   - dependency between consecutive measurements;
   - homogeneity of the Markov process.



Gaussian Graphical Model for Time-Course data

Assumption

A microarray can be represented as a multivariate Gaussian vector $X = (X(1), \ldots, X(p)) \in \mathbb{R}^p$, following a first order vector autoregressive process $VAR(1)$:

$$X_t = \mathbf{\Theta} X_{t-1} + \mathbf{b} + \varepsilon_t, \quad t \in [1, n]$$

where we are looking for $\mathbf{\Theta} = (\theta_{ij})_{i,j \in \mathcal{P}}$.

Graphical interpretation

conditional dependency between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

non null partial correlation between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

$\theta_{ij} \neq 0$

Gaussian Graphical Model for Time-Course data

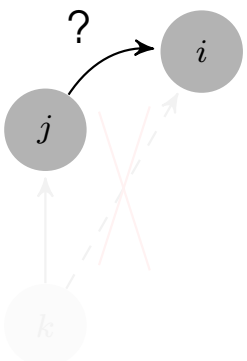
Assumption

A microarray can be represented as a multivariate Gaussian vector $X = (X(1), \ldots, X(p)) \in \mathbb{R}^p$, following a first order vector autoregressive process $VAR(1)$:

$$X_t = \mathbf{\Theta} X_{t-1} + \mathbf{b} + \varepsilon_t, \quad t \in [1, n]$$

where we are looking for $\mathbf{\Theta} = (\theta_{ij})_{i,j \in \mathcal{P}}$.

Graphical interpretation



if and only if

conditional dependency between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

non null partial correlation between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

$\theta_{ij} \neq 0$

Gaussian Graphical Model for Time-Course data

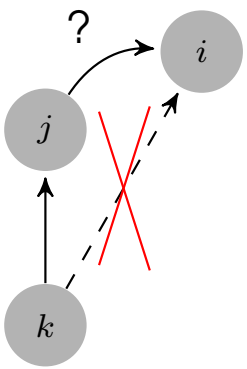
Assumption

A microarray can be represented as a multivariate Gaussian vector $X = (X(1), \ldots, X(p)) \in \mathbb{R}^p$, following a first order vector autoregressive process $VAR(1)$:

$$X_t = \mathbf{\Theta} X_{t-1} + \mathbf{b} + \varepsilon_t, \quad t \in [1, n]$$

where we are looking for $\mathbf{\Theta} = (\theta_{ij})_{i,j \in \mathcal{P}}$.

Graphical interpretation



if and only if

conditional dependency between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

non null partial correlation between $X_{t-1}(j)$ and $X_t(i)$

$\updownarrow$

$\theta_{ij} \neq 0$

Gaussian Graphical Model for Time-course data

Let

- $\mathbf{X}$ be the $n \times p$ matrix whose kth row is X_k,
- $\mathbf{S} = n^{-1}\mathbf{X}_{\setminus n}^\intercal \mathbf{X}_{\setminus n}$ be the within time covariance matrix,
- $\mathbf{V} = n^{-1}\mathbf{X}_{\setminus n}^\intercal \mathbf{X}_{\setminus 0}$ be the across time covariance matrix.

The log-likelihood

$$\mathcal{L}_{\text{time}}(\mathbf{\Theta}; \mathbf{S}, \mathbf{V}) = n \operatorname{Trace}(\mathbf{V}\mathbf{\Theta}) - \frac{n}{2}\operatorname{Trace}(\mathbf{\Theta}^\intercal \mathbf{S}\mathbf{\Theta}) + c.$$

$\rightsquigarrow$ Maximum Likelihood Estimator $\widehat{\mathbf{\Theta}}^{MLE} = \mathbf{S}^{-1}\mathbf{V}$

- not defined for $n < p$;
- even if $n > p$, requires multiple testing.

Charbonnier, Chiquet, Ambroise, *SAGMB* 2010

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}} \mathcal{L}_{\text{time}}(\mathbf{\Theta}; \mathbf{S}, \mathbf{V}) - \lambda \cdot \sum_{i,j \in \mathcal{P}} \mathbf{P}^{\mathbf{Z}}_{ij} |\mathbf{\Theta}_{ij}|$$

where λ is an overall tuning parameter and $\mathbf{P}^{\mathbf{Z}}$ is a (non-symmetric) matrix of weights depending on the underlying clustering $\mathbf{Z}$.

It performs

1. *regularization* (needed when $n \ll p$),
2. *selection* (specificity of the ℓ_1-norm),
3. *cluster-driven inference* (penalty adapted to $\mathbf{Z}$).

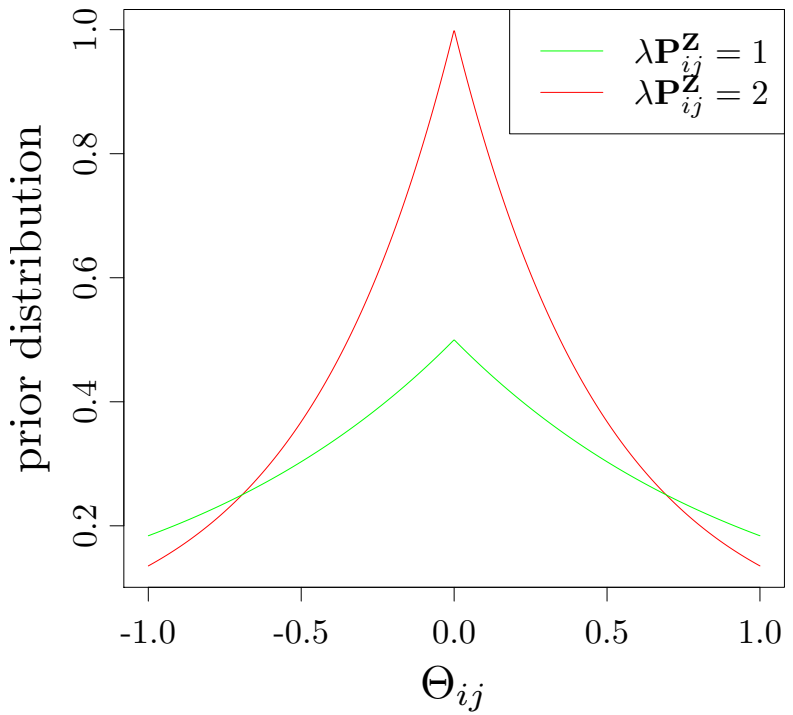
Structured regularization

"Bayesian" interpretation of ℓ_1 regularization

Laplacian prior on $\mathbf{\Theta}$ depends on the clustering $\mathbf{Z}$

$$\mathbb{P}(\mathbf{\Theta}|\mathbf{Z}) \propto \prod_{i,j} \exp\left\{-\lambda \cdot \mathbf{P}^{\mathbf{Z}}_{ij} \cdot |\mathbf{\Theta}_{ij}|\right\}.$$

$\mathbf{P_Z}$ summarizes prior information on the position of edges



How to come up with a latent clustering?

Biological expertise

- Build $\mathbf{Z}$ from prior biological information
  - transcription factors vs. regulatees,
  - number of potential binding sites,
  - KEGG pathways, ...
- Build the weight matrix from $\mathbf{Z}$.

Inference: Erdös-Rényi **Mix**ture for **Net**works (Daudin et al., 2008)

- Spread the nodes into $\mathcal{Q}$ classes;
- Connexion probabilities depends upon node classes:

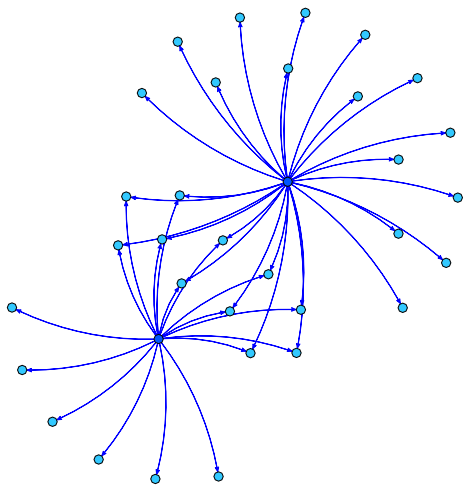
$$\mathbb{P}(i \to j | i \in \text{class } q, j \in \text{class } \ell) = \pi_{q\ell}.$$

- Build $P_{\mathbf{Z}} \propto 1 - \pi_{q\ell}$.

Algorithm

Suppose you want to recover a clustered network:

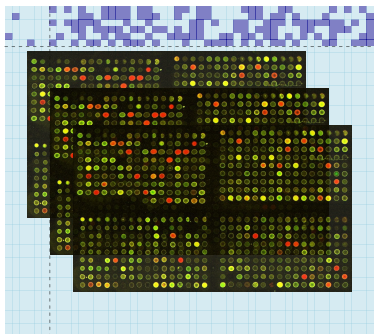
Target Adjacency Matrix



Target Network

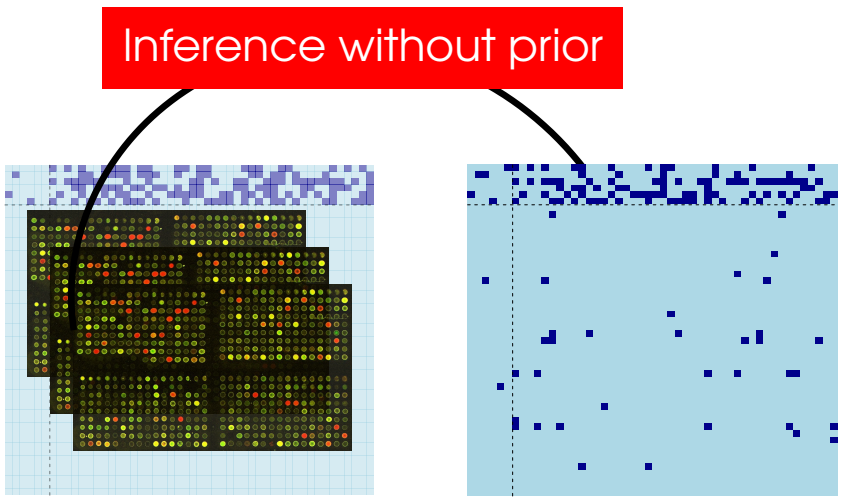
Algorithm

Start with microarray data

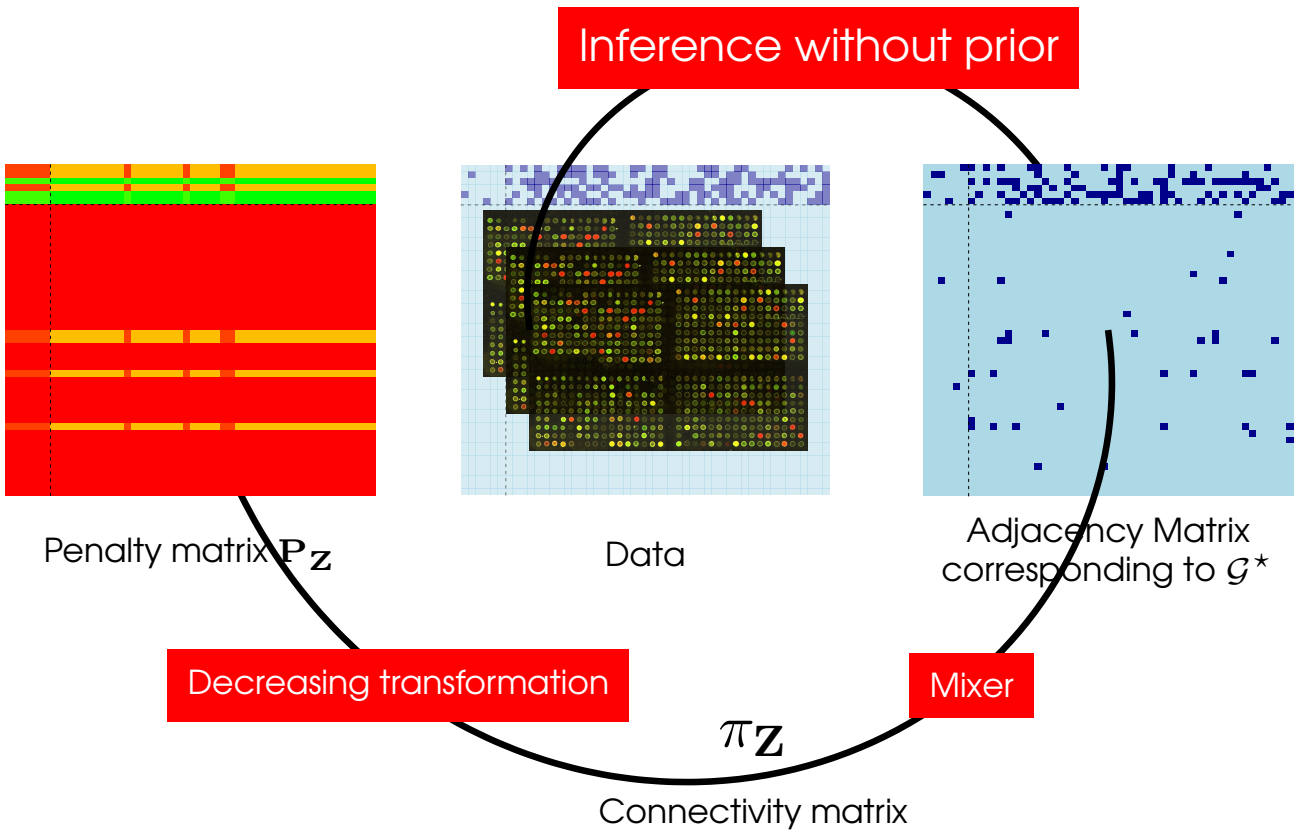


Data

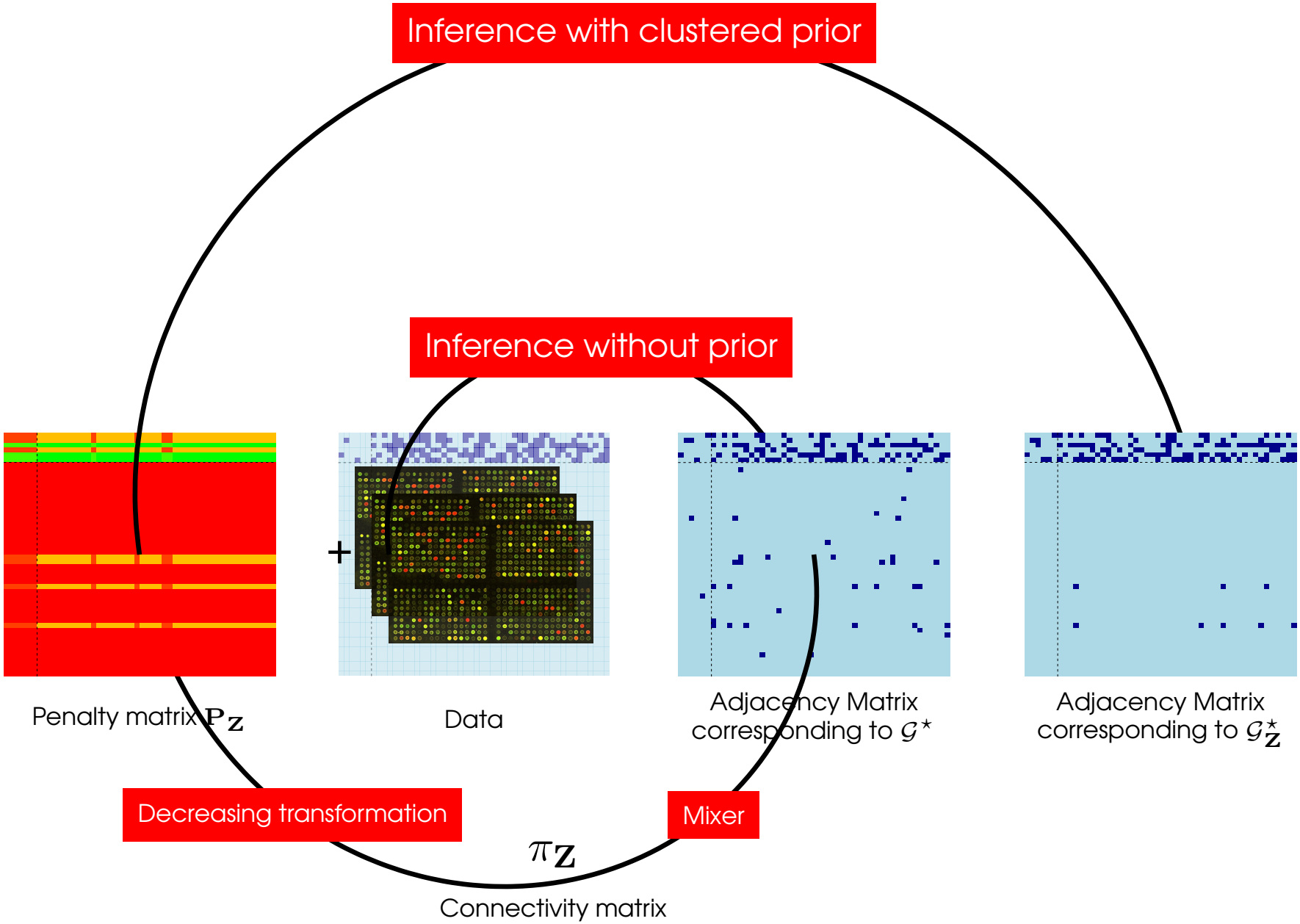
Algorithm


Inference without prior
Data
Adjacency Matrix
corresponding to $\mathcal{G}^\star$

Algorithm


Inference without prior
Penalty matrix $\mathbf{P_Z}$
Data
Adjacency Matrix corresponding to $\mathcal{G}^\star$
Decreasing transformation
Mixer
$\pi_{\mathbf{Z}}$
Connectivity matrix

Algorithm


Inference with clustered prior
Inference without prior
Penalty matrix $\mathbf{P_Z}$
Data
Adjacency Matrix corresponding to $\mathcal{G}^\star$
Adjacency Matrix corresponding to $\mathcal{G}^\star_{\mathbf{Z}}$
Decreasing transformation
Mixer
$\pi_{\mathbf{Z}}$
Connectivity matrix

Degrees of freedom of the Lasso (Zou et al. 2008)

$$\mathrm{df}(\hat{\beta}^{\lambda}) = \sum_{k} \mathbf{1}(\hat{\beta}_k^{\lambda} \neq 0)$$

Straightforward extensions to the graphical framework

$$\mathrm{BIC}(\lambda) = \mathcal{L}(\hat{\mathbf{\Theta}}_{\lambda}; \mathbf{X}) - \mathrm{df}(\hat{\mathbf{\Theta}}_{\lambda}) \frac{\log n}{2}$$

$$\mathrm{AIC}(\lambda) = \mathcal{L}(\hat{\mathbf{\Theta}}_{\lambda}; \mathbf{X}) - \mathrm{df}(\hat{\mathbf{\Theta}}_{\lambda})$$

- Rely on asymptotic approximations,
  but still relevant on simulated small samples.

Inference methods

1. Lasso *(Tibshirani)*
2. Adaptive Lasso *(Zou et al.)*

   Weights inversely proportional to an initial Lasso estimate.
3. *KnwCl*

   Weights structured according to true clustering.
4. *InfCl*

   Weights structured according to inferred clustering.
5. Renet-VAR *(Shimamura et al.)*

   Edge estimation based on a recursive elastic net.
6. G1DBN *(Lèbre et al.)*

   Edge estimation based on dynamic Bayesian networks followed by statistical testing of edges.
7. Shrinkage *(Opgen-Rhein et al.)*

   Edge estimation based on shrinkage followed by multiple testing local false discovery rate correction.

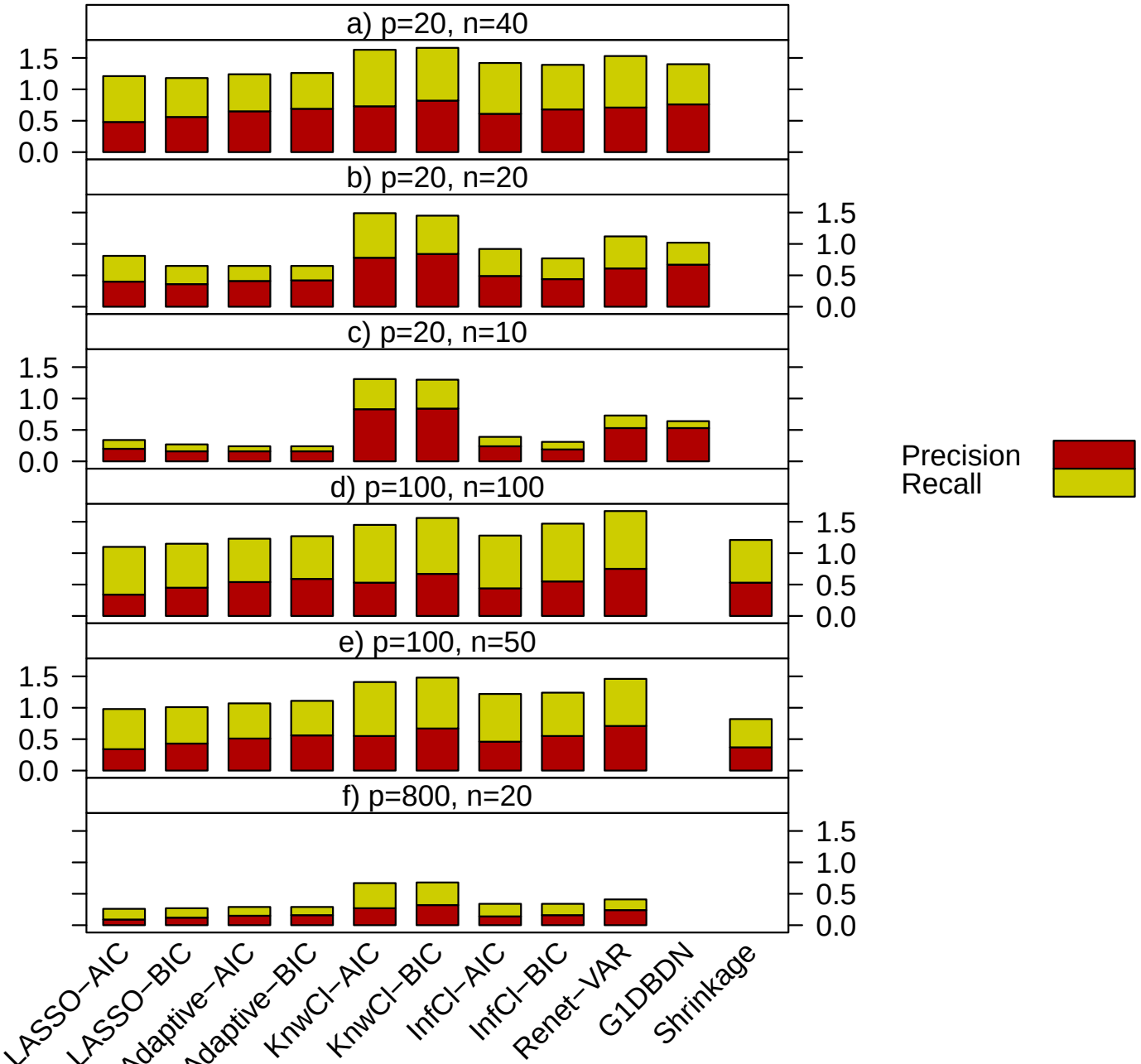
Simulations: time-course data with star-pattern

Simulation settings

- 2 classes, hubs and leaves, with proportions $\boldsymbol{\alpha} = (0.1, 0.9)$,
- $K = 2p$ edges, among which:
  - 85% from hubs to leaves,
  - 15% between hubs.

p genes	n arrays	samples
20	40	500
20	20	500
20	10	500
100	100	200
100	50	200
800	20	100

Simulations: time-course data with star-pattern



Reasonnable computing time

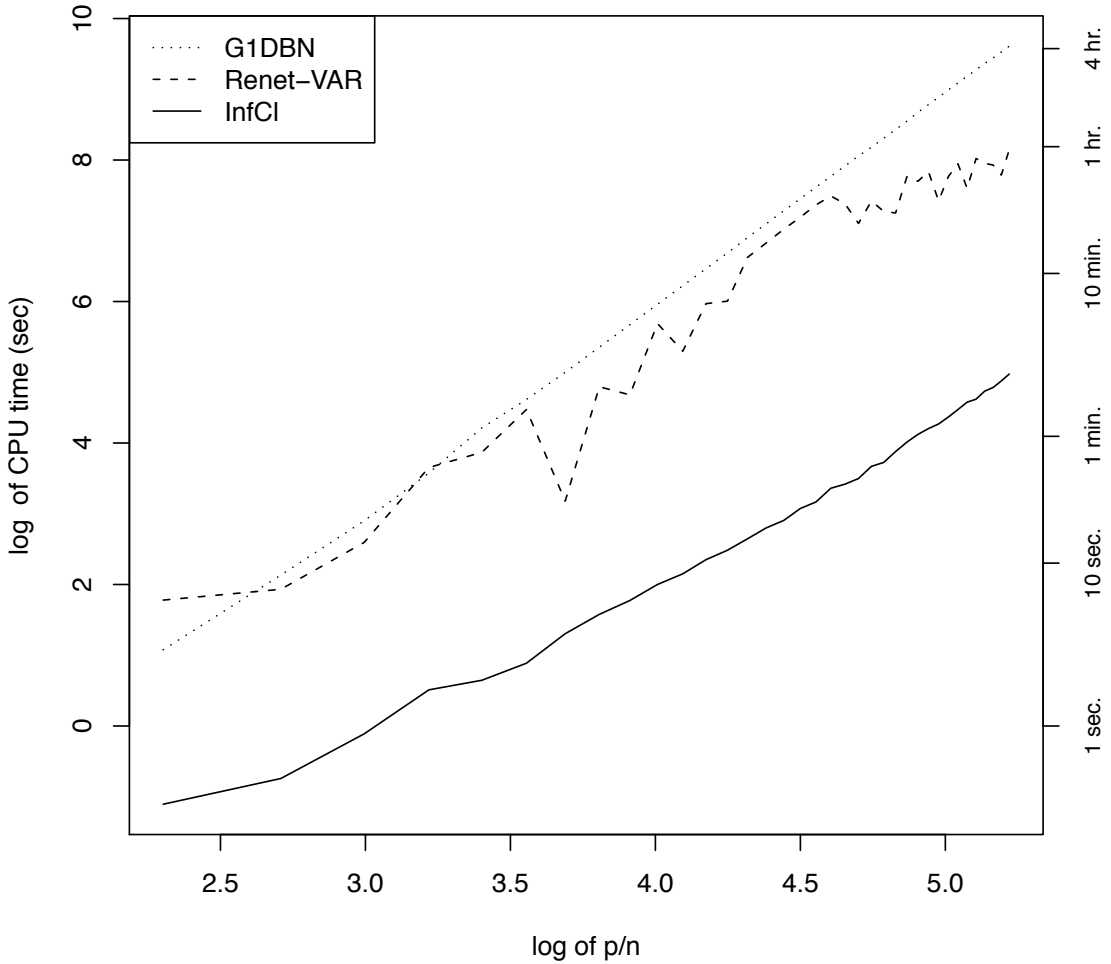


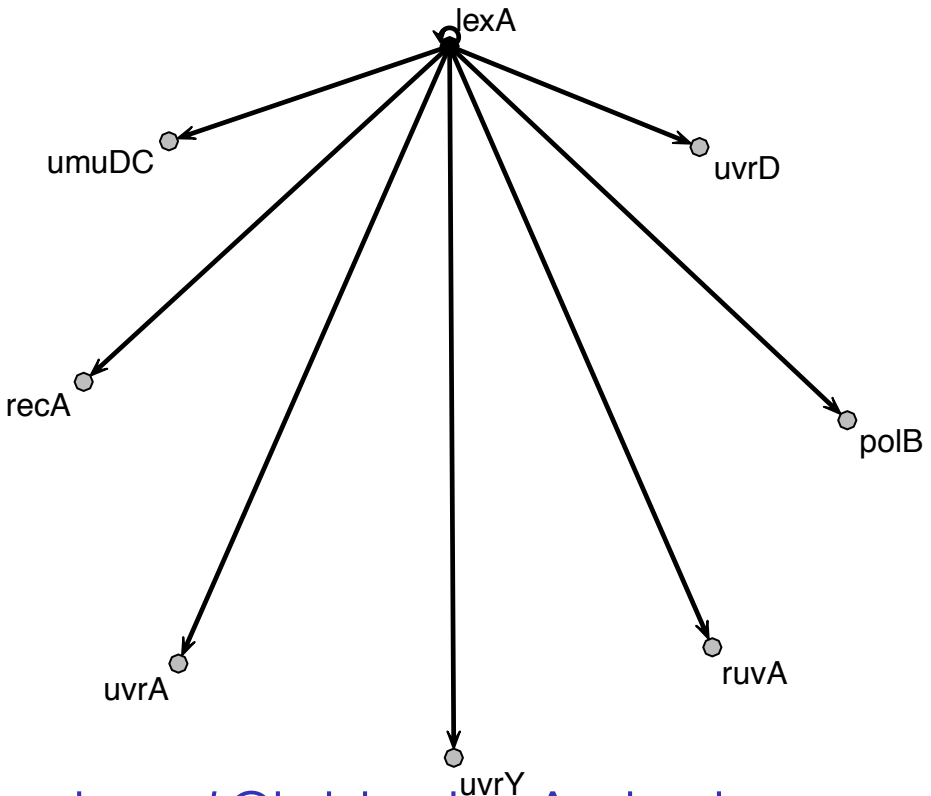
Figure: Computing times on the log-log scale for Renet-VAR, G1DBN and InfCI (including inference of classes). Intel Dual Core 3.40 GHz processor.

E.coli S.O.S data

Assigning numbers to the arrows: Parameterizing a gene regulation network by using accurate expression kinetics

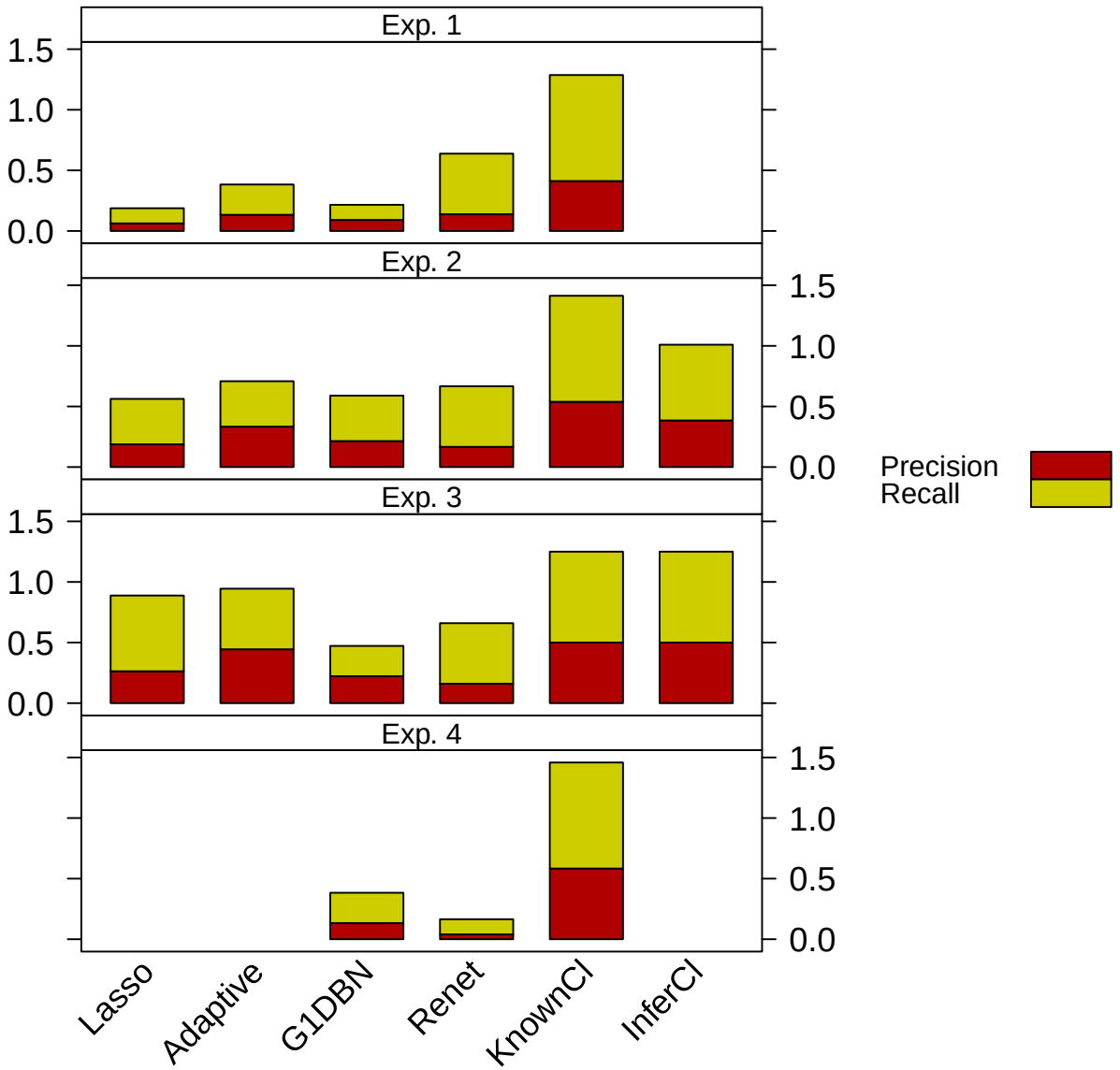
Michal Ronen[†], Revital Rosenberg[†], Boris I. Shraiman[‡], and Uri Alon[†§¶]

- 8 major genes involved in the S.O.S. response
- 50 time points



E. coli S.O.S DNA repair network

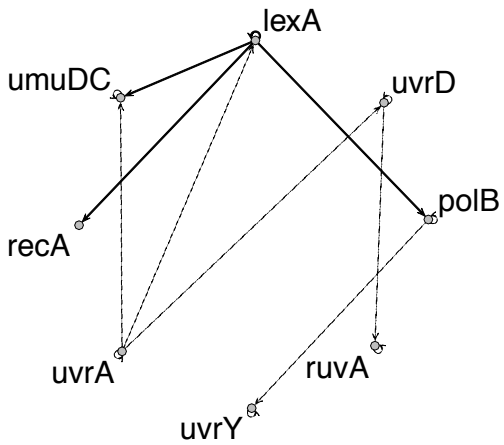
Precision and Recall rates



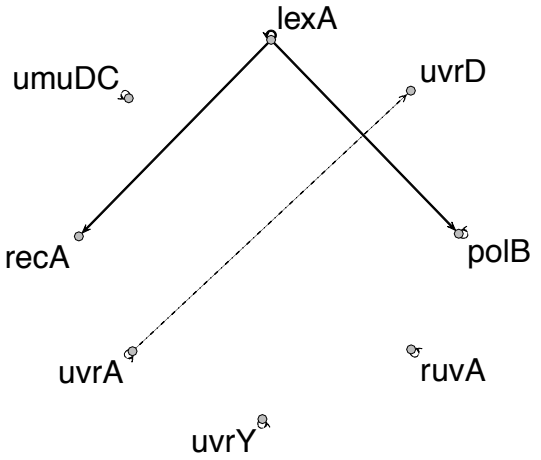
E. coli S.O.S DNA repair network

Inferred networks

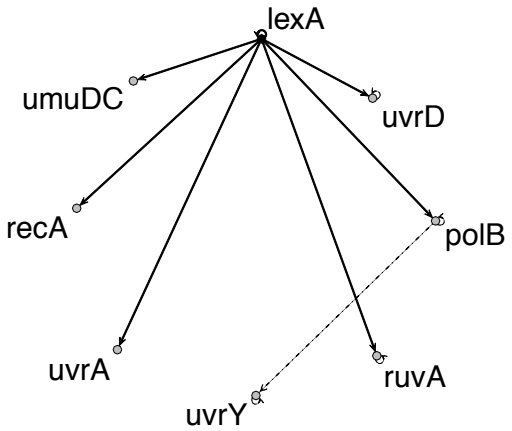
Lasso



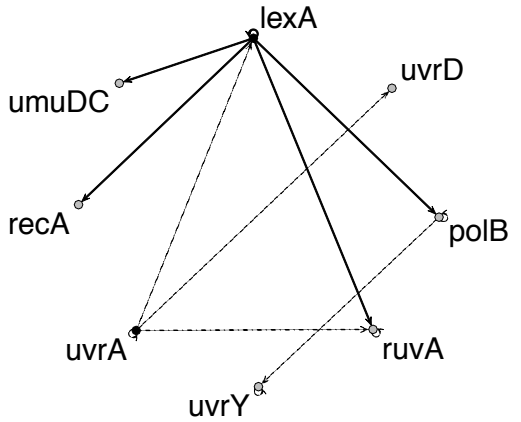
Adaptive-Lasso



Known Classification



Inferred Classification



Conclusions

To sum-up

- cluster-driven inference of gene regulatory networks from time-course data,
- expert-based or inferred latent structure,
- embedded in the `SIMoNe` `R` package along with similar algorithms dealing with steady-state or multitask data.

Perspectives

- inference of truely dynamic networks,
- use of additive biological information to refine the inference,
- comparison of inferred networks.

Publications


Ambroise, Chiquet, Matias, 2009.
Inferring sparse Gaussian graphical models with latent structure
Electronic Journal of Statistics, 3, 205-238.


Chiquet, Smith, Grasseau, Matias, Ambroise, 2009.
SIMoNe: Statistical Inference for MOdular NEtworks *Bioinformatics*, 25(3), 417-418.


Charbonnier, Chiquet, Ambroise, 2010.
Weighted-Lasso for Structured Network Inference from Time Course Data., *SAGMB*, 9.


Chiquet, Grandvalet, Ambroise, 2010. *Statistics and Computing*.
Inferring multiple Gaussian graphical models.

Publications

- Ambroise, Chiquet, Matias, 2009.
  Inferring sparse Gaussian graphical models with latent structure
  Electronic Journal of Statistics, 3, 205-238.
- Chiquet, Smith, Grasseau, Matias, Ambroise, 2009.
  SIMoNe: Statistical Inference for MOdular NEtworks *Bioinformatics*, 25(3), 417-418.
- Charbonnier, Chiquet, Ambroise, 2010.
  Weighted-Lasso for Structured Network Inference from Time Course Data., *SAGMB*, 9.
- Chiquet, Grandvalet, Ambroise, 2010. *Statistics and Computing*.
  Inferring multiple Gaussian graphical models.
- Working paper: Chiquet, Charbonnier, Ambroise, Grasseau.
  SIMoNe: An R package for inferring Gausssian networks with latent structure, *Journal of Statistical Softwares*.
- Working paper: Chiquet, Grandvalet, Ambroise, Jeanmougin.
  Biological analysis of breast cancer by multitasks learning.