Construction d'un réseau orienté stable de facteurs de transcription

Yann Vasseur [(1)] & Gilles Celeux [(1,2)] & Marie-Laure Martin-Magniette [(3,4,5)] & Guillem Rigaill [(3,4)]

[1] *Laboratoire de Mathématiques, UMR 8628, Université Paris-Sud, F-91405 Orsay, France.*

[2] *INRIA Saclay, Ile-de-France, Université Paris-Sud, F-91405 Orsay, France.*

[3] *Institute of Plant Sciences Paris-Saclay IPS2, CNRS, INRA, Université Paris-Sud, Université Evry, Université Paris-Saclay, Bâtiment 630, 91405 Orsay, France.*

[4] *Institute of Plant Sciences Paris-Saclay IPS2, Paris Diderot, Sorbonne Paris-Cité, Bâtiment 630, 91405, Orsay, France.*

[5] *UMR MIA-Paris, AgroParisTech, INRA, Université Paris-Saclay, 75005, Paris, France.*

Résumé. L'objectif est d'inférer un réseau de régulation de facteurs de transcriptions (FTs), régulateurs clés de l'expression des gènes, d'*Arabidopsis thaliana*, à partir de données d'expression. Les 1937 FTs sont les variables statistiques et les 2670 données sont les observations. Nous représentons ce réseau par un modèle graphique gaussien dont il faut déterminer les arêtes. Pour ce faire, nous cherchons pour chaque variable (FT) l'ensemble des variables explicatives (FTs régulateurs) à l'aide de régressions linéaires pénalisées traitées de manière indépendante. Une approche préliminaire qui consiste à sélectionner un ensemble de variables du chemin de régularisation via un critère de vraisemblance pénalisée s'avère être instable et fournit trop de variables explicatives. Pour contrecarrer cela, nous mettons en compétition deux procédures fondées sur le rééchantillonnage dans le but de disposer, non plus d'un, mais de plusieurs chemins de régularisation par variable régressée. La première procédure choisit un ensemble de variables par sous-échantillon et ne conserve que les variables les plus fréquentes à l'aide d'un seuil. La seconde confronte directement les sous-échantillons entre eux et garde celui de plus grande vraisemblance. La performance de ces procédures a été évaluée sur des jeux simulés de même structure que le jeu réel.

Mots-clés. Inférence de réseaux biologiques, haute dimension, modèles graphiques gaussiens, LASSO, rééchantillonnage, critères de vraisemblance pénalisée.

Abstract. The goal is to infer a regulatory network of transcription factors (TFs), key regulators of gene expression, of *Arabidopsis thaliana*, using transcriptome data. Our variables are the 1937 TFs and our observed values are the 2670 data. The network is represented by a graphical gaussian model which edges have to be indentified. For each variable (TF) we look for the set of relevant variables (regulators of TFs) using

independant penalized linear regressions. A first approach which consists in selecting a set of variables from the regularization path with the help of a penalized likehood criterion is unstable and provides too much relevant variables. To solve this problem, two procedures based on resampling are confronted so that we dispose of several regularization paths per regressed variable rather than an only one. The first procedure selects a set of variables per sub-sample and only keeps the most frequent variables with the use of a threshold. The second one directly confronts sub-samples among themselves and keeps the one having the highest likehood. The performance of these procedures were evaluated on simulated data with the same structure than our real data.

Keywords. Biological network inference, high dimension, gaussian graphical models, LASSO, resampling, penalized likehood criteria.

1 Introduction

1.1 Contexte

Le but de cette contribution est d'inférer un réseau de régulation orienté entre les facteurs de transcription (FTs) d'*Arabidopsis thaliana* (At), plante modèle. Les FTs sont des gènes qui contrôlent l'expression des autres gènes, représentant des acteurs clés des réseaux de régulation. Nous avons n données d'expression pour chacun des p FTs d'At. Ce problème de haute dimension s'inscrit dans un cadre statistique raisonnable ($p \sim n$).

1.2 Modélisation

L'objectif est de déterminer, pour chaque FT, l'ensemble de ses FTs régulateurs. D'un point de vue statistique, nous considérons les $p = 1937$ FTs comme étant nos variables et l'ensemble $\mathcal{I}$ des $n = 2670$ données d'expression comme étant nos observations. La matrice d'observations X de taille $n \times p$ est centrée et réduite en colonnes.

On modélise le réseau à inférer par un graphe $\mathcal{G}$ où les FTs correspondent à ses nœuds et les liens de régulations à ses arêtes. Une arête dirigée du noeud j' vers le noeud j signifie que j' régule j. $\mathcal{G}$ sera un modèle graphique gaussien (GGM) dont il faut déterminer la matrice d'adjacence $A_{\mathcal{G}}$. Nous adopterons les notations suivantes :

- X suit une loi normale multidimensionnelle $\mathcal{N}(0, \Sigma)$
- Un noeud j est représenté par une variable gaussienne $X_j \sim \mathcal{N}(0, 1)$.
- $X^{\mathcal{M}}$ (resp. $X^{-\mathcal{M}}$) la matrice de taille $n \times |\mathcal{M}|$ (resp. $n \times (p - |\mathcal{M}|)$) correspondant à la restriction de X aux variables X_c pour $c \in \mathcal{M}$ (resp. pour $c \in \mathcal{M}^C$). Si $\mathcal{M} = \{j\}$, on posera $X^j = X^{\{j\}}$ (resp. $X^{-j} = X^{-\{j\}}$).

Notons qu'un GGM est généralement utilisé pour fonder un graphe non orienté via l'estimation de sa matrice de précision symétrique $K = (\Sigma)^{-1}$. Le graphe que nous désirons construire étant orienté, nous ne chercherons pas à estimer K.

1.3 Méthodologie

Le cadre gaussien induit la propriété suivante :

$$X_{j'} \text{ variable explicative de } X_j \ (j' \text{ régule } j) \quad \Leftrightarrow \quad \theta_{j,j'} \neq 0$$

$\theta_{j,j'}$ est le coefficient de $X_{j'}$ dans le modèle de régression de X_j sur les autres variables :

$$X^j = X^{-j}\Theta_j + \epsilon_j \text{ avec } \begin{cases} \Theta_j = \left\{\theta_{j,j'}; j' \in \{1,...,p\} \backslash j\right\}^T \\ \epsilon_j = \{\epsilon_{j,1},..,\epsilon_{j,n}\}^T \text{ tq } \{\epsilon_{j,i}\}_{i\in\{1,..,n\}} \text{ i.i.d de loi } \mathcal{N}(0,\sigma^2) \end{cases}.$$

On désignera par support de j, l'ensemble S_j des variables explicatives à la variable régressée X_j ($S_j = \{j' \text{ tq } \theta_{j,j'} \neq 0\}$). Estimer $A_{\mathcal{G}}$ revient à estimer le support des p variables. Pour cela, Meinshausen.N et Bühlmann.P [1] ont proposé la procédure Gauss-LASSO qui effectue p régressions linéaires pénalisées indépendantes d'une variable sur les autres.

2 Procédure Gauss-LASSO + BIC

Pour un X_j, Gauss-LASSO fournit un chemin de régularisation. Le support estimé correspondra à celui dont la pénalité associée λ_j est choisie par BIC. En voici le détail :

1. Régression pénalisée par LASSO : $\widehat{\Theta}_j^{Las}(\lambda_j) = \underset{\Theta}{\text{argmin}}\left(\frac{1}{2}\|X^j - X^{-j}\Theta\|_2^2 + \lambda_j\|\Theta\|_1\right)$.
   - Collection de modèles obtenue : $\mathcal{M}_j^{Las} = \left\{X^j = X^{-j}\widehat{\Theta}_j^{Las}(\lambda_j) + \widehat{\epsilon}_j^{Las}\right\}_{\lambda_j}$.
   - Ensemble de supports associés : $\mathcal{S}_j^{Las} = \left\{\widehat{S}_j^{Las}(\lambda_j) = \left\{j' \text{ tel que } \widehat{\theta}_{j,j'}^{Las}(\lambda_j) \neq 0\right\}\right\}_{\lambda_j}$.
2. Réestimation des coefficients dans les supports par moindres carrés ordinaires :
$$\forall \lambda_j, \left(\widehat{\theta}_{j,j'}^{GL}(\lambda_j)\right)_{j'\in\widehat{S}_j^{Las}(\lambda_j)} = \underset{\Theta}{\text{argmin}}\|X^j - X^{\widehat{S}_j^{Las}(\lambda_j)}\Theta\|_2^2.$$
   Cette étape ne modifie pas les supports mais la vraisemblance des modèles associés :
   - Nouvelle collection de modèle : $\mathcal{M}_j^{GL} = \left\{M_j^{GL}(\lambda_j)\right\}_{\lambda_j} = \left\{X^j = X^{-j}\widehat{\Theta}_j^{GL}(\lambda_j) + \widehat{\epsilon}_j^{GL}\right\}_{\lambda_j}$.
   - Ensemble de supports associés inchangé : $\mathcal{S}_j^{GL} = \mathcal{S}_j^{Las}$.
3. Sélection par le critère BIC d'un λ_j et par conséquent du support de $\mathcal{S}_j^{GL}$ associé :
$$\lambda_j^{BIC} = \underset{\lambda_j}{\text{argmin}}\left(-2V_{max}\left(M_j^{GL}(\lambda_j)\right) + \left|\widehat{S}_j^{GL}(\lambda_j)\right|\log(n)\right).$$
   où $V_{max}(M)$ est la log-vraisemblance maximisée du modèle M en ses paramètres.

Les p supports $\left\{\widehat{S}_j^{GL}(\lambda_j^{BIC})\right\}_j$ contiennent en moyenne 196 variables. La sélection est trop peu parcimonieuse. De plus, en appliquant la procédure à des matrices d'observations simulées de taille $n \times p$ de loi $\mathcal{N}(0,\widehat{\Sigma})$ où $\widehat{\Sigma}$ est la matrice de covariance estimée sur le jeu réel X, on s'aperçoit que les supports estimés sont très variables d'un jeu à l'autre : la procédure s'avère très instable. Pour stabiliser les supports estimés, nous avons mis en place des procédures fondées sur le rééchantillonnage.

3 Stabilisation par rééchantillonnage

Le but du rééchantillonnage est, en perturbant les données, de proposer à X_j non plus un ensemble de supports candidats comme le fait le Gauss-LASSO, mais plusieurs. Pour diversifier au maximum ces ensembles de supports, nous proposons de rééchantillonner par Sample-Splitting (SS), inspiré de [3] :

- Tirage de $m/2$ sous-ensembles $\mathcal{J}_1, , ..., \mathcal{J}_{m/2}$ du jeu $\mathcal{I}$, tels que $|\mathcal{J}_1| = ... = |\mathcal{J}_{m/2}| = \lfloor \frac{n}{2} \rfloor$.
- Rajout des $m/2$ ensembles complémentaires $\left(\forall k, \mathcal{J}_{m/2+k} = (\mathcal{J}_k)^C\right)$ dans $\mathcal{I}$.
- Restriction de X à chacun des m jeux créés : $\forall k \in \{1, .., m\}$, on pose ${}^{(k)}X = X_{|\mathcal{J}_k}$.
- Création de m sous-modèles pour X_j : $\forall k \in \{1, .., m\}$, ${}^{(k)}X^j = {}^{(k)}X^{-j}.\Theta_j + \epsilon_j$.
- Application de Gauss-LASSO sur chacun d'eux. Obtention de m collections de modèles ${}^{(k)}\mathcal{M}_j^{GL} = \left\{{}^{(k)}M_j^{GL}({}^{(k)}\lambda_j)\right\}_{{}^{(k)}\lambda_j}$ et ensembles de supports ${}^{(k)}\mathcal{S}_j^{GL} = \left\{{}^{(k)}\widehat{S}_j^{GL}({}^{(k)}\lambda_j)\right\}_{{}^{(k)}\lambda_j}$.

Pour gagner en stabilité, nous exploitons par la suite ces ensembles de supports candidats avec deux procédures, Gauss-LASSO *stabilisé* et Gauss-LASSO *enrichi*, construites différemment dont nous calibrerons les paramètres et comparerons les résultats.

4 Gauss-LASSO stabilisé

Comme Bolasso [4], nous nous inspirons du principe de stabilité dans la sélection [2] :

- $\forall k \in \{1, .., m\}$, choix dans ${}^{(k)}\mathcal{S}_j^{GL}$, du support ${}^{(k)}\widehat{S}_j^{GL}\left({}^{(k)}\lambda_j^{BIC}\right)$ désigné par BIC.
- Calcul du score de chaque variable explicative candidate, à savoir sa fréquence d'apparition dans ces m supports : $\forall j' \in \{1, .., p\}\backslash j$, $\mathbb{S}(j, j') = \frac{1}{m}\sum_{k=1}^{m} \mathbb{1}_{j' \epsilon {}^{(k)}\widehat{S}_j^{GL}({}^{(k)}\lambda_j^{BIC})}$.
- Création du support final avec les variables candidates aux scores plus élevés qu'un seuil s_j à calibrer : $\widehat{S}_j^{GLstab}(s_j) = \left\{j' \in \{1, .., p\}\backslash j, \text{ tq } \mathbb{S}(j, j') \geq s_j\right\}$.

Notons ici que le SS permet de mieux contrôler le nombre de variables sélectionnées à tort qu'un rééchantillonnage classique : une variable au score élevé est sélectionnée sur plusieurs sous-échantillons disjoints et s'avère donc être effectivement stable.

Il reste à calibrer s_j. Une bonne calibration doit rendre la procédure suffisamment parcimonieuse pour qu'elle permette un bon compromis entre complexité et prédiction. Nous calculons, en fonction de s_j, l'erreur de prédiction des variables explicatives sélectionnées par une procédure de validation croisée (10-*fold*). L'erreur de prédiction va être utilisée pour établir un bon compromis complexité-prédiction. Son calcul étant lourd, il a uniquement été réalisé pour cinq variables représentatives du panel. Nous présenterons des moyens peu coûteux de calibrer le seuil de chacune des p variables régressées, fondés sur les log-vraisemblances des modèles associés aux supports candidats $\{\widehat{S}_j^{GLstab}(s_j)\}_{s_j}$, estimant des supports $\{\widehat{S}_j^{GLstab}(\widehat{s_j})\}_j$ de tailles réduites de moitié comparé à Gauss-LASSO+BIC et, pour les cinq variables représentatives, présentant des erreurs de prédiction raisonnables.

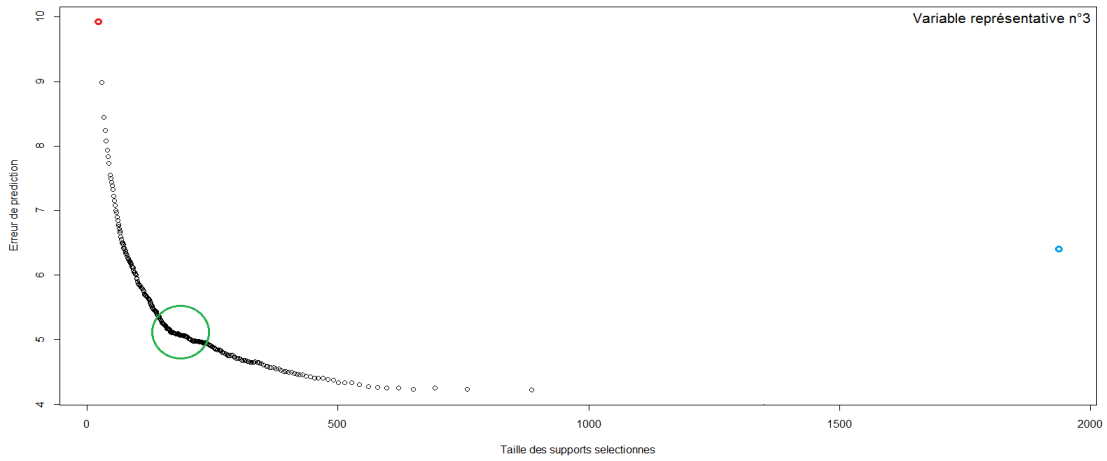

Figure 1 - Tracé, pour une variable représentative, de l'évolution de l'erreur de prédiction (ordonnée) des variables de $\widehat{S}_j^{GLstab}(s_j)$ et de leur nombre (abscisse) lorsque s_j varie. Le point associé à $s_j = 1$ est en rouge et celui associé à $s_j = 0$ en bleu. Les seuils résultant en un bon compromis prédiction-complexité escompté sont ceux entourés en vert.

Néanmoins, nous nous efforcerons de calibrer les seuils plus précisément pour obtenir des supports de taille minimale dans la zone entourée en vert et ainsi fournir une procédure plus stable.

5 Gauss-LASSO enrichi

Cette procédure enrichit la collection de modèles $\mathcal{M}_j^{GL}$ résultant de Gauss-LASSO en la collection regroupant l'ensemble des modèles issus du rééchantillonnage :

- Création de la collection $\mathcal{M}_j^{GLenri} = \bigcup_{k=1}^{m} {}^{(k)}\mathcal{M}_j^{GL}$ et de l'ensemble $\mathcal{S}_j^{GLenri} = \bigcup_{k=1}^{m} {}^{(k)}\mathcal{S}_j^{GL}$.
- Regroupement des modèles de même dimension en sous-collections : $\forall D \geq 1$, $\mathcal{M}_j^{GLenri}(D) = \left\{M \in \mathcal{M}_j^{GLenri}, \left|\widehat{S}_M\right| = D\right\}$ avec $\widehat{S}_M \in \mathcal{S}_j^{GLenri}$ support associé à M.
- Choix dans chaque sous-collection du modèle de plus grande log-vraisemblance : $\forall D \geq 1, {}^{(l)}M_j^{max}(D) = \underset{\mathcal{M}_j^{GLenri}(D)}{\operatorname{argmax}} \{V_{max}(M)\}$, avec deux cas possibles ($l = 1$ ou 2) :
  - $V_{max}(M) = V_{max}^{(k)}(M)$ calculée sur le sous-échantillon de sélection $\mathcal{J}_k$ de M ($l = 1$).
  - $V_{max}(M) = V_{max}^{(n)}(M)$ calculée sur le jeu entier $\mathcal{I}$ ($l = 2$).
- Établissement de la meilleure collection de modèles $\mathcal{M}_j^{max}(l) = \left\{{}^{(l)}M_j^{max}(D)\right\}_{D\geq 1}$ et de l'ensemble des supports associés $\mathcal{S}_j^{max}(l) = \left\{{}^{(l)}\widehat{S}_j^{max}(D)\right\}_{D\geq 1}$.
- Sélection du support final $\widehat{S}_j^{GLenri}(l)$ de $\mathcal{S}_j^{max}(l)$ par heuristique de pente ([5], [6]).

Contrairement à Gauss-LASSO *stabilisé*, elle ne retient donc pas un support dans chaque ensemble issu du rééchantillonnage mais crée le meilleur ensemble de supports possible.

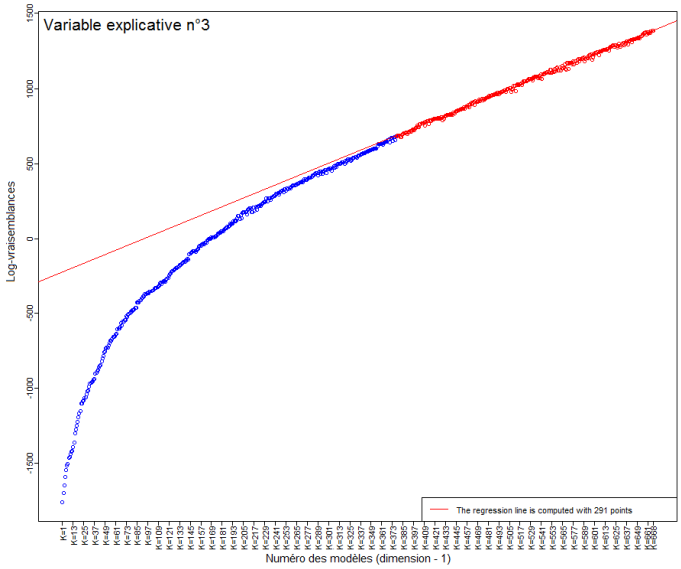

Figure 2 - Tracé , pour une variable représentative, de l'évolution de V_{max} (évaluée sur les sous-jeux) des modèles de la meilleure collection $\mathcal{M}_j^{max}(l = 1)$ en fonction de la taille de leur support. La courbe présente un comportement linéaire pour des tailles de supports élevées propice à l'application de l'heuristique de pente. Notons ${}^{(1)}\kappa_j$ la pente de cette partie linéaire.

Nous observons ce même comportement linéaire dans le cas où V_{max} est évaluée sur $\mathcal{I}$. Notons ${}^{(2)}\kappa_j$ la pente de la courbe correspondante.

Le choix de l'heuristique de pente comme critère de vraisemblance pénalisée est motivé par son caractère non asymptotique contrairement à BIC. Elle adapte, en effet, la pénalité à la collection de modèles ($\text{pen}_{\text{HP}}(M) = 2 \times {}^{(l)}\kappa_j \times \left|\widehat{S}_M\right|$ avec $\widehat{S}_M$ support associé au modèle M). Les supports $\widehat{S}_j^{GLenri}(l)$ ainsi estimés par Gauss-LASSO *enrichi* dépendent fortement du jeu de données sur lequel les log-vraisemblances sont calculées. Nous illustrerons le fait que ces supports sont stables lorsque la log-vraisemblance de chaque modèle est évaluée sur le sous-jeu $\mathcal{J}_k$ associé à son sous-échantillon de sélection.

Enfin, nous illustrerons le fait que pour une variable X_j, les supports estimés par Gauss-LASSO *stabilisé* et *enrichi* sont de taille comparable et ont un grand nombre de variables explicatives en commun.

Bibliographie

[1] Meinshausen.N et Bühlmann.P (2006), High dimensional graphs and variable selection with the Lasso, *Annals of Statistics, 34 (3), 1436 - 1462.*
[2] Meinshausen.N et Bühlmann.P (2010), Stability Selection, *Journal of the Royal Statistical Society : Series B ; 72 : 417-73.*
[3] Haury.A et al. (2012), TIGRESS : Trustful Inference of Gene REgulation using Stability Selection, *BMC Systems Biology*
[4] Bach.F (2008), Bolasso : Model Consistent Lasso Estimation through the Bootstrap, *Proceedings of the 25th international conference on Machine learning*
[5] Birgé.L et Massart.P (2001), Gaussian model selection, *Journal of the European Mathematical Society*, 3(3) : 203-268
[6] Birgé.L et Massart.P (2006), Minimal penalties for gaussian model selection, *Probability Theory and Related Fields*, 138(1-2) : 33-73