An attempt to couple network inference and differential analysis

Pierre Gutierrez's MsC research training period

Pierre Gutierrez, Guillem Rigaill and Julien Chiquet

Statistique et Génome, CNRS & Université d'Évry Val d'Essonne

NETBIO – 2012, Nov. the 19th

http://stat.genopole.cnrs.fr/

Agenda

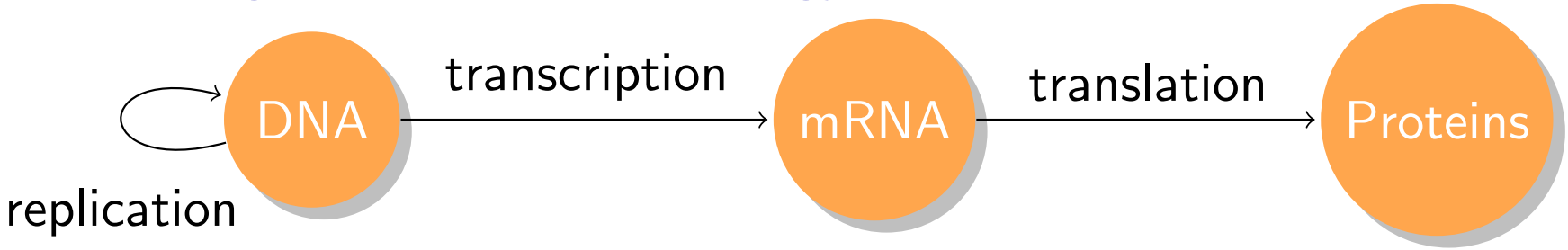
Motivations

- Biostatistical context
- Statistical issues

Current research leads and progress

What are we looking at?

Central dogma of molecular biology



Proteins

- are building blocks of any cellular functionality,
- are encoded by the genes,
- *do* interact (at the protein and gene level – regulations).

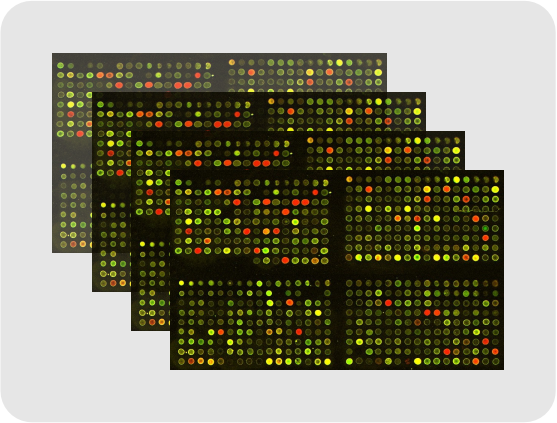
Basic biostatistical issues

1. Selecting some genes of interest (biomarkers)
   - Differential analysis
2. Looking for interactions between them (pathway analysis).
   - Network inference

How is this measured? (1)

Microarray technology: parallel measurement of many biological features

signal processing



Matrix of features $n \ll p$

Expression levels of p probes are simultaneously monitored for n individuals

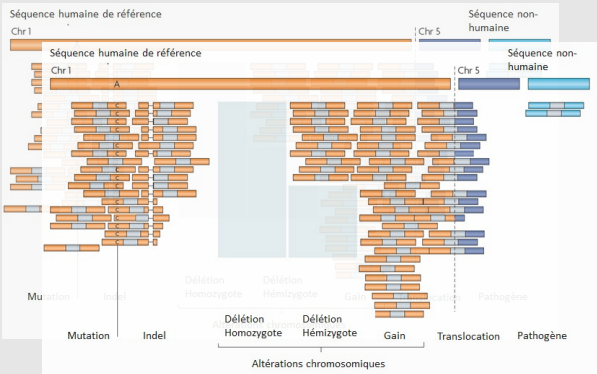
pretreatment

$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \dots & x_n^p \end{pmatrix}$$

How is this measured? (2)

Next Generation Sequencing: parallel measurement of even many more biological features

assembling



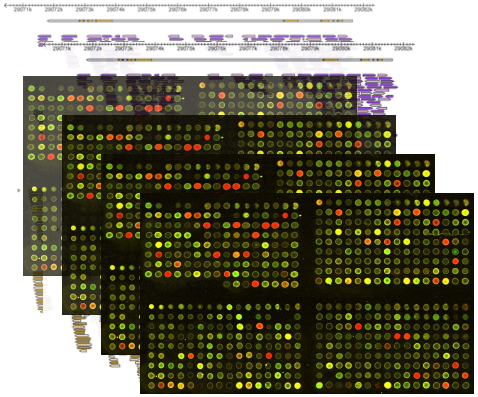
Matrix of features $n \lll p$

Expression counts are extracted from small repeated sequences and monitored for n individuals

pretreatment

$$\mathbf{X} = \begin{pmatrix} k_1^1 & k_1^2 & k_1^3 & \dots & k_1^p \\ \vdots & & & & \\ k_n^1 & k_n^2 & k_1^2 & \dots & k_n^p \end{pmatrix}$$

The problem at hand



Inference

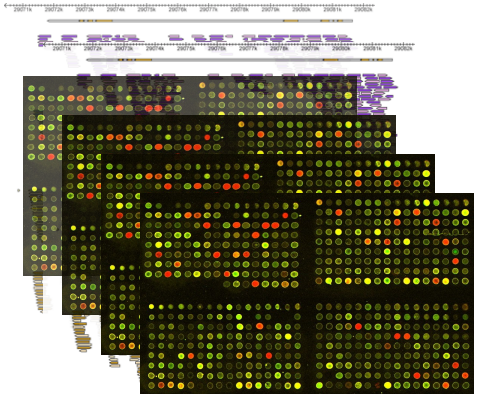
$\approx$ 10s/100s microarray/sequencing experiments

$\approx$ 1000s probes ("genes")

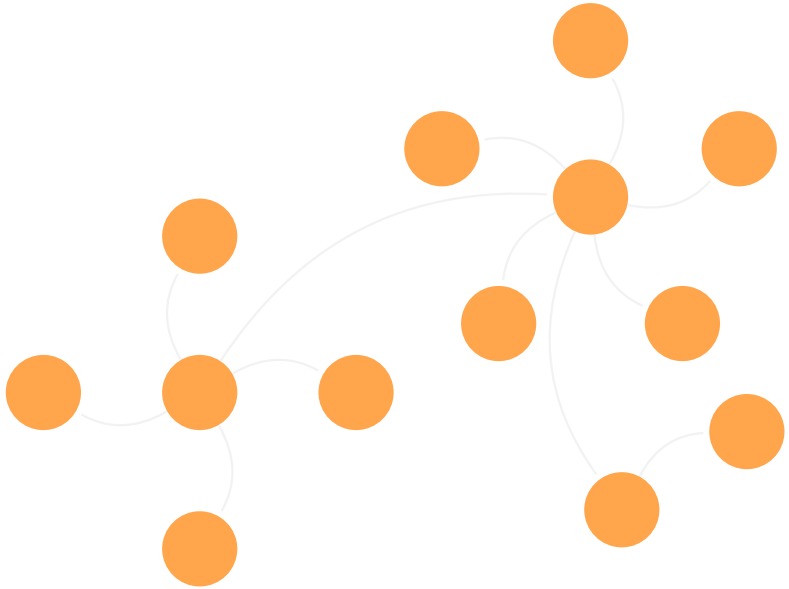
Questions

- What genes are interacting?
- What is the underlying interaction?

The problem at hand



Inference



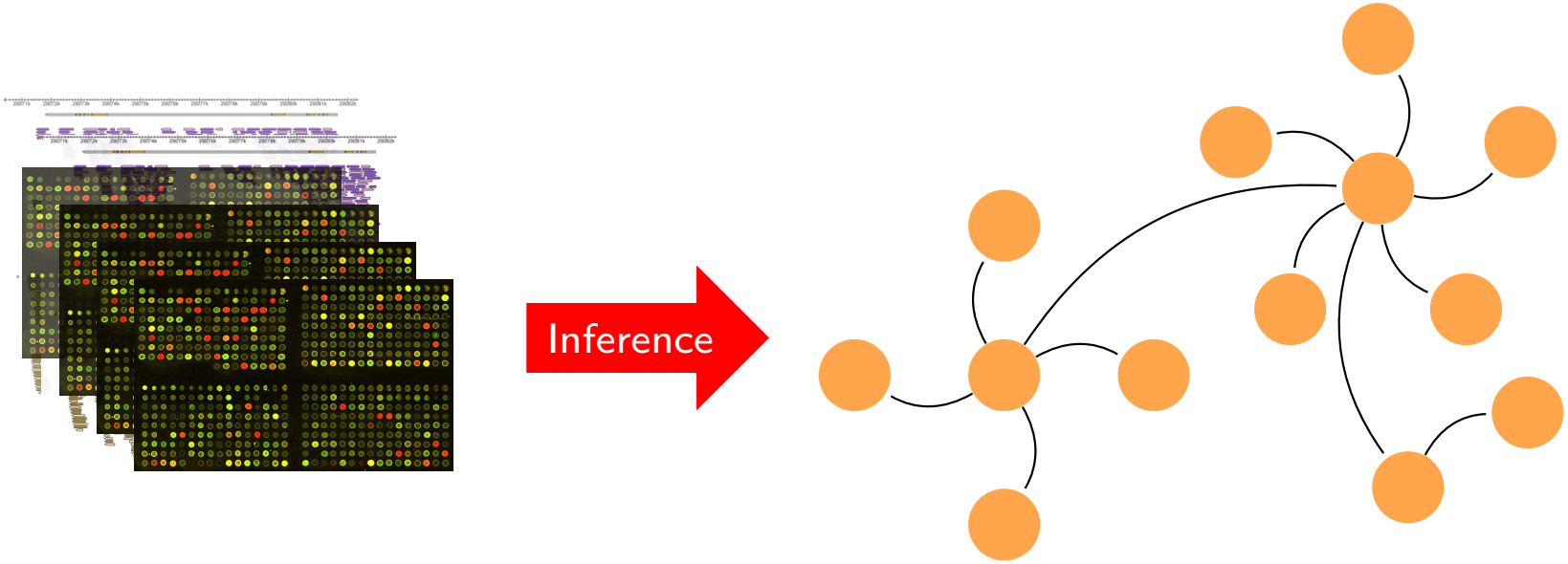
$\approx$ 10s/100s microarray/sequencing experiments

$\approx$ 1000s probes ("genes")

Questions

1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?

The problem at hand



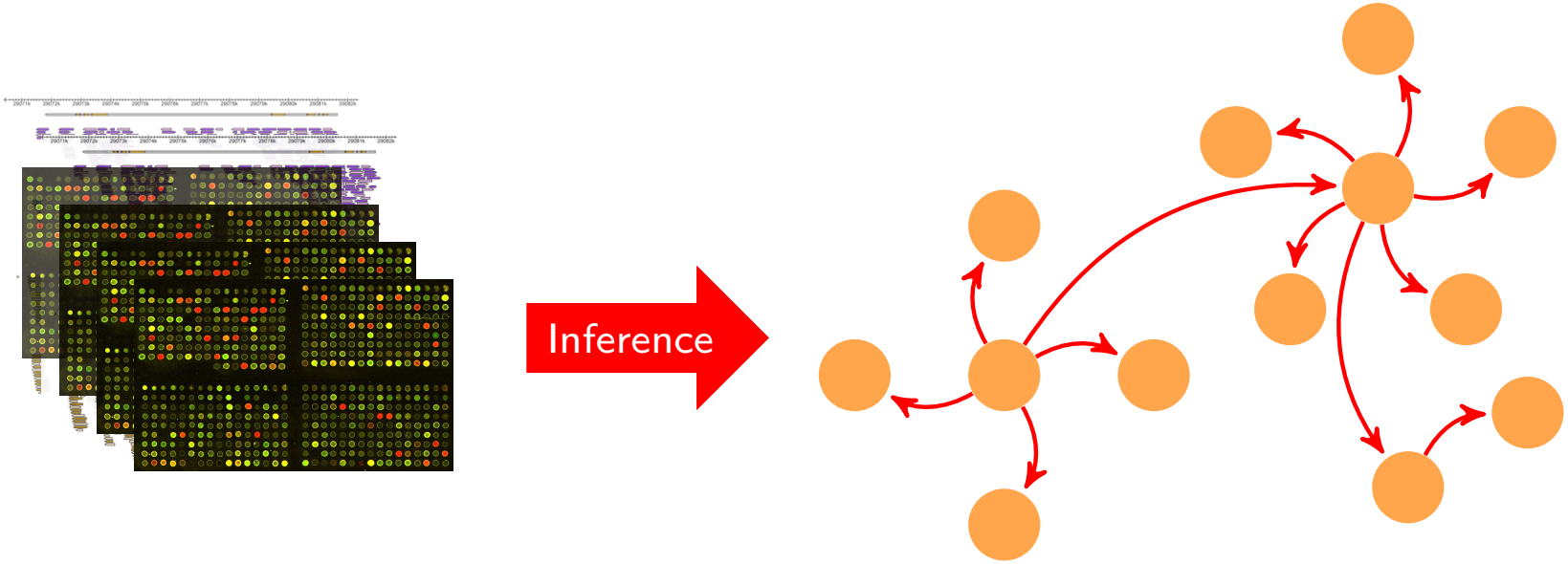
$\approx$ 10s/100s microarray/sequencing experiments

$\approx$ 1000s probes ("genes")

Questions

1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?

The problem at hand



$\approx$ 10s/100s microarray/sequencing experiments

$\approx$ 1000s probes ("genes")

Questions

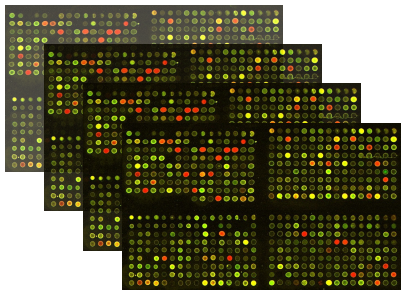
1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?

Handling the scarcity of data (1)

By reducing the number of parameters

Assumption

Connections will only appear between informative genes



Candidate genes

select p key genes $\mathcal{P}$

p "reasonable" compared to n

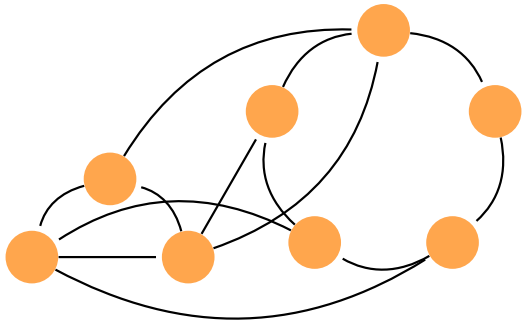
typically, $n \in [p/5; 5p]$

the learning dataset

n size–p vectors of expression

$(X_1, \ldots, X_n)$ with $X_i \in \mathbb{R}^p$

inference



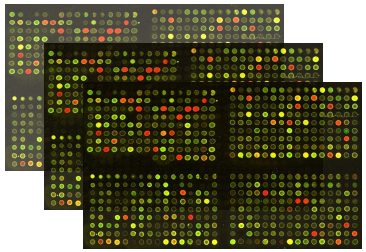
Handling the scarcity of data (2)

By taking as many observations as possible into account

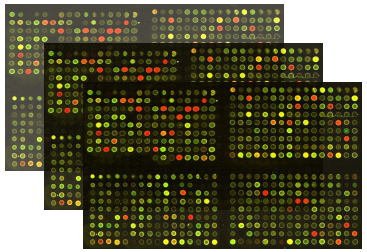
Multitask learning

How should we merge the data?

Condition 1



Condition 2



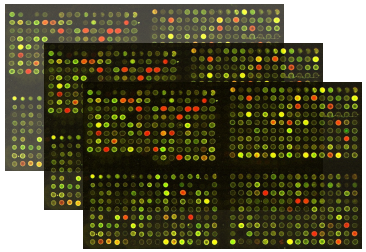
Handling the scarcity of data (2)

By taking as many observations as possible into account

Multitask learning

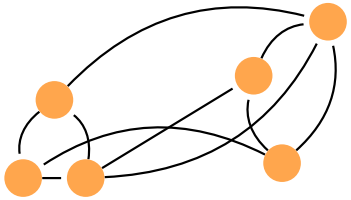
by inferring each network independently

Condition 1

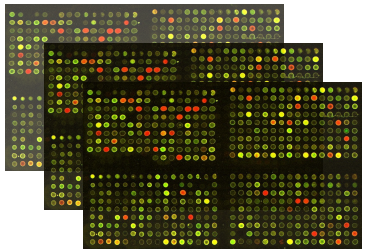


$(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), X_i^{(1)} \in \mathbb{R}^{p_1}$

inference

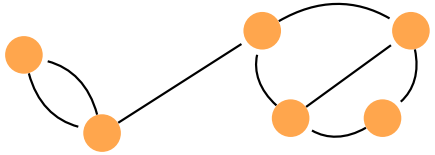


Condition 2



$(X_1^{(2)}, \ldots, X_{n_2}^{(2)}), X_i^{(2)} \in \mathbb{R}^{p_2}$

inference



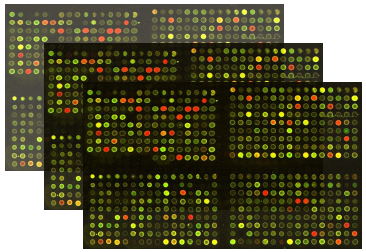
Handling the scarcity of data (2)

By taking as many observations as possible into account

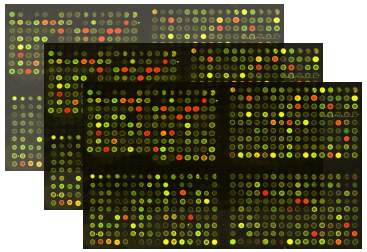
Multitask learning

by pooling all the available data

Condition 1

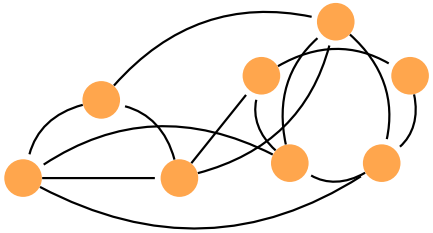


Condition 2



$(X_1, \ldots, X_n), X_i \in \mathbb{R}^p$, with $n = n_1 + n_2$.

inference



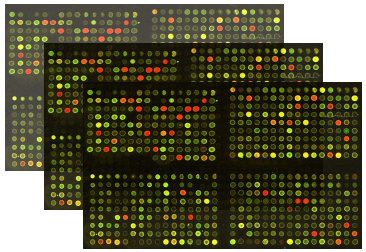
Handling the scarcity of data (2)

By taking as many observations as possible into account

Multitask learning

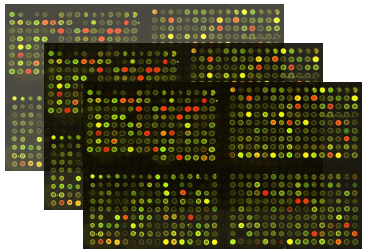
by breaking the separability

Condition 1



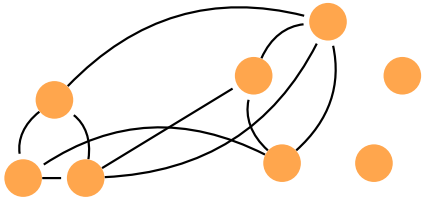
$(X_1^{(1)}, \ldots, X_{n_1}^{(1)}), X_i^{(1)} \in \mathbb{R}^{p_1}$

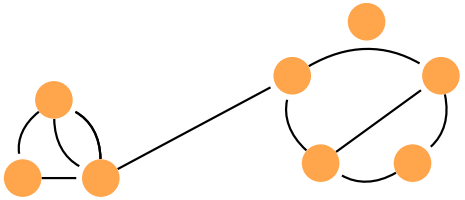
Condition 2



$(X_1^{(2)}, \ldots, X_{n_2}^{(2)}), X_i^{(2)} \in \mathbb{R}^{p_2}$

inference





Multiple network inference and differential analysis

Differential analysis studies

Conditions 1 and 2 typically stand for

- stress experiments,
- case/control studies,
- placebo/treatment studies, ...

Current network inference strategy

To handle scarcity of data in that context, we

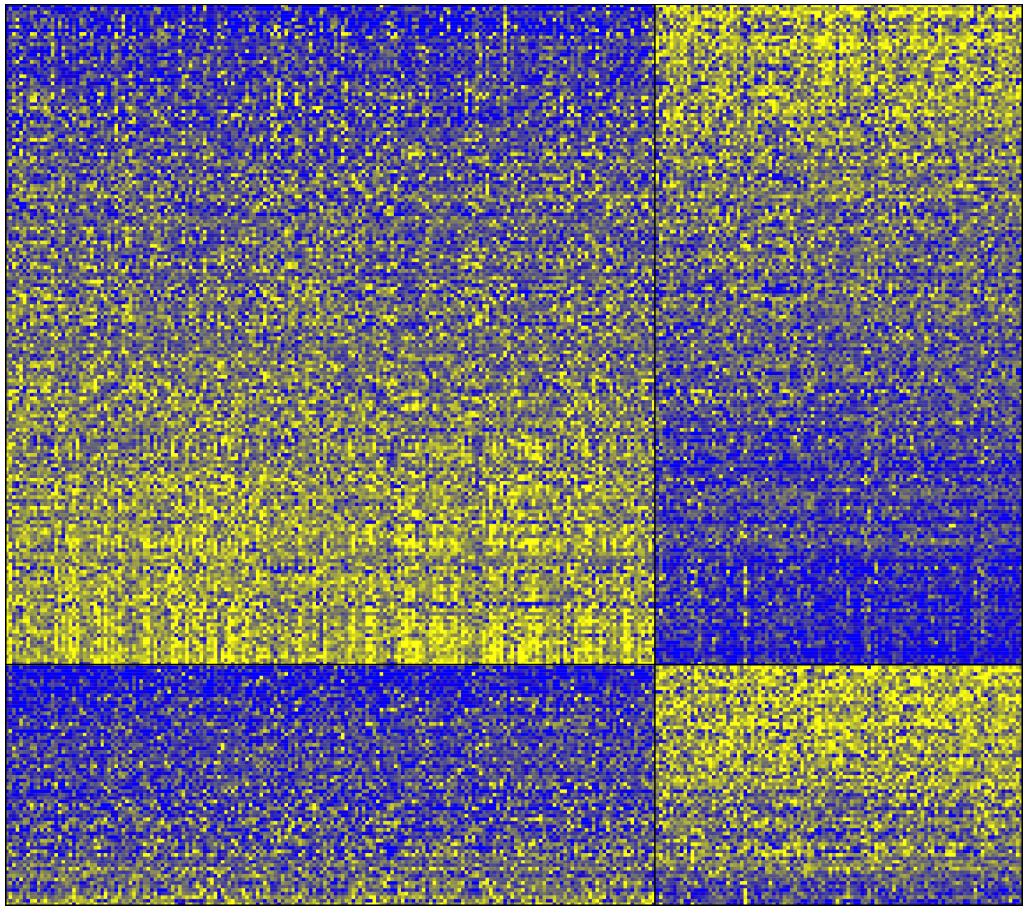
1. perform a differential analysis to select a set of candidate genes,
2. perform joint network inference on this restricted set of genes.

J. Chiquet, Y. Granvalet and C. Ambroise
Infering multiple graphical structures

Multiple network inference and differential analysis

Illustration on the Loi dataset

- $n_R = 68$ tamoxifen-resistant tumors
- $n_R = 187$ tamoxifen-sensible tumors
- Expression matrix $\mathbf{X}$ has 255 rows (patients) and 15,537 columns (genes),
- $\mathbf{X}$ has been ordered and cut with BH multiple-testing procedure at 5%.



$\rightsquigarrow$ Multiple network inference is performed on this restricted matrix.

Issues

Why doing this?

The underlying statistical models (GGM or linear model) are known not to perform well[1] in ultra-high dimension ($n \lll p$). See e.g.

N. Verzelen.

Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons

$\rightsquigarrow$ We *have* to limit the number of genes in the networks.

Perspectives

1. How this 2-step procedure affects the inferred networks?
2. Can we do better by performing simultaneously differential analysis *and* network inference?

[1] meaning completely 'useless'

Issues

Why doing this?

The underlying statistical models (GGM or linear model) are known not to perform well[1] in ultra-high dimension ($n \lll p$). See e.g.

N. Verzelen.

Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons

$\rightsquigarrow$ We *have* to limit the number of genes in the networks.

Perspectives

1. How this 2-step procedure affects the inferred networks?
2. Can we do better by performing simultaneously differential analysis *and* network inference?

[1] meaning completely 'useless'

Agenda

Motivations

- Biostatistical context
- Statistical issues

Current research leads and progress

Penalized Regression

Network inference

- Network inference = Inverse covariance matrix inference
- Assumption : sparse matrix => Penalized Regression (convex problem)

N. Meinshausen and P. Buhlmann

High-dimensional graphs and variable selection with the lasso

Differential Analysis

- First objective : can we formulate differential analysis as a penalized regression ?
- => Our solution : Fused Anova (convex)
- Having these two penalities, can we merge them to have a unified problem ?

Penalized Regression

Objectives

- Formulating Differential Analysis as a penalized Regression
- Including the effect of a known network
- Infering the network while performing the differential analysis

Fused Anova Model

Fused Anova

- Penalised Regression using the fused Lasso penality

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \frac{1}{2}\sum_k n_k \left(Y_\bullet^{(k)} - \beta_k\right)^2 + \lambda \sum_{k\neq\ell} \left(\omega_{k\ell}|\beta_k - \beta_\ell|\right)$$

  - K number of groups
  - n_k number of individuals for group k
  - $Y_\bullet^{(k)}$ the mean of group k
  - λ penalty coefficient
  - $\omega_{k\ell}$ weights

Fused Anova Model

Fused Anova

- Penalised Regression using the fused Lasso penality

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^K} \frac{1}{2}\sum_k n_k \left(Y_\bullet^{(k)} - \beta_k\right)^2 + \lambda \sum_{k\neq\ell} \left(\omega_{k\ell}|\beta_k - \beta_\ell|\right)$$

- Similar to the Clusterpath and CAS-ANOVA

T.B. Hocking, A. Joulin, F. Bach and J-P. Vert
Clusterpath: An Algorithm for Clustering using Convex Fusion Penalties

H. D. Bondell and B. J. Reich
Simultaneous factor selection and collapsing levels in ANOVA

Fused Anova Model

Properties

- Simple designs => fast and easy to implement path algorithm

  H. Hoefling
  A path algorithm for the Fused Lasso Signal Approximator

- For two groups : statistic t $= \lambda_{fuse}$
  - Default weights ($\omega_{k\ell} = n_k n_\ell$) => same ROC curve performances than the t-test
  - Other weights can do better but loose part of the algorithm efficiency
- For more than two groups :
  - Do not need to run all pairwise tests
  - The hierarchy is directly generated for each variable

Further studies

Including the effect of a known network

L. Jacob, P. Neuvial and S. Dudoit

More Power via Graph-Structured Tests for differential Analysis of Gene Networks

F. Rapaport, A. Zinovyev, M. Dutreix, E. Barillot and J. P. Vert

Classification of microarray data using gene networks

Our problem would thus be :

$$\arg\min_{\boldsymbol{B}} \operatorname{tr}\left(\left(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B}^T\right)\boldsymbol{\Omega}(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{B})\right) + \lambda \boldsymbol{W}||\boldsymbol{D}\boldsymbol{B}||_1$$

Further studies

Coupling Network Inference and Differential Analysis

A. J. Rothman, E. Levina and J. Zhu

Sparse Multivariate Regression with Covariance Estimation

K. Sohn and S. Kim

Joint Estimation of Structured Sparsity and Output Structure in Multiple-Output Regression via Inverse-Covariance Regularization

Conclusion

Near Future work

- Fused Anova performance testing
- Work on its statistical properties
- Including the effect of a known network
- Implementation in R and C

Thank You for your attention