III. Inférence de réseaux de co-expression de gènes

Catherine Matias

CNRS - Laboratoire de Probabilités et Modèles Aléatoires, Paris
catherine.matias@math.cnrs.fr
http://cmatias.perso.math.cnrs.fr/

ENSAE - 2014/2015

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

Quel type de réseaux ? Quelles caractéristiques ? (1/3)

Réseaux d'interaction de protéines (protein interaction network, PIN)

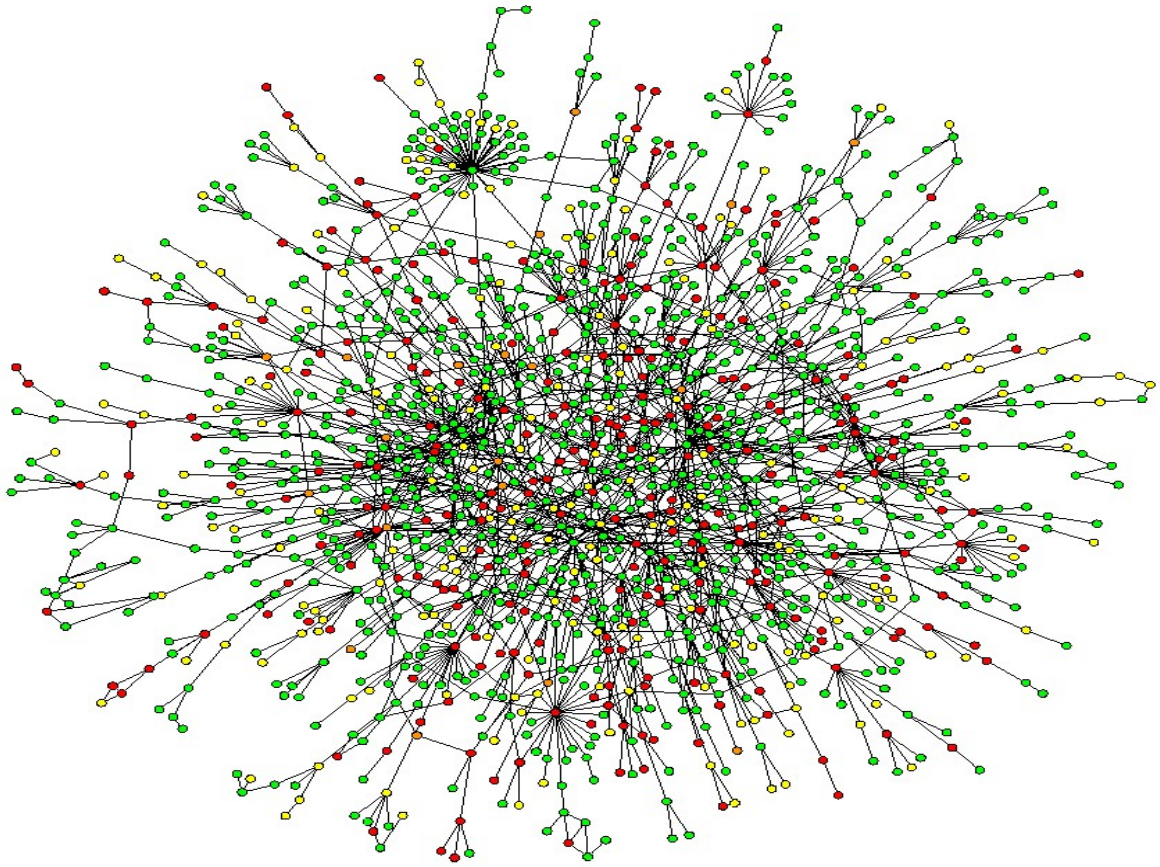


FIGURE : Yeast Protein Interaction Network. Source :
http ://www.bordalierinstitute.com/images/yeastProteinInteractionNetwork.jpg

Quel type de réseaux ? Quelles caractéristiques ? (1/3)

Réseaux d'interaction de protéines (protein interaction network, PIN)

Les protéines adoptent une conformation dans l'espace (pas nécessairement fixe), qui leur permet d'interagir entre elles.

- Un PIN décrit les interactions physiquement possibles entre des protéines (formation de complexes protéiques, cascades de phosphorylation . . .)
- Bases de données publiques qui contiennent les interactions répertoriées dans la littérature.
- Beaucoup de ces interactions sont obtenues par expériences double-hybride (Y2H), et donc des faux positifs.

Quel type de réseaux ? Quelles caractéristiques ? (2/3)

Réseaux métaboliques

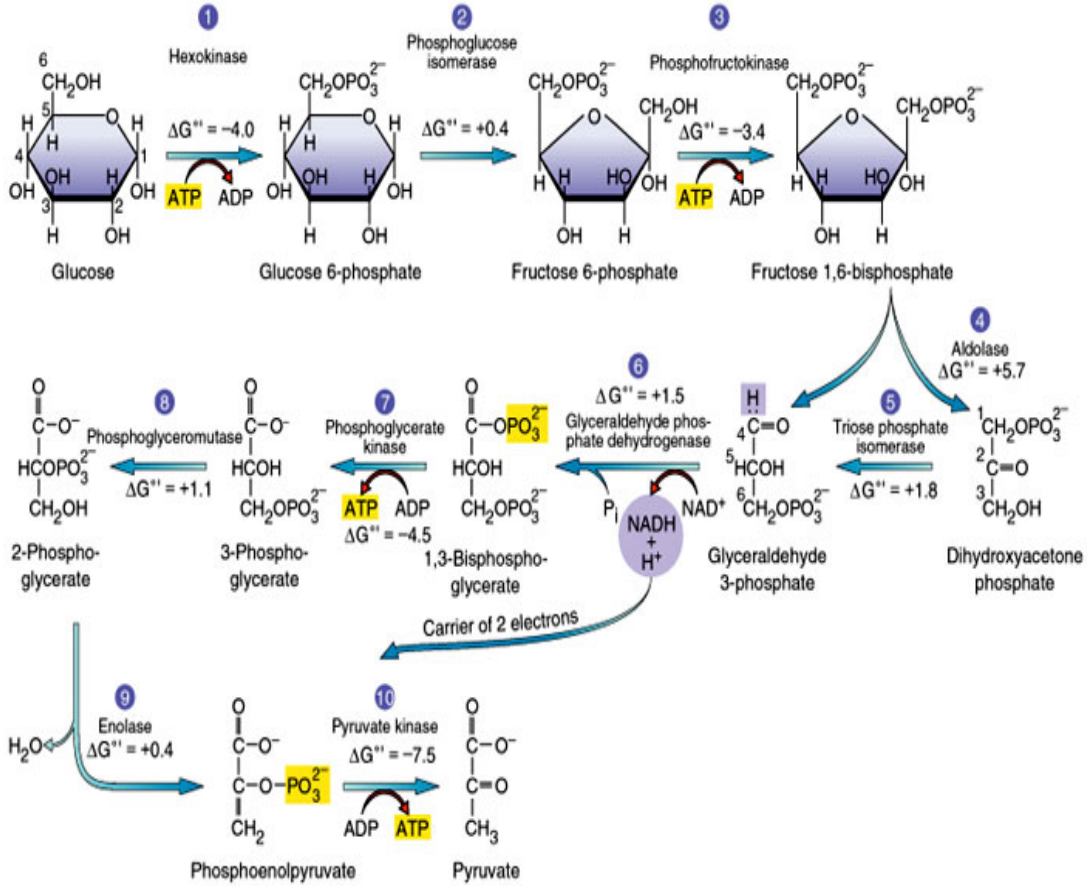


FIGURE : Voie métabolique de la glycolyse.

Quel type de réseaux ? Quelles caractéristiques ? (2/3)

Réseaux métaboliques

- Décrit réactions chimiques entre des métabolites (petites molécules) transformant un substrat en un certain produit.
- La plupart des réactions ont besoin d'être catalysées par des enzymes et sont considérées comme réversibles.
- Les réseaux métaboliques sont majoritairement inférés à partir de génomique comparative, ce qui induit beaucoup de faux négatifs.
- Modélisés via des hypergraphes

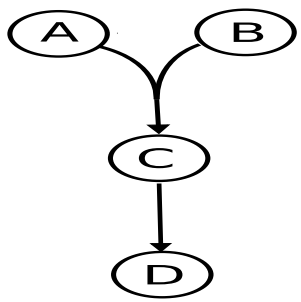


FIGURE : Hypergraphe orienté modélisant un réseau métabolique. Source : V. Lacroix.

Quel type de réseaux ? Quelles caractéristiques ? (3/3)

Réseaux de régulation génique

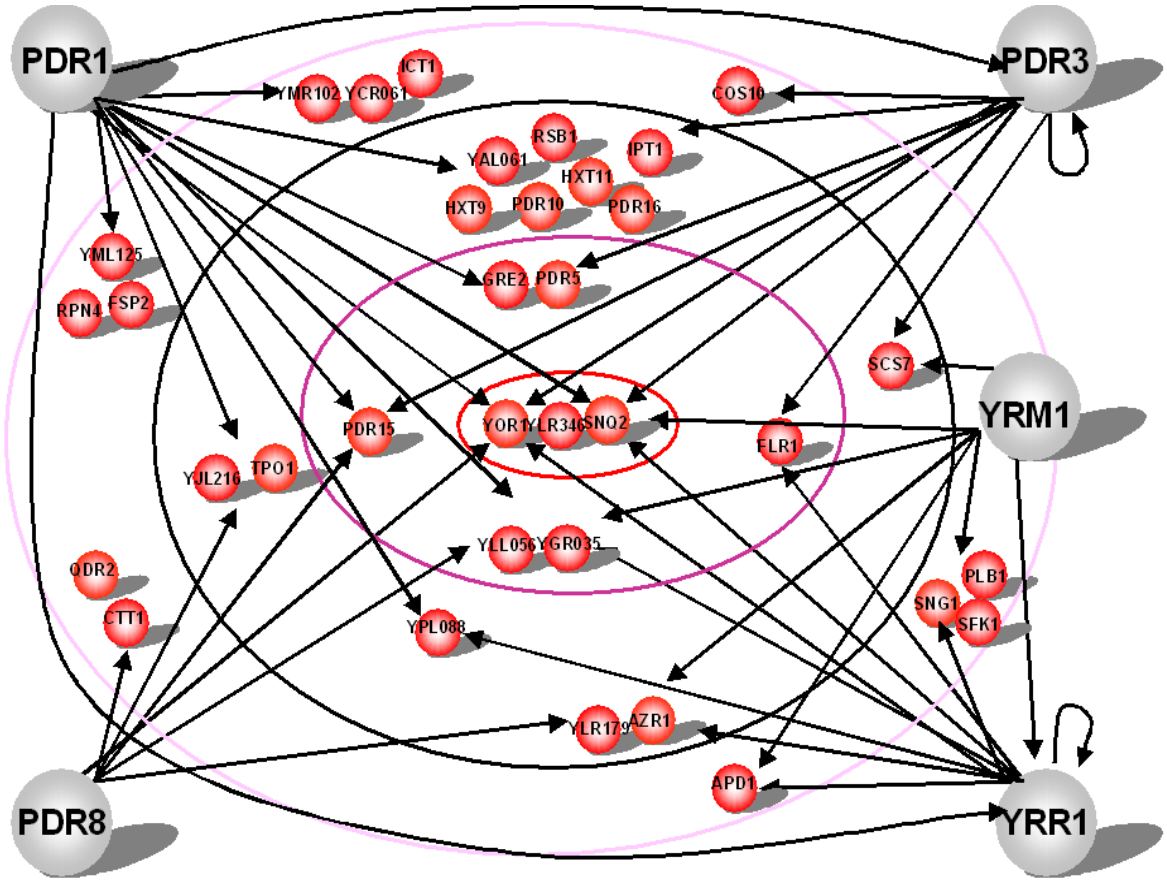


FIGURE : Réseau de régulation PDR chez *S. cerevisiae*. Source : Lab. Génomique de la levure, ENS.

Quel type de réseaux ? Quelles caractéristiques ? (3/3)

Réseaux de régulation génique

- Décrit les régulations (inhibitions ou activations) de l'expression des gènes, par d'autres gènes (via leurs produits).
- Graphes orientés, avec labels positifs ou négatifs.
- Vision statique ou dynamique (dans le temps).
- La plupart sont statistiquement estimés à partir de données d'expression
  $\rightsquigarrow$ On parle alors de réseaux de co-expression.

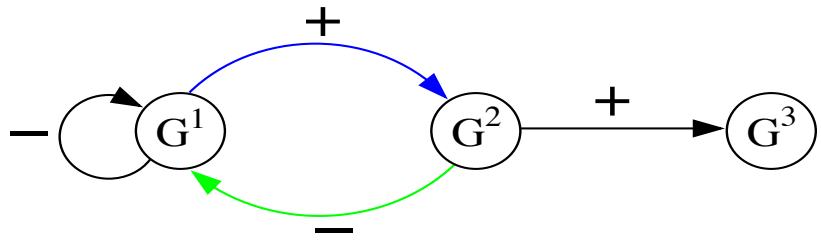


FIGURE : Exemple d'un motif de régulation. Source : S. Lèbre.

Défis posés par les réseaux biologiques

- Analyser de très grands jeux de données (milliers de nœuds et d'arêtes).
- Données fortement bruitées.
- Identifier des structures (motifs, groupes, etc).
- Les réseaux observés résultent d'un échantillonnage des réseaux existants : induit un biais (échantillonnage non uniforme des nœuds et des arêtes).

Objet de ce cours : les réseaux de co-expression et leur inférence à partir de données d'expression.

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

Problématique

Questions initiales

- Comment estimer un réseau de régulation génique entre des gènes ?
- À partir de quelles données ?

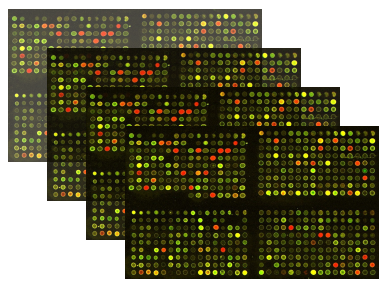
Réponses (partielles)

- On dispose de données d'expression des gènes,
- On cherche à en déduire des relations de causalité entre expressions de gènes.

Données de transcriptome

- Il existe différents types de technologies, appelées « puces », qui permettent la mesure simultanée d'une très grande quantité de matériel biologique (*cf. Cours 3-4 : Transcriptomique*).
- Ici, on s'intéresse à des « puces à ADN », qui mesurent le transcriptome, *i.e.* les ARN présents dans un ensemble de cellules à un instant et dans des conditions expérimentales données.
- Ainsi, on mesure l'expression d'un ensemble de gènes dans un ensemble de cellules.

But : Inférer à partir de ces données, un réseau de régulation génique.



n expériences (puces) de mesures d'expression de p gènes.

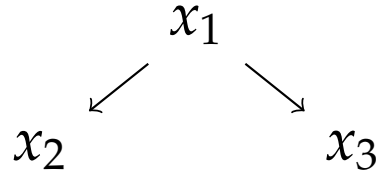
Réseaux de co-expression

1er postulat biologique : Les gènes qui interagissent sont co-exprimés

$\rightsquigarrow$ un réseau de co-expression doit traduire les dépendances entre les expressions des gènes.

Dépendances simple et conditionnelle (relations directe et indirecte)

Ex : DAG



$Cov(x_2, x_3) \neq 0$ pourtant la corrélation partielle $\rho_{2,3|1} = 0$.

- La dépendance simple est une notion trop *faible* pour représenter l'interaction des gènes. La dépendance conditionnelle est plus adaptée.
- Réseau de co-expression des gènes = graphe de dépendance conditionnelle entre variables.
- Les modèles graphiques (dirigés ou non) explicitent les relations de dépendance conditionnelle entre les variables.

Approches classiques

Se donner une classe de modèles et

- Soit estimer toutes les corrélations partielles et leur appliquer un seuil (pbm du choix du seuil).
- Soit sélectionner un graphe parmi tous les $2^{p(p-1)/2}$ graphes de dépendances possibles (sélection de modèle).

Problème de la grande dimension

- On dispose souvent de données de puces de l'ordre de quelques dizaines de réplicats (n nb d'obs.), pour quelques milliers de gènes mesurés (p nb. de variables).
- Impossible d'estimer toutes les corrélations partielles ($p(p-1)/2$ paramètres) ou de sélectionner un graphe parmi les $2^{p(p-1)/2}$ graphes possibles sans une hypothèse supplémentaire.

2nd postulat biologique : « Sparsity » ou parcimonie

Postulat

- Peu de gènes interagissent entre eux.
- Formulation statistique : on a un pbm « sparse », *i.e.* beaucoup de zéros et peu de corrélations non nulles à estimer.

$\rightsquigarrow$ Hypothèse permet d'inférer des réseaux où $p \sim n$ et le nb réel d'interactions possibles $\ll n$ (et non pas de l'ordre de p^2).

Approches adaptées aux données

1) Corrélations partielles d'ordre limité m

- On remplace $\rho_{i,j|\mathcal{P}\setminus\{i,j\}}$ par $\{\rho_{i,j|k}; k \in \mathcal{P}\setminus\{i,j\}\}$ (cas $m = 1$) ou plus généralement par $\{\rho_{i,j|\mathcal{P}_m}; \mathcal{P}_m \subset \mathcal{P}\setminus\{i,j\}, |\mathcal{P}_m| = m\}$
- $(i,j) \in E$ ssi $\rho_{i,j} \neq 0$ et $\forall k \in \mathcal{P}\setminus\{i,j\}, \rho_{i,j|k} \neq 0$.
- Test multiple des hypothèses $H_0(i,j) : \rho_{i,j} = 0$ ou $\exists k, \rho_{i,j|k} = 0$ vs $H_1(i,j) : \rho_{i,j} \neq 0$ et $\forall k, \rho_{i,j|k} \neq 0$.

A. Wille & P. Bühlmann.
Low-order conditional independence graphs for inferring genetic networks. SAGMB, 2006.

R. Castelo & A. Roverato.
A robust procedure for Gaussian graphical model search from microarray data with p larger than n. JMLR, 2006.

S. Lèbre.
Inferring Dynamic Genetic Networks with Low Order Independencies. SAGMB, 2009.

Approches adaptées aux données

2) Modèles graphiques gaussiens (GGMs) régularisés

- Pénalisation ℓ_1 de la vraisemblance des observations qui permet d'obtenir un estimateur sparse.
- Généralisation des méthodes lasso pour la régression $\rightsquigarrow$ Glasso (graphical lasso).


O. Banerjee, L. El Ghaoui & A. d'Aspremont.
Model selection through sparse maximum likelihood estimation for multivariate Gaussian, JMLR, 2008.


C. Charbonnier, J. Chiquet & C. Ambroise.
Weighted-Lasso for Structured Network Inference from Time Course Data, SAGMB, 2010.

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

Modèles graphiques (dirigés ou non dirigés)

Un modèle graphique est un modèle probabiliste dans lequel un graphe représente la structure de dépendance de la distribution.

Caractéristiques

$\mathbb{P}$ une distribution sur $\mathcal{X}^{|\mathcal{P}|}$ et $\mathcal{G} = (\mathcal{P}, E)$ un graphe t.q.

- l'ensemble $\mathcal{P} = \{1, \dots, p\}$ des nœuds indexe un ensemble de v.a. $\{x_i\}_{i \in \mathcal{P}}$ à valeurs dans $\mathcal{X}^{|\mathcal{P}|}$,
- L'ensemble des arêtes E décrit les relations de dépendance entre les v.a. $\{x_i\}_{i \in \mathcal{P}}$ sous la loi $\mathbb{P}$.
  Plus précisément,
  - Soit $\mathcal{G}$ est acyclique et dirigé (DAG), alors $\mathbb{P}$ se factorise selon $\mathcal{G}$, *i.e.* $\mathbb{P}(\{x_i\}_{i \in \mathcal{P}}) = \prod_{i \in \mathcal{P}} \mathbb{P}(x_i | pa(x_i, \mathcal{G}))$, où $pa(x_i, \mathcal{G})$ sont les var. parents de x_i dans $\mathcal{G}$.
  - Soit $\mathcal{G}$ est non dirigé, alors si $(i, j) \notin E$, on a $x_i \perp\!\!\!\perp x_j \mid x_{\mathcal{P} \setminus \{i,j\}}$.

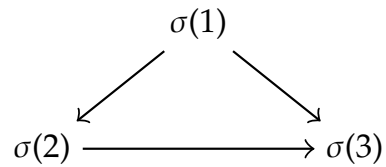
Modèles graphiques (dirigés ou non dirigés)

Exemples

- Champs de Markov (non dirigés)
- Modèles graphiques gaussiens (non dirigés)
- Réseaux bayésiens (dirigés). Ex : Chaînes de Markov, ou chaînes de Markov cachés.

Remarques

- Si $(\mathbb{P}, \mathcal{G})$ est un modèle graphique dirigé alors le graphe moral (*cf. Cours 1*) de $\mathcal{G}$ est non dirigé et code les indépendances conditionnelles des v.a. sous $\mathbb{P}$.
- Si $\mathbb{P}$ se factorise sous un DAG $\mathcal{G}$, alors $\mathcal{G}$ n'est pas unique en général.
  Ex : sans contrainte sur $\mathbb{P}$, on a
  $\mathbb{P}(x_1, x_2, x_3) = \mathbb{P}(x_3|x_1, x_2)\mathbb{P}(x_2|x_1)\mathbb{P}(x_1) = \mathbb{P}(x_{\sigma(3)}|x_{\sigma(1)}, x_{\sigma(2)})\mathbb{P}(x_{\sigma(2)}|x_{\sigma(1)})\mathbb{P}(x_{\sigma(1)})$, pour toute permutation σ.



Modèles graphiques

Références sur les modèles graphiques

C. M. Bishop.
Pattern recognition and machine learning. (Chap. 8)
Information Science and Statistics. Springer, 2006.

S. L. Lauritzen.
Graphical models, volume 17 of *Oxford Statistical Science Series*.
Oxford University Press, New York, 1996.

Modèles graphiques gaussiens (GGMs)

Modèle gaussien

- $\mathcal{P} = \{1, \ldots, p\}$ ensemble des variables,
- $X = (x_1, \ldots, x_p)$ vecteur de $\mathbb{R}^p$ de loi $\mathcal{N}_p(\mathbf{0}, \Sigma)$, avec $\Sigma > 0$ (définie positive),
- $K = (K_{ij})_{(i,j)\in\mathcal{P}^2} := \Sigma^{-1}$ la matrice de concentration.

Proposition

La corrélation partielle entre x_i et x_j conditionnellement à l'ensemble des autres variables vérifie

$$\rho_{ij|\mathcal{P}\setminus\{i,j\}} = -K_{ij}/\sqrt{K_{ii}K_{jj}}.$$

Preuve

Lemme

Si $X = (x_1, x_2) \sim \mathcal{N}_p(\xi, \Sigma)$ *avec* $p = p_1 + p_2$, $\xi^\intercal = (\xi_1, \xi_2)^\intercal$ *et* $K = \Sigma^{-1} = \begin{pmatrix} K_{11} & K_{12} \\ K_{21} & K_{22} \end{pmatrix}$, *alors, si* K_{11} *inversible, on a*

$$x_1 \mid x_2 \sim \mathcal{N}_{p_1}(\xi_{1|2}, \Sigma_{1|2})$$
$$\text{où } \xi_{1|2} = \xi_1 - K_{11}^{-1} K_{12}(x_2 - \xi_2) \text{ et } \Sigma_{1|2} = K_{11}^{-1}.$$

Démonstration : Écrire $f(x_1|x_2) \propto f(x_1, x_2)$.

Conséquence

$(x_i, x_j) \mid x_{\mathcal{P}\setminus\{i,j\}}$ suit une loi gaussienne de matrice de covariance $\begin{pmatrix} K_{ii} & K_{ij} \\ K_{ji} & K_{jj} \end{pmatrix}^{-1} = \frac{1}{K_{ii}K_{jj} - K_{ij}^2} \begin{pmatrix} K_{jj} & -K_{ij} \\ -K_{ji} & K_{ii} \end{pmatrix}$.

Modèles graphiques gaussiens (GGMs)

Interprétation graphique

Le graphe $\mathcal{G} = (\mathcal{P}, E)$ tel que

$$\text{arête } (i,j) \notin E \Leftrightarrow K_{ij} = 0 \Leftrightarrow \rho_{ij|\mathcal{P}\setminus\{i,j\}} = 0 \Leftrightarrow x_i \perp\!\!\!\perp x_j | x_{\mathcal{P}\setminus\{i,j\}},$$

est appelé le graphe de concentration. C'est un graphe non orienté qui code les indép. cond. des $\{x_i\}_{i\in\mathcal{P}}$.

Conséquences

À partir d'un échantillon de la loi $\mathcal{N}_p(\mathbf{0}, \Sigma)$, on peut considérer 2 problèmes distincts :

- estimer K (estimation paramétrique),
- estimer la position des 0 de K (estimation de structure).

Inférence statistique dans un GGM

Observations

$X^1, \dots, X^n$ i.i.d. de loi $\mathcal{N}_p(\mathbf{0}, \Sigma)$. On note :
$\mathbf{X} = (X^1, \dots, X^n)^\intercal = (X_1, \dots, X_p)$ matrice $n \times p$ tq
$(X^k)^\intercal = (x_1^k, \dots, x_p^k)^\intercal$: $k$ème ligne de $\mathbf{X}$,
$X_k = (x_k^1, \dots, x_k^n)$ est la $k$ème colonne de $\mathbf{X}$,
$\bar{X} = \frac{1}{n} \sum_{i=1}^n X^i$ et $\bar{\mathbf{X}}$ matrice $n \times p$ dont toutes les lignes sont identiques et égales à $\bar{X}^\intercal$.

- La matrice de covariance Σ est estimée par la covariance empirique
  $S_n = \frac{1}{n-1} \sum_{i=1}^n (X^i - \bar{X})(X^i - \bar{X})^\intercal = \frac{1}{n-1} (\mathbf{X} - \bar{\mathbf{X}})^\intercal (\mathbf{X} - \bar{\mathbf{X}})$.
- S_n est une matrice $p \times p$ tq $rang(S_n) \leq n \wedge p$ donc si $n < p$, S_n n'est jamais inversible.
- Donc si $n < p$, on ne peut pas estimer K ou sa structure à partir de S_n.
- Mettre en œuvre des techniques de régularisation (ex : pénalisation ℓ_1) pour obtenir un estimateur inversible de Σ.

Liens GGMs / régression

$$x_k|pa(x_k) \sim \mathcal{N}(\xi_k, \sigma_k^2) \text{ avec } \xi_k = \sum_{j \in pa(X_k)} \theta_{kj} x_j,$$
$$i.e\ x_k = \sum_{j \in pa(X_k)} \theta_{kj} x_j + \epsilon_k, \text{ où } \epsilon_k \sim \mathcal{N}(0, \sigma_k^2).$$

- Chercher la structure du graphe de concentration revient à chercher les coeffs non nuls de la régression de chaque variable sur les autres.
- Lorsque le pbm est sparse, on fait de la régression pénalisée pour obtenir un estimateur dont certains coeffs sont exactement 0.
- On va donc adapter la régression pénalisée au contexte des GGMs.

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

Régression pénalisée

Moindres carrés ordinaires (OLS - modèle standardisé)

$$\text{Argmin}_\beta \|y - X\beta\|_2^2,$$

où $y = (y_1, \ldots, y_n)^\intercal$ réponse centrée, X matrice de taille $n \times p$ de régresseurs dont les colonnes sont normalisées et
$\beta = (\beta_1, \ldots, \beta_p)$ vecteur de paramètres.

Moindres carrés pénalisés

$$\text{Argmin}_\beta \|y - X\beta\|_2^2 + \lambda \|\beta\|_\alpha^\alpha,$$

où λ paramètre de pénalité, $\alpha \geq 0$ et $\|\beta\|_\alpha^\alpha = \sum_{k=1}^p |\beta_k|^\alpha$.
Ex :

- $\alpha = 0$, $\|\beta\|_0^0 = \sharp\{k; \beta_k \neq 0\}$: pénalité ℓ_0 (sparsity),
- $\alpha = 1$, $\|\beta\|_1 = \sum_k |\beta_k|$, pénalité lasso (sparsity),
- $\alpha = 2$, $\|\beta\|_2^2 = \sum_k |\beta_k|^2$, pénalité ridge (shrinkage).

Digression : Shrinkage et pénalité ℓ_2

En dimension $p = 1$,

$$\min_\beta \sum_{i=1}^n (y_i - x_i\beta)^2 \text{ sous la contrainte } \beta^2 \leq c,$$

équivalent à (via multiplicateur de Lagrange)

$$\min_\beta \left\{ \sum_{i=1}^n (y_i - x_i\beta)^2 + \lambda_c \beta^2 \right\}.$$

En différenciant, on obtient la solution $\hat{\beta}_c$

$$-2 \sum_{i=1}^n x_i (y_i - x_i \hat{\beta}_c) + 2\lambda_c \hat{\beta}_c = 0$$

$$i.e. \quad \hat{\beta}_c = \frac{1}{n \sum_i x_i^2 + \lambda_c} \sum_{i=1}^n x_i y_i \ ,$$

à comparer avec $\hat{\beta}^{ols} = \frac{1}{n \sum_i x_i^2} \sum_{i=1}^n x_i y_i$.

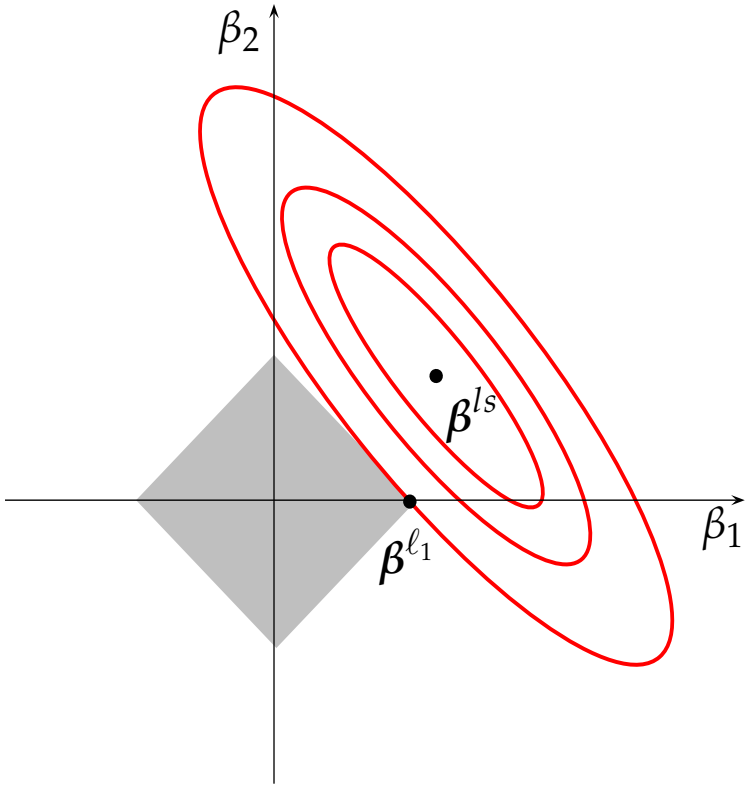
- La pénalisation ℓ_2 (régression ridge) induit un rétrécissement (shrinkage) des paramètres.
- On a $\lim_{c \to \infty} \hat{\beta}_c = \hat{\beta}^{ols}$ et $\lim_{c \to 0} \hat{\beta}_c = 0$ ou encore $\lim_{\lambda \to 0} \hat{\beta}_c = \hat{\beta}^{ols}$ et $\lim_{\lambda \to \infty} \hat{\beta}_c = 0$.

Régression pénalisée

Choix de la norme $\|\cdot\|_\alpha$ de pénalisation

- Pour $\alpha \geq 1$, le pbm de minimisation est convexe,
- Pour $\alpha \leq 1$, la solution du pbm est sparse.

$\rightsquigarrow \alpha = 1$: pbm convexe et solution sparse.



$$\text{Argmin}_\beta \|y - X\beta\|_2^2 + \lambda_c \|\beta\|_\alpha$$

$$\Leftrightarrow$$

$$\begin{cases} \text{Argmin}_\beta \|y - X\beta\|_2^2, \\ \text{tq } \|\beta\|_\alpha \leq c \end{cases}$$

FIGURE : Source : J. Chiquet

Régression lasso ($\alpha = 1$)

$$\text{Argmin}_\beta \|y - X\beta\|_2^2 + \lambda\|\beta\|_\alpha$$

Paramètre de pénalité

- Si $\lambda = 0$: pas de pénalité, on retrouve la solution OLS.
- Quand λ augmente, on obtient des coordonnées de $\hat{\beta}$ exactement égales à 0.
- Pour $\lambda = \lambda_{\max}$ on a $\hat{\beta}^{\text{lasso}}_{\lambda_{\max}} = \mathbf{0}$.

N.B. : Pas de solution analytique pour $\hat{\beta}^{\text{lasso}}$ en général.


R. Tibshirani.
The Lasso : Least Absolute Shrinkage and Selection Operator. J. R. Stat. Soc., Ser. B, 1996.


S. Chen , D. Donoho , M. Saunders.
Atomic decomposition by basis pursuit. SIAM J. on Scientific Computing, 1999.

Lasso versus OLS : cas particulier du design orthogonal

Lemme

Si $X^\intercal X = Id$, *alors*

$$\forall k, \quad \hat{\beta}_k^{lasso} = sign(\hat{\beta}_k^{ols}) \max(0, |\hat{\beta}_k^{ols}| - \lambda/2) \text{ (seuillage doux).}$$

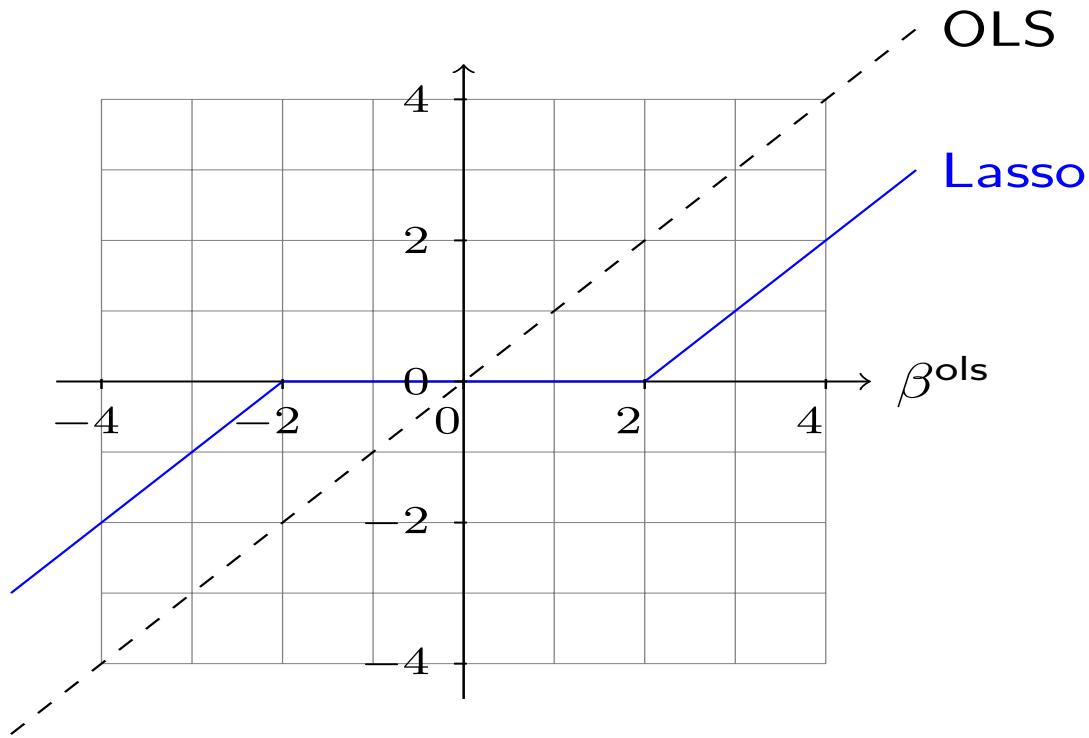


FIGURE : Source : J. Chiquet

N.B. : en grande dimension, on n'a jamais $X^\intercal X = Id$, et c'est rarement le cas en pratique même si $n > p$.

Lasso versus OLS : cas particulier du design orthogonal

Démonstration pour $p = 1$

On veut calculer : $\text{Argmin}_\beta \sum_{i=1}^n (y_i - x_i\beta)^2 + \lambda|\beta|$.

En différenciant, on obtient

$$-2\sum_{i=1}^n x_i(y_i - x_i\hat{\beta}^{\text{lasso}}) + \lambda \text{sign}(\hat{\beta}^{\text{lasso}}) = 0.$$

où $\text{sign}(u)\begin{cases} = +1 & \text{si } u > 0 \\ = -1 & \text{si } u < 0 \\ \in [-1;1] & \text{si } u = 0 \end{cases}$.

Or $\hat{\beta}^{\text{ols}} = (X^\intercal X)^{-1} X^\intercal y = X^\intercal y = \sum_{i=1}^n x_i y_i$

et $\sum_{i=1}^n x_i^2 = X^\intercal X = 1$, d'où

$$-2\hat{\beta}^{\text{ols}} + 2\hat{\beta}^{\text{lasso}} + \lambda \text{sign}(\hat{\beta}^{\text{lasso}}) = 0.$$

On vérifie alors que l'expression donnée est bien solution.

Algorithme Lars et chemins de régularisation

B. Efron, T. Hastie, I. Johnstone, & R. Tibshirani.
Least angle regression. Annals of Statistics, 2004.

Algorithme efficace pour le problème lasso

Solution LARS (least angle regression) est un ensemble de courbes qui donnent la valeur des $\hat{\beta}_k^{\text{lasso}}$ en fonction de la pénalité λ.

- chemin de solutions linéaire par morceaux, qui part de 0 et va jusqu'à la valeur de l'estimateur OLS.
- (presque) le même coût de calcul que l'OLS
- bien adapté à la validation croisée (choix de λ).

Algorithme Lars et chemins de régularisation

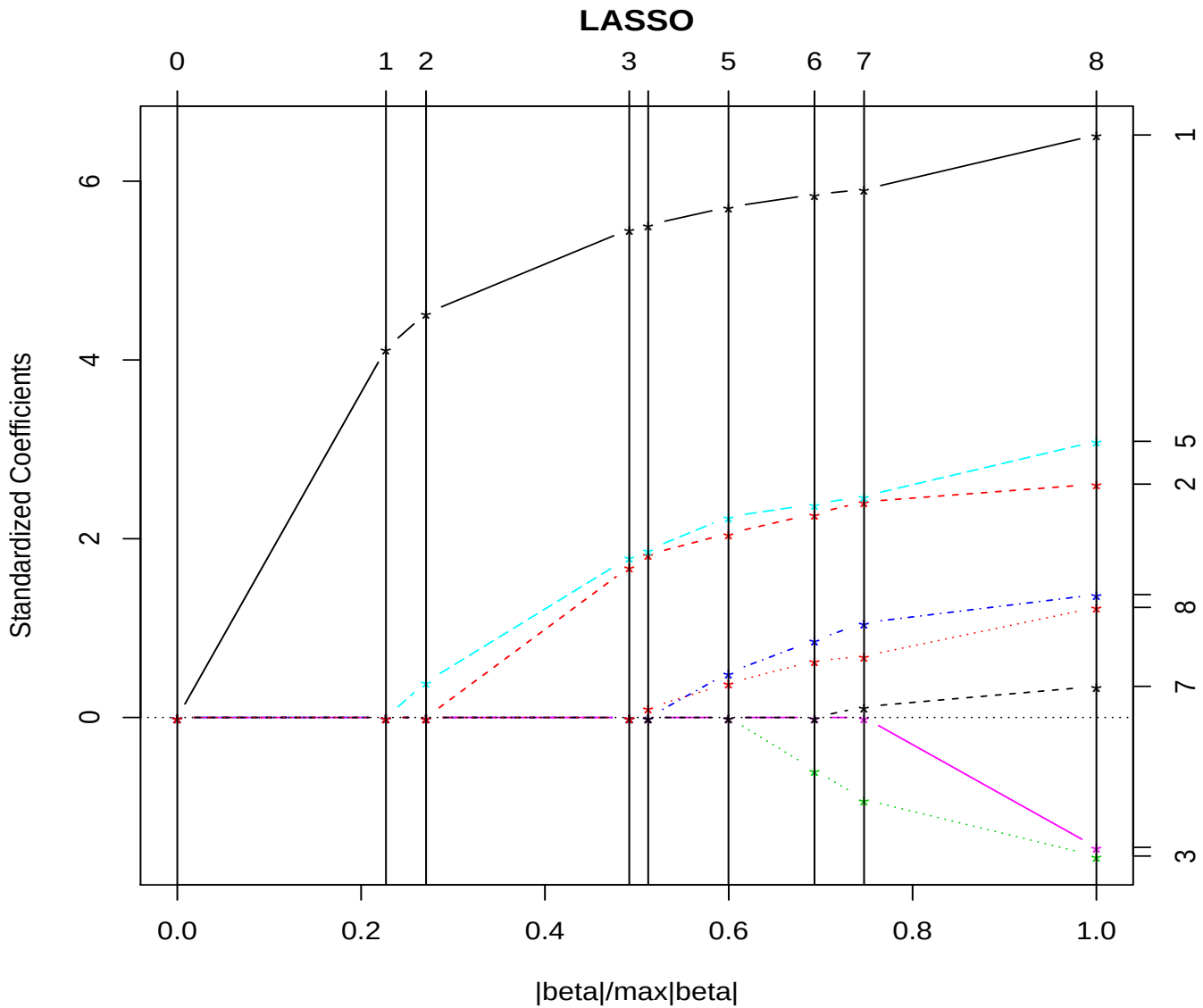


FIGURE : Solution LARS Source : J. Chiquet

Choix du paramètre de pénalité

Critères de sélection de modèle

$$
\begin{aligned}
BIC(\lambda) &= \|y - X\hat{\beta}_\lambda\|_2^2 - \mathrm{df}(\hat{\beta}_\lambda)\frac{\log n}{2} \\
AIC(\lambda) &= \|y - X\hat{\beta}_\lambda\|_2^2 - \mathrm{df}(\hat{\beta}_\lambda),
\end{aligned}
$$

où $\mathrm{df}(\hat{\beta}_\lambda)$: nb de coeffs non nuls dans $\hat{\beta}_\lambda$.

Validation croisée

- Données découpées en V morceaux $D_1, \ldots, D_V$, de taille n/V,
- Pour $i = 1, \ldots, V$, on estime un paramètre $\hat{\beta}^i_\lambda$ sur les données $\cup_{k \neq i} D_k$ et on calcule l'erreur de test sur D_i, $err^i(\lambda) = \sum_{k \in D_i} (y_k - (X\hat{\beta}^i_\lambda)_k)^2$,
- on moyenne les erreurs : $err^{CV}(\lambda) = \frac{1}{V} \sum_i err^i(\lambda)$,
- Enfin, $\hat{\lambda} = \mathrm{Argmin}_\lambda \, err^{CV}(\lambda)$.

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

Régression pénalisée dans un GGM

Rappel

$X^1, \ldots, X^n$ i.i.d. de loi $\mathcal{N}_p(\mathbf{0}, \Sigma)$ avec $K = \Sigma^{-1}$. On note :
$\mathbf{X} = (X^1, \ldots, X^n)^\intercal = (X_1, \ldots, X_p)$ matrice $n \times p$ tq
$(X^k)^\intercal = (x_1^k, \ldots, x_p^k)^\intercal$: $k$ème ligne de $\mathbf{X}$,
$X_k = (x_k^1, \ldots, x_k^n)$ est la $k$ème colonne de $\mathbf{X}$.

1ère approche : [MB06]

Considérer p pbms de régression distincts : pour chaque $k \in \mathcal{P}$,

$$x_k = \sum_{j \neq k} x_j \beta_j + \epsilon$$

où $\beta_j = -K_{kj}/K_{kk}$, résolus par lasso

$$\hat{\beta} = \text{Argmin}_\beta \, \|X_k - \mathbf{X}_{\setminus k}\beta\|_2^2 + \lambda \|\beta\|_1,$$

où $\mathbf{X}_{\setminus k}$: matrice $\mathbf{X}$ privée de sa $k$ème colonne.

N. Meinshausen & P. Bühlmann.
High dimensional graphs and variable selection with the lasso. Ann. of Stats, 2006.

Régression pénalisée dans un GGM

Inconvénients

- Estimateur $\hat{K}$ obtenu non symétrique !

  Dans la régression de x_k sur x_j, on peut avoir $\hat{\beta}_j = 0$ alors que lorsque x_j est régressé sur x_k, $\hat{\beta}_k \neq 0$.
- Étape de post-symétrisation par règles OU ou ET.
- Peut-on estimer K (ou sa structure) directement ?

Remarque

Dans le cas gaussien, le critère des moindres carrés correspond à la maximisation de la vraisemblance.
$\rightsquigarrow$ remplacer un pbm de moindre carré par une maximisation de vraisemblance pénalisée :

$$\text{Argmax}_{K,K>0}\{\mathcal{L}(K) - \lambda\|K\|_1\},$$

où $\|K\|_1 = \sum_{ij} |K_{ij}|$.

Régression pénalisée et vraisemblance pénalisée

Interprétation de l'approche [MB06]

On considère la pseudo-log-vraisemblance des observations

$$\tilde{\mathcal{L}}(K) := \sum_{i=1}^{n} \sum_{k=1}^{p} \log \mathbb{P}(x_k^i | x_{\mathcal{P}\setminus\{k\}}^i)$$

N.B. : Critère fondé sur l'obs. d'un n-échantillon de la mesure $\mu(x_1, \ldots, x_p) = \prod_{k=1}^{p} \mathbb{P}(x_k | x_{\mathcal{P}\setminus\{k\}})$ qui n'est pas une proba (en général).

Lemme

On note $\|K\|_1 = \sum_{ij} |K_{ij}|$. *La solution du pbm*

$$\widehat{\mathbf{K}} = Argmax_{K, K_{ij} \neq K_{ji}} \{\tilde{\mathcal{L}}(K) - \lambda \|\mathbf{K}\|_1\},$$

a la même structure (de zéros) que la solution des p problèmes de régression pénalisée indépendants proposée par [MB06].

Vraisemblance dans un GGM

On utilise ici $S_n := \frac{1}{n}\mathbf{X}^\intercal\mathbf{X} = \frac{1}{n}\sum_{i=1}^n X_i X_i^\intercal$.

Proposition

La log-vraisemblance des observations s'écrit

$$\ell_n(K) = \frac{n}{2}\left\{\log\det(K) - \mathrm{tr}(S_n K)\right\} + cte.$$

Démonstration :

$$\log f(X^1,\ldots,X^n) = \frac{n}{2}\log(\det K) - \frac{1}{2}\sum_{i=1}^n (X^i)^\intercal K(X^i) + c$$

$$\text{Or, } \sum_{i=1}^n (X^i)^\intercal K(X^i) = \sum_{i=1}^n\sum_{k;l=1}^p X_{il}X_{lp}X_{ip} = \mathrm{tr}(\mathbf{X}K\mathbf{X}^\intercal) = \mathrm{tr}(\mathbf{X}^\intercal\mathbf{X}K).$$

Enjeux posés par la vraisemblance pénalisée d'un GGM

Vraisemblance pénalisée

$$\hat{K} := \underset{K, K>0}{\text{Argmax}} \left\{ \frac{n}{2} \Big[\log\det(K) + \text{tr}(S_n K) \Big] - \lambda \|K\|_1 \right\},$$

où $\|K\|_1 = \sum_{ij} |K_{ij}|$.

Enjeux

- Comment résoudre ce problème de maximisation ?
- Comment gérer la contrainte K symétrique définie positive ?

⇝ Algo Glasso (graphical lasso) proposé par [BGA08].

 O. Banerjee, L. El Ghaoui & A. d'Aspremont. Model selection through sparse maximum likelihood estimation for multivariate Gaussian, JMLR, 2008.

Sommaire

Introduction aux réseaux biologiques

Réseaux de co-expression

Modèles graphiques

Introduction à la régression pénalisée

Graphical Lasso

Illustration

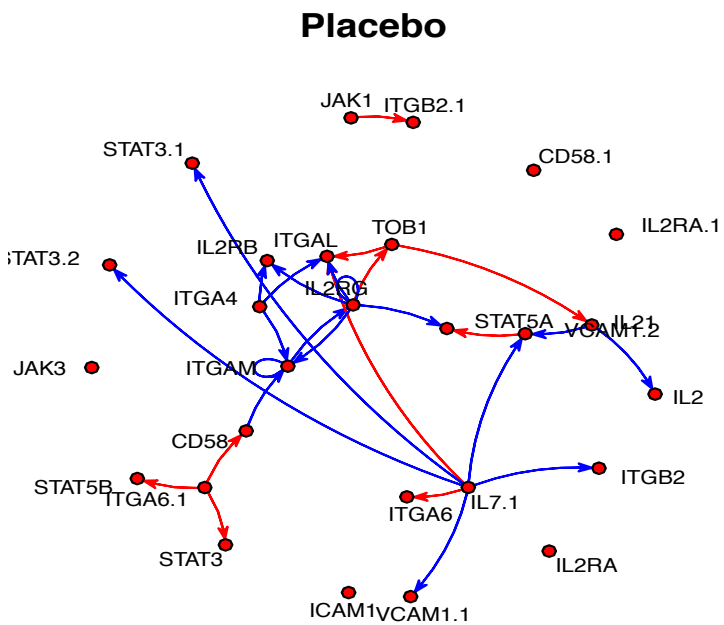
Illustration : Sclérose en plaques (Charbonnier, Chiquet, Ambroise & Corvol) I

Données

- 31 gènes mesurés, pré-identifiés comme impliqués dans la physiopathologie de la SEP,
- Données dynamiques avec 2 groupes de patients : placebo/traités,
- Au total : 100 puces, 25 patients (environ 75%/25% traités/placebo), 4 points de temps.

Comparaison des réseaux inférés dans les groupes traités/placebo.

Illustration : Sclérose en plaques (Charbonnier, Chiquet, Ambroise & Corvol) II



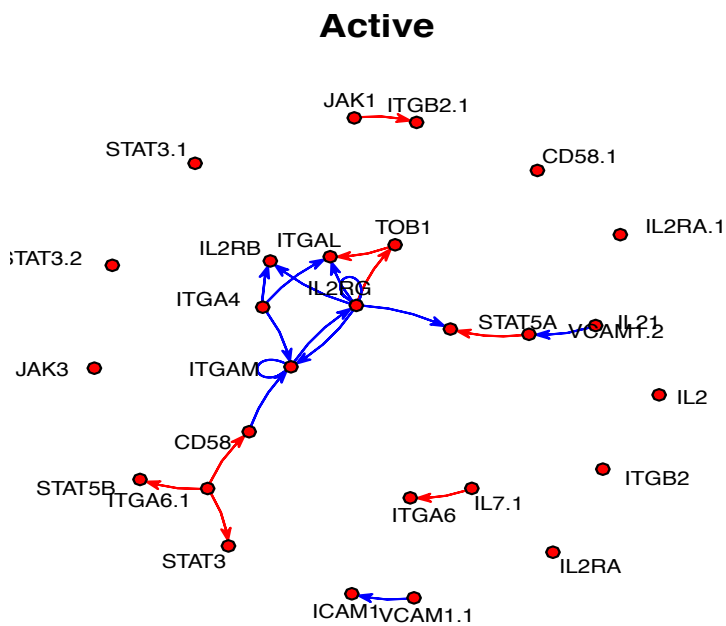
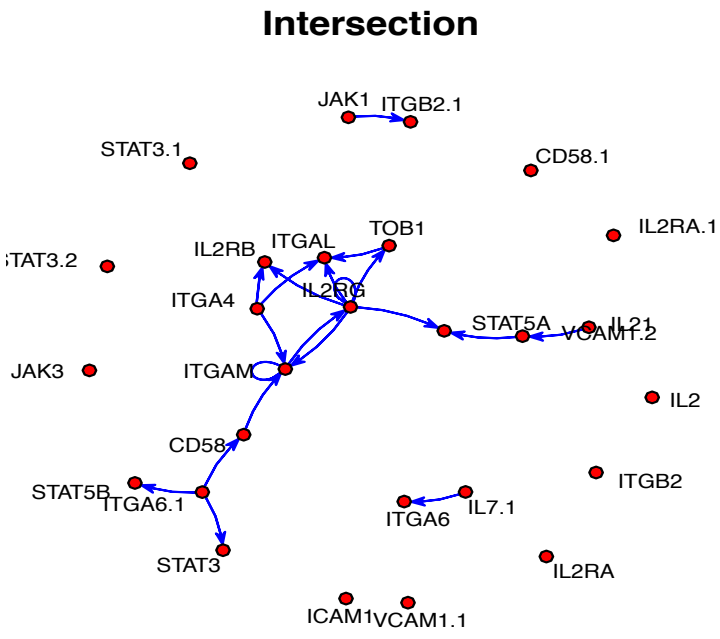


Illustration : Sclérose en plaques (Charbonnier, Chiquet, Ambroise & Corvol) III



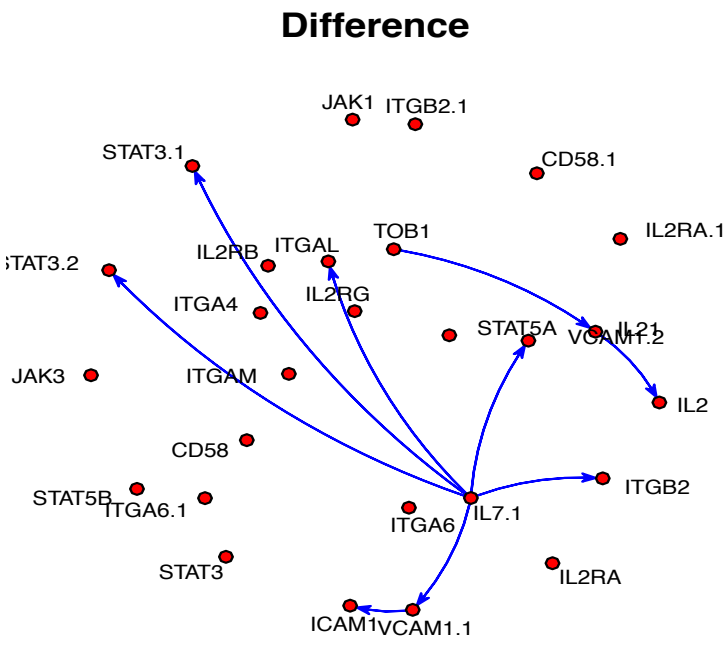


Illustration : Sclérose en plaques (Charbonnier, Chiquet, Ambroise & Corvol) IV

Conclusion

- Plus d'infos que dans une analyse différentielle.
- Réseaux de co-expression traduisent comment les expressions de gènes varient ensemble.
- But à atteindre : faire en sorte que les arêtes traduisent effectivement des régulations.