Joint gene network inference with multiple samples: a bootstrapped consensual approach

Nathalie Villa-Vialaneix
http://www.nathalievilla.org
nathalie.villa@univ-paris1.fr

Réunion annuelle du groupe NETBIO
Paris, 10 septembre 2013

Joint work with Matthieu Vignes, Nathalie Viguerie and Magali SanCristobal

Outline

Framework

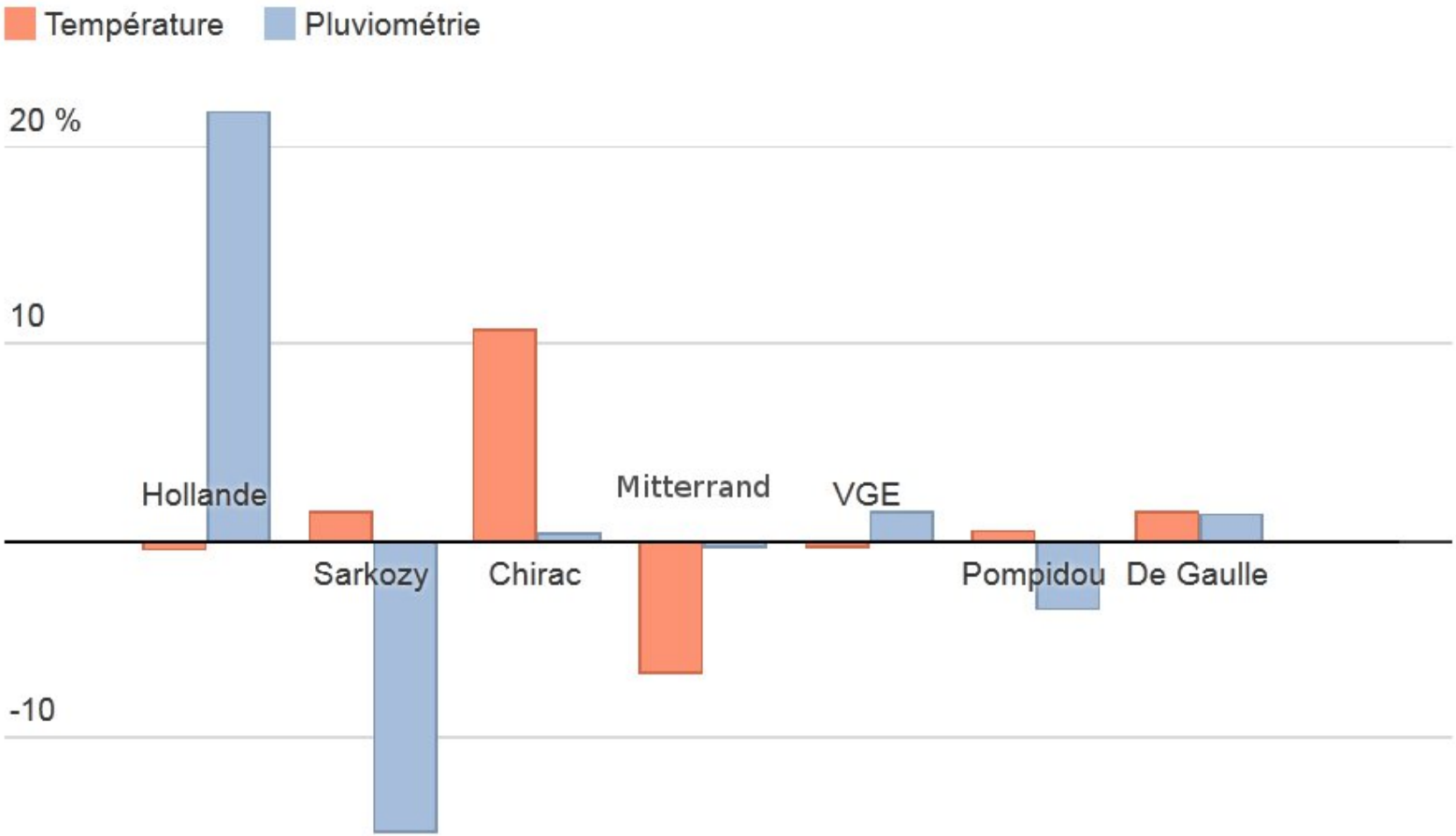
Data: large scale gene expression data

$$\underbrace{\begin{matrix}\text{individuals}\\ n \simeq 30/50\end{matrix}}\quad \underbrace{\left\{ X = \begin{pmatrix} \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & X_i^j & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \right.}_{\text{variables (genes expression), } p \simeq 10^{3/4}}$$

What we want to obtain: a graph/network with

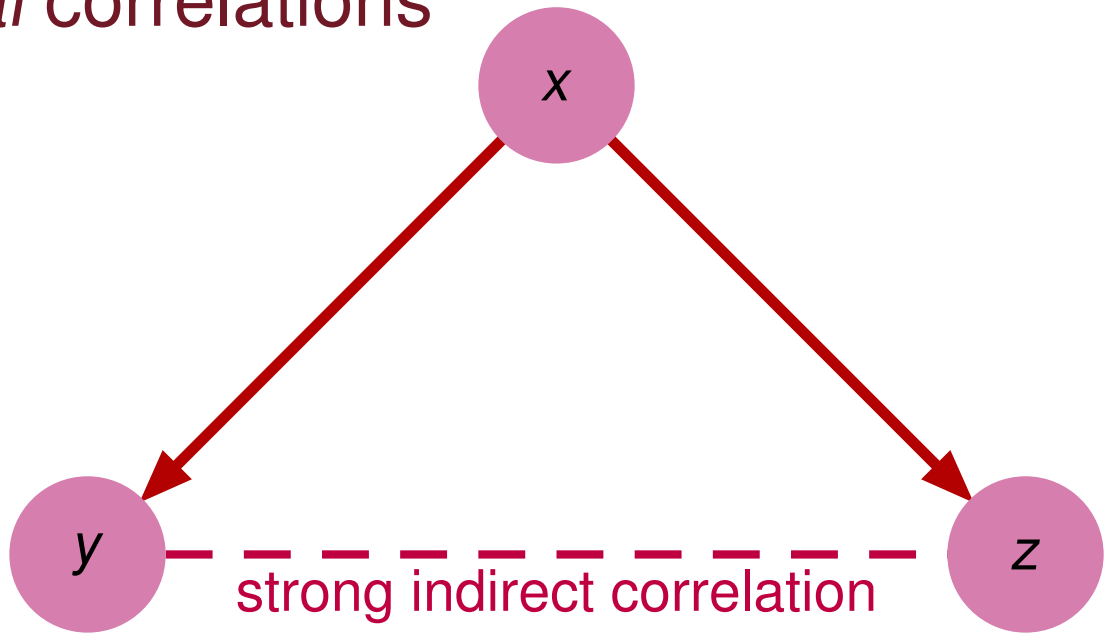
- nodes: genes;
- edges: strong links between gene expressions.

Using *partial* correlations



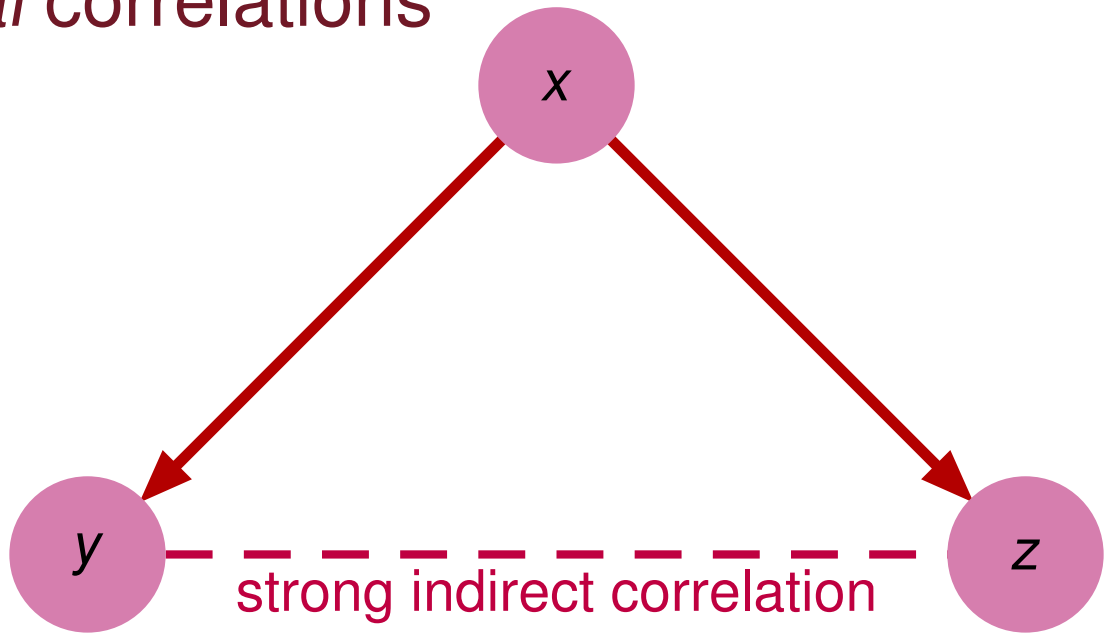
correlation is not causality...

Using *partial* correlations




```
set.seed(2807); x <- rnorm(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y) [1] 0.998826
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z) [1] 0.998751
cor(y,z) [1] 0.9971105
```

Using *partial* correlations




```
set.seed(2807); x <- rnorm(100)
y <- 2*x+1+rnorm(100,0,0.1); cor(x,y) [1] 0.998826
z <- 2*x+1+rnorm(100,0,0.1); cor(x,z) [1] 0.998751
cor(y,z) [1] 0.9971105
# Partial correlation
cor(lm(x z)$residuals,lm(y z)$residuals) [1] 0.7801174
cor(lm(x y)$residuals,lm(z y)$residuals) [1] 0.7639094
cor(lm(y x)$residuals,lm(z x)$residuals) [1] -0.1933699
```

Theoretical framework

Gaussian Graphical Models (GGM) **[Schäfer and Strimmer, 2005, Meinshausen and Bühlmann, 2006, Friedman et al., 2008]**
gene expressions: $X \sim \mathcal{N}(0, \Sigma)$
Sparse approach: partial correlations are estimated by using linear models and a sparse penalty: $\forall\, j$

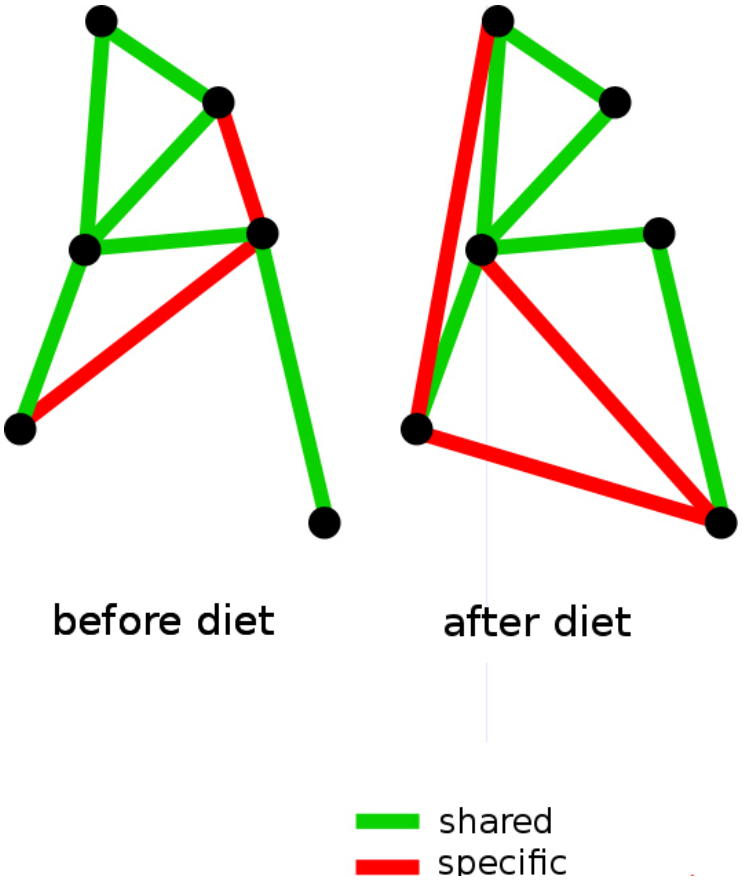
$$X^j = \beta_j^T X^{-j} + \epsilon \qquad ; \qquad \arg\max_{(\beta_{jj'})_{j'}} \left(\log \mathrm{ML}_j - \lambda \sum_{j' \neq j} |\beta_{jj'}| \right)$$

In the **Gaussian framework**: $\beta_{jj'} = -\frac{S_{jj'}}{S_{jj}}$ where $S = \Sigma^{-1}$ (concentration matrix) is related to partial correlations by $\pi_{jj'} = -\frac{S_{jj'}}{\sqrt{S_{jj} S_{j'j'}}}$.

Outline

Motivation for multiple networks inference

Pan-European project Diogenes[1] **(with Nathalie Viguerie, INSERM):** gene expressions (lipid tissues) from 204 obese women **before** and **after** a low-calorie diet (LCD).



- **Assumption**: A common functioning exists regardless the condition;
- Which genes are linked **independently from/depending on** the condition?

[1] http://www.diogenes-eu.org/; see also **[Viguerie et al., 2012]**

Naive approach: independent estimations

Notations: p genes measured in k samples, each corresponding to a specific condition: $(X_j^c)_{j=1,\dots,p} \sim \mathcal{N}(0, \Sigma^c)$, for $c = 1, \dots, k$.
For $c = 1, \dots, k$, n_c independent observations $(X_{ij}^c)_{i=1,\dots,n_c}$ and $\sum_c n_c = n$.

Independent inference

Estimation $\forall\, c = 1, \dots, k$ and $\forall\, j = 1, \dots, p$,

$$X_j^c = \mathbf{X}_{\setminus j}^c \beta_j^c + \epsilon_j^c$$

are estimated (independently) by maximizing pseudo-likelihood:

$$\mathcal{L}(S|\mathbf{X}) = \sum_c \sum_j \log \mathbb{P}\left(X_j^c | \mathbf{X}_{\setminus j}^c,\ S_j^c\right),\ S \text{ concentration matrix}$$

Related papers

Problem: previous estimation does not use the fact that the different networks should be somehow alike!

Previous proposals

- **[Chiquet et al., 2011]** replace Σ^c by $\widetilde{\Sigma}^c = \frac{1}{2}\Sigma^c + \frac{1}{2}\overline{\Sigma}$ and add a sparse penalty;
- **[Chiquet et al., 2011]** LASSO and Group-LASSO type penalties to force consistent or sign-coherent edges between conditions;
- **[Danaher et al.,]** add a sparse penalty and the penalty $\sum_{c \neq c'} \|S^c - S^{c'}\|_{L^1}$;
- **[Mohan et al., 2012]** add a group-LASSO like penalty $\sum_{c \neq c'} \sum_j \|S^c_j - S^{c'}_j\|_{L^2}$ that focuses on differences due to a few number of **nodes** only.

Consensus LASSO

Proposal: Infer multiple networks by forcing them toward a consensual network: i.e., explicitly **constraining the differences** between conditions to be under control but **with a L^2 penalty** to allow for more differences than with Group-LASSO type penalties.

Original optimization:

$$\max_{(\beta_{jk}^c)_{k\neq j, c=1,\ldots,C}} \sum_c \left(\log \mathrm{ML}_j^c - \lambda \sum_{k\neq j} |\beta_{jk}^c| \right).$$

Consensus LASSO

Proposal: Infer multiple networks by forcing them toward a consensual network: i.e., explicitly **constraining the differences** between conditions to be under control but **with a L^2 penalty** to allow for more differences than with Group-LASSO type penalties.

Original optimization:

$$\max_{(\beta_{jk}^c)_{k\neq j, c=1,\ldots,C}} \sum_c \left(\log \mathrm{ML}_j^c - \lambda \sum_{k\neq j} |\beta_{jk}^c| \right).$$

[Ambroise et al., 2009, Chiquet et al., 2011]: is equivalent to minimize p problems having dimension $k(p-1)$:

$$\frac{1}{2}\beta_j^T \widehat{\Sigma}_{\backslash j \backslash j} \beta_j + \beta_j^T \widehat{\Sigma}_{j \backslash j} + \lambda ||\beta_j||_{L^1}$$

with $\widehat{\Sigma}_{\backslash j \backslash j}$ is the block diagonal matrix $\mathbb{D}\mathrm{iag}\left(\widehat{\Sigma}^1_{\backslash j \backslash j}, \ldots, \widehat{\Sigma}^k_{\backslash j \backslash j}\right)$ and similarly for $\widehat{\Sigma}_{j \backslash j}$.

Consensus LASSO

Proposal: Infer multiple networks by forcing them toward a consensual network: i.e., explicitly **constraining the differences** between conditions to be under control but **with a L^2 penalty** to allow for more differences than with Group-LASSO type penalties.

Add a constraint to force inference toward a "consensus" β^{cons}

$$\frac{1}{2}\beta_j^T\widehat{\Sigma}_{\backslash j\backslash j}\beta_j + \beta_j^T\widehat{\Sigma}_{j\backslash j} + \lambda\|\beta_j\|_{L^1} + \mu\sum_c w_c\|\beta_j^c - \beta_j^{\text{cons}}\|_{L^2}^2$$

with:

- w_c: real number used to weight the conditions ($w_c = 1$ or $w_c = \frac{1}{\sqrt{n_c}}$);
- μ regularization parameter;
- β_j^{cons} whatever you want...?

Choice of a consensus

$\beta_j^{\text{cons}} = \sum_c \frac{n_c}{n} \beta_j^c$ is a good choice because:

- the consensual penalty is then **quadratic** with respect to β_j;

Choice of a consensus

$\beta_j^{\text{cons}} = \sum_c \frac{n_c}{n} \beta_j^c$ is a good choice because:

- the consensual penalty is then **quadratic** with respect to β_j;
- thus, solving the optimization problem is **equivalent to maximizing**

$$\frac{1}{2}\beta_j^T S_j(\mu)\beta_j + \beta_j^T \widehat{\Sigma}_{j\backslash j} + \lambda \sum_c \frac{1}{n_c} \|\beta_j^c\|_1$$

with $S_j(\mu) = \widehat{\Sigma}_{j\backslash j} + 2\mu A^T A$ with A a matrix that does not depend on j.

Choice of a consensus

$\beta_j^{\text{cons}} = \sum_c \frac{n_c}{n} \beta_j^c$ is a good choice because:

- the consensual penalty is then **quadratic** with respect to β_j;
- thus, solving the optimization problem is **equivalent to maximizing**

$$\frac{1}{2}\beta_j^T S_j(\mu)\beta_j + \beta_j^T \widehat{\Sigma}_{j\backslash j} + \lambda \sum_c \frac{1}{n_c} \|\beta_j^c\|_1$$

with $S_j(\mu) = \widehat{\Sigma}_{j\backslash j} + 2\mu A^T A$ with A a matrix that does not depend on j.

$$\text{Convex part} + L^1\text{-norm penalty}$$

similar to standard LASSO problems: use of an "active set" approach as described in **[Osborne et al., 2000, Chiquet et al., 2011]**

Bootstrap estimation

Bootstrapped Consensus Lasso

Require: List of genes: $\{1, \ldots, p\}$; Gene expressions: X; Condition ids: $c_i \in \{1, ..., C\}$
Initialize $\forall j, j' \in \{1, \ldots, p\}$, $N^c(j, j') \leftarrow 0$; μ fixed
for $b = 1 \rightarrow P$ **do**
    Take a bootstrap sample B_b
    Estimate $(\beta_j^{c,b,\lambda})_{j,c,\lambda}$ from the previous method for several λ (decreasing order)
    Find $\left\{\left(\sum_{j,j',c} \mathbb{I}_{\beta_j^{c,\lambda,b} \neq 0}\right) > T_1\right\}$ **return** $(\beta_j^{c,b})_{j,c} := (\beta_j^{c,\lambda_{\max},b})_{j,c}$
    if $\beta_j^{c,b} \neq 0$ **then**
        $N^c(j, j') \leftarrow N^c(j, j') + 1$
    end if
end for
Select edges with $N^c(j, j') > T_2$ (T_2 chosen)

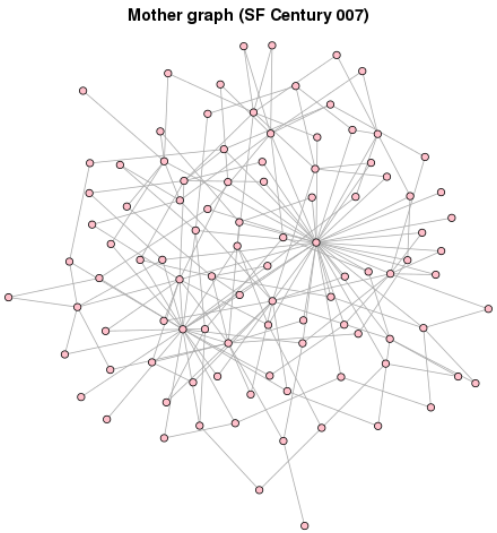
Outline

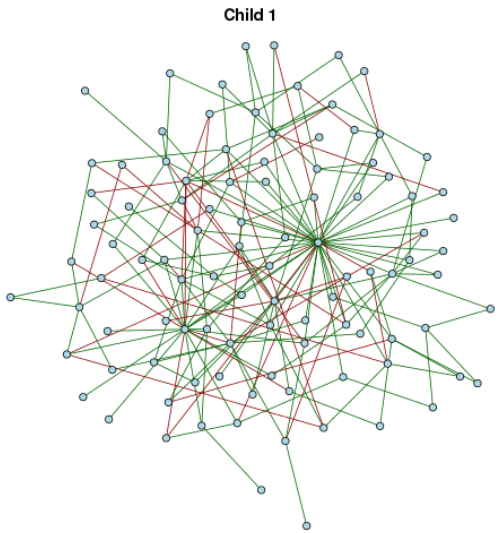
Simulated data

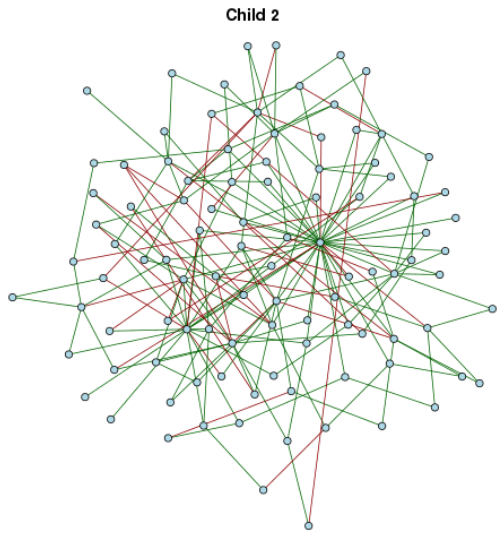
Expression data with known co-expression network

- original network (scale free) taken from `http://www.comp-sys-bio.org/AGN/data.html` (100 nodes, $\sim$ 200 edges, loops removed);
- rewire a ratio r of the edges to generate k "children" networks (sharing approximately $100(1-2r)\%$ of their edges);
- generate "expression data" with a random Gaussian process from each chid.

An example with $k = 2$, $r = 5\%$



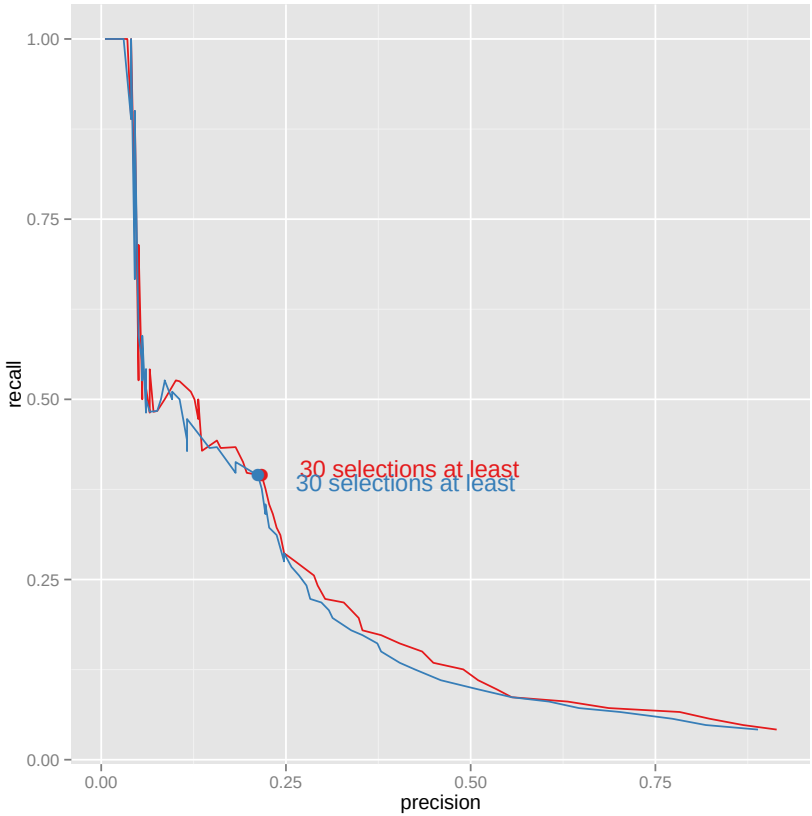


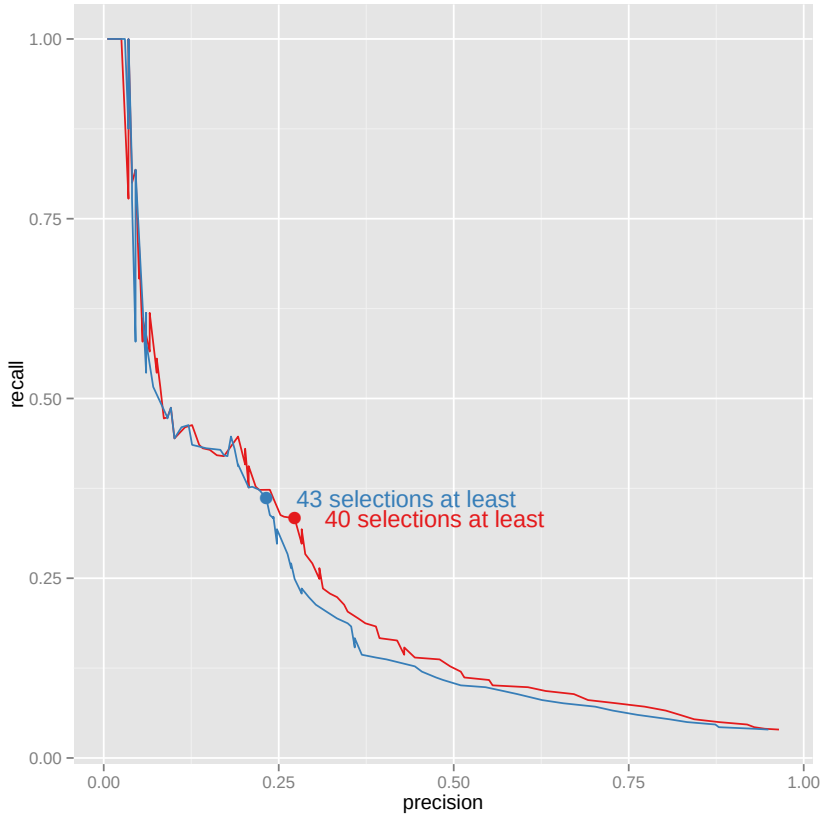


Choice for T_2

Data: $r = 0.05$, $k = 2$ and $n_1 = n_2 = 20$

100 bootstrap samples, $\mu = 1$, $T_1 = 250$ **or** 500





Dots correspond to best $F = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}}$

$\Rightarrow$ Best F corresponds to selecting a number of edges approximately equal to the number of edges of the original network.

Choice for T_1 and μ

	μ 0.1/1	T_1 $\{250, 300, 500\}$	% of improvement of bootstrapping
network sizes		**rewired edges: 5%**	
20-20	1	500	30.69
20-30	0.1	500	11.87
30-30	1	300	20.15
50-50	1	300	14.36
20-20-20-20-20	1	500	86.04
30-30-30-30	0.1	500	42.67
network sizes		**rewired edges: 20%**	
20-20	0.1	300	-17.86
20-30	0.1	300	-18.35
30-30	1	500	-7.97
50-50	0.1	300	-7.83
20-20-20-20-20	0.1	500	10.27
30-30-30-30	1	500	13.48

Comparisons (best/worst case F for different parameters)

Method	gLasso	cLasso	gLasso+boot	cLasso+boot
n_c	**rewired edges: 5%**			
20-20	0.19	0.22 (0.18)	0.27 (0.26)	**0.29 (0.27)**
20-30	0.26	0.30 (0.26)	0.31 (0.29)	**0.33 (0.32)**
30-30	0.28	0.31 (0.27)	0.35 (0.31)	**0.38 (0.36)**
50-50	0.36	0.43 (0.36)	0.47 (0.46)	**0.49 (0.49)**
20-20-20-20-20	0.19	0.23 (0.18)	0.39 (0.38)	**0.43 (0.40)**
30-30-30-30	0.30	0.36 (0.29)	0.49 (0.48)	**0.51 (0.50)**
n_c	**rewired edges: 20%**			
20-20	0.21	**0.23 (0.19)**	0.18 (0.17)	0.19 (0.17)
20-30	**0.26**	**0.26 (0.25)**	0.20 (0.19)	0.22 (0.20)
30-30	0.28	**0.31 (0.29)**	0.27 (0.27)	0.29 (0.28)
50-50	0.42	**0.43 (0.41)**	0.38 (0.37)	0.40 (0.38)
20-20-20-20-20	0.20	0.22 (0.20)	0.22 (0.20)	**0.24 (0.24)**
30-30-30-30	0.27	0.29 (0.27)	0.30 (0.30)	**0.33 (0.31)**

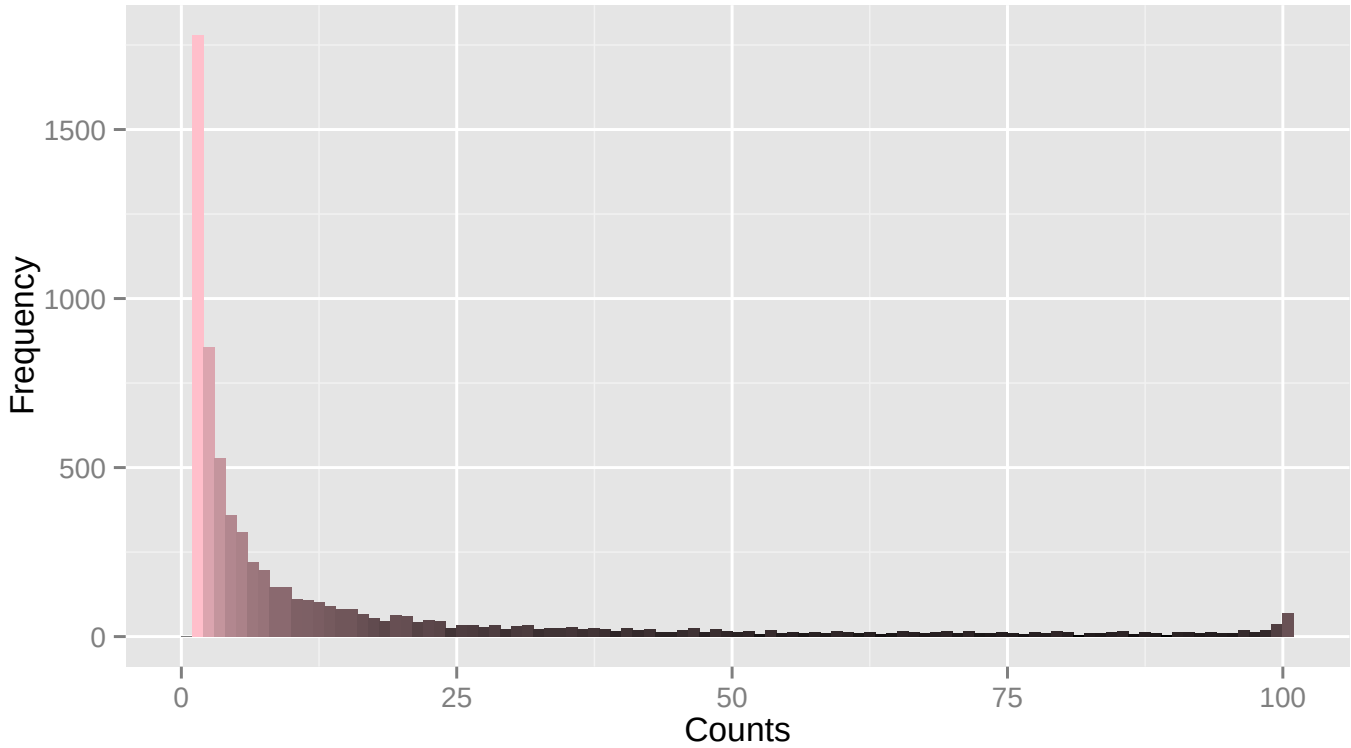
Not shown here but when the % of rewired edges is larger (20%), **intertwinned Lasso** has better performances (they are not improved by bootstrapping).

Real data

204 obese women ; expression of 221 genes before and after a LCD

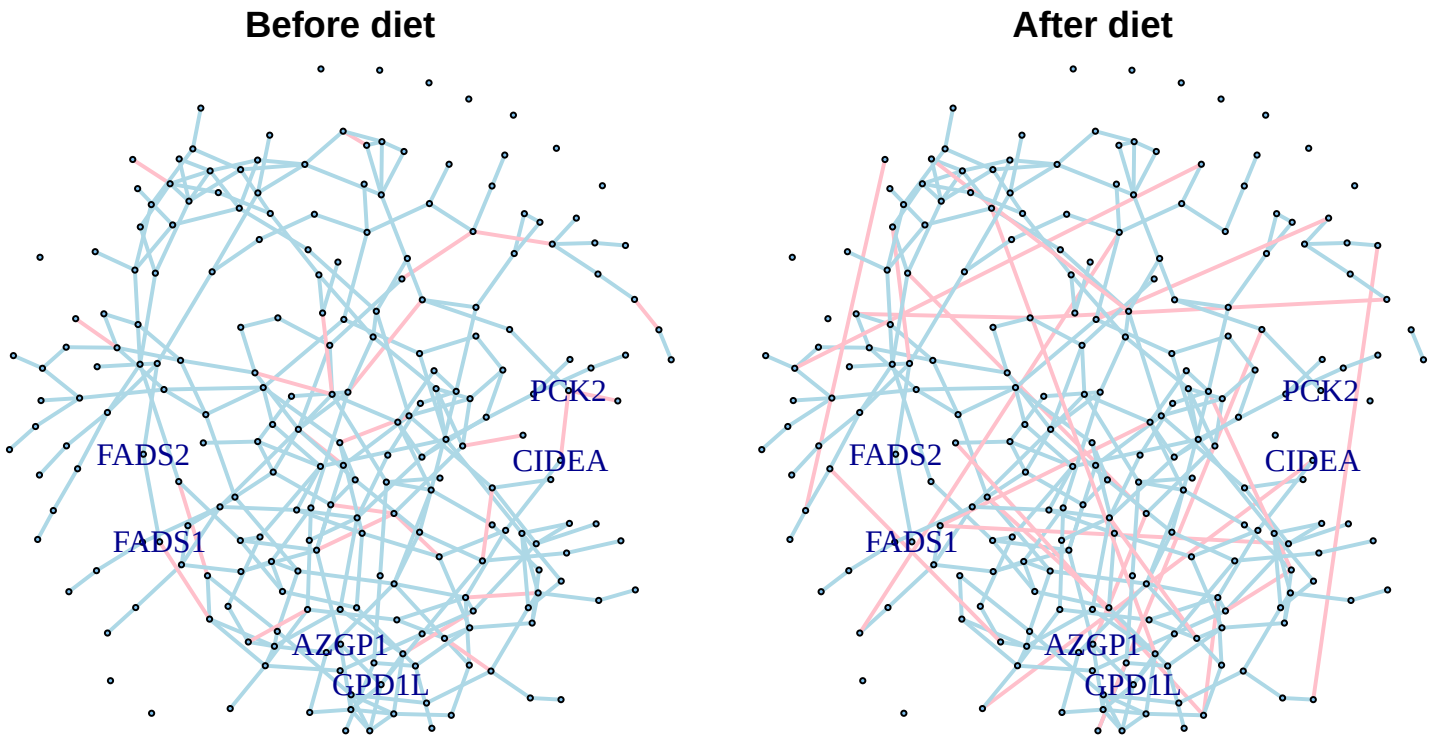
$\mu = 1$; $T_1 = 1000$ (target density: 4%)

Distribution of the number of times an edge is selected over 100 bootstrap samples



(70% of the pairs of nodes are never selected) $\Rightarrow T_2 = 80$

Networks



densities about 1.3% - some interactions (both shared and specific) make sense for the biologist

Thank you for your attention...

Programs available in the R package **therese** (on R-Forge). Joint work with

Magali SanCristobal
(LGC, INRA de Toulouse)

Matthieu Vignes
(MIAT, INRA de Toulouse)

Nathalie Viguerie
(I2MC, INSERM Toulouse)

References

Ambroise, C., Chiquet, J., and Matias, C. (2009).
Inferring sparse Gaussian graphical models with latent structure.
Electronic Journal of Statistics, 3:205–238.

Chiquet, J., Grandvalet, Y., and Ambroise, C. (2011).
Inferring multiple graphical structures.
Statistics and Computing, 21(4):537–553.

Danaher, P., Wang, P., and Witten, D.
The joint graphical lasso for inverse covariance estimation accross multiple classes.
Preprint arXiv 1111.0324v3. Submitted for publication.

Friedman, J., Hastie, T., and Tibshirani, R. (2008).
Sparse inverse covariance estimation with the graphical lasso.
Biostatistics, 9(3):432–441.

Meinshausen, N. and Bühlmann, P. (2006).
High dimensional graphs and variable selection with the lasso.
Annals of Statistic, 34(3):1436–1462.

Mohan, K., Chung, J., Han, S., Witten, D., Lee, S., and Fazel, M. (2012).
Structured learning of Gaussian graphical models.
In *Proceedings of NIPS (Neural Information Processing Systems) 2012*, Lake Tahoe, Nevada, USA.

Osborne, M., Presnell, B., and Turlach, B. (2000).
On the LASSO and its dual.
Journal of Computational and Graphical Statistics, 9(2):319–337.

Schäfer, J. and Strimmer, K. (2005).
An empirical bayes approach to inferring large-scale gene association networks.
Bioinformatics, 21(6):754–764.

Viguerie, N., Montastier, E., Maoret, J., Roussel, B., Combes, M., Valle, C., Villa-Vialaneix, N., Iacovoni, J., Martinez, J., Holst, C., Astrup, A., Vidal, H., Clément, K., Hager, J., Saris, W., and Langin, D. (2012).
Determinants of human adipose tissue gene expression: impact of diet, sex, metabolic status and *cis* genetic regulation.
PLoS Genetics, 8(9):e1002959.