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

Matrix of features $n \ll p$
Expression levels of $p$ probes are simultaneously monitored for $n$ individuals

$X = \begin{pmatrix}
  x_1^1 & x_2^1 & x_3^1 & \ldots & x_p^1 \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  x_n^1 & x_n^2 & x_n^3 & \ldots & x_n^p
\end{pmatrix}$
Matrix of features $n \ll p$

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

$X = \begin{pmatrix} k^1_1 & k^2_1 & k^3_1 & \ldots & k^p_1 \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ k^1_n & k^2_n & k^2_1 & \ldots & k^p_n \end{pmatrix}$
The problem at hand

≈ 10s/100s microarray/sequencing experiments
≈ 1000s probes ("genes")

Questions

1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?
The problem at hand

≈ 10s/100s microarray/sequencing experiments
≈ 1000s probes ("genes")

Questions

1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?
The problem at hand

≈ 10s/100s microarray/sequencing experiments
≈ 1000s probes ("genes")

Questions

1. Which nodes (subset of relevant genes)?
2. Which edges (significant interactions)?
The problem at hand

≈ 10s/100s microarray/sequencing experiments
≈ 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

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 $X_i \in \mathbb{R}^p$
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**

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

**Condition 2**

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

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

Network 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

(\(X_1^{(2)}, \ldots, X_{n_2}^{(2)}\), \(X_i^{(2)}\) \(\in \mathbb{R}^{p_2}\))
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 $X$ has 255 rows (patients) and 15,537 columns (genes),
- $X$ has been ordered and cut with BH multiple-testing procedure at 5%.

⇝ 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 \((n \ll p)\). See e.g. N. Verzelen.

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

\(\implies\) 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 \((n \ll p)\). See e.g.

\[ N. \text{ Verzelen.} \]

Minimax risks for sparse regressions: Ultra-high-dimensional phenomenons
\[ \rightsquigarrow \text{We have to limit the number of genes in the networks.} \]

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
Network inference

Network inference = Inverse covariance matrix inference

Assumption : sparse matrix $\Rightarrow$ 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 ?

$\Rightarrow$ Our solution : Fused Anova (convex)

Having these two penalities, can we merge them to have a unified problem ?
Objectives

- Formulating Differential Analysis as a penalized Regression
- Including the effect of a known network
- Inferring the network while performing the differential analysis
Fused Anova Model

Fused Anova

- Penalised Regression using the fused Lasso penalty

\[
\min_{\beta \in \mathbb{R}^K} \frac{1}{2} \sum_k n_k \left( Y_k' - \beta_k \right)^2 + \lambda \sum_{k \neq \ell} (\omega_{k\ell} |\beta_k - \beta_\ell|)
\]

- \(K\) number of groups
- \(n_k\) number of individuals for group \(k\)
- \(Y_k'\) the mean of group \(k\)
- \(\lambda\) penalty coefficient
- \(\omega_{k\ell}\) weights
Fused Anova Model

Fused Anova

- Penalised Regression using the fused Lasso penalty

\[
\min_{\beta \in \mathbb{R}^K} \frac{1}{2} \sum_k n_k \left( Y^{(k)}_k - \beta_k \right)^2 + \lambda \sum_{k \neq \ell} (\omega_{k\ell} | \beta_k - \beta_{\ell} |)
\]

- 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
Properties

- Simple designs $\implies$ 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 $(\omega_{k\ell} = n_k n_\ell) \implies$ 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:

$$\arg \min_B \text{tr} \left( (Y - XB^T) \Omega (Y - XB) \right) + \lambda W \| DB \|_1$$
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