Applying Multivariate Regressions in Large Dimension Data

Trung Ha, 1st year PhD student.
Ecole doctorale des Genomes Aux Organismes, Univ. d'Evry
Supervisors : Marie-Laure Martin Magniette, Julien Chiquet, Guillem Rigaill.

September-2014

Introduction

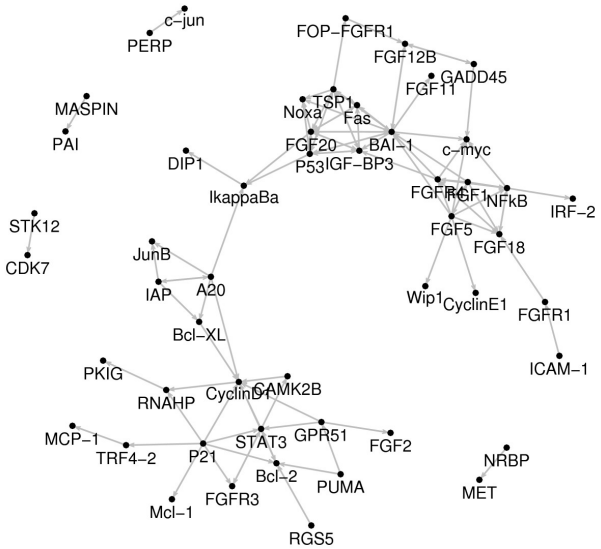
Biological Context

Object

- Reconstruct the biological network between genes based on data of genes expression levels by using linear regression models.

Example

- Each node is one gene.
- Presence/Absence of an edge represents the relation between 2 genes.



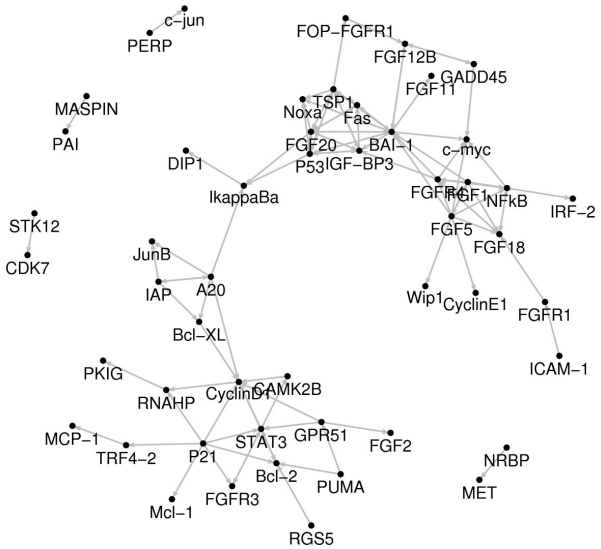
Biological Context

Object

- Reconstruct the biological network between genes based on data of genes expression levels by using linear regression models.

Example

- Each node is one gene.
- Presence/Absence of an edge represents the relation between 2 genes.



Previous Works

Graphical Gaussian Model (GGM)

Hypothesis

- The level expression of 1 gene is a random variable X_j, with $j = 1, .., p$.
- Suppose the data follows one multivariate normal distribution:

$$(X_1, .., X_p) \sim N(\beta, \Sigma).$$

Remark

- Genes expressions might be shifted by 2 nonindependent phenomenons:
  - 1. Its average expression level of genes.
  - 2. Its relations with others genes.

Graphical Gaussian Model (GGM)

Hypothesis

- The level expression of 1 gene is a random variable X_j, with $j = 1, .., p$.
- Suppose the data follows one multivariate normal distribution:

$$(X_1, .., X_p) \sim N(\beta, \Sigma).$$

Remark

- Genes expressions might be shifted by 2 nonindependent phenomenons:
  - 1. Its average expression level of genes.
  - 2. Its relations with others genes.

Disadvantages

Main Principles

- Raw data must be normalized by empirical mean values before using.
- Data follows multivariate normal distribution:

$$(X_1, .., X_p) \sim N(\beta = 0, \Sigma).$$

Potential problems due to the high dimension framework

- Empirical means could be not a good estimators in some sorts of data.
- For a gene, its expression and its relations with other genes may be linked.

Disadvantages

Main Principles

- Raw data must be normalized by empirical mean values before using.
- Data follows multivariate normal distribution:

$$(X_1, .., X_p) \sim N(\beta = 0, \Sigma).$$

Potential problems due to the high dimension framework

- Empirical means could be not a good estimators in some sorts of data.
- For a gene, its expression and its relations with other genes may be linked.

Our Model

COULD WE ?

- Get better estimators for β ?

$$\Downarrow$$

Graphical Gaussian Model

- Suppose the data follows multivariate normal distribution:

$$(X_1, .., X_p) \sim N(\beta, \Sigma).$$

- In the case of real data with K different conditions of p genes:

$$(X_1, .., X_p)^k \sim N(\beta^k, \Sigma^k).$$

COULD WE ?

- Get better estimators for β ?

$$\Downarrow$$

Graphical Gaussian Model

- Suppose the data follows multivariate normal distribution:

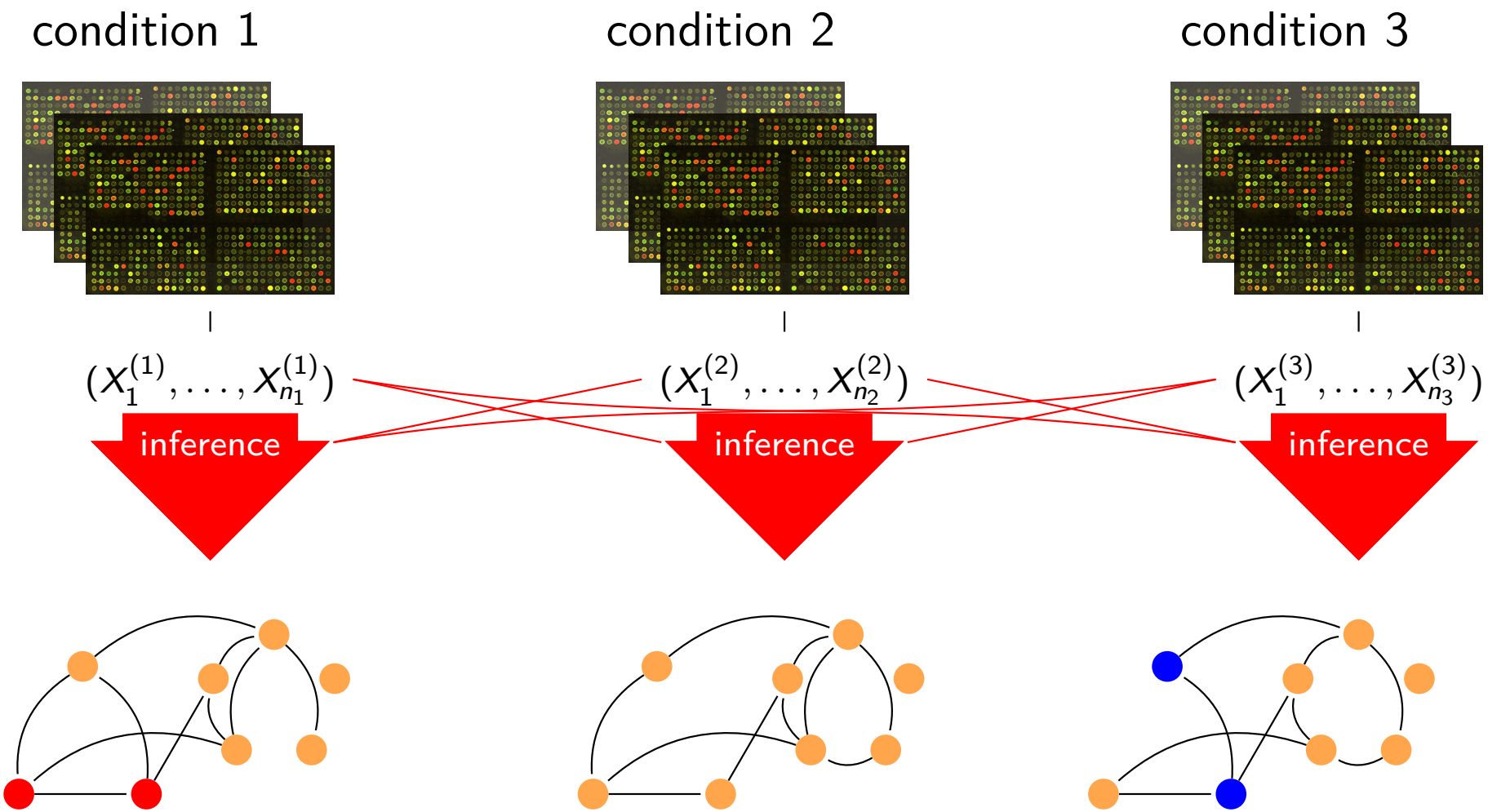
$$(X_1, .., X_p) \sim N(\beta, \Sigma).$$

- In the case of real data with K different conditions of p genes:

$$(X_1, .., X_p)^k \sim N(\beta^k, \Sigma^k).$$

Merge several experimental conditions

By breaking the separability



How to estimate those parameters ?

Linear Regression Model

- Adapt the previous works of Meinshausen and Buhlmann in 2006 (Neighborhood selection). GGM becomes:

$$X_{ij}^k = \beta_j^k + \sum_{a=1}^{p} \theta_{ja}^k (X_{ia}^k - \beta_a^k) + \epsilon_j^k$$
$$\epsilon_j^k \sim N(0, \sigma^2)$$

- Note that the adjacency matrix of θ^k and $(\Sigma^{-1})^k$ are the same:

$$\theta_{ij}^k \neq 0 \Longleftrightarrow (\Sigma^{-1})_{ij}^k \neq 0$$
$$\theta_{ij}^k = 0 \Longleftrightarrow (\Sigma^{-1})_{ij}^k = 0$$

Model

Notations

- X_{ij}^k is the expression level of gene j with replication i, in the condition k.
- β_j^k is the mean expression level of gene j in the condition k.
- θ_{ja}^k explains the relation between genes a and gene j.

Model

$$X_{ij}^k = \beta_j^k + \sum_{a=1}^{p} \theta_{ja}^k (X_{ia}^k - \beta_a^k) + \epsilon_j^k,$$

Criterion

Minimize

$$E = L + \lambda_1 F(\theta^k) + \lambda_2 (\sum_j \omega_{12} |\beta_j^1 - \beta_j^2|) + \lambda_3 \sum_k \|\beta^k\|_1$$

$$L := \sum_{i,j,k} \|X_{ij}^k - \beta_j^k - \sum_{a=1}^{p} \theta_{ja}^k (X_{ia}^k - \beta_a^k)\|_2^2$$

Penalties

- 1st Penalty : Fewer edges or taking similarity networks into account.
- 2nd Penalty : Fused β.
- 3rd Penalty : Control the Magnitude of β.

Choices of F

1st Penalty- Tibshirani et al(1996) - Chiquet et al (2011)

- Lasso:

$$\lambda_1 \sum_{j \neq a} \sum_{k} |\theta_{ja}^k|$$

- Group Lasso:

$$\lambda_1 \sum_{j \neq a} (\sum_{k} (\theta_{ja}^k)^2))^{1/2}$$

- Cooperative Lasso:

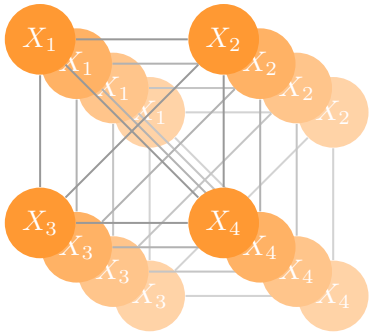
$$\lambda_1 \sum_{j \neq a} (\sum_{k} (-\theta_{ja}^k, 0)_+^2))^{1/2} + \lambda_1 \sum_{j \neq a} (\sum_{k} (\theta_{ja}^k, 0)_+^2))^{1/2}$$

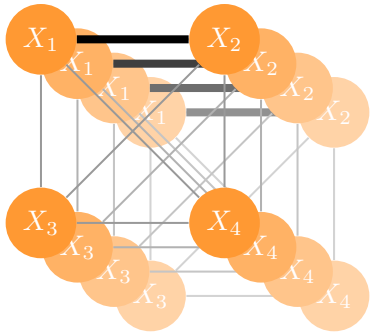
Grouping effects induced

Potential groups

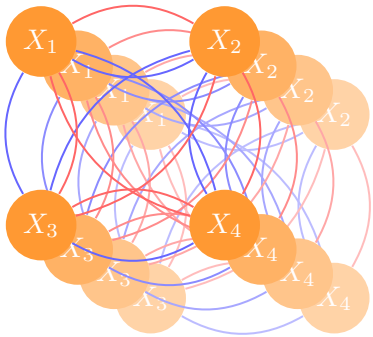
Group(s) induced by edges $(1, 2)$

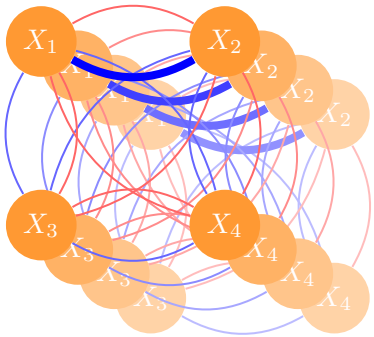
Group-Lasso

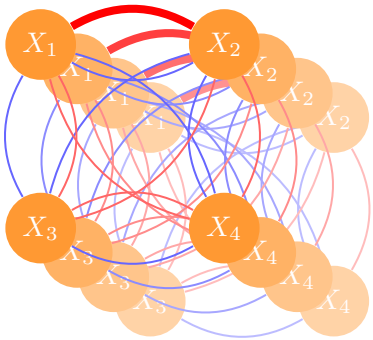




Cooperative-Lasso







Estimate the parameters

Minimize

$$E = L + \lambda_1 \sum_k \|\theta^k\|_1 + \lambda_2 (\sum_j \omega_{12} |\beta_j^1 - \beta_j^2|) + \lambda_3 \sum_k \|\beta^k\|_1$$

$$L := \sum_{i,j,k} \|X_{ij}^k - \beta_j^k - \sum_{a=1}^{p} \theta_{ja}^k (X_{ia}^k - \beta_a^k)\|_2^2$$

Algorithm

While (not converge) **do**

- Fixed all $\beta^{(k)}$, find $\theta^{(k)}$ [*simone* - Chiquet et al, *glasso* - Tibshirani et al].
- Fixed all $\theta^{(k)}$, find $\beta^{(k)}$ [*genlasso* - Arnold et al].

end

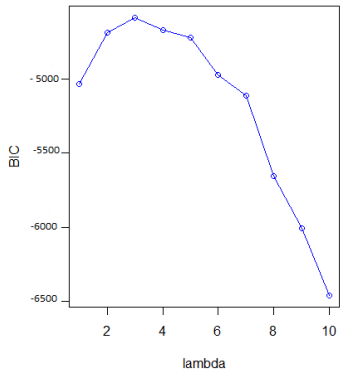
Choice for $\lambda_1, \lambda_2, \lambda_3$

BIC criterion

$$\text{BIC}(\lambda_1, \lambda_2, \lambda_3) = 2\text{Log-likelihood} - df \times \text{Log(nK)}$$

$$df = \sharp\{(j,k)|\beta_j^k \neq 0\} + \sharp\{(i,j,k)|\theta_{i,j}^k \neq 0\}/2$$

- Making a 3 dimensions grid of triplet $(\lambda_1, \lambda_2, \lambda_3)$.
- Choose the triplet which maximize BIC.



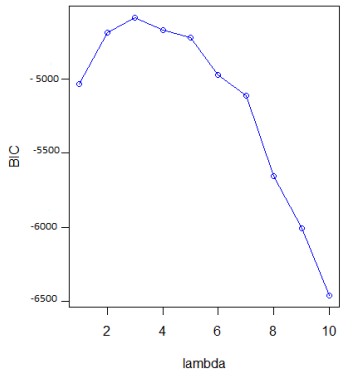
Choice for $\lambda_1, \lambda_2, \lambda_3$

BIC criterion

$$\text{BIC}(\lambda_1, \lambda_2, \lambda_3) = 2\text{Log-likelihood} - df \times \text{Log(nK)}$$

$$df = \sharp\{(j,k)|\beta_j^k \neq 0\} + \sharp\{(i,j,k)|\theta_{i,j}^k \neq 0\}/2$$

- Making a 3 dimensions grid of triplet $(\lambda_1, \lambda_2, \lambda_3)$.
- Choose the triplet which maximize BIC.



Numerical Experiments

Simulation data

Details

- We consider the case with only 2 conditions. For all conditions, we choose the same number of replications $n = \{30, 60, 100, 200\}$. Number of variables, or genes $p = 100$ always.
- Each data file contains two matrices $n \times p$ corresponding with 2 conditions.

scenarios

1. Two simulated data have same θ(s), same $\beta^k(s)$.
2. Two simulated data have same θ(s), a percentage of $\beta^k(s)$ is different.

Simulation data

Details

- We consider the case with only 2 conditions. For all conditions, we choose the same number of replications $n = \{30, 60, 100, 200\}$. Number of variables, or genes $p = 100$ always.
- Each data file contains two matrices $n \times p$ corresponding with 2 conditions.

scenarios

1. Two simulated data have same θ(s), same $\beta^k(s)$.
2. Two simulated data have same θ(s), a percentage of $\beta^k(s)$ is different.

Measures of Quality

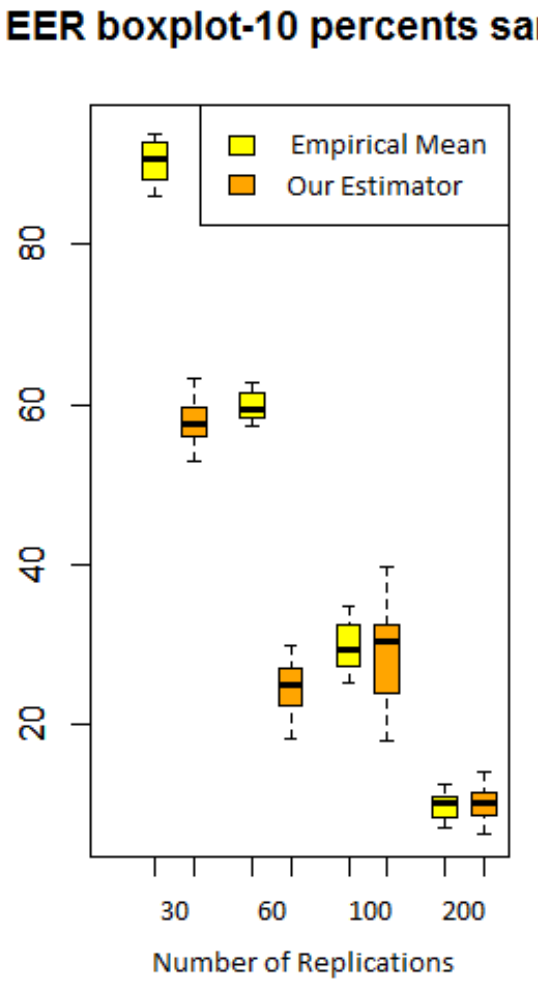
- Relative Error:

$$RE(\hat{\beta}^{(1)}, \hat{\beta}^{(2)}) = \frac{100}{K \times p} \sum_j \sum_k \frac{|\hat{\beta}_j^{(k)} - \beta_j^{(k)true}|}{|\beta_j^{(k)true}|}$$

- Mean Square Error:

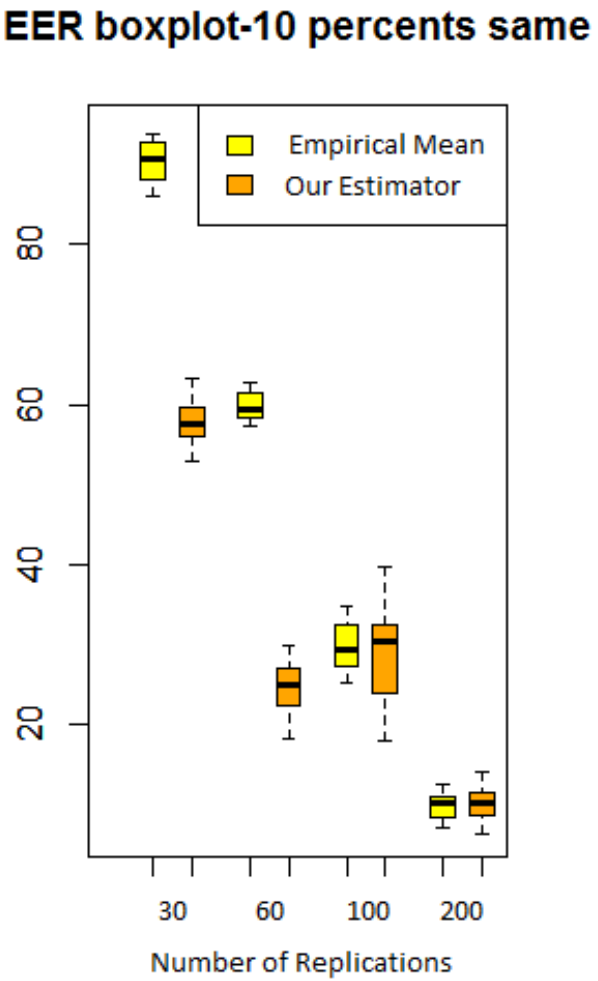
$$MSE(\hat{\beta}^{(1)}, \hat{\beta}^{(2)}) = \frac{1}{K \times p} \sum_k \sum_j |\hat{\beta}_j^{(k)} - \beta_j^{(k)true}|^2$$

Comparing our mean with empirical mean



- Our estimator is better than the empirical mean, especially in case of few replications.
- They all tends to the true values in case of many replications.

Comparing our mean with empirical mean



- Our estimator is better than the empirical mean, especially in case of few replications.
- They all tends to the true values in case of many replications.

Comparing our network with other methods

Algorithm

While (not converge) **do**

- Fixed all $\beta^{(k)}$, find $\theta^{(k)}$ [*simone* - Chiquet et al, *glasso* - Tibshirani et al].
- Fixed all $\theta^{(k)}$, find $\beta^{(k)}$ [*genlasso* - Arnold et al].

end

- We usually got the same adjacency matrix.
- However the magnitude of θ(s) are different.

Outlook

Conclusion

- We propose a new way to estimate average level expression of genes while using GGM and linear regression model for gene expression data.
- In term of mean expression, we got some good results. However, we have not improved results on the networks yet.

Perspective

- Finding different reactions of networks and genes in different conditions.
- Theoretical results on consistency of our estimators $(\hat{\beta}, \hat{\theta})$ are in progress.

Outlook

Conclusion

- We propose a new way to estimate average level expression of genes while using GGM and linear regression model for gene expression data.
- In term of mean expression, we got some good results. However, we have not improved results on the networks yet.

Perspective

- Finding different reactions of networks and genes in different conditions.
- Theoretical results on consistency of our estimators $(\hat{\beta}, \hat{\theta})$ are in progress.

Reference

- Tibshirani, R. Regression shrinkage and selection via the lasso. J. Roy. Statist, (1996).
- Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, Keith Knight. Sparsity and smoothness via the fused lasso. Journal of the Royal Statistical Society, (2005)
- Julien Chiquet, Yves Grandvalet, Christophe Ambroise. Inferring Multiple Graphical Models. (2011)
- Nicolai Meinshausen, Peter Buhlmann. High dimensional graphs and variable selection with the LASSO. The Annals of Statistics, (2006).

THANK YOU !