Comparison of network inference packages and methods for multiple networks inference

Nathalie Villa-Vialaneix

http://www.nathalievilla.org

nathalie.villa@univ-paris1.fr

1ères Rencontres R - BoRdeaux, 3 Juin 2012

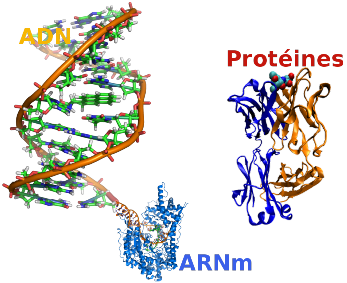
Joint work with **Nicolas Edwards**, **Laurence Liaubet**, **Nathalie Viguerie** & **Magali SanCristobal**

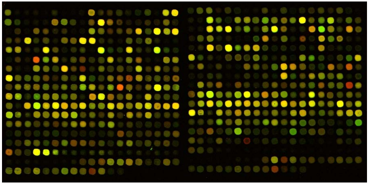
Plan

1. From transcriptomic data to network
2. Network inference and multiple networks inference using R
3. Simulations

Transcriptome

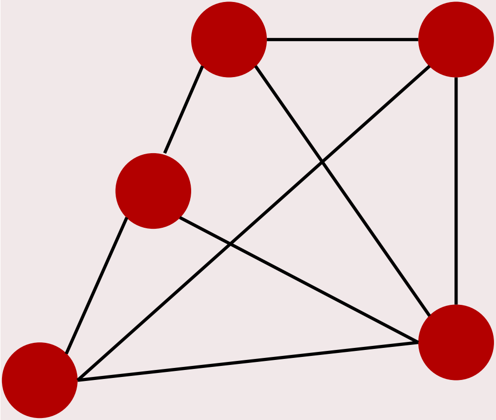
- **DNA** contains the genetic instructions used in the development and functioning of living organims
- Molecular unit of the DNA, **genes**, are not all identically **expressed** in a given cell: it is assessed by means of the quantity of the corresponding mRNA
- Genes expression can be measured by microarray, RT PCR...: **transcriptomic data**





Modelling multiple interactions between genes with a network

Co-expression networks



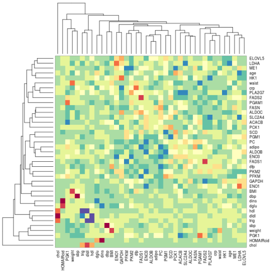
- **nodes**: genes
- **edges**: “direct” co-expression between two genes

Modelling multiple interactions between genes with a network

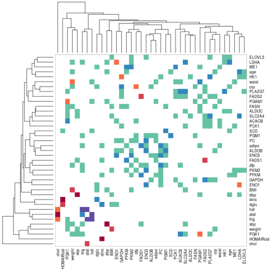
Co-expression networks

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

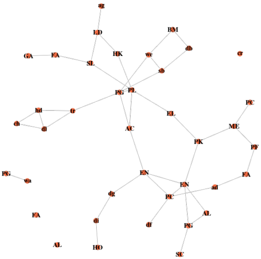
Method:



“Correlations”



Thresholding



Graph

Multiple networks inference

Transcriptomic data coming from several different conditions.

Examples:

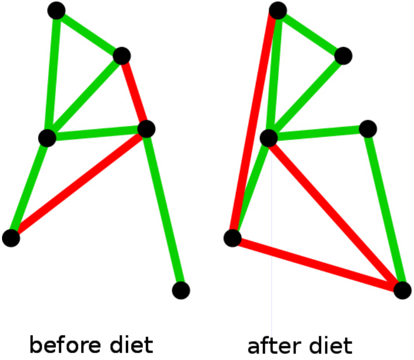
- genes expression from pig muscle in **Landrace** and **Large white** breeds;
- genes expression from obese humans **after** and **before** a diet.

Multiple networks inference

Transcriptomic data coming from several different conditions.

Examples:

- genes expression from pig muscle in **Landrace** and **Large white** breeds;
- genes expression from obese humans **after** and **before** a diet.



before diet

after diet

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

Plan

1. From transcriptomic data to network

2. Network inference and multiple networks inference using R

3. Simulations

Theoretical framework

Gaussian Graphical Models (GGM) $X \sim \mathcal{N}(0, \Sigma)$ Seminal work **[Schäfer and Strimmer, 2005]**, **GeneNet**: estimation of the **partial correlations**

$$\pi_{jj'} = \mathrm{Cor}(X^j, X^{j'} | X^k, k \neq j, j')$$

(by using the inverse of $\widehat{\Sigma} + \lambda\mathbb{I}$) and edges selection by a Bayesian test based on a mixture model.

Theoretical framework

Gaussian Graphical Models (GGM) $X \sim \mathcal{N}(0, \Sigma)$ Edges selection by sparse penalty: **graphical LASSO** **[Meinshausen and Bühlmann, 2006, Friedman et al., 2008]**, **glasso**:

$$X^j = \sum_{k \neq j} \beta_{jk} X^k + \epsilon.$$

where $(\beta_{jk})_{jk}$ are estimated by

$$\max_{(\beta_{jk})_{k \neq j}} \left(\log \mathrm{ML}_j - \lambda \sum_{k \neq j} |\beta_{jk}| \right).$$

β_{jk} is related to $S = \Sigma^{-1}$ by $\beta_{jk} = -\frac{S_{jk}}{S_{jj}}$.

Theoretical framework

Gaussian Graphical Models (GGM) $X \sim \mathcal{N}(0, \Sigma)$ Edges selection by sparse penalty: **graphical LASSO** **[Meinshausen and Bühlmann, 2006, Friedman et al., 2008]**, **glasso**:

$$X^j = \sum_{k \neq j} \beta_{jk} X^k + \epsilon.$$

where $(\beta_{jk})_{jk}$ are estimated by

$$\max_{(\beta_{jk})_{k \neq j}} \left(\log \mathrm{ML}_j - \lambda \sum_{k \neq j} |\beta_{jk}| \right).$$

β_{jk} is related to $S = \Sigma^{-1}$ by $\beta_{jk} = -\frac{S_{jk}}{S_{jj}}$.

Other related packages: **parcor** (different regularization methods for GGM, CV selection), **GGMselect** (network selection among a family): not used here

Multiple networks

Independent estimations: if $c = 1, \ldots, C$ are different samples (or "conditions", e.g., breeds or before/after diet...)

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

Multiple networks

Independent estimations: if $c = 1, \ldots, C$ are different samples (or "conditions", e.g., breeds or before/after diet...)

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

Joint estimations:

Implemented in the package **simone**, **[Chiquet et al., 2011]**

GroupLasso Consensual network between conditions (enforces identical edges by a group LASSO penalty)

CoopLasso Sign-coherent network between conditions (prevents edges that corresponds to partial correlations having different signs; thus allows one to obtain a few differences between the conditions)

Intertwined In GLasso replace $\widehat{\Sigma}^c$ by $1/2\widehat{\Sigma}^c + 1/2\overline{\Sigma}$ where $\overline{\Sigma} = \frac{1}{C}\sum_c \widehat{\Sigma}^c$

Multiple networks

Independent estimations: if $c = 1, \dots, C$ are different samples (or "conditions", e.g., breeds or before/after diet...)

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

Joint estimations: Additional tested approaches:

- Use the fact that individuals are paired (if concerned) to compute the partial correlations: $\widehat{\mathbf{X}}_i^c = 1/2\mathbf{X}_i^c + 1/2\overline{\mathbf{X}_i}$ with $\overline{\mathbf{X}_i} = \sum_c \widehat{\mathbf{X}}_i^c$ (implemented with **GeneNet** and **simone**)
- Combine the partial correlations instead of the correlations as in **Intertwined** (implemented from independent estimations obtained using **simone**, called "therese")

Tested packages and features

	Indep.	Joint	Selection?	Inputs	Outputs
GeneNet	[1]	No	confidence threshold	X	$(\pi_{ij})_{ij}$
glasso	[2,3]	No	none (but LASSO path is available)	$\widehat{\Sigma}$	$(S_{ij})_{ij}$
simone	[2,3]	Yes	number of edges AIC, BIC (LASSO path)	X	$(S_{ij})_{ij}$

with

[1] **[Schäfer and Strimmer, 2005]**

[2] **[Meinshausen and Bühlmann, 2006]**

[3] **[Friedman et al., 2008]**

not shown: CV selection is not included in **glasso** and **simone**, but it can be implemented (be careful to the internal scaling and to the outputs)

Plan

1. From transcriptomic data to network

2. Network inference and multiple networks inference using R

3. Simulations

Data

Datasets coming from

The ANR project “DéLiSus” (“caractérisations génétique et phénotypique fines de populations porcines françaises”, genetic and phenotypic variability of French pigs)

The pan-European project “DiOGenes” (Diet, Obesity and Genes: new insight on obesity problems and routes to prevention)

Datasets description

Real datasets

“DiOGenes” dataset:

- **variables**: 39 variables (genes expressions and clinical variables)
- **conditions**: before/after a diet (paired individuals: 204 obese women)

“DeLiSus” dataset:

- **variables**: expression of 123 genes
- **conditions**: two breeds (33 “Landrace” and 51 “Large white”)

Datasets description

Real datasets

"DiOGenes" dataset:

- **variables**: 39 variables (genes expressions and clinical variables)
- **conditions**: before/after a diet (paired individuals: 204 obese women)

"DeLiSus" dataset:

- **variables**: expression of 123 genes
- **conditions**: two breeds (33 "Landrace" and 51 "Large white")

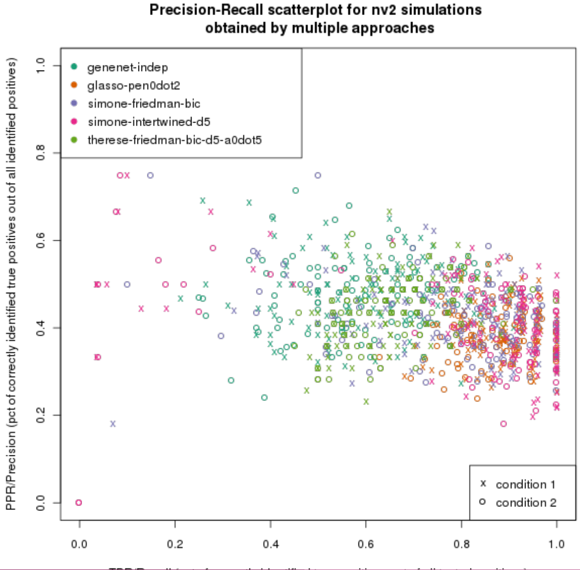
Simulated dataset

To compare methods, a dataset was simulated from a GGM (with **simone**):

- **underlying network**: 39 variables with 5 groups of preferential attachment and a density equal to approximatly 3-4%.
- **children networks**: two networks obtained by randomly permuting 10% of the edges;
- **variables**: 2 × 204 observations of a GGM coming from these networks (observations are not pairwise).

Simulation results and conclusions

All methods

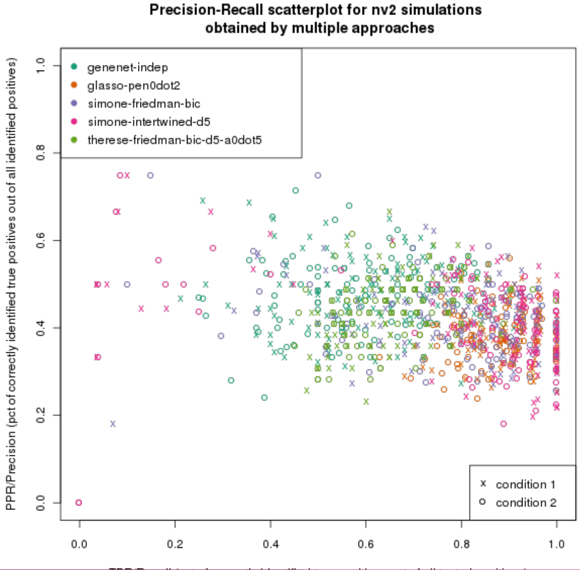


$$\textbf{Precision} = \frac{\text{tp}}{\text{p}}$$

$$\textbf{Recall} = \frac{\text{tp}}{\text{tp}+\text{fn}}$$

Simulation results and conclusions

All methods



Precision$= \frac{\text{tp}}{\text{p}}$

Recall$= \frac{\text{tp}}{\text{tp}+\text{fn}}$

- **glasso** performs well (with very low variability) but no real solution for tuning;
- **simone** performs well (especially joint methods), with an automatic tuning but large variability;
- "therese" has a low variability but no real solution for tuning;
- **GeneNet** has a low recall and a low variability.

Simulation results and conclusions

Numerical performances

Graph densities

True density: 3.57% (on average)

- **GeneNet** (automatic): 4.38%
- **glasso** (manual): 8.14%
- **simone** (indep, BIC): 6.65% and **simone** (joint, BIC): 5.87%
- “therese” (semi manual): 5.26%

Simulation results and conclusions

Numerical performances

Graph densities

True density: 3.57% (on average)

- **GeneNet** (automatic): 4.38%
- **glasso** (manual): 8.14%
- **simone** (indep, BIC): 6.65% and **simone** (joint, BIC): 5.87%
- "therese" (semi manual): 5.26%

Shared edges between conditions

Truth: 20.28% (on average)

- **GeneNet** (automatic): 15.95%
- **glasso** (manual): 32.74%
- **simone** (indep, BIC): 26.69% and **simone** (joint, BIC): 31.15%
- "therese" (semi manual): 30.92%

“DiOGenes” dataset (39 variables, 204 obese women, fixed density 5%)

	Density	Transitivity	% shared
[1] **GeneNet**	0.06	0.22	0.68
[2] **GeneNet** (paired)	0.09	0.24	0.84
[3] **simone** (indep., Fried.)	0.05	0.52	0.76
[4] **simone**, CoopLasso	0.06	0.30	1.00
[5] **simone**, GroupLasso	0.06	0.30	1.00
[6] **simone**, intertwined	0.05	0.37	0.97
[7] **simone**, paired	0.04	0.52	0.94
[8] “therese”	0.05	0.46	0.82

	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
[1]	1.00	0.98	0.45	0.61	0.61	0.53	0.42	0.42
[2]		1.00	0.58	0.66	0.66	0.66	0.55	0.58
[3]			1.00	0.79	0.79	0.84	1.00	0.92
[4]				1.00	1.00	0.95	0.76	0.76
[5]					1.00	0.95	0.76	0.76
[6]						1.00	0.82	0.79
[7]							1.00	0.97
[8]								1.00

“DeLiSus” dataset (restricted dataset with 84 genes (51 pigs))

	Density	Transitivity	% shared
[1] **GeneNet**	0.00	0.71	0.46
[2] **simone**, MB-AND	0.05	0.08	0.17
[3] **simone**, Fried.	0.05	0.19	0.22
[4] **simone**, intertwined	0.05	0.09	0.52
[5] **simone**, CoopLasso	0.06	0.09	0.88
[6] **simone**, GroupLasso	0.04	0.07	0.99
[7] “therese”	0.05	0.17	0.66

	[1]	[2]	[3]	[4]	[5]	[6]	[7]
[1]	1.00	0.00	0.00	0.00	0.00	0.00	0.00
[2]		1.00	0.71	0.76	0.64	0.56	0.57
[3]			1.00	0.67	0.55	0.53	0.78
[4]				1.00	0.80	0.67	0.58
[5]					1.00	0.84	0.60
[6]						1.00	0.74
[7]							1.00

Conclusion

- **simulations**: BIC is not always relevant ⇒ target density, CV, **GGMselect**...? Joined methods produce more shared edges between conditions

Conclusion

- **simulations**: BIC is not always relevant ⇒ target density, CV, **GGMselect**...? Joined methods produce more shared edges between conditions
- **real life datasets**
  - **low dimension case**: large consensus between methods; joined methods are too similar (except maybe paired **GeneNet** and "therese")
  - **larger dimension case**: methods are less consensual; GroupLasso and CoopLasso still produce too much shared edges
  - **very large dimension** *(not shown)*: 464 gene expressions for $51 + 33$ pigs gave very bad performances: on real dataset, some methods were unable to produce results (and BIC selected graphs with no edge); hence, on simulated datasets with the same sample size and dimension, the recall was always very low.

Collaboration

Any questions?...

Co-authors

Nathalie Villa-Vialaneix
(SAMM, U. Paris 1)

Nicolas Edwards
(LGC, INRA Tlse)

Laurence Liaubet
(LGC, INRA Tlse)

Nathalie Viguerie
(ORL, INSERM)

Magali SanCristobal
(LGC, INRA Tlse)

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.

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