An introduction to graph analysis and modeling
Topology Inference

MSc in Statistics for Smart Data – ENSAI

Autumn semester 2017

http://julien.cremeriefamily.info

Introduction

first two courses: Analysis of an existing, observed network

- $\rightsquigarrow$ basic characterization
- $\rightsquigarrow$ find an underlying organization (clustering)

Today: reconstruct (infer) a network from external data

Become familiar with

- Gaussian graphical models,
- sparse inference methods (ℓ_1-regularization *a.k.a* the Lasso).

Canonical example: Genomic data

We consider examples from life science, but everything said extends to

- Sociology, Astrophysics, Stock exchange, Insurance data, ...
- ...any multivariate data.

Introduction

first two courses: Analysis of an existing, observed network

$\rightsquigarrow$ basic characterization

$\rightsquigarrow$ find an underlying organization (clustering)

Today: reconstruct (infer) a network from external data

Become familiar with

- Gaussian graphical models,
- sparse inference methods (ℓ_1-regularization *a.k.a* the Lasso).

Canonical example: Genomic data

We consider examples from life science, but everything said extends to

- Sociology, Astrophysics, Stock exchange, Insurance data, ...
- ...any multivariate data.

Introduction

first two courses: Analysis of an existing, observed network

$\rightsquigarrow$ basic characterization

$\rightsquigarrow$ find an underlying organization (clustering)

Today: reconstruct (infer) a network from external data

Become familiar with

- Gaussian graphical models,
- sparse inference methods (ℓ_1-regularization *a.k.a* the Lasso).

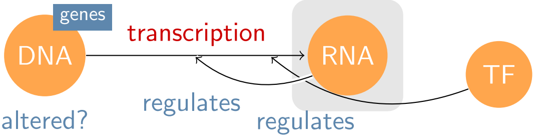
Canonical example: Genomic data

We consider examples from life science, but everything said extends to

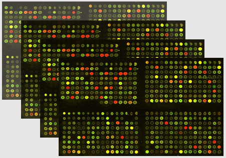
- Sociology, Astrophysics, Stock exchange, Insurance data, ...
- ...any multivariate data.

Omic: parallel measurement of many biological features

Focus e.g. on *transcription*, looking toward *gene regulatory networks*



signal processing



Matrix of features $n \ll p$

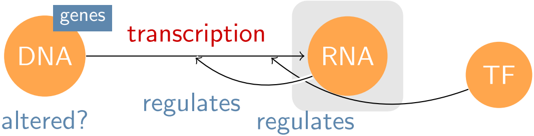
Expression levels of p probes are simultaneously monitored for n individuals

pretreatment

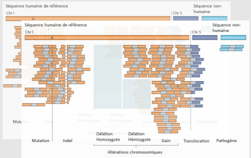
$$\mathbf{X} = \begin{pmatrix} x_1^1 & x_1^2 & x_1^3 & \dots & x_1^p \\ \vdots & & & & \\ x_n^1 & x_n^2 & x_1^2 & \dots & x_n^p \end{pmatrix}$$

Omic: parallel measurement of many biological features

Focus e.g. on *transcription*, looking toward *gene regulatory networks*



assembling



Matrix of features $n \lll p$

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

pretreatment

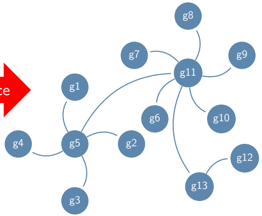
$$\mathbf{X} = \begin{pmatrix} k_1^1 & k_1^2 & k_1^3 & \dots & k_1^p \\ \vdots & & & & \\ k_n^1 & k_n^2 & k_1^2 & \dots & k_n^p \end{pmatrix}$$

Network inference: a challenging problem

$\approx$ 10s/1,000s assays

$\approx$ 1,000s/1,000,000s features

Inference



1. Nodes are fixed
   - restricted to a set of interest
2. Edges (interactions) are inferred
   - based upon statistical concepts

Statistical question

1. Variable selection (which edges?)

Main statistical challenges

1. (Ultra) High dimensionality ($n < p$, $n \lll p$)
2. Heterogeneity/structure of the data

Outline

1. Network and data modeling
   - Statistical dependence
   - Gaussian Graphical models
2. Network Inference
   - Inducing sparsity for edge selection
   - Limitations and extensions of sparse GGM
   - Example: plasmodium data set

Outline

References

- Graphical Models in Applied Multivariate Statistics, Joe Whittaker
- Graphical Models, S. Lauritzen

Outline

Modeling relationship between variables

Independence

Definition (Independence of events)

Two events A and B are independent if and only if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B),$$

which is usually denoted by $A \perp\!\!\!\perp B$. Equivalently,

- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$,
- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A|B^c)$

Example (class vs party)

	party	
class	Labour	Tory
working	0.42	0.28
bourgeoisie	0.06	0.24

	party	
class	Labour	Tory
working	0.60	0.40
bourgeoisie	0.20	0.80

Table: Joint probability (left) vs. conditional probability (right)

Modeling relationship between variables

Independence

Definition (Independence of events)

Two events A and B are independent if and only if

$$\mathbb{P}(A, B) = \mathbb{P}(A)\mathbb{P}(B),$$

which is usually denoted by $A \perp\!\!\!\perp B$. Equivalently,

- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A)$,
- $A \perp\!\!\!\perp B \Leftrightarrow \mathbb{P}(A|B) = \mathbb{P}(A|B^c)$

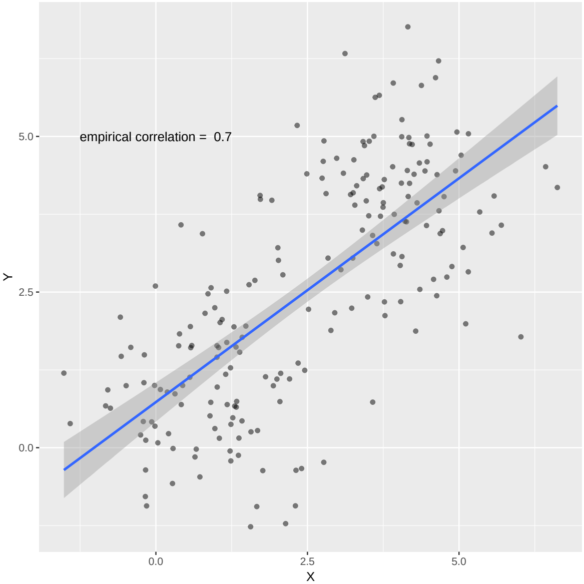
Example (class vs party)

	party	
class	Labour	Tory
working	0.42	0.28
bourgeoisie	0.06	0.24

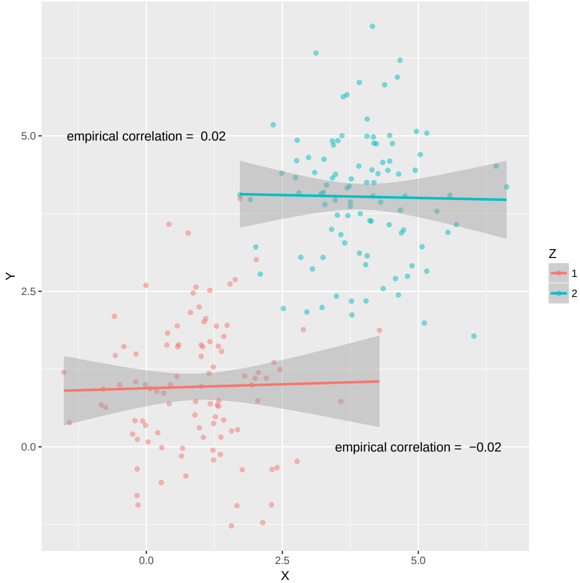
	party	
class	Labour	Tory
working	0.60	0.40
bourgeoisie	0.20	0.80

Table: Joint probability (left) vs. conditional probability (right)

Limits of correlation for network reconstruction



Limits of correlation for network reconstruction



Modeling relationships between variables (2)

Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

Definition (Conditional independence of events)

Two events A and B are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight".

Modeling relationships between variables (2)

Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

Definition (Conditional independence of events)

Two events A and B are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight".

Modeling relationships between variables (2)

Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

Definition (Conditional independence of events)

Two events A and B are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight". Estimating[1] $\mathbb{P}(A, B)$, $\mathbb{P}(A)$ and $\mathbb{P}(B)$ in a sample would lead to

$$\mathbb{P}(A, B) \neq \mathbb{P}(A)\mathbb{P}(B)$$

Modeling relationships between variables (2)

Conditional independence

Generalizing to more than two events requires strong assumptions (mutual independence). Better handle with

Definition (Conditional independence of events)

Two events A and B are conditionally independent if and only if

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C),$$

which is usually denoted by $A \perp\!\!\!\perp B|C$

Example (Does QI depends on weight?)

Consider the events $A =$ "having low QI", $B =$ "having low weight". But in fact, introducing $C =$ "having a given age",

$$\mathbb{P}(A, B|C) = \mathbb{P}(A|C)\mathbb{P}(B|C)$$

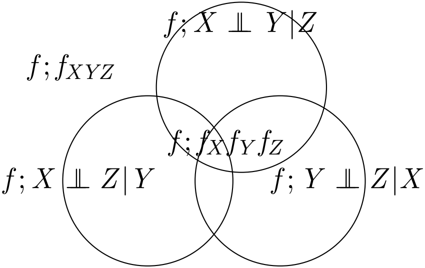
(Conditional) independence of random vectors

A "natural" generalization

Definition

Consider 3 random variables X, Y, Z with distribution f_X, f_Y, f_Z, jointly f_{XY}, f_{XYZ}. Then,

- X and Y are independent iif $f_{XY}(x, y) = f_X(x) f_Y(y)$;
- X and Y are conditionally independent on Z, $z : f_Z(z) > 0$ iif $f_{XY|Z}(x, y; z) = f_{X|Z}(x; z) f_{Y|Z}(y; z)$.



Outline

Graphical models

Definition

A graphical model gives a graphical (intuitive) representation of the dependence structure of a probability distribution, by linking

1. a random vector (or a set of random variables.) $X = \{X_1, \dots, X_p\}$ with distribution $\mathbb{P}$,

2. a graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ where
   - $\mathcal{P} = \{1, \dots, p\}$ is the set of nodes associated to each variable,
   - $\mathcal{E}$ is a set of edges describing the dependence relationship of $X \sim \mathbb{P}$.

Definition

The conditional independence graph of a random vector X is the undirected graph $\mathcal{G} = \{\mathcal{P}, \mathcal{E}\}$ with the set of node $\mathcal{P} = \{1, \dots, p\}$ and where

$$(i, j) \notin \mathcal{E} \Leftrightarrow X_i \perp X_j | \mathcal{P} \backslash \{i, j\}.$$

Graphical models

Definition

A graphical model gives a graphical (intuitive) representation of the dependence structure of a probability distribution, by linking

1. a random vector (or a set of random variables.) $X = \{X_1, \dots, X_p\}$ with distribution $\mathbb{P}$,

2. a graph $\mathcal{G} = (\mathcal{P}, \mathcal{E})$ where
   - $\mathcal{P} = \{1, \dots, p\}$ is the set of nodes associated to each variable,
   - $\mathcal{E}$ is a set of edges describing the dependence relationship of $X \sim \mathbb{P}$.

Definition

The conditional independence graph of a random vector X is the undirected graph $\mathcal{G} = \{\mathcal{P}, \mathcal{E}\}$ with the set of node $\mathcal{P} = \{1, \dots, p\}$ and where

$$(i, j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | \mathcal{P} \backslash \{i, j\}.$$

Conditional Independence Graphs

An example

Let X_1, X_2, X_3, X_4 be four random variables with joint probability density function $f_X(x) = \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4)$ with u a given constant.

Apply the factorization property

$$
\begin{aligned}
f_X(x) &= \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4) \\
&= \exp(u) \ \cdot \ \exp(x_1 + x_1 x_2) \ \cdot \ \exp(x_2 x_3 x_4)
\end{aligned}
$$

Graphical representation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$ such as $\mathcal{P} = \{1, 2, 3, 4\}$ and

$$\mathcal{E} =$$

Conditional Independence Graphs

An example

Let X_1, X_2, X_3, X_4 be four random variables with joint probability density function $f_X(x) = \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4)$ with u a given constant.

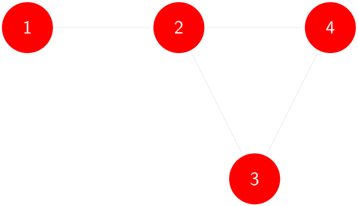
Apply the factorization property

$$
\begin{aligned}
f_X(x) &= \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4) \\
&= \exp(u) \ \cdot \ \exp(x_1 + x_1 x_2) \ \cdot \ \exp(x_2 x_3 x_4)
\end{aligned}
$$

Graphical representation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$ such as $\mathcal{P} = \{1, 2, 3, 4\}$ and

$$\mathcal{E} = \{?\}$$



Conditional Independence Graphs

An example

Let X_1, X_2, X_3, X_4 be four random variables with joint probability density function $f_X(x) = \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4)$ with u a given constant.

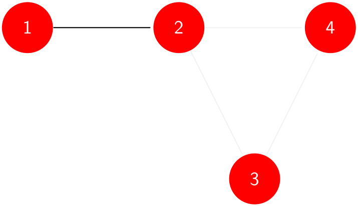
Apply the factorization property

$$\begin{aligned} f_X(x) &= \exp(u + x_1 + x_1 x_2 + x_2 x_3 x_4) \\ &= \exp(u) \; \cdot \; \exp(x_1 + x_1 x_2) \; \cdot \; \exp(x_2 x_3 x_4) \end{aligned}$$

Graphical representation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$ such as $\mathcal{P} = \{1, 2, 3, 4\}$ and

$$\mathcal{E} = \{(1, 2)\}$$



Conditional Independence Graphs

An example

Let X_1, X_2, X_3, X_4 be four random variables with joint probability density function $f_X(x) = \exp(u + x_1 + x_1x_2 + x_2x_3x_4)$ with u a given constant.

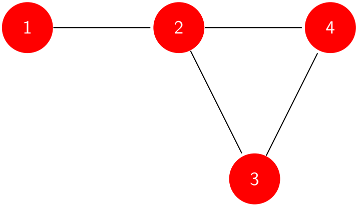
Apply the factorization property

$$
\begin{aligned}
f_X(x) &= \exp(u + x_1 + x_1x_2 + x_2x_3x_4) \\
&= \exp(u) \;\cdot\; \exp(x_1 + x_1x_2) \;\cdot\; \color{red}{\exp(x_2x_3x_4)}
\end{aligned}
$$

Graphical representation

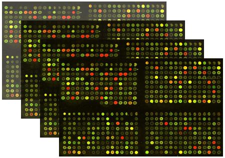
$\mathcal{G} = (\mathcal{P}, \mathcal{E})$ such as $\mathcal{P} = \{1, 2, 3, 4\}$ and

$$\mathcal{E} = \{(2,3), (3,4), (2,4)\}$$



The Gaussian case

The data



Inference

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

Assuming $f_X(\mathbf{X})$ multivariate Gaussian

Greatly simplifies the inference:

- ⇝ naturally links independence and conditional independence to the covariance and partial covariance,
- ⇝ gives a straightforward interpretation to the graphical modeling previously considered.

Why Gaussianity helps?

Case of 2 variables or size-2 random vector

Definitions (Let X, Y be two real random variables.)

$$\mathrm{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

$$\rho_{XY} = \mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathbb{V}(X) \ \cdot \ \mathbb{V}(Y)}}.$$

Proposition

- $\mathrm{cov}(X, X) = \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$,
- $\mathrm{cov}(X + Y, Z) = \mathrm{cov}(X, Z) + \mathrm{cov}(X, Z)$,
- $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + \mathrm{cov}(X, Y)$.
- $X \perp\!\!\!\perp Y \Rightarrow \mathrm{cov}(X, Y) = 0$.
- $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{cov}(X, Y) = 0$ *when* X, Y *are Gaussian.*

Why Gaussianity helps?

Case of 2 variables or size-2 random vector

Definitions (Let X, Y be two real random variables.)

$$\mathrm{cov}(X, Y) = \mathbb{E}\Big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\Big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

$$\rho_{XY} = \mathrm{cor}(X, Y) = \frac{\mathrm{cov}(X, Y)}{\sqrt{\mathbb{V}(X) \cdot \mathbb{V}(Y)}}.$$

Proposition

- $\mathrm{cov}(X, X) = \mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)]$,
- $\mathrm{cov}(X + Y, Z) = \mathrm{cov}(X, Z) + \mathrm{cov}(X, Z)$,
- $\mathbb{V}(X + Y) = \mathbb{V}(X) + \mathbb{V}(Y) + \mathrm{cov}(X, Y)$.
- $X \perp\!\!\!\perp Y \Rightarrow \mathrm{cov}(X, Y) = 0$.
- $X \perp\!\!\!\perp Y \Leftrightarrow \mathrm{cov}(X, Y) = 0$ *when* X, Y *are Gaussian.*

The bivariate Gaussian distribution

The Covariance Matrix

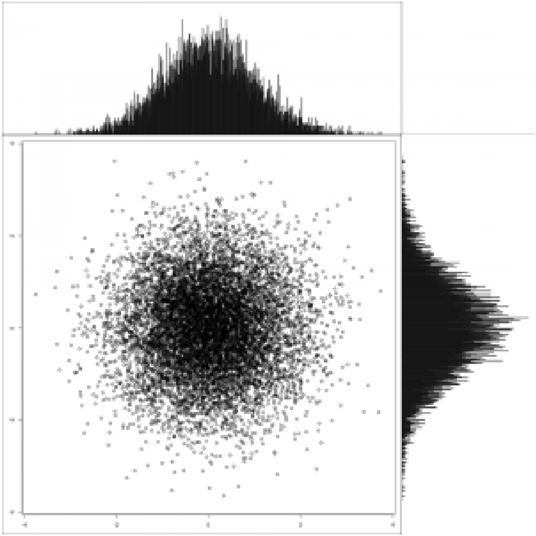
Let

$$X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

with unit variance and $\rho_{XY} = 0$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

The shape of the 2-D distribution evolves accordingly.



The bivariate Gaussian distribution

The Covariance Matrix

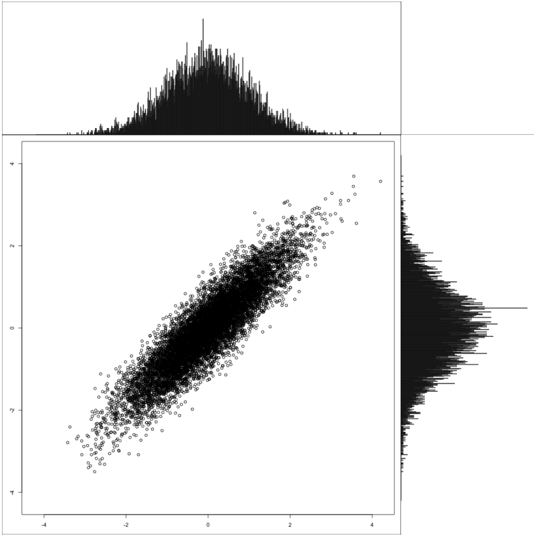
Let

$$X \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}),$$

with unit variance and $\rho_{XY} = 0.9$

$$\mathbf{\Sigma} = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

The shape of the 2-D distribution evolves accordingly.



Generalization: multivariate Gaussian vector

Now need partial covariance and partial correlation

Let X, Y, Z be real random variables.

Definitions

$$\mathrm{cov}(X, Y|Z) = \mathrm{cov}(X, Y) - \mathrm{cov}(X, Z)\mathrm{cov}(Y, Z)/\mathbb{V}(Z).$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho^2_{XZ}}\sqrt{1-\rho^2_{YZ}}}.$$

$\rightsquigarrow$ Give the interaction between X and Y once removed the effect of Z.

Proposition

When X, Y, Z are jointly Gaussian, then

$$\mathrm{cov}(X, Y|Z) = 0 \Leftrightarrow \mathrm{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp Y|Z.$$

Generalization: multivariate Gaussian vector

Now need partial covariance and partial correlation

Let X, Y, Z be real random variables.

Definitions

$$\mathrm{cov}(X, Y|Z) = \mathrm{cov}(X, Y) - \mathrm{cov}(X, Z)\mathrm{cov}(Y, Z)/\mathbb{V}(Z).$$

$$\rho_{XY|Z} = \frac{\rho_{XY} - \rho_{XZ}\rho_{YZ}}{\sqrt{1-\rho_{XZ}^2}\sqrt{1-\rho_{YZ}^2}}.$$

$\rightsquigarrow$ Give the interaction between X and Y once removed the effect of Z.

Proposition

When X, Y, Z are jointly Gaussian, then

$$\mathrm{cov}(X, Y|Z) = 0 \Leftrightarrow \mathrm{cor}(X, Y|Z) = 0 \Leftrightarrow X \perp\!\!\!\perp Y|Z.$$

Gaussian Graphical Model: canonical settings

Experiments in comparable Gaussian conditions

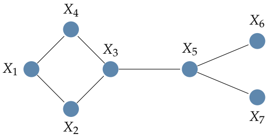
1. $X \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, with $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ the precision matrix.
2. a sample $(X^1, \dots, X^n)$ of exp. stacked in an $n \times p$ data matrix $\mathbf{X}$.

Conditional independence structure

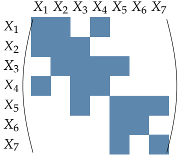
$$(i, j) \notin \mathcal{E} \Leftrightarrow X_i \perp\!\!\!\perp X_j | X_{\setminus \{i,j\}} \Leftrightarrow \Omega_{ij} = 0.$$

Graphical interpretation

$\mathcal{G} = (\mathcal{P}, \mathcal{E})$



$\boldsymbol{\Theta}$



$\rightsquigarrow$ "Covariance" selection

Gaussian vector and linear regression (I)

Proposition (Gaussian vector and conditioning)

$$Z = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma}), \quad \mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}, \quad \mathbf{\Omega} = \mathbf{\Sigma}^{-1} = \begin{pmatrix} \mathbf{\Omega}_{11} & \mathbf{\Omega}_{12} \\ \mathbf{\Omega}_{21} & \mathbf{\Omega}_{22} \end{pmatrix}.$$

Then,

$$Z_2 | Z_1 = \mathbf{z} \sim \mathcal{N}\left(-\mathbf{\Omega}_{22}^{-1}\mathbf{\Omega}_{21}\mathbf{z}, \mathbf{\Omega}_{22}^{-1}\right)$$

and

$$\mathbf{\Omega}_{22}^{-1} = \mathbf{\Sigma}_{22} - \mathbf{\Sigma}_{21}\mathbf{\Sigma}_{11}^{-1}\mathbf{\Sigma}_{12}.$$

Corollary

Partial correlations are related to the inverse of the covariance matrix:

$$\mathrm{cor}(Z_i, Z_j | Z_k, k \neq i, j) = -\frac{\Omega_{ij}}{\sqrt{\Omega_{ii}\Omega_{jj}}}$$

Gaussian vector and linear regression (II)

Consider the linear model

$$Y = X^{\intercal}\boldsymbol{\beta} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \sigma). \tag{1}$$

Other interpretation for the regression coefficients

If $(X^T, Y)^T \sim \mathcal{N}(\mathbf{0}, \mathbf{\Sigma})$ with a block-wise decomposion of $\mathbf{\Sigma}$ and $\mathbf{\Omega} = \mathbf{\Sigma}^{-1}$ then by condition $Y|X$ we get

$$Y = \sum_{j=1}^{p} X_j \mathrm{cor}(X_j, Y|X_k, k \neq j) \sqrt{\frac{(\mathbf{\Omega}_{XX})_{jj}}{\omega_{YY}}} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 1/\omega_{YY}). \tag{2}$$

By comparing (1) to (2) then β_j is related to the partial correlation between X_j and Y, i.e. describes effect of X_j on Y once effect of other predictors have been removed.

Gaussian Graphical Model and Linear Regression

Linear regression viewpoint

Variable X_i is linearly explained by the other variables:

$$X_i | X_{\setminus i} = -\sum_{j \neq i} \frac{\Theta_{ij}}{\Theta_{ii}} X_j + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma_i), \quad \varepsilon_i \perp X$$

Conditional on its neighborhood, other variables do not give additional insights

$$X_i | X_{\setminus i} = \sum_{j \in \text{neighbors}(i)} \beta_j X_j + \varepsilon_i \quad \text{with } \beta_j = -\frac{\Theta_{ij}}{\Theta_{ii}}.$$

$\leadsto$ "Neighborhood" selection

Outline

References

- Habilitation, J. Chiquet, Chapter 2
  https://tel.archives-ouvertes.fr/tel-01288976/

- The Element of Statistical Learning Hastie, Tibshirani, Friedman, chapter 17.

Some families of methods for network reconstruction

Test-based methods

- Tests the nullity of each entries
- Combinatorial problem when $p > 30 \ldots$

Bayesian methods

- Compute the posterior probability of each edge
- Usually more computationally demanding
- For special graphs, computation gets easier

Sparsity-inducing regularization methods

- induce sparsity with the ℓ_1-norm penalization
- Use results from convex optimization
- Versatile and computationally efficient

Outline

Inference: maximum likelihood estimator

The natural approach for parametric statistics

Let $X \sim f_X(x; \mathbf{\Omega})$, where $\mathbf{\Omega}$ are the model parameters.

Maximum likelihood estimator

$$\hat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} \ell(\mathbf{\Omega}; \mathbf{X})$$

where ℓ is the log likelihood, a function of the parameters:

$$\ell(\mathbf{\Omega}; \mathbf{X}) = \log \prod_{i=1}^{n} f_X(\mathbf{x}_i; \mathbf{\Omega}),$$

where $\mathbf{x}_i$ is the ith row of $\mathbf{X}$.

- This a convex optimization problem,
- We just need to detect non zero coefficients in $\mathbf{\Omega}$

The multivariate Gaussian log-likelihood

Let $\mathbf{S} = n^{-1}\mathbf{X}^\intercal\mathbf{X}$ be the empirical variance-covariance matrix: $\mathbf{S}$ is a sufficient statistic of $\mathbf{\Omega}$.

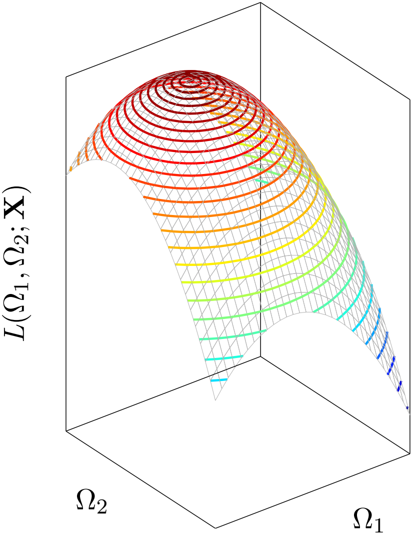
The log-likelihood

$$\ell(\mathbf{\Omega}; \mathbf{S}) = \frac{n}{2} \log\det(\mathbf{\Omega}) - \frac{n}{2} \mathrm{Trace}(\mathbf{S}\mathbf{\Omega}) + \frac{n}{2} \log(2\pi).$$

- $\rightsquigarrow$ The MLE $= \mathbf{S}^{-1}$ of $\mathbf{\Omega}$ is not defined for $n < p$ and never sparse.
- $\rightsquigarrow$ The need for regularization is huge.

A Geometric View of Shrinkage

Constrained Optimization



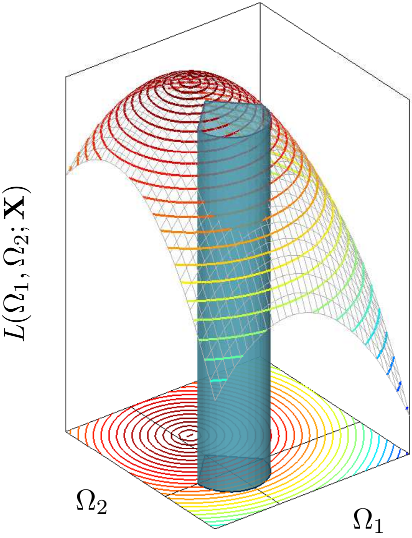
We basically want to solve a problem of the form

$$\underset{\Omega_1, \Omega_2}{\text{maximize}}\, \ell(\Omega_1, \Omega_2; \mathbf{X})$$

where ℓ is typically a concave likelihood function.

A Geometric View of Shrinkage

Constrained Optimization



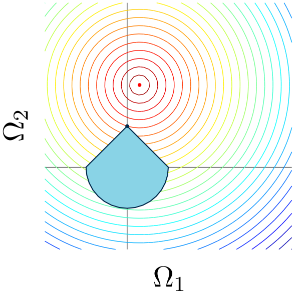
$$\begin{cases} \underset{\Omega_1,\Omega_2}{\text{maximize}} & \ell(\Omega_1, \Omega_2; \mathbf{X}) \\ \text{s.t.} & \Omega(\Omega_1, \Omega_2) \leq c \end{cases},$$

where Ω defines a domain that *constrains* $\boldsymbol{\beta}$.

How shall we define Ω ?

A Geometric View of Shrinkage

Constrained Optimization



$$\left\{\begin{array}{ll}\underset{\Omega_1,\Omega_2}{\text{maximize}} & \ell(\Omega_1,\Omega_2;\mathbf{X}) \\ \text{s.t.} & \Omega(\Omega_1,\Omega_2) \leq c\end{array}\right. ,$$

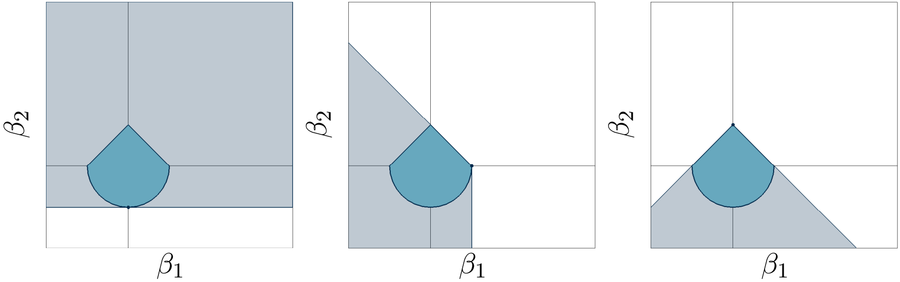
where Ω defines a domain that *constrains* $\boldsymbol{\beta}$.

How shall we define Ω ?

A Geometric View of Sparsity

Dual and Polar Cones

Generalizes normals



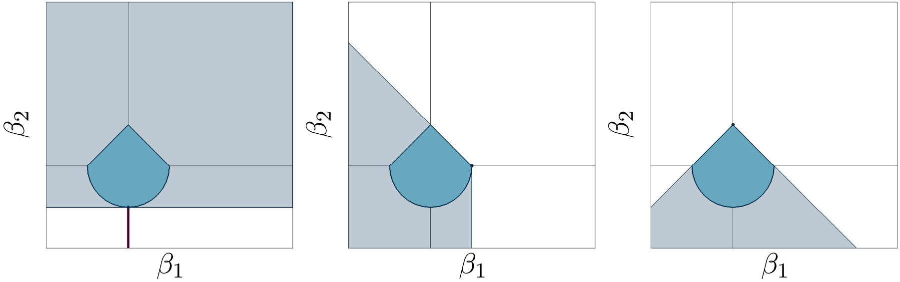
Let C be a convex set,

- $C^\star(x_0) = \left\{y | y^T(x - x_0) \geq 0, x \in C\right\}$ is the dual cone in x_0,
- $N_C(x_0) = \left\{y | y^T(x - x_0) \leq 0, x \in C\right\}$ is the polar (or normal) cone,

A Geometric View of Sparsity

Dual and Polar Cones

Generalizes normals



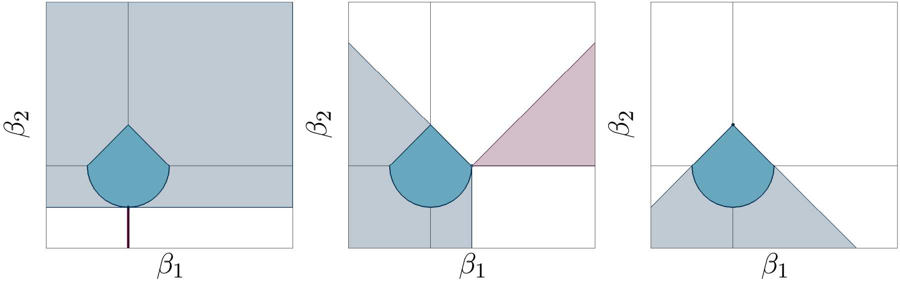
Let C be a convex set,

- $C^\star(x_0) = \{y | y^T(x - x_0) \geq 0, x \in C\}$ is the dual cone in x_0,
- $N_C(x_0) = \{y | y^T(x - x_0) \leq 0, x \in C\}$ is the polar (or normal) cone,

A Geometric View of Sparsity

Dual and Polar Cones

Generalizes normals



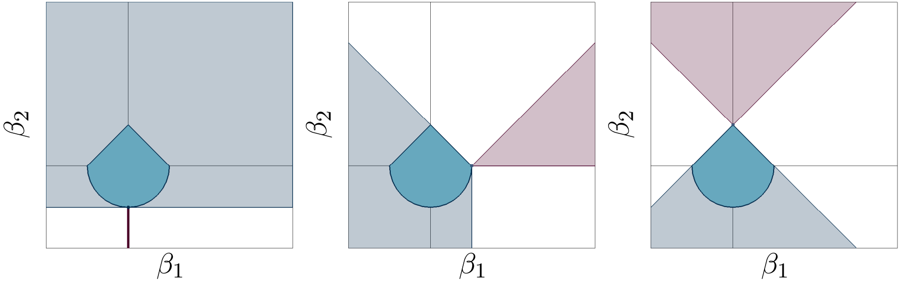
Let C be a convex set,

- $C^\star(x_0) = \{y | y^T(x - x_0) \geq 0, x \in C\}$ is the dual cone in x_0,
- $N_C(x_0) = \{y | y^T(x - x_0) \leq 0, x \in C\}$ is the polar (or normal) cone,

A Geometric View of Sparsity

Dual and Polar Cones

Generalizes normals



Let C be a convex set,

- $C^\star(x_0) = \{y | y^T(x - x_0) \geq 0, x \in C\}$ is the dual cone in x_0,
- $N_C(x_0) = \{y | y^T(x - x_0) \leq 0, x \in C\}$ is the polar (or normal) cone,

Shape of cones $\Rightarrow$ sparsity pattern

The Lasso

Least Absolute Shrinkage and Selection Operator

Idea

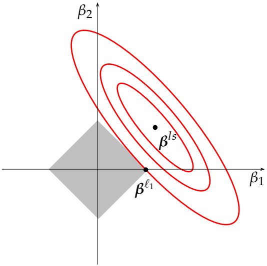
Suggest an admissible set that induces sparsity (force several entries to exactly zero in $\hat{\boldsymbol{\beta}}$).

Lasso as a regularization problem

The Lasso estimate of $\boldsymbol{\beta}$ is the solution to

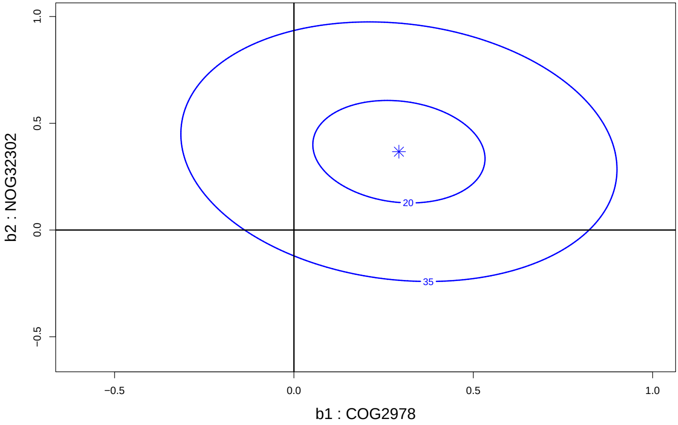
$$\hat{\boldsymbol{\theta}}^{\text{lasso}} = \arg\min_{\boldsymbol{\theta}} -\ell(\boldsymbol{\theta}), \quad \text{s.t. } \sum_{j=1}^{p} |\Omega_j| \leq s,$$

where s is a shrinkage factor.



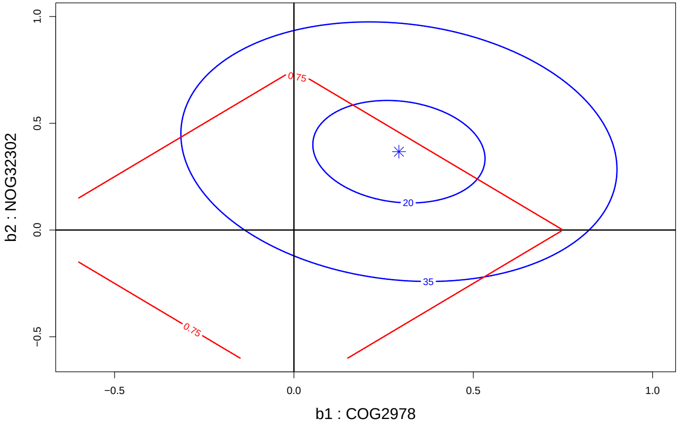
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{no constraints}$$



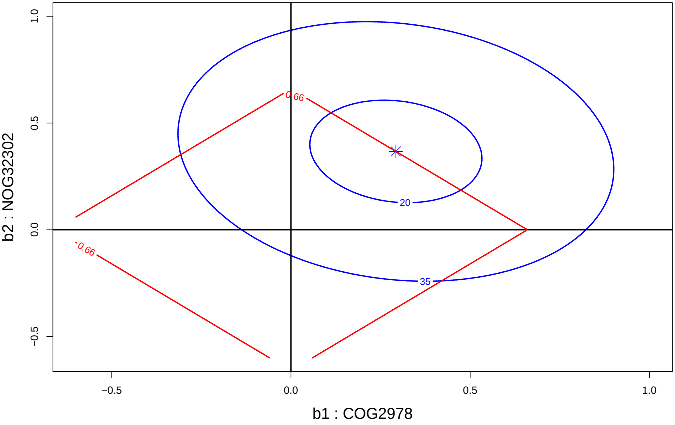
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{s.t. } |\Omega_1| + |\Omega_2| < 0.75$$



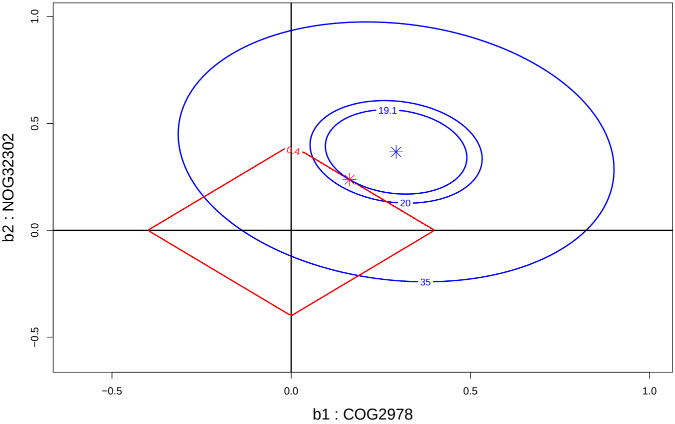
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{s.t. } |\Omega_1| + |\Omega_2| < 0.66$$



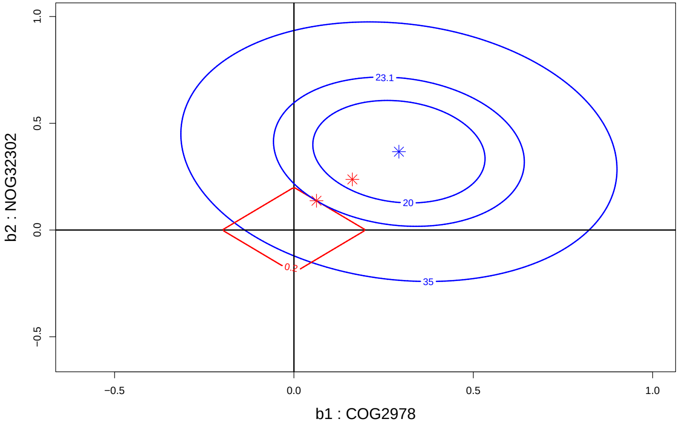
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{s.t. } |\Omega_1| + |\Omega_2| < 0.4$$



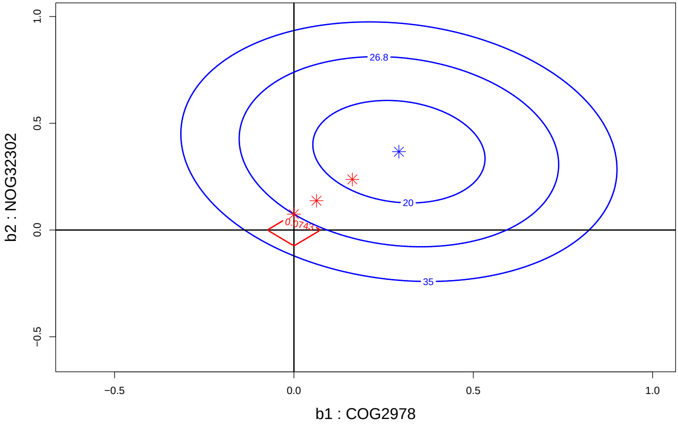
Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{s.t. } |\Omega_1| + |\Omega_2| < 0.2$$



Insights: 2-dimensional example with the square loss

$$\sum_{i=1}^{n}(y_i - x_i^1\Omega_1 - x_i^2\Omega_2)^2, \qquad \text{s.t. } |\Omega_1| + |\Omega_2| < 0.0743$$



Application to GGM

A penalized likelihood approach

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}\in\mathbb{S}_+} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda \mathrm{pen}_{\ell_1}(\mathbf{\Theta})$$

where

- ℓ is the model log-likelihood,
- pen_{ℓ_1} is a penalty function tuned by $\lambda > 0$.
  1. *regularization* (needed when $n \ll p$),
  2. *selection* (sparsity induced by the ℓ_1-norm).

Gold standard penalized approaches

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\boldsymbol{\Theta}}_\lambda = \arg\max_{\boldsymbol{\Theta}\in\mathbb{S}_+} \ell(\boldsymbol{\Theta};\mathbf{X}) - \lambda\|\boldsymbol{\Theta}\|_1$$

+ symmetric, positive-definite
– solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).
- R-packages **glasso**, **quic**, **huge**.

Neighborhood Selection (Meinshausen & Bülhman, 2006)

– not symmetric, not positive-definite
+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).
- R-package **huge**.

Gold standard penalized approaches

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\boldsymbol{\Theta}}_\lambda = \arg\max_{\boldsymbol{\Theta}\in\mathbb{S}_+} \ell(\boldsymbol{\Theta};\mathbf{X}) - \lambda\|\boldsymbol{\Theta}\|_1$$

+ symmetric, positive-definite
− solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).
- R-packages **glasso**, **quic**, **huge**.

Neighborhood Selection (Meinshausen & Bülhman, 2006)

For variable j, solve $\hat{\boldsymbol{\beta}}_j = \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^{p-1}} \frac{1}{2}\|\mathbf{X}_j - \mathbf{X}_{\backslash j}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_{\ell_1}.$

− not symmetric, not positive-definite
+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).
- R-package **huge**.

Gold standard penalized approaches

Penalized likelihood (Banerjee *et al.*, Yuan and Lin, 2008)

$$\hat{\mathbf{\Theta}}_\lambda = \arg\max_{\mathbf{\Theta}\in\mathbb{S}_+} \ell(\mathbf{\Theta};\mathbf{X}) - \lambda\|\mathbf{\Theta}\|_1$$

+ symmetric, positive-definite
− solved by the "Graphical-Lasso" ($\mathcal{O}(p^3)$, *Friedman et al, 2007*).
- R-packages **glasso**, **quic**, **huge**.

Neighborhood Selection (Meinshausen & Bülhman, 2006)

$$\hat{\mathbf{B}}^{\text{ns}} = \arg\min_{\mathbf{B}\in\mathbb{R}^{p\times p},\mathrm{diag}(\mathbf{B})=\mathbf{0}_p} \frac{1}{2}\mathrm{tr}(\mathbf{B}^\top\mathbf{S}_n\mathbf{B}) - \mathrm{tr}(\mathbf{B}^\top\mathbf{S}_n) + \lambda\|\mathbf{B}\|_{\ell_1}.$$

− not symmetric, not positive-definite
+ p Lasso solved with Lars-like algorithms ($\mathcal{O}(npd)$ for d neighbors).
- R-package **huge**.

Outline

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j\in\mathcal{P}}(\mathrm{degree_j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2 \log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d \log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d\log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Practical implications of theoretical results

Selection consistency (Ravikumar, Wainwright, 2009-2012)

Denote $d = \max_{j\in\mathcal{P}}(\mathrm{degree_j})$. Consistency for an appropriate λ and

- $n \approx \mathcal{O}(d^2\log(p))$ for the graphical Lasso and Clime.
- $n \approx \mathcal{O}(d\log(p))$ for neighborhood selection (sharp).

(Irrepresentability) conditions are not strictly comparable...

Ultra high-dimension phenomenon (Verzelen, 2011)

Minimax risk for sparse regression with d-sparse models: useless when

$$\frac{d\log(p/d)}{n} \geq 1/2, \qquad (\text{e.g.}, n = 50, p = 200, d \geq 8).$$

Good news! when n is small, we don't need to solve huge problems because they can't but fail.

Model selection

Cross-validation

Optimal in terms of prediction, not in terms of selection

Information based criteria

- GGMSelect (Girault *et al*, '12) selects among a family of candidates.
- Adapt IC to sparse high dimensional problems, e.g.

$$\text{EBIC}_\gamma(\widehat{\boldsymbol{\Omega}}_\lambda) = -2\text{loglik}(\widehat{\boldsymbol{\Omega}}_\lambda; \mathbf{X}) + |\mathcal{E}_\lambda|(\log(n) + 4\gamma \log(p)),$$

Resampling/subsampling

Keep edges frequently selected on an range of λ after sub-samplings

- Stability Selection (Meinshausen and Bühlman, 2010, Bach 2008)
- Stability approach to Regularization Selection (StaRS) (Liu, 2010).

Outline

The plasmodium data

```r
library(Matrix)
load("plasmodium_expression.Rdata")
dim(Y)
```

```
## [1] 3490   46
```

```r
head(Y)[, 1:5]
```

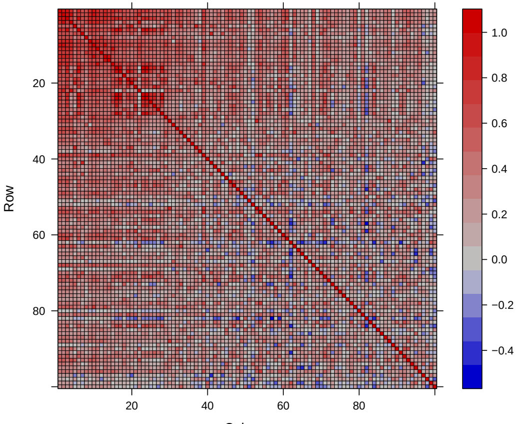
```
##                TP1    TP2    TP3    TP4    TP5
## MAL13P1.100 0.4510 0.6532 1.0760 0.5515 0.4238
## MAL13P1.102 1.5320 1.8920 0.8803 1.0300 0.9328
## MAL13P1.103 0.5218 0.5213 0.5328 0.3719 0.3258
## MAL13P1.105 0.5515 0.5527 0.8627 0.4541 0.4299
## MAL13P1.107 0.5630 0.4463 1.0760 0.4035 0.2082
## MAL13P1.112 0.5390 0.5393 0.5642 0.5326 0.4469
```

The plasmodium data

Gene to Gene empirical covariance

Covariance between the 100 most variable genes.

```
genes.subset <- order(apply(Y,1,var))[1:100]
image(Matrix(cor(t(Y[genes.subset, ]))), userRaster=TRUE)
```



Network between the genes I

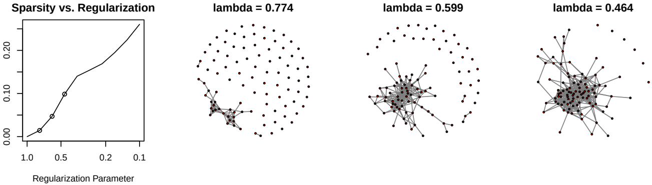
Sparse Estimation

Regulatory network between the 100 most variable genes.

```
library(huge)
huge.out <- huge(as.matrix(t(Y[genes.subset, ])), method="glasso", cov.output=TRUE)

## Conducting the graphical lasso (glasso) with lossless screening....in progress:0%
Conducting the graphical lasso (glasso) with lossless screening....in progress:9%
Conducting the graphical lasso (glasso) with lossless screening....in progress:19%
Conducting the graphical lasso (glasso) with lossless screening....in progress:30%
Conducting the graphical lasso (glasso) with lossless screening....in progress:40%
Conducting the graphical lasso (glasso) with lossless screening....in progress:50%
Conducting the graphical lasso (glasso) with lossless screening....in progress:60%
Conducting the graphical lasso (glasso) with lossless screening....in progress:70%
Conducting the graphical lasso (glasso) with lossless screening....in progress:80%
Conducting the graphical lasso (glasso)....done.

plot(huge.out)
```

Network between the genes II

Sparse Estimation



Network between the genes I

Inverse covariance

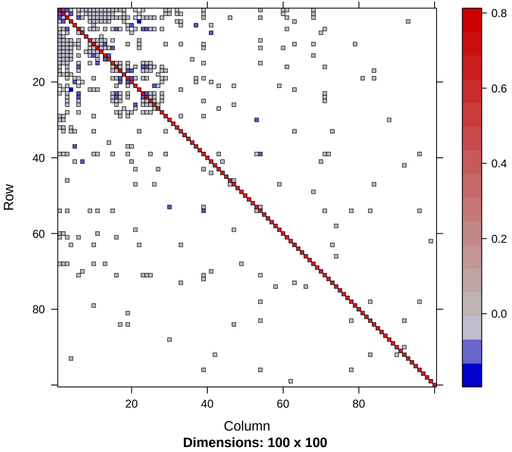
```
library(huge)
huge.out$df
```

```
##  [1]    0   71  233  488  693  763  836  963 1110 1289
```

```
image(Matrix(huge.out$icov[[3]]))
```

Network between the genes II

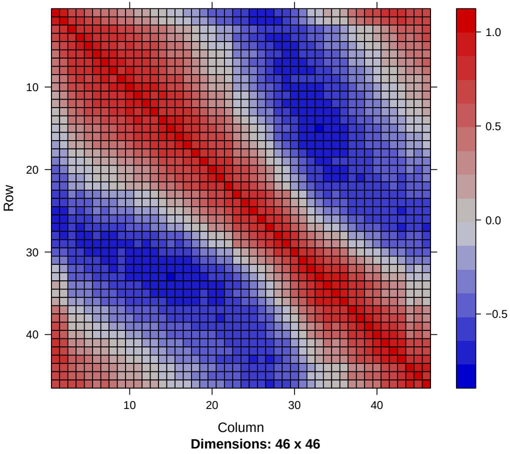
Inverse covariance



The plasmodium data

Covariance between conditions

```r
image(Matrix(cor(Y)))
```



Covariance structure between the conditions

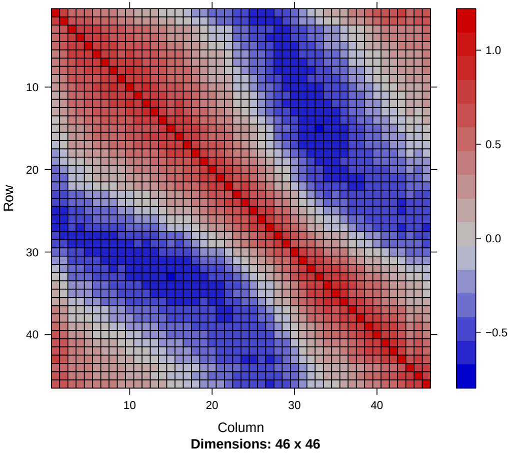
Sparse Estimation

```
library(huge)
huge.out <- huge(as.matrix(Y), method="glasso", cov.output=TRUE)

## Conducting the graphical lasso (glasso) with lossless screening....in progress:0%
Conducting the graphical lasso (glasso) with lossless screening....in progress:9%
Conducting the graphical lasso (glasso) with lossless screening....in progress:19%
Conducting the graphical lasso (glasso) with lossless screening....in progress:30%
Conducting the graphical lasso (glasso) with lossless screening....in progress:40%
Conducting the graphical lasso (glasso) with lossless screening....in progress:50%
Conducting the graphical lasso (glasso) with lossless screening....in progress:60%
Conducting the graphical lasso (glasso) with lossless screening....in progress:70%
Conducting the graphical lasso (glasso) with lossless screening....in progress:80%
Conducting the graphical lasso (glasso)....done.

sel.out  <- huge.select(huge.out)

## Conducting extended Bayesian information criterion (ebic) selection....done
```

Covariance structure between the conditions

Sparse Estimation

```
image(sel.out$opt.cov)
```



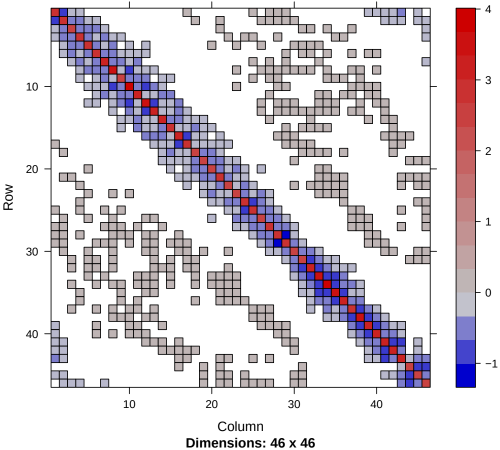
Covariance structure between the conditions I

Sparse Estimation of the inverse covariance

```
sum(abs(sel.out$opt.icov) != 0)

## [1] 760

ncol(sel.out$opt.icov) ** 2

## [1] 2116

image(sel.out$opt.icov)
```

Covariance structure between the conditions II

Sparse Estimation of the inverse covariance



Covariance structure between the conditions I

Associated network

```
plot(huge.out)
```

Covariance structure between the conditions II

Associated network