Consensual gene co-expression network inference with multiple samples

Nathalie Villa-Vialaneix[(1,2)]

http://www.nathalievilla.org

nathalie.villa@univ-paris1.fr

Joint work with Magali SanCristobal, Matthieu Vignes

(1) INRA SCIENCE & IMPACT (2) SAMM

45e Journées de Statistique, Toulouse, 27 mai 2013

SFdS SOCIETE FRANÇAISE DE STATISTIQUE

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

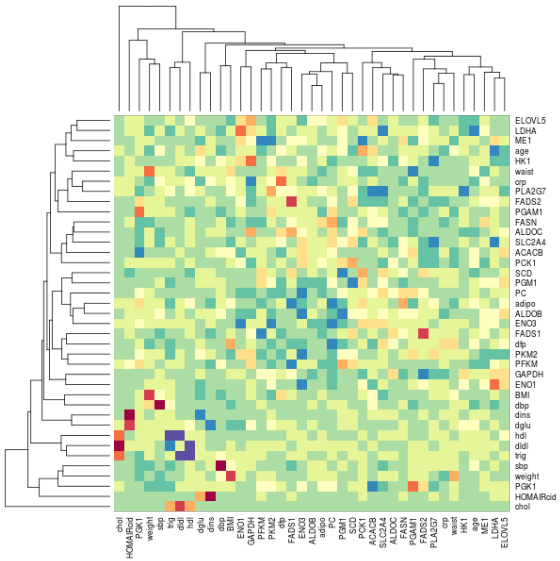
- nodes: genes;
- edges: “significant” and direct co-expression between two genes (track transcription regulations).

Modeling multiple interactions between genes with a network

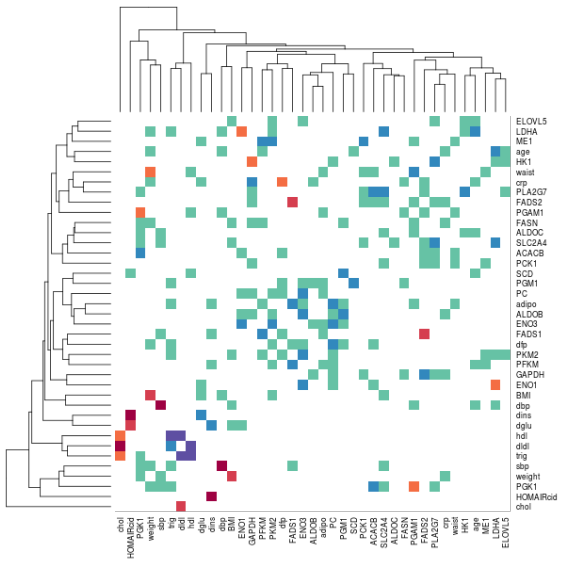
Co-expression networks

- **nodes**: genes
- **edges**: “direct” co-expression between two genes

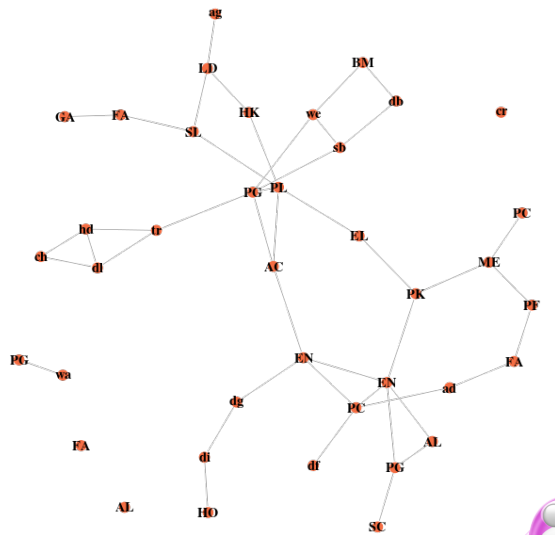
Method:



“Correlations”

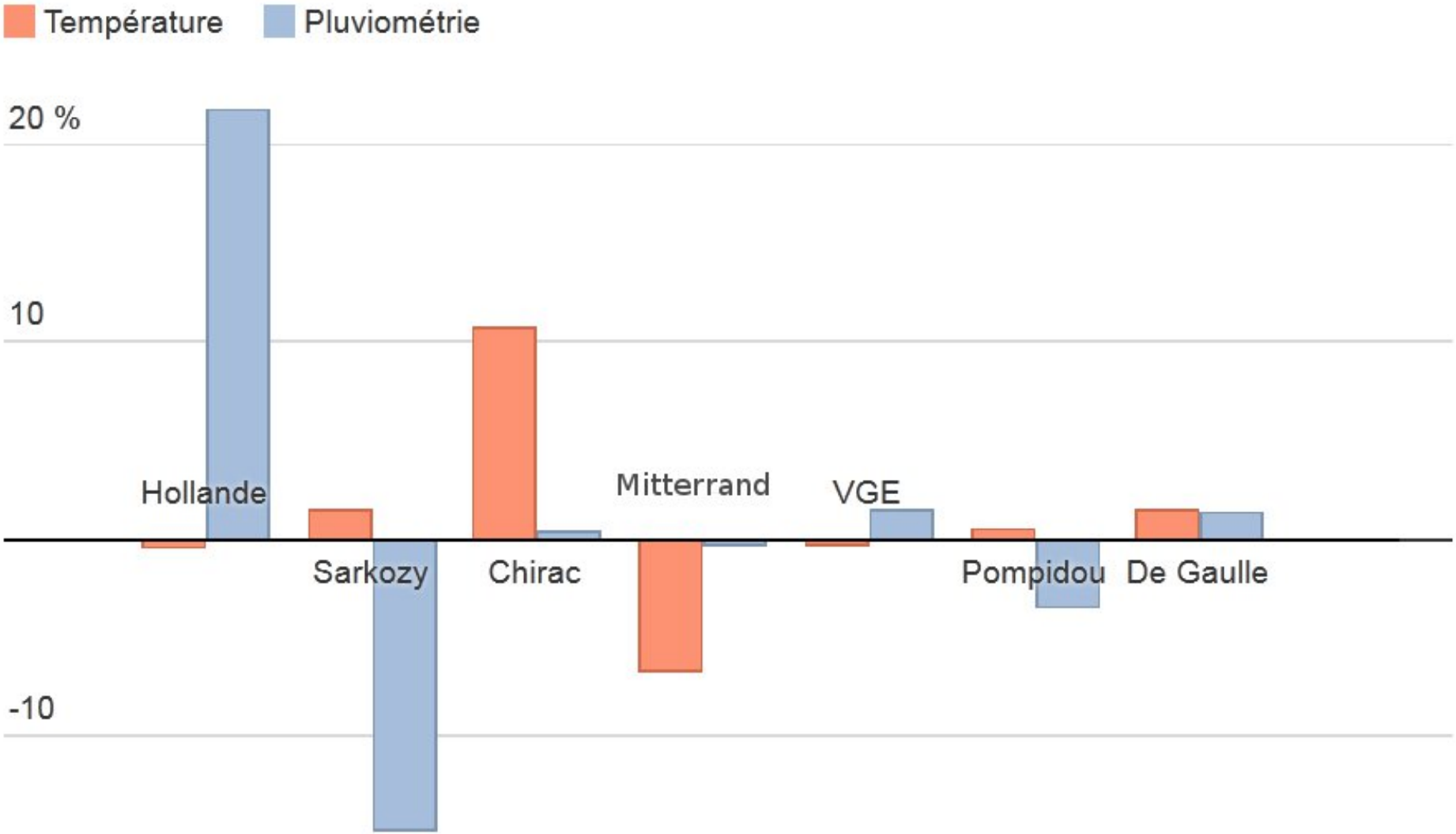


Thresholding

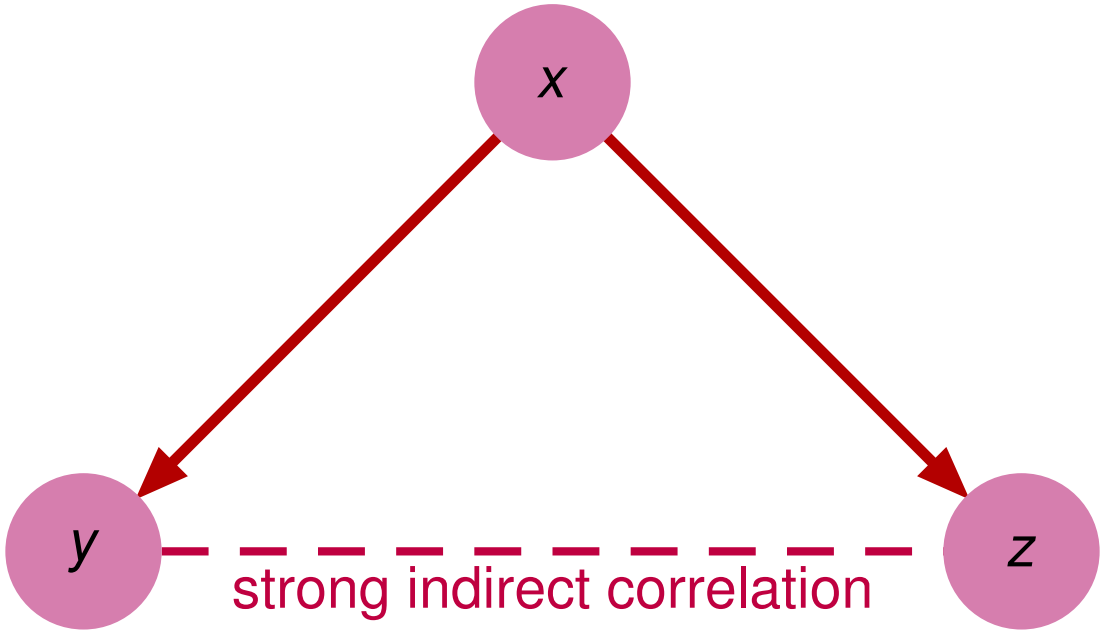


Graph

But correlation is not causality...

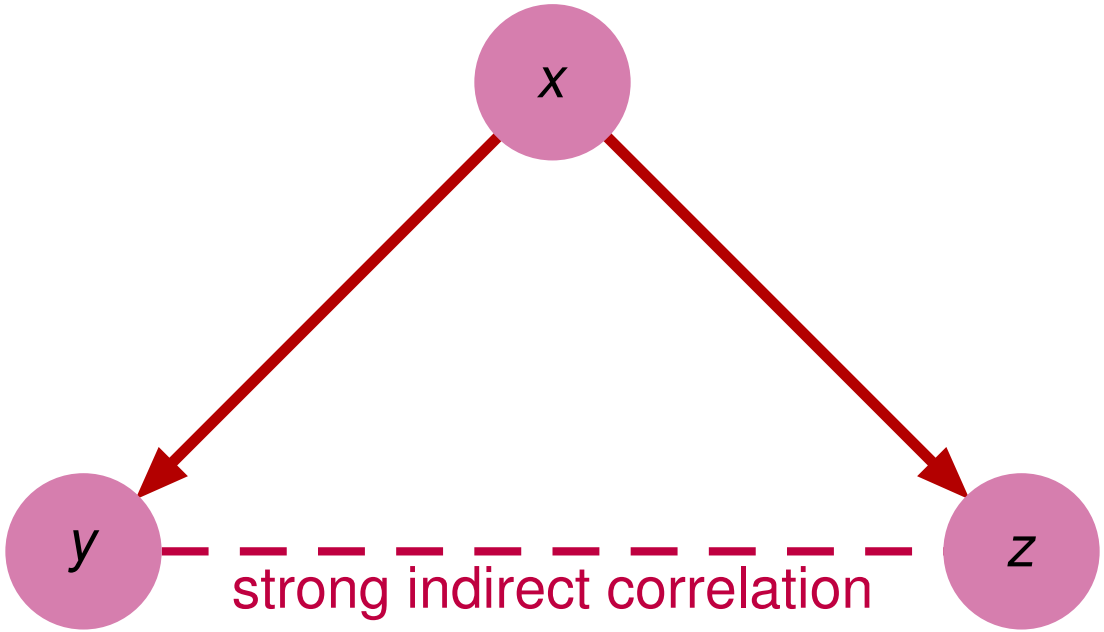


But correlation is not causality...




```
set.seed(2807); x <- runif(100)
y <- 2*x+1 + rnorm(100,0,0.1); cor(x,y); [1] 0.9870407
z <- -x+2 + rnorm(100,0,0.1); cor(x,z); [1] -0.9443082
cor(y,z) [1] -0.9336924
```

But correlation is not causality...




```
set.seed(2807); x <- runif(100)
y <- 2*x+1 + rnorm(100,0,0.1); cor(x,y); [1] 0.9870407
z <- -x+2 + rnorm(100,0,0.1); cor(x,z); [1] -0.9443082
cor(y,z) [1] -0.9336924
# Partial correlation
cor(lm(y x)$residuals,lm(z x)$residuals) [1] -0.03071178
```

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: $\forall\, j$

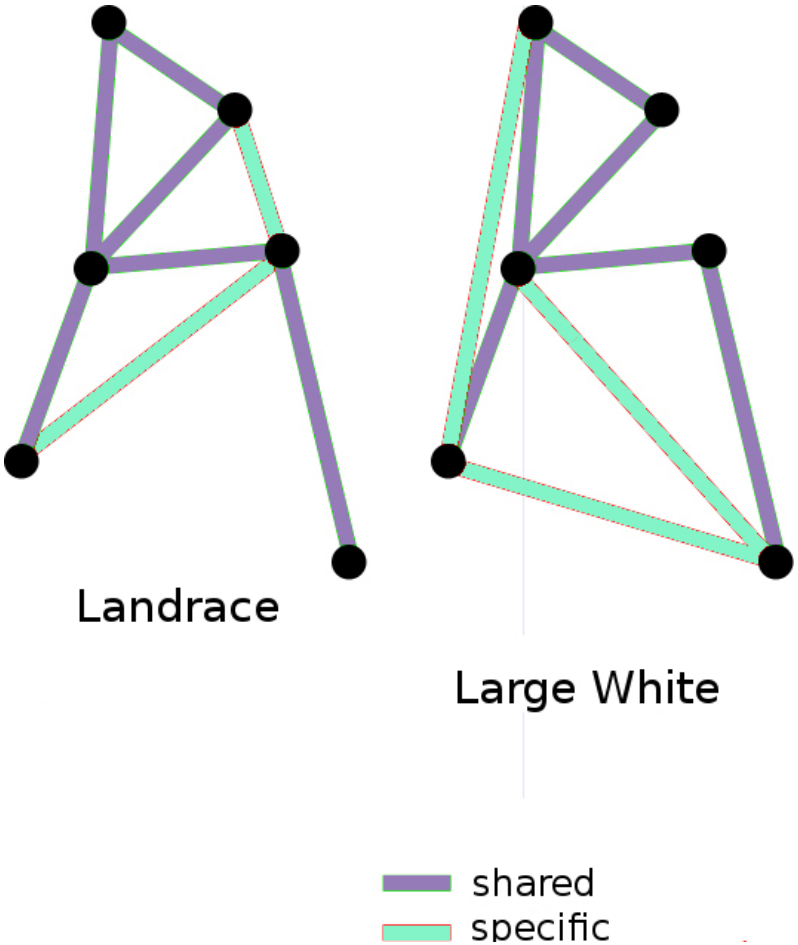
$$X^j = \beta_j^T X^{-j} + \epsilon \qquad ; \qquad \arg\max_{(\beta_{jj'})_{j'}} \left(\log \mathrm{ML}_j - \lambda \sum_{j' \neq j} |\beta_{jj'}| \right)$$

In the Gaussian framework: $\beta_{jj'} = -\frac{s_{jj'}}{s_{jj}}$ where $S = \Sigma^{-1}$ related to partial correlations by $\pi_{jj'} = -\frac{s_{jj'}}{\sqrt{s_{jj} s_{j'j'}}}$.

Outline

Motivation for multiple networks inference

Projet DeLiSus (with Laurence Liaubet): gene expressions from pig muscle in **Landrace** and **Large white** breeds;



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

Similar references: **[Chiquet et al., 2011, Mohan et al., 2012]**

Consensus LASSO

Proposal: Infer multiple networks by forcing them toward a consensual network.

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.

Original optimization:

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

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

$$\max_{(\beta_{jk}^c)_{k\neq j, c=1,\ldots,C}} \sum_c \left(\log \mathrm{ML}_j^c - \lambda \sum_{k\neq j} |\beta_{jk}^c| - \mu \sum_c w_c \|\beta_j^c - \beta_j^{\mathrm{cons}}\|^2 \right)$$

Choice of a consensus

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

- $\frac{\partial \beta_j^{\text{cons}}}{\partial \beta_j^c}$ **exists**;

Choice of a consensus

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

- $\frac{\partial \beta_j^{\text{cons}}}{\partial \beta_j^c}$ **exists**;
- 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 $S_j(\mu) = \widehat{\Sigma}_{j\backslash j} + 2\mu A^T A$ ($\widehat{\Sigma}_{j\backslash j}$: *jth* row of empirical covariance matrix deprived from its *jth* column; $\widehat{\Sigma}_{j\backslash j}$: empirical covariance matrix deprived from its *jth* row and column; A: a matrix that does not depend on j).

Standard LASSO problem that can be solved using a sub-gradient method.

Bootstrap estimation

Boostraped Consensus Lasso

Require: List of genes: $\{1, \ldots, p\}$; Gene expressions: X; Condition ids: $c_i \in \{1, ..., C\}$
Initialize $\forall\, j, j' \in \{1, \ldots, p\}$, $N^c(j, j') \leftarrow 0$
for $b = 1 \rightarrow P$ **do**
  Take a bootstrap sample B_b
  Estimate $(\beta_j^c)_{j,c}$ from the previous method$^{(*)}$
  if $\beta_j^c \neq 0$ **then**
    $N^c(j, j') \leftarrow N^c(j, j') + 1$
  end if
end for
Select edges with $N^c(j, j') > T$ (T chosen)

$^{(*)}$ μ fixed, a full regularization path is used for λ, the 10% first edges are kept

Outline

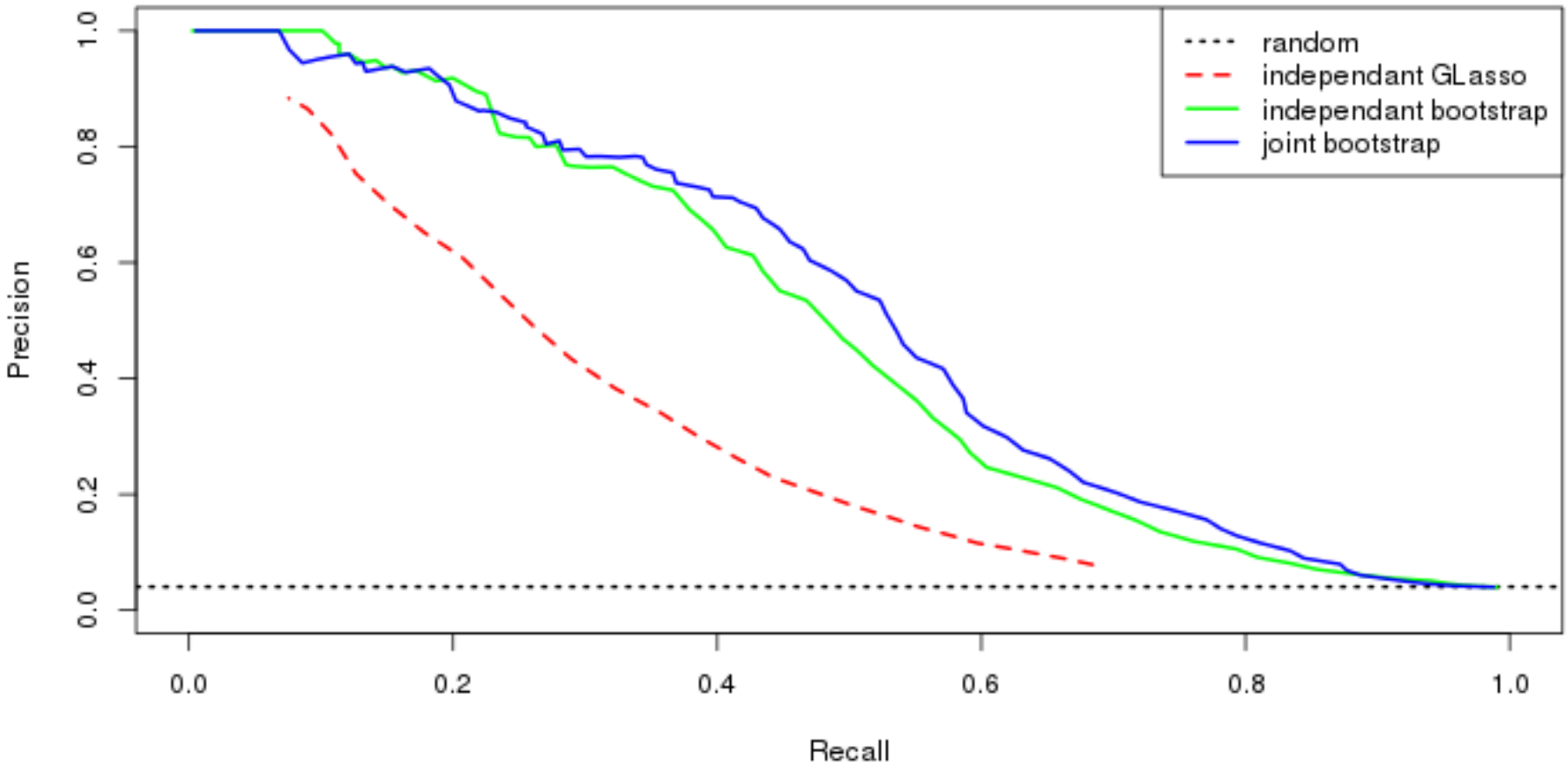
Simulated data

Expression data with known co-expression network

- 9 networks (scale free) taken from http://www.comp-sys-bio.org/AGN/data.html (100 nodes, $\sim$ 200 edges, loops removed);
- rewire 5% or 10% of the edges to generate two "children" networks (sharing approximately 90% to 80% of their edges);
- generate "expression data" with a random Gaussian process from each chid.

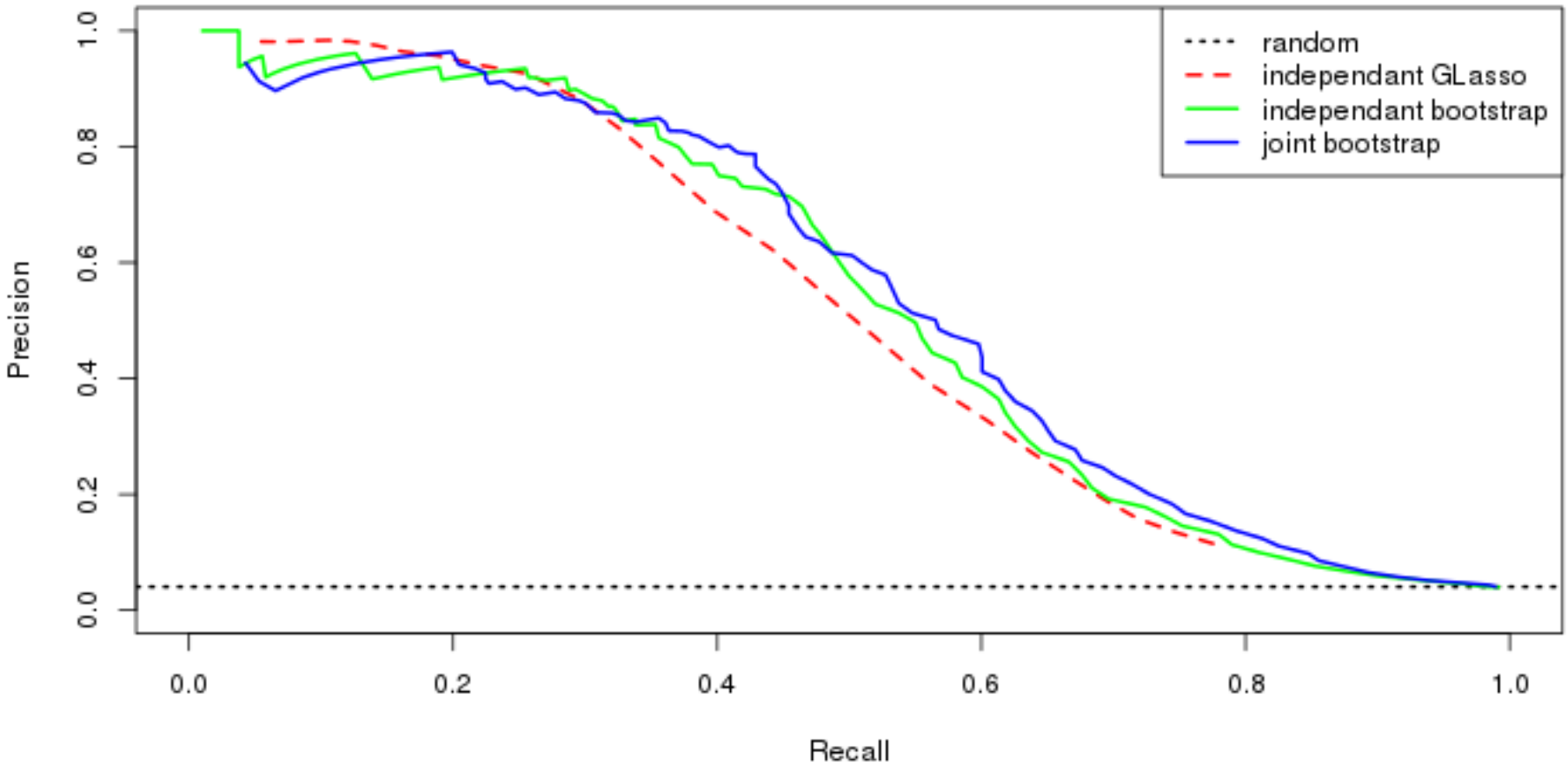
Precision/recall curve 1

% or rewired edges: 5%, sample sizes: 50×2, $\mu = 0.1$. Average global precision and recall over the 9 network pairs



Precision/recall curve 2

% or rewired edges: 10%, sample sizes: 100×2, $\mu = 0.1$. Average global precision and recall over the 9 network pairs



Numeric comparison

Calcul of $F = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$

Conclusions:

- the boostrap approach increases F;
- the best F obtained is with the bootstrap consensus Lasso except for 1% rewired edges and sample sizes 100×100;
- differences between the F sequences with/without the consensus regularization are significant (Wilcoxon test, 1%) except for 1% rewired edges and sample sizes 100×100.

Perspectives

- Use an **out-of-bag approach to tune** μ;
- **Test on real datasets** to search for specific/common regulations between species; but biological validation is long/hard to perform, especially on pigs (which genome is badly annotated).

Thank you for your attention...

Joint work with

Magali SanCristobal
(LGC, INRA de Toulouse)

Matthieu Vignes
(MIAT, INRA de Toulouse)

Laurence Liaubet
(LGC, INRA de Toulouse)

References

Chiquet, J., Grandvalet, Y., and Ambroise, C. (2011).

Inferring multiple graphical structures.

Statistics and Computing, 21(4):537–553.

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.

Schäfer, J. and Strimmer, K. (2005).

An empirical bayes approach to inferring large-scale gene association networks.

Bioinformatics, 21(6):754–764.