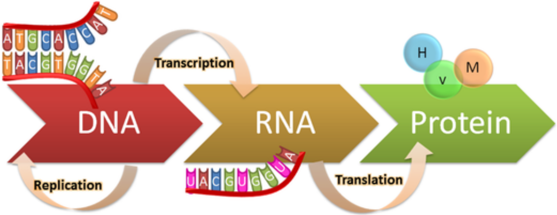
Network inference in genomics under censoring

Veronica Vinciotti
Brunel University London

Leeds, 27 February 2019

Joint work with Luigi Augugliaro and Antonino Abbruzzo

Expression Data: Complex Data from Different Platforms



- A number of platforms to measure expression (mRNA) levels:
  - microarray hybridization
  - massively parallel/next-generation sequencing (RNA-seq)
  - quantitative real-time reverse transcription-PCR (RT-qPCR)
- Observations are on nodes/variables, *not* on *edges*/relationships.
- Typically many variables, few units ("$\boldsymbol{p} >> \boldsymbol{n}$")

Aim: Recover/infer the underlying regulatory network from data

Sparse Gaussian Graphical Models

A popular tool for inference of networks from biological data.

GGM in genomics:

- $\boldsymbol{X} = (X_1, \ldots, X_p)^\top$: a p-dimensional vector of random variables
- A graph $G = (\Gamma, E)$, where Γ is the set of p genes and $E \subset \Gamma \times \Gamma$ the set of genomic interactions
- A normality assumption: $\boldsymbol{X} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with density

$$\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Theta}) = (2\pi)^{-p/2} |\boldsymbol{\Theta}|^{1/2} \exp\{-1/2(\boldsymbol{x} - \boldsymbol{\mu})^\top \boldsymbol{\Theta}(\boldsymbol{x} - \boldsymbol{\mu})\}.$$

- The precision matrix $\boldsymbol{\Theta} = \boldsymbol{\Sigma}^{-1}$ provides the structure of the conditional independence graph (non-zeros $\leftrightarrow$ edges)
- If $p > n$ and the network is expected to be sparse, $\boldsymbol{\Theta}$ can be estimated under an ℓ_1 penalty.
- Friedman et al. (2008) developed an efficient computational algorithm to perform ℓ_1 optimization (graphical lasso).

Various Extensions

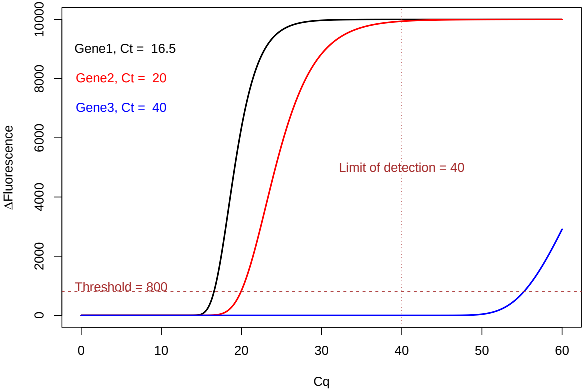
Various extensions to the graphical lasso have been proposed for different types of data:

- Hierarchical graphical models
- Dynamic graphical models
- Copula graphical models
- ...

This talk: Sparse Gaussian graphical models under missing data

- Missing-at-Random (Städler and Bühlmann, 2012)
- Censoring (Augugliaro, Abbruzzo and Vinciotti, 2018)

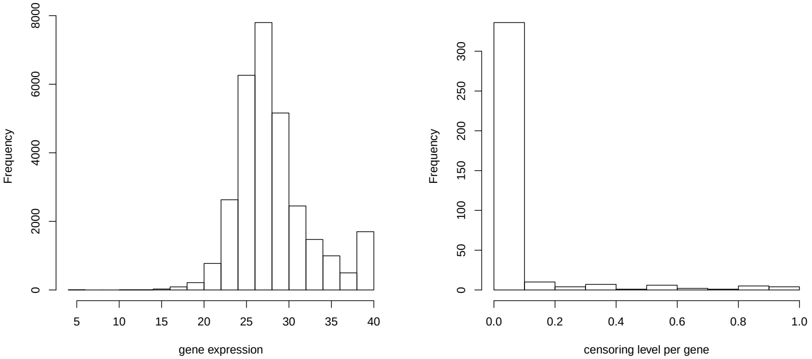
Motivation: qPCR data are censored



- Repeated cycles of DNA amplification followed by expression measurements, with a max of (typically) 40 cycles.
- The cycle at which expression reaches a fixed threshold is reported.

Example: Multidrug Resistance Gene Expression

376 multidrug resistance genes in 80 tumor specimens collected at initial surgery to debulk primary serous carcinoma (Gillet et al 2012)



Here the data are right-censored, but we will develop the method under general censoring mechanisms.

The Censoring Mechanism

Let $\boldsymbol{l} = (l_1, \dots, l_p)^\top$ and $\boldsymbol{u} = (u_1, \dots, u_p)^\top$, with $l_h < u_h$ for $h = 1, \dots, p$ the left and right censoring, respectively.

So X_h is observed if it is inside the interval $[l_h, u_h]$, censored from below if $X_h < l_h$ or censored from above if $X_h > u_h$.

Let $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ encode the censoring patterns, with hth element given by

$$R(X_h; l_h, u_h) = I(X_h > u_h) - I(X_h < l_h),$$

where $I(\cdot)$ is the indicator function.

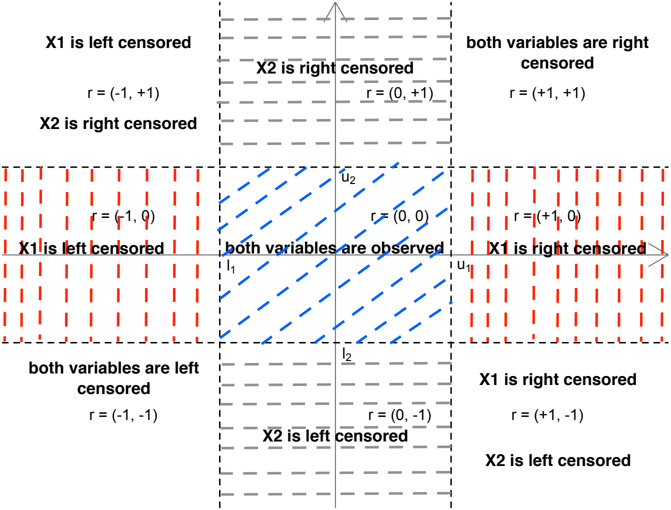
$R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ is a discrete random vector with support set $\{-1, 0, 1\}^p$ and

$$P(R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}) = \int_{D_{\boldsymbol{r}}} \phi(\boldsymbol{x}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x},$$

where $D_{\boldsymbol{r}} = \{\boldsymbol{x} \in \mathcal{R}^p : R(\boldsymbol{x}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r}\}$ and ϕ the density of $\boldsymbol{X}$.

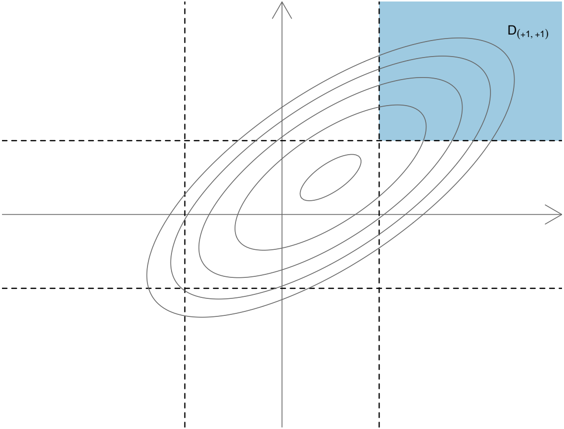
The Censoring Mechanism: Simple Case

If $p = 2$, then $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ assumes values $\boldsymbol{r} \in \{-1, 0, 1\}^2$.



The Censoring Mechanism: Simple Case

If both X_1 and X_2 are right censored, then



$$P(R(\boldsymbol{X};\boldsymbol{l},\boldsymbol{u})=(1,1))=\int_{D_{\boldsymbol{r}}}\phi(\boldsymbol{x};\boldsymbol{\mu},\Theta)\mathrm{d}\boldsymbol{x}=\int_{u_1}^{+\infty}\int_{u_2}^{+\infty}\phi(\boldsymbol{x};\boldsymbol{\mu},\Theta)\mathrm{d}x_1\mathrm{d}x_2.$$

The Density Function under Censoring

Denote with $\boldsymbol{o} = \{h \in \mathcal{I} : r_h = 0\}$, where $\mathcal{I} = \{1, \ldots, p\}$. Then the subvector of the non-censored data in $\boldsymbol{x}$ is denoted by $\boldsymbol{x_o} = (x_h)_{h \in \boldsymbol{o}}$ and, consequently, the vector of the observed data is $(\boldsymbol{x_o}^\top, \boldsymbol{r}^\top)^\top$.

The joint probability distribution of $\{\boldsymbol{X}_o^\top, R(\boldsymbol{X}, \boldsymbol{l}, \boldsymbol{u})\}$ is obtained by integrating $\boldsymbol{X_c}$ out of the joint distribution of $\{\boldsymbol{X}^\top, R(\boldsymbol{X}, \boldsymbol{l}, \boldsymbol{u})\}$, i.e.

$$\begin{aligned} \varphi(\boldsymbol{x}_o, \boldsymbol{r}; \boldsymbol{\mu}, \Theta) &= \int \phi(\boldsymbol{x}_o, \boldsymbol{x}_{c^-}, \boldsymbol{x}_{c^+}; \boldsymbol{\mu}, \Theta) P(R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u}) = \boldsymbol{r} \,|\, \boldsymbol{X} = \boldsymbol{x}) \mathrm{d}\boldsymbol{x}_{\boldsymbol{c}^-} \mathrm{d}\boldsymbol{x}_{\boldsymbol{c}^+} \\ &= \left\{ \int_{D_{\boldsymbol{c}}} \phi(\boldsymbol{x_o}, \boldsymbol{x_c}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x_c} \right\} I(\boldsymbol{l_o} \leq \boldsymbol{x_o} \leq \boldsymbol{u_o}), \end{aligned}$$

where

$$\boldsymbol{c} = \underbrace{\{h \in \mathcal{I} : r_h = -1\}}_{\boldsymbol{c}^-} \cup \underbrace{\{h \in \mathcal{I} : r_h = +1\}}_{\boldsymbol{c}^+}, \; D_{\boldsymbol{c}} = (-\infty, \boldsymbol{l}_{\boldsymbol{c}^-}] \times [\boldsymbol{u}_{\boldsymbol{c}^+}, +\infty).$$

The Censored Gaussian Graphical Model

Definition

Let $\boldsymbol{X}$ be a p-dimensional random vector following a multivariate Gaussian distribution whose density $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \Theta)$ factorizes according to an undirected graph $G = \{V, E\}$ and let $R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u})$ be a p-dimensional random censoring-data indicator defined by the censoring values $\boldsymbol{l}$ and $\boldsymbol{u}$.
The censored Gaussian Graphical Model (cGGM) is defined to be the set

$$\{\boldsymbol{X}, R(\boldsymbol{X}; \boldsymbol{l}, \boldsymbol{u}), \varphi(\boldsymbol{x_o}, \boldsymbol{r}; \boldsymbol{\mu}, \Theta), G\}.$$

This definition includes:

- the GGM (Lauritzen, 1996), if $\boldsymbol{l} = -\infty, \boldsymbol{u} = +\infty$,
- the left censored GGM, if $\boldsymbol{l} < \infty, \boldsymbol{u} = +\infty$,
- the right censored GGM, if $\boldsymbol{l} = -\infty, \boldsymbol{u} < \infty$,
- the censored GGM, if $\boldsymbol{l} < \infty, \boldsymbol{u} < \infty$.

Inference for cGGM

Consider a sample of n independent observations drawn from the cGGM. The observed log-likelihood function can be written as

$$\ell(\boldsymbol{\mu}, \Theta) = \sum_{i=1}^{n} \log \int_{D_{c_i}} \phi(\boldsymbol{x}_{io_i}, \boldsymbol{x}_{ic_i}; \boldsymbol{\mu}, \Theta) \mathrm{d}\boldsymbol{x}_{ic_i} = \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{i\boldsymbol{o}_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta),$$

where $\boldsymbol{o}_i = \{h \in \mathcal{I} : r_{ih} = 0\}$, $\boldsymbol{r}_i$ is the realization of $R(\boldsymbol{X}_i; \boldsymbol{l}_i, \boldsymbol{u}_i)$.

Under a lasso penalty, the estimator for a ℓ_1-penalized cGGM is

$$\{\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho}\} = \arg \max_{\boldsymbol{\mu}, \Theta \succ 0} \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|.$$

cGGMs Inference: Algorithm

Theorem

Necessary and sufficient conditions for $\{\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho\}$ *to be the solution of the maximization problem*

$$\max_{\boldsymbol{\mu},\Theta \succ 0} \sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \boldsymbol{\mu}, \Theta) - \rho \sum_{h \neq k} |\theta_{hk}|$$

are

$$\begin{aligned} \bar{x}_h(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - \hat{\mu}_h^\rho &= 0 \\ \hat{\sigma}_{hk}^\rho(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - s_{hk}(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho) - \rho \hat{v}_{hk} &= 0 \end{aligned}$$

where $\hat{v}_{hk}$ *denotes the subgradient of the absolute value function at* $\hat{\theta}_{hk}^\rho$, *i.e.,* $\hat{v}_{hk} = sign(\hat{\theta}_{hk}^\rho)$ *if* $\hat{\theta}_{hk}^\rho \neq 0$ *and* $|\hat{v}_{hk}| \leq 1$ *if* $\hat{\theta}_{hk}^\rho = 0$.

$\bar{x}_h(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho)$ and $s_{hk}(\hat{\boldsymbol{\mu}}^\rho, \widehat{\Theta}^\rho)$: 1st and 2nd moments of a truncated Gaussian distribution.

cGGM Inference: Moments of Truncated Gaussian

For any $i = 1, \ldots, n$, and $h, k = 1, \ldots .p$, let

$$x_{i,h}(\boldsymbol{\mu}, \Theta) = \begin{cases} x_{ih} & \text{if } r_{ih} = 0 \\ E_{c_i|o_i}(X_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}) & \text{otherwise,} \end{cases}$$

$$x_{i,hk}(\boldsymbol{\mu}, \Theta) = \begin{cases} x_{ih}x_{ik} & \text{if } r_{ih} = 0 \text{ and } r_{ik} = 0 \\ x_{ih}E_{c_i|o_i}(X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}) & \text{if } r_{ih} = 0 \text{ and } r_{ik} \neq 0 \\ E_{c_i|o_i}(X_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})x_{ik} & \text{if } r_{ih} \neq 0 \text{ and } r_{ik} = 0 \\ E_{c_i|o_i}(X_{ih}X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}) & \text{if } r_{ih} \neq 0 \text{ and } r_{ik} \neq 0, \end{cases}$$

where $E_{c_i|o_i}(\cdot \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$ denotes the expected value computed using the conditional distribution of $\boldsymbol{X}_{ic_i}$ given $\boldsymbol{x}_{io_i}$ truncated over D_{c_i}. Then

$$\bar{x}_h(\boldsymbol{\mu}, \Theta) = \frac{\sum_{i=1}^{n} x_{i,h}(\boldsymbol{\mu}, \Theta)}{n}; \quad \bar{\boldsymbol{x}}(\boldsymbol{\mu}, \Theta) = \{\bar{x}_1(\boldsymbol{\mu}, \Theta), \ldots, \bar{x}_p(\boldsymbol{\mu}, \Theta)\}^\top,$$

$$s_{hk}(\boldsymbol{\mu}, \Theta) = \frac{\sum_{i=1}^{n} x_{i,hk}(\boldsymbol{\mu}, \Theta)}{n} - \bar{x}_h(\boldsymbol{\mu}, \Theta)\bar{x}_k(\boldsymbol{\mu}, \Theta); \; S(\boldsymbol{\mu}, \Theta) = \{s_{hk}(\boldsymbol{\mu}, \Theta)\}$$

cGGM: essentially glasso within an EM algorithm

E-step

Denoting by $\{\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}\}$ an initial estimate, compute the conditional expectations $x_{i,h}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$ and $x_{i,hk}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})$, for $i = 1, \ldots, n$.

M-step

Estimate Θ by maximizing the following objective function

$$Q(\Theta \mid \widehat{\Theta}^{\rho}_{ini}) = \log\det\Theta - \mathrm{tr}\{\Theta S(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini})\} - \rho \sum_{h,k} |\theta_{hk}|.$$

This leads to

$$\begin{aligned} \bar{x}_h(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}) - \hat{\mu}^{\rho}_{h} &= 0 \\ \hat{\sigma}^{\rho}_{hk}(\hat{\boldsymbol{\mu}}^{\rho}, \widehat{\Theta}^{\rho}) - s_{hk}(\hat{\boldsymbol{\mu}}^{\rho}_{ini}, \widehat{\Theta}^{\rho}_{ini}) - \rho\hat{v}_{hk} &= 0 \end{aligned}$$

which are the stationary conditions of a standard graphical lasso problem.

cGGM: Computational Cost

Although the M-step can be efficiently solved using graphical lasso implementations, the calculations of moments of a truncated normal can be time consuming.

Following Guo et al (2015), we consider a mean field approximation:

Approximate EM

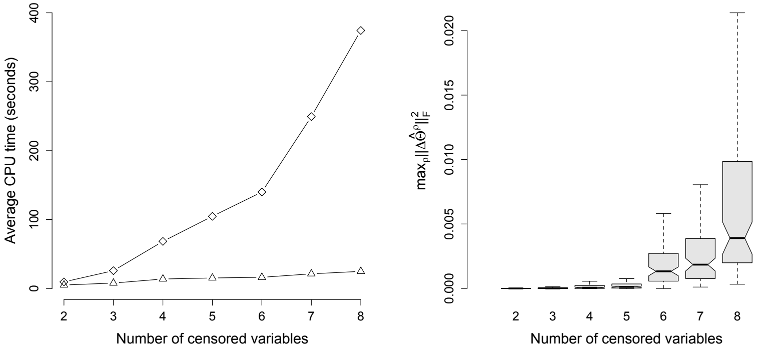
$$E_{c_i|o_i}(X_{ih}X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i}) \approx E_{c_i|o_i}(X_{ih} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})E_{c_i|o_i}(X_{ik} \mid \boldsymbol{X}_{ic_i} \in D_{c_i})$$

This reduces computational time dramatically, as only conditional mean and variance are needed.

Simulation 1: Computational Cost of Approximate EM

$p = 10$, $n = 100$, P(Censoring)=0.25 (marginally) in a randomly drawn set $\mathcal{D}$ of the 10 variables.

x-axis: $|\mathcal{D}|$, y-axis: largest Frobenius distance between Θ estimated using full and approximate EM.



Simulation 2: Comparison with Existing Methods

1. Fix n, p, a threshold level $k = 40$ and a censoring level c

2. Generate Θ using `huge.generator` (with varying sparsity levels) and $\boldsymbol{\mu} = (\mathbf{40}_c, runif(p-c, 10, 35))$, i.e. P(Censoring)=0.5 marginally for the censored variables

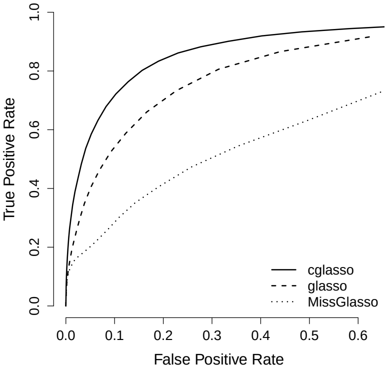
3. Generate data $X \sim N(\boldsymbol{\mu}, \Theta^{-1})$ (`mvrnorm`) and transform the data into censored data $X[X > k] = k$

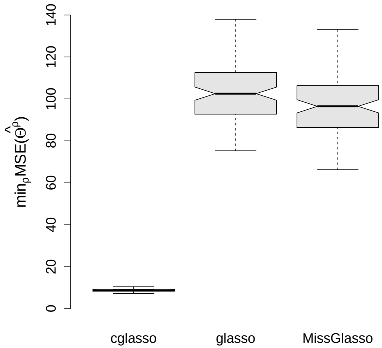
4. For a set of ρ values, compare our method (`cglasso`) with
   - `glasso`: leave the data as they stand and use `glassopath` to find an estimator of Θ
   - `missGlasso`: treat the censored as missing at random and estimate both $\boldsymbol{\mu}$ and Θ as a function of ρ (Städler and Bühlmann, 2012)

5. Repeat each simulation 100 times

Case 1: Different Levels of Censoring

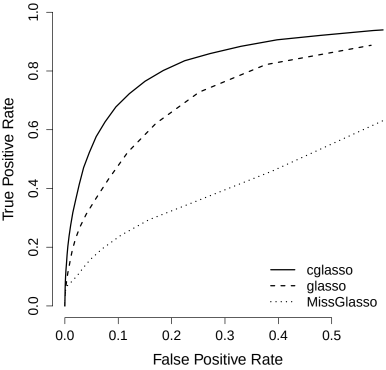
$n = 100,\ p = 50,\ P(\theta_{hk} \neq 0) = 0.06,$ 50% censoring level

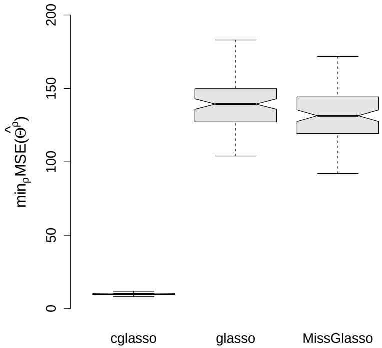




Case 1: Different Levels of Censoring

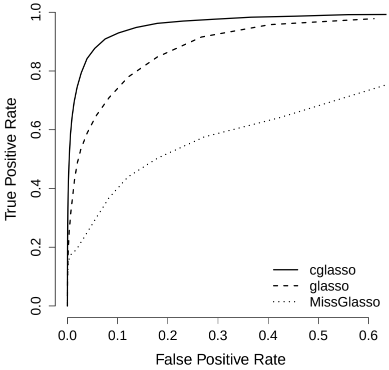
$n = 100,\ p = 50,\ P(\theta_{hk} \neq 0) = 0.06,$ 70% censoring level

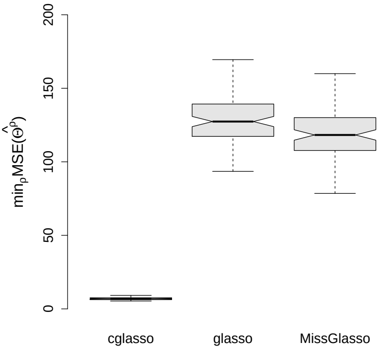




Case 2: Different Levels of Sparsity

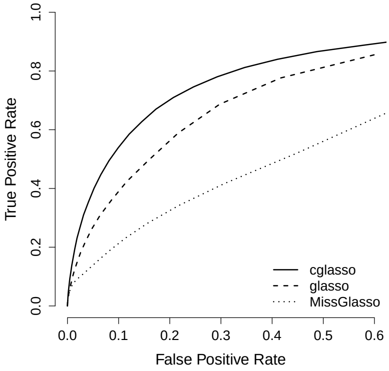
$n = 100$, $p = 50$, 60% of variables censored, $P(\theta_{hk} \neq 0) = 0.02$

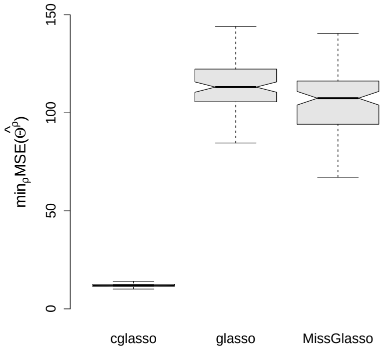




Case 2: Different Levels of Sparsity

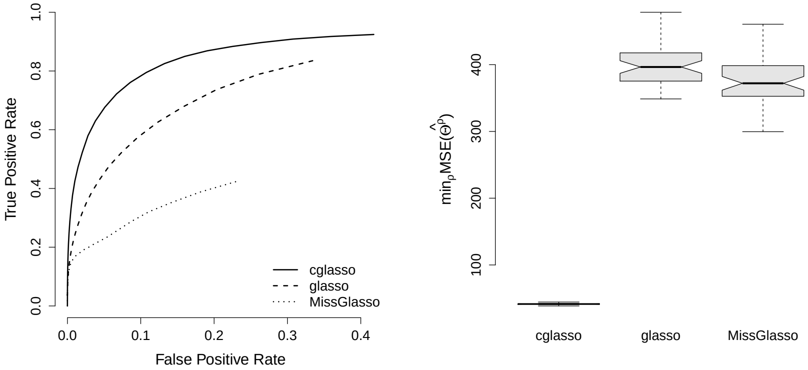
$n = 100$, $p = 50$, 60% of variables censored, $P(\theta_{hk} \neq 0) = 0.10$





Case 3: High-dimensional

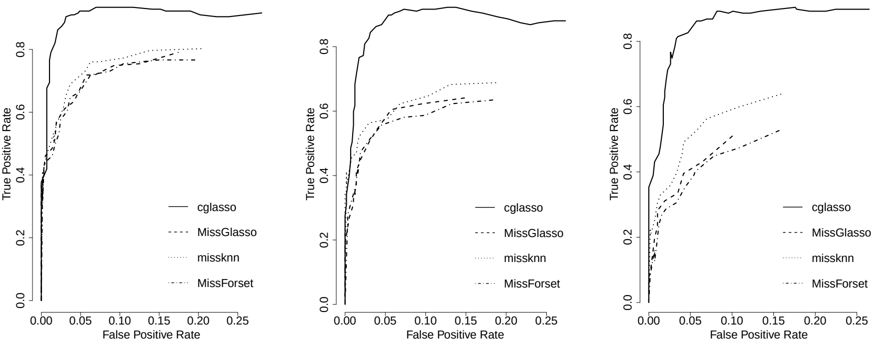
$n = 100$, $p = 200$, 50% of variables censored, $P(\theta_{hk} \neq 0) = 0.015$



Simulation 3: “Real” Biological Data

Expression data on *Arabidopsis thaliana* from Wille et al (2004):

- $n = 118$ experiments on $p = 39$ genes
- Fully observed, but we create a dataset where observations are made artificially censored (3 cases: 10%, 20%, 30%)



(a) 10% (b) 20% (c) 30%

Model Selection: Extended BIC

The tuning parameter ρ controls the sparsity of the network. Using the eBIC, one needs to calculate:

$$\mathrm{BIC}_{\gamma}(\widehat{\mathcal{E}}^{\rho}) = -2\sum_{i=1}^{n} \log \varphi(\boldsymbol{x}_{io_i}, \boldsymbol{r}_i; \hat{\boldsymbol{\mu}}, \widehat{\Theta}(\widehat{\mathcal{E}}^{\rho})) + a(\rho)(\log n + 4\gamma \log p),$$

where

$\widehat{\Theta}(\widehat{\mathcal{E}}^{\rho})$: MLE of the Gaussian graphical model specified by $\widehat{\mathcal{E}}^{\rho} = \{(\hat{\theta}^{\rho}_{hk} \neq 0\}$
$a(\rho)$: number of nonzero off-diagonal estimates of $\widehat{\Theta}^{\rho}$.

Since the log-likelihood is not a direct output of the EM-algorithm, we use the following approximate measure (Ibrahim et al, 2008):

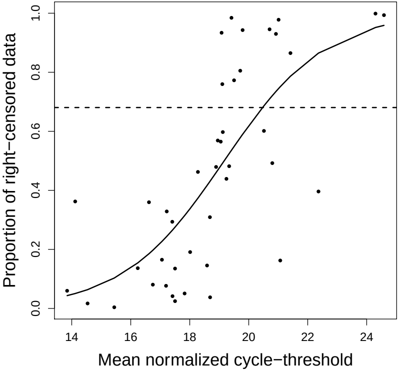
$$\overline{\mathrm{BIC}}_{\gamma}(\widehat{\mathcal{E}}^{\rho}) = -n[\log\det\widehat{\Theta}^{\rho} - \mathrm{tr}\{\Theta S(\hat{\boldsymbol{\mu}}, \widehat{\Theta}(\widehat{\mathcal{E}}^{\rho}))\}] + a(\rho)(\log n + 4\gamma \log p),$$

i.e. substitute the exact log-lik with the Q-function used in the M-Step.

Mechanisms of Early Blood Development from qPCR Data

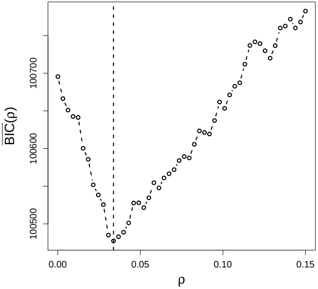
Single-cell experiments from Moignard et al (2015):

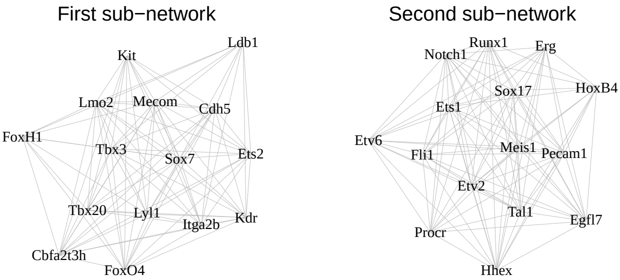
- $n = 770$ endothelial mouse cells; $p = 42$ genes (33 TFs, 9 markers)
- Threshold for censoring is set at 25
- Data normalized based on 4 housekeeping genes (Pipelers et al 2017)



Retain only genes with $< 70\%$ censoring $\rightarrow$ 30 genes for the analysis

Inferred Network has 2 Distinct Sub-networks



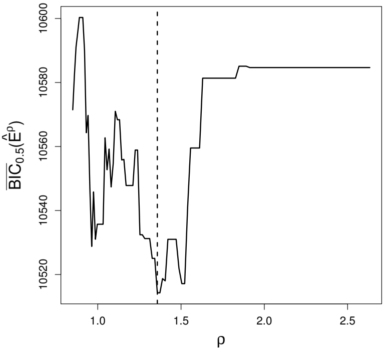

First sub−network

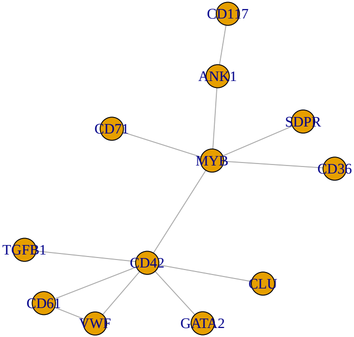
Second sub−network

Second application: $p > n$

Single-cell experiments from Psaila et al (2016):

- $n = 48$ human MK-MEP cells; $p = 87$ genes
- Threshold for censoring is set at 40
- Data normalized based on 2 housekeeping genes (Pipelers et al 2017)

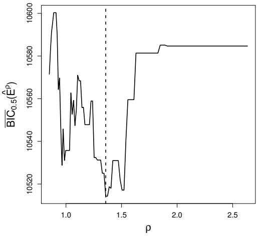


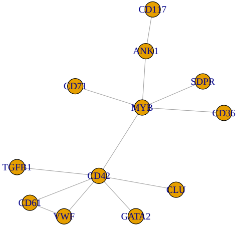


Implementation: R package cglasso

Main functions: cglasso, ebic, plot

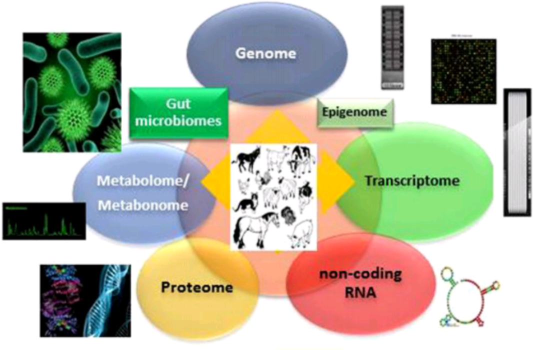
```
out <- cglasso(MKMEP, nrho = 200, rho.min.ratio = 0.35)
out.e <- ebic(out)
plot(out.e, type = "l")
out.graph <- to_graph(out, nrho = which.min(out.e$value_gof))
plot(out.graph)
```





mglasso function also available for inference under a missing-at-random mechanism (Städler and Bühlmann, 2012)

Extension: Conditional Censored Gaussian Graphical Model



Can we predict one data type from another?

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\boldsymbol{B} + \epsilon \\ \epsilon &\sim N(\mathbf{0}, \boldsymbol{\Theta}^{-1}) \end{aligned}$$

We consider closely the case of:

censored response $\mathbf{Y}$ + high dimensionality both in $\mathbf{X}$ and $\mathbf{Y}$

Conditional Censored Gaussian Graphical Model: Inference

There are two “networks” now: $\boldsymbol{B}, \boldsymbol{\Theta}$

Under censoring and sparsity, we wish to optimize

$$\sum_{i=1}^{n} \log \varphi(\boldsymbol{y}_{io_i}, \boldsymbol{r}_i, \boldsymbol{x}_i; \boldsymbol{B}, \boldsymbol{\Theta}) - \rho \sum_{h \neq k} |\theta_{hk}| - \lambda \sum_{j,l} |b_{jl}|.$$

We have developed an efficient EM algorithm that:

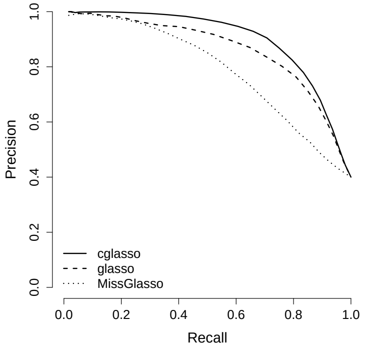
E-step: calculates summary statistics based on $\boldsymbol{B}$ and $\boldsymbol{\Theta}$
M-step: alternates estimation of $\boldsymbol{B}$ with estimation of $\boldsymbol{\Theta}$ on the residuals of the model

All good ... but two tuning parameters now (ρ, λ)...

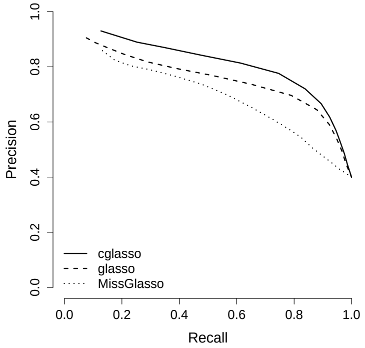
Conditional cGGM: Simulation

$n = 100$, $p = 50$, $q = 5$, $P(\theta_{hk} \neq 0) = 0.06$, 50% censoring level in $\boldsymbol{Y}$, 50% of $\boldsymbol{Y}$ variables censored, 2 non-zero values of $\boldsymbol{B}$ per row

Precision recall curves based on a 30 × 30 grid of values of ρ and λ:



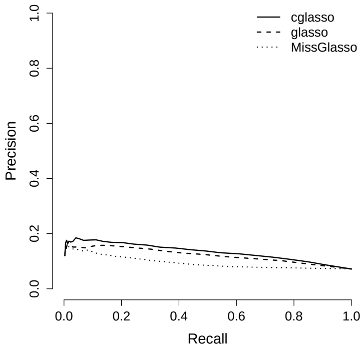
$\rho = \rho_1$ (sparse)



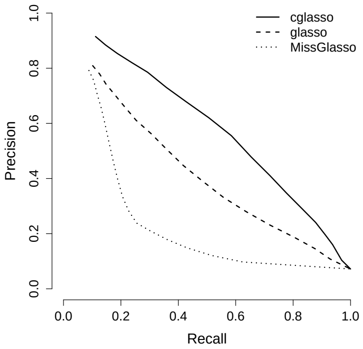
$\rho = \rho_{30}$ (dense)

Recovery of $\boldsymbol{B}$ is not too affected by estimation of the precision matrix

But the reverse is not true!



$\lambda = \lambda_1$ (sparse)



$\lambda = \lambda_{30}$ (dense)

Better estimates of $\boldsymbol{B}$ lead to better recovery of the network $\boldsymbol{\Theta}$

Related Work: Probit Models

Probit with Correlated Random Effects

$$\begin{array}{rcl} \mathbf{Y}_r^* & = & \mathbf{X}_r\beta + \mathbf{Z}_r\mathbf{u}_r + \epsilon_r, \\ \mathbf{Y}_r = 1 & \text{if} & \mathbf{Y}_r^* \geq 0, \text{ 0 otherwise,} \\ & \text{with} & \\ \mathbf{u_r} & \sim & N(0, \mathbf{\Sigma}_{G\times G}), \quad r = 1, \ldots, R. \end{array}$$

In the context of a credit risk application:

- Y: firm's default ($p \sim 60000$)
- G=13 industrial sectors (e.g. agriculture, manufacturing, ...)
- R=59 geographical regions
- Dependencies are captured at the higher level of industrial sectors

Correlated random effects → Hierarchical graphical model

The mixed model imposes block constraints on the covariance/precision matrix. In particular $\mathbf{Y}^* \sim N(\mathbf{X}\beta, \mathbf{\Sigma})$, with

$$\underset{p\times p}{\mathbf{\Sigma}} = \begin{pmatrix} \mathbf{\Sigma}_1 & 0 & \dots & 0 \\ 0 & \mathbf{\Sigma}_2 & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & \dots & 0 & \mathbf{\Sigma}_R \end{pmatrix} \quad \underset{N_r\times N_r}{\mathbf{\Sigma}_r} = \begin{pmatrix} \sigma_1 & \sigma_{12} & \dots & \sigma_{1G} \\ \sigma_{21} & \sigma_2 & \dots & \sigma_{2G} \\ \dots & \dots & \dots & \dots \\ \sigma_{G1} & \sigma_{G2} & \dots & \sigma_G \end{pmatrix}$$

where σ_{ij} are rectangular blocks of size given by # companies in sector i × # companies in sector j and N_r is the number of companies in region r.

$\mathbf{\Sigma}_r$ can be conveniently written in terms of $\mathbf{\Sigma}_G$:

$$\mathbf{\Sigma}_r = \mathbf{Z}_r \mathbf{\Sigma}_G \mathbf{Z}_r' + \mathbf{I}_{N_r}.$$

We have developed also in this case an efficient EM algorithm (Tosetti and Vinciotti (2018) arXiv: 1808.06798).

Conclusions

- Biological data from RT-qPCR data is naturally censored
- We have developed penalised censored Gaussian graphical models for network inference under censoring
- The method can be applied in the presence of any censoring
- R package `cglasso` on CRAN
- Main reference: *Augugliaro, Abbruzzo, Vinciotti (2019) L_1-Penalised Censored Gaussian Graphical Model. Biostatistics.*
- Possible extensions to multivariate regression models for integration of data from multiple sources and under different patterns of missingness

Partly funded by **COSTNET**
European Cooperation for Statistics
of Network Data Science
COST Action CA15109