

# Estimation robuste du support dans une régression linéaire gaussienne en grande dimension pour identifier des interactions entre les facteurs de transcription chez *Arabidopsis thaliana*.

Perrine Lacroix, doctorante

Marie-Laure  
Martin-Magniette



Pascal Massart

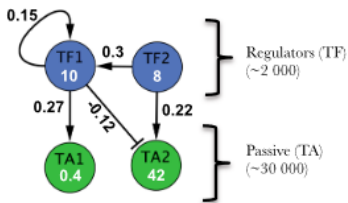
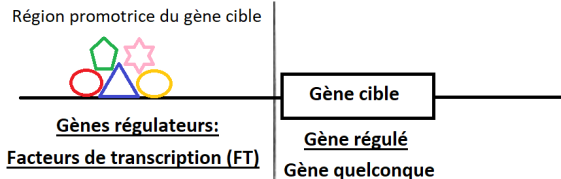


Mélina Gallopin



Présentation NETBIO, 16 Octobre 2019

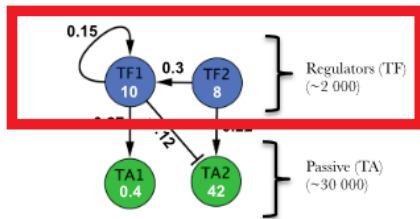
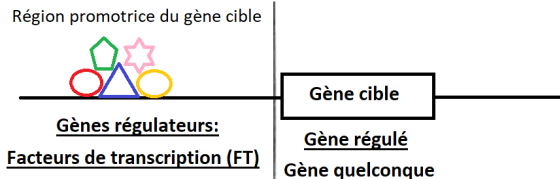
# Les facteurs de transcription



- Un FT peut avoir plusieurs cibles : FT ou non.
- Un FT peut être cible de plusieurs FT.

Objectif biologique : déterminer le réseau de régulation des FT

# Les facteurs de transcription



- Un FT peut avoir plusieurs cibles : FT ou non.
- Un FT peut être cible de plusieurs FT.

**Objectif biologique : déterminer le réseau de régulation des FT**

## Données disponibles

### Puces à ADN :

Acquisition de l'expression de milliers de gènes simultanément.

Données transcriptomiques :

- $p = 1935$  FT
- $n = 1335$  données d'expression **continues** de chaque FT dans des conditions d'expérience très variées

$$(p > n)$$

### Arabidopsis thaliana



# Du problème biologique à la modélisation mathématique

$n$  valeurs d'expression de  $p$  FT :  $(x_1, \dots, x_p) \in \mathbb{R}^{p \times n}$

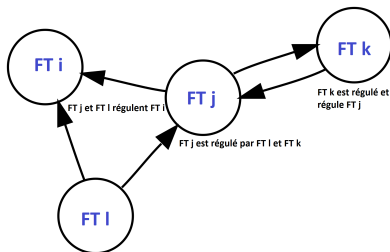
Les expériences sont des réalisations i.i.d de  $\mathcal{N}(\mu_p, \Sigma_{p \times p})$

# Du problème biologique à la modélisation mathématique

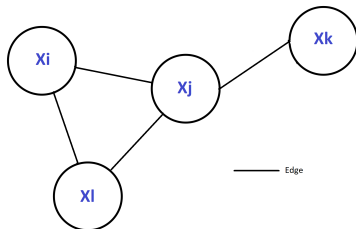
$n$  valeurs d'expression de  $p$  FT :  $(x_1, \dots, x_p) \in \mathbb{R}^{p \times n}$

Les expériences sont des réalisations i.i.d de  $\mathcal{N}(\mu_p, \Sigma_{p \times p})$

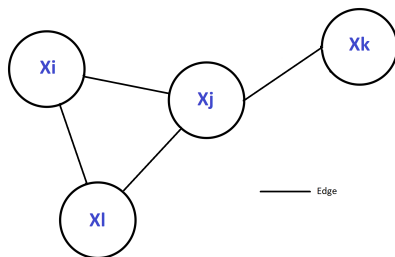
Réseau biologique



Graphe non-orienté



## Inférence du graphe



- Régression de chaque gène  $i$  en fonction des  $p - 1$  restants

$$x_i = X_{-i}\beta_i + \varepsilon$$

où  $X_{-i} = (x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_p)$

- $(\beta_i)_j \neq 0 \Leftrightarrow$  **corrélation partielle** entre  $x_i$  et  $x_j$   
 $\Leftrightarrow$  arête de  $x_j$  à  $x_i$ .

## Cadre : Régression en grande dimension

$$Y = X\beta + \epsilon$$

régression linéaire gaussienne...

- $X$  matrice de taille  $n \times p$  fixée
- $\epsilon = (\epsilon_1, \dots, \epsilon_n)$  i.i.d  $\sim \mathcal{N}(0, \sigma^2 I_n)$ ,  $\sigma^2$  inconnue
- $\beta = (\beta_1, \dots, \beta_p)$  inconnu

... en grande dimension

- grande dimension :  $p \gg n$
- sparsité :  $\#(\text{supp}(\beta))$  petit
- Parmi les variables explicatives  $X_1, \dots, X_p$ , quelles sont celles qui expliquent  $Y$ ?  $\rightarrow$  Estimation du support de  $\beta$

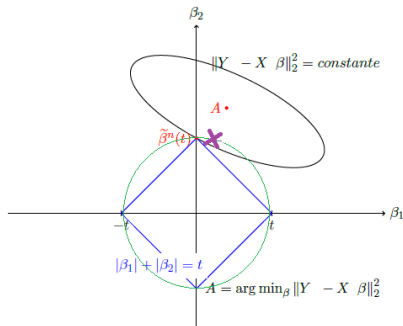


# Pénalité Lasso et Gestion des structures entre les variables

$$Y = X\beta + \epsilon$$

→ **Estimateur Lasso :**

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|^2 + \lambda |\beta|_1 \}$$



→ **Estimateur Elastic-Net :**

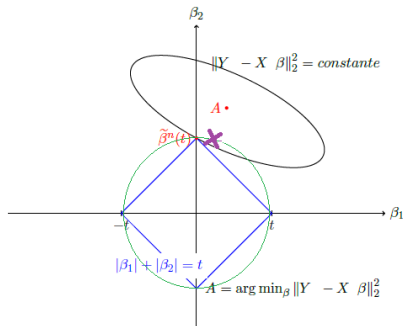
$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|^2 + \lambda ((1 - \alpha)|\beta|_1 + \alpha \|\beta\|_2) \}, \quad \alpha > 0$$

# Pénalité Lasso et Gestion des structures entre les variables

$$Y = X\beta + \epsilon$$

→ **Estimateur Lasso :**

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|^2 + \lambda |\beta|_1 \}$$



→ **Estimateur Elastic-Net :**

$$\operatorname{argmin}_{\beta \in \mathbb{R}^p} \{ \|Y - X\beta\|^2 + \lambda ((1 - \alpha)|\beta|_1 + \alpha\|\beta\|_2) \}, \quad \alpha > 0$$

## Mise en œuvre

- En pratique :
  - 1 Choisir une grille de paramètres de régularisation  $\Lambda$
  - 2 Pour tout  $\lambda \in \Lambda$ , résolution de l'équation LASSO ou l'équation Elastic-Net
  - 3 Obtention d'une collection de supports  $m_\lambda$  et de sous-espaces :  $S_\lambda = \text{Vect}\{X_j, j \in m_\lambda\}$

### Question

Quel  $\lambda \in \Lambda$  choisir à partir de la collection de modèles

$$(\hat{\beta}_\lambda, m_\lambda, S_\lambda)_{\lambda \in \Lambda}$$

→ Compromis biais-variance

## Mise en œuvre

- En pratique :
  - 1 Choisir une grille de paramètres de régularisation  $\Lambda$
  - 2 Pour tout  $\lambda \in \Lambda$ , résolution de l'équation LASSO ou l'équation Elastic-Net
  - 3 Obtention d'une collection de supports  $m_\lambda$  et de sous-espaces :  $S_\lambda = \text{Vect}\{X_j, j \in m_\lambda\}$

### Question

Quel  $\lambda \in \Lambda$  choisir à partir de la collection de modèles

$$(\hat{\beta}_\lambda, m_\lambda, S_\lambda)_{\lambda \in \Lambda}$$

→ Compromis biais-variance

## Sélection de modèles

### Question

Quel  $\lambda \in \Lambda$  choisir à partir de la collection de modèles

$$(\hat{\beta}_\lambda, m_\lambda, S_\lambda)_{\lambda \in \Lambda}$$

$n \rightarrow \infty$

$$\frac{E_{\beta_0}[\|X\beta_0 - X\hat{\beta}_{\hat{\lambda}}\|^2]}{\inf_{\lambda \in \Lambda} \{E_{\beta_0}[\|X\beta_0 - X\hat{\beta}_\lambda\|^2]\}} \xrightarrow[n \rightarrow \infty]{p.s.} 1$$

$n$  fixé

**Inégalité non asymptotique :**

$$E_{\beta_0}[\|X\beta_0 - X\hat{\beta}_{\hat{\lambda}}\|^2] \leq C_n \inf_{\lambda \in \Lambda} \{E_{\beta_0}[\|X\beta_0 - X\hat{\beta}_\lambda\|^2]\} + R_n$$

# Critères de pénalisation asymptotiques

## Idée générale

Contrôle des résidus quadratiques + Volonté de parcimonie

→ Compromis entre

ajustement du modèle et complexité de l'estimation.

Le but est de minimiser en  $\lambda$  : → **eBIC**

$$\underbrace{n \log\left(\frac{\|Y - X\hat{\beta}_\lambda\|^2}{n}\right)}_{\text{"Résidus quadratiques"}} + \underbrace{\dim(S_\lambda) \log(n)}_{\text{pénalité BIC}} - \underbrace{\gamma \log\left(\binom{p}{\dim(S_\lambda)}\right)}_{\text{Complexité du sous-espace } S_\lambda}$$

## Critères de pénalisation non asymptotiques

Le but est de minimiser en  $\lambda$  :

$$\|Y - X\hat{\beta}_\lambda\|^2 + K\sigma^2 \dim(S_\lambda) f(\lambda)$$

- Quel  $K > 1$  choisir ? Quelle fonction  $f$  choisir ? Que faire en pratique pour une **variance inconnue** ?

## Critères de pénalisation non asymptotiques

Le but est de minimiser en  $\lambda$  :

$$\|Y - X\hat{\beta}_\lambda\|^2 + K\sigma^2 \dim(S_\lambda) f(\lambda)$$

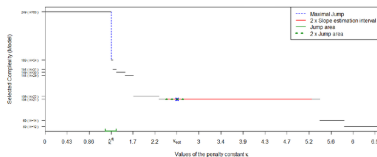
- Quel  $K > 1$  choisir ? Quelle fonction  $f$  choisir ? Que faire en pratique pour une **variance inconnue** ?

- **LinSelect** : Contrôle théorique non-asymptotique explicite :
  - $K \simeq 1.1$
  - $f$  complexe mais explicite
  - $\hat{\sigma}_\lambda$  explicite et dépend du chemin disponible

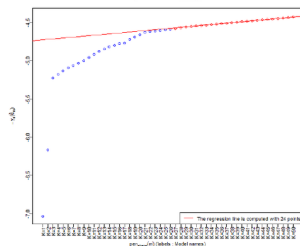


# Critères de pénalisation non asymptotiques

## Dimension Jump



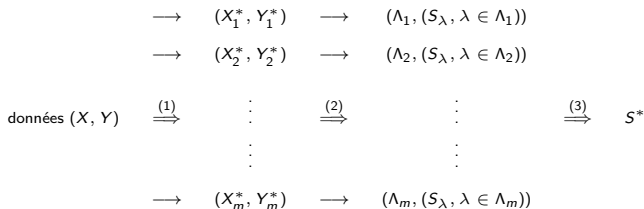
## Heuristique de pente



- Trouver  $\kappa := K\sigma^2$  directement :
  - 1 **dimension jump** : par l'étude de  $\kappa \rightarrow \dim(S_{\hat{\lambda}(\kappa)})$
  - 2 **heuristique de pente** : par l'analyse du graphe  $\{(\dim(S_\lambda), \log(\text{Vrais}(S_\lambda))), \lambda \in \Lambda\}$   
 $\rightarrow f$  non explicite pour la régression grande dimension  
 (absence de propriétés sur les estimateurs)

# Principe et critères

## - Création d'un chemin stable



(1) Choix de la procédure : **sous-échantillonnage** : **Stability Selection**, **ré-échantillonnage** : **Bolasso**

(2) Procédures Lasso indépendantes les unes des autres

(3) Choix du support final : intersection des supports ? variables les plus fréquentes (-> seuil ?)

- Sélection d'un  $\hat{\lambda}$  directement : **critère ESCV** : fonction stabilité + cross validation.

# Principe et critères

## - Création d'un chemin stable

$$\begin{array}{ccccccc}
 & \longrightarrow & (X_1^*, Y_1^*) & \longrightarrow & (\Lambda_1, (S_\lambda, \lambda \in \Lambda_1)) & & \\
 & \longrightarrow & (X_2^*, Y_2^*) & \longrightarrow & (\Lambda_2, (S_\lambda, \lambda \in \Lambda_2)) & & \\
 \text{données } (X, Y) & \xRightarrow{(1)} & \vdots & \xRightarrow{(2)} & \vdots & \xRightarrow{(3)} & S^* \\
 & & \vdots & & \vdots & & \\
 & \longrightarrow & (X_m^*, Y_m^*) & \longrightarrow & (\Lambda_m, (S_\lambda, \lambda \in \Lambda_m)) & & 
 \end{array}$$

(1) Choix de la procédure : **sous-échantillonnage** : **Stability Selection**, **ré-échantillonnage** : **Bolasso**

(2) Procédures Lasso indépendantes les unes des autres

(3) Choix du support final : intersection des supports ? variables les plus fréquentes (-> seuil ?)

- Sélection d'un  $\hat{\lambda}$  directement : **critère ESCV** : fonction stabilité + cross validation.

# Simulation des données

## Données FRANK : dépendances respectant des propriétés biologiques



$$\mathbf{TF} : X_{\text{TF}}(t) = e^{V(t)} + \varepsilon_{\text{TF}}(t),$$

$$V(t) = V(t-1) + AV(t-1)$$

$$\mathbf{TA} : X_{\text{TA}}(t) = e^{W(t)} + \varepsilon_{\text{TA}}(t),$$

$$W(t) = W(t-1) + BW(t-1)$$

### Etude des TFs

- Le plus connecté :  
degré = de 20 à 50
- Le moins connecté :  
degré = 1

$$\longrightarrow p = 200, n = 150,$$

40 itérations

# Simulation des données

## Données FRANK : dépendances respectant des propriétés biologiques



$$\mathbf{TF} : X_{\text{TF}}(t) = e^{V(t)} + \varepsilon_{\text{TF}}(t),$$

$$V(t) = V(t-1) + AV(t-1)$$

$$\mathbf{TA} : X_{\text{TA}}(t) = e^{W(t)} + \varepsilon_{\text{TA}}(t),$$

$$W(t) = W(t-1) + BW(t-1)$$

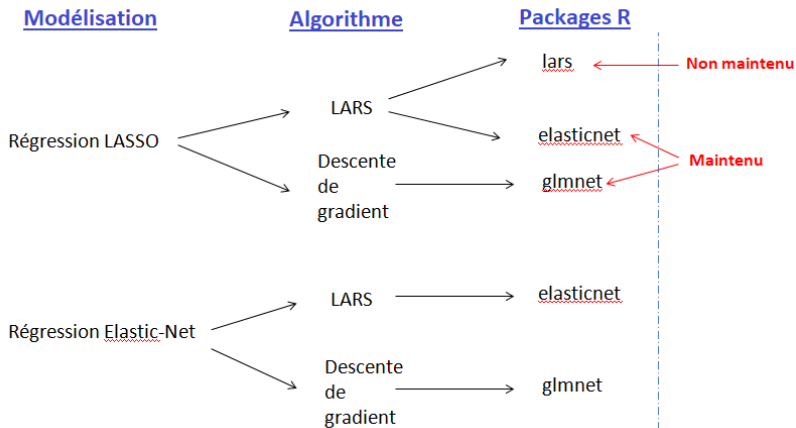
### Etude des TFs

- Le plus connecté :  
degré = de 20 à 50
- Le moins connecté :  
degré = 1

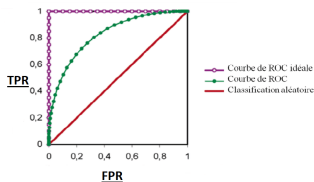
$$\longrightarrow p = 200, n = 150,$$

40 itérations

# Génération des chemins de régularisation



# Aires des courbes ROC



- **Sensibilité (TPR) :**  $\frac{TP}{TP+FN}$  : trouver les bonnes variables
- **Anti-spécificité (FPR) :**  $\frac{FP}{TN+FP} = 1 - \frac{TN}{TN+FP}$  : sélectionner les mauvaises variables

	FRANK-max	FRANK-min	FRANK-max	FRANK-min
lasso (d.coordonnée)	0.849 (0.071)	0.778 (0.071)	0.025 (0.014)	0.000 (0.000)
lasso (lars)	0.857 (0.071)	0.779 (0.071)	0.022 (0.011)	0.000 (0.000)
elastic-net (d.coordonnée)	0.866 (0.069)	0.778 (0.069)	0.024 (0.013)	0.000 (0.000)
elastic-net (lars)	<b>0.915</b> (0.054)	<b>0.808</b> (0.054)	<b>0.015</b> (0.010)	<b>0.000</b> (0.000)

Même seuillage. Aire idéale : 1

Le tout début de la courbe. Aire idéale : 0

## Non asymptotiques $\gg$ Asymptotiques

**Prévision : Risque quadratique** : pour  $(\tilde{Y}, \tilde{X})$  un jeu test

$$\frac{1}{|\tilde{Y}|} \sum_{i=1}^{|\tilde{Y}|} \left( \tilde{Y}_i - (\tilde{X} \hat{\beta}_\lambda)_i \right)^2$$

**Description :**

- **Sensibilité** :  $\frac{TP}{TP+FN}$  : sélectionner les bonnes variables

- **Spécificité** :  $\frac{TN}{TN+FP}$  : ne pas sélectionner les mauvaises variables

- , non asymptotiques plus performants que asymptotiques pour toutes les métriques
- , mêmes conclusions pour les 2 types de nœud étudiés
- pas de différence significative entre elastic-net et lasso
- d. coordonnées souvent significativement plus performant que lars (risque et spécificité)



## Critères non asymptotiques

	max	min	max	min	max	min
<b>heur/dim_Id</b>						
lasso (d.coordonnée)	0.174	0.650	<b>1.000</b>	<b>0.901</b>	0.696	<b>1.573</b>
lasso (lars)	0.503	0.675	0.675	<b>0.780</b>	0.931	2.120
elastic-net (d.coordonnée)	0.176	<b>0.775</b>	<b>1.000</b>	0.713	<b>0.693</b>	2.330
elastic-net (lars)	0.490	0.675	0.621	0.718	1.162	4.831
<b>heur/dim_f</b>						
lasso (d.coordonnée)	<b>0.264</b>	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	<b>0.571</b>	1.041
lasso (lars)	0.036	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.944	<b>1.039</b>
elastic-net (d.coordonnée)	0.262	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.572	<b>1.039</b>
elastic-net (lars)	0.036	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.944	<b>1.039</b>
<b>LinSelect</b>						
lasso (d.coordonnée)	<b>0.273</b>	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	<b>0.561</b>	<b>1.039</b>
lasso (lars)	0.270	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.562	<b>1.039</b>
elastic-net (d.coordonnée)	0.261	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.572	<b>1.039</b>
elastic-net (lars)	0.255	<b>0.600</b>	<b>1.000</b>	<b>0.998</b>	0.573	<b>1.039</b>

**Sensibilité / Spécificité / Risque quadratique**

## La stabilité

**Prévision : Risque quadratique :** pour  $(\tilde{Y}, \tilde{X})$  un jeu test

$$\frac{1}{|\tilde{Y}|} \sum_{i=1}^{|\tilde{Y}|} \left( \tilde{Y}_i - (\tilde{X} \hat{\beta}_\lambda)_i \right)^2$$

- Description :**
- **Sensibilité :**  $\frac{TP}{TP+FN}$  : sélectionner les bonnes variables
  - **Spécificité :**  $\frac{TN}{TN+FP}$  : ne pas sélectionner les mauvaises variables
  - Sensibilité  $\searrow$  % critères pénalisés : support plus petit
  - Spécificité  $\nearrow$  % critères pénalisés : sélection de meilleure qualité
  - Bolasso : meilleur compromis % 3 métriques
  - Elastic net plus performant que lasso
  - descente par coordonnées plus performant que lars (risque et spécificité)

# Conclusion et Perspectives

## Résumé

- non asymptotique  $\gg$  asymptotique
- Elastic-net plus performant que lasso

## Perspectives

Comment choisir  $\lambda$  pour obtenir un estimateur du support de  $\beta$   
**robuste et non-asymptotique ?**

- Combiner heuristique de pente/dimension jump et bolasso pour  $\nearrow$  sensibilité et  $\nearrow$  spécificité
- Trouver un compromis entre **sensibilité** et **spécificité**
- Données réelles : pré-processing, construction du réseau, etc...

- Birgé, L. and Massart, P. (2006). Minimal penalties for gaussian model selection. *Probability Theory and Related Fields*.
- Buhlmann, P. and Meinshausen, N. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3) :1436–1462.
- Carré, C., Mas, A., and Krouk, G. (2017). Reverse engineering highlights potential principles of large gene regulatory network design and learning. *Systems Biology and Applications*, pages 3–17.
- Giraud, C., Huet, S., and Verzelen, N. (2012). High-dimensional regression with unknown variance. *Statistical Science*, 27(4) :500–518.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Serie B (Methodological)*, 58(1) :267–288.

**MERCI !**