

Présentation de projet : Méthodes de statistique en grande dimension pour l'analyse de données de biologie moléculaire

Louis Lacoste

13 Février, 2024

Présentation des données

Contexte biologique

Le jeu de données choisi étudie l'influence de 2 paramètres sur la **capacité germinative** des graines à l'aide de données de métabolomique pour des graines fraîchement récoltées (*Freshly Harvested*).

La température, variable qualitative à 3 niveaux :

- *Low*
- *Medium*
- *Elevated*

Le stade d'imbibition, variable qualitative à 3 modalités :

- DS (*Dry seed*)
- EI après 6h d'imbibition (*Early imbibition*) correspondant à la fin de la prise d'eau,
- LI après 20h d'imbibition (*Late imbibition*)

Extrait des données

Nous présentons ici les 5 premières colonnes du jeu de données¹ :

temperature	imbibition	m_Alanine	m_Arginine	m_Asparagine
Low	DS	0.7670471	0.0293251	0.4197357
Low	DS	0.7360741	0.0349146	0.4447546
Low	DS	0.8128032	0.0464299	0.4347309
Medium	DS	0.4299879	0.0210281	0.5830220
Medium	DS	0.5130800	0.0119001	0.5458675
Medium	DS	0.4696609	0.0197695	0.5696875

¹Les 4 premières colonnes présentent différentes informations pour identifier les conditions expérimentales et la répétition.

Statistiques descriptives

Statistiques descriptives

Vérifications élémentaires

Il y a 0 colonnes dont la moyenne est nulle, 0 colonnes dont l'écart-type est nul et 0 NAs.

Méthode de statistique en grande dimensions

Principe

Matrices X et Y

$$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & & & & & & & & \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \end{bmatrix}$$

Calcul des résidus

En calculant les résidus du modèle linéaire on obtient une *p-value* de 0.051846 qui est à peine au-dessus du seuil. Malgré tout nous allons voir si le blanchiement permettrait d'améliorer cela.

Principe du blanchiement

Le principe du blanchiement est de supprimer les corrélations existants entre les colonnes.

Pour cela il faut estimer $\Sigma^{-1/2}$ et alors le modèle se ré-écrit :

$$\mathbf{Y}\Sigma^{-1/2} = \mathbf{X}\mathbf{B}\Sigma^{-1/2} + \mathbf{E}\Sigma^{-1/2}$$

Puis on peut appliquer le critère LASSO et la *stability selection* sur le modèle vectorisé :

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}$$

avec

$$\mathcal{Y} = \text{vec}(\mathbf{Y}\Sigma^{-1/2}), \mathcal{X} = (\Sigma^{-1/2}) \otimes \mathbf{X}, \mathcal{B} = \text{vec}(\mathbf{B}), \mathcal{E} = \text{vec}(\mathbf{E}\Sigma^{-1/2})$$

Il faut donc estimer $\Sigma^{-1/2}$ avec un estimateurs $\hat{\Sigma}^{-1/2}$.

Pour cela le package R `MultiVarSel` (Perrot-Dockès, Lévy-Leduc, and Chiquet (n.d.)) implémente 3 méthode pour blanchir les données en utilisant 3 structures de dépendances :

- $AR1$
- $ARMA(p, q)$
- Non paramétrique²

²Suppose uniquement que le processus est stationnaire.

Blanchiment des données

Et ainsi la méthode qui blanchit le mieux ces données est la méthode *non paramétrique*³. Nous récupérons à la fin de cette étape la matrice $\hat{\Sigma}^{-\frac{1}{2}}$ permettant de blanchir les données.

Table 2: Tableau de résultats des tests de Portmanteau pour les différentes méthodes

	Pvalue	Decision
AR1	0.127	WHITE NOISE
nonparam	0.722	WHITE NOISE
ARMA 1 1	0.13	WHITE NOISE
no_whitening	0.052	WHITE NOISE

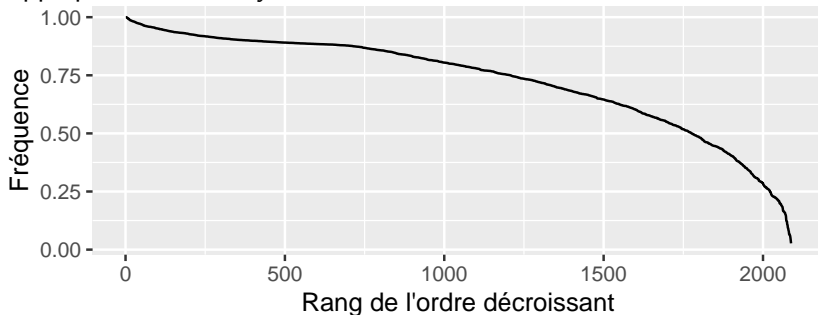
³Ce qui est empiriquement régulièrement le cas.

Sélection de variable

La *stability selection*

Le principe est

Voici un graphique des fréquences obtenues par ordre décroissant en appliquant la *stability selection*⁴



⁴Nous avons fait 5000 réplicats en utilisant le *cluster* Migale.

Sur le graphique, on observe une cassure de la fréquence aux alentours de la 750e fréquence par ordre décroissant.

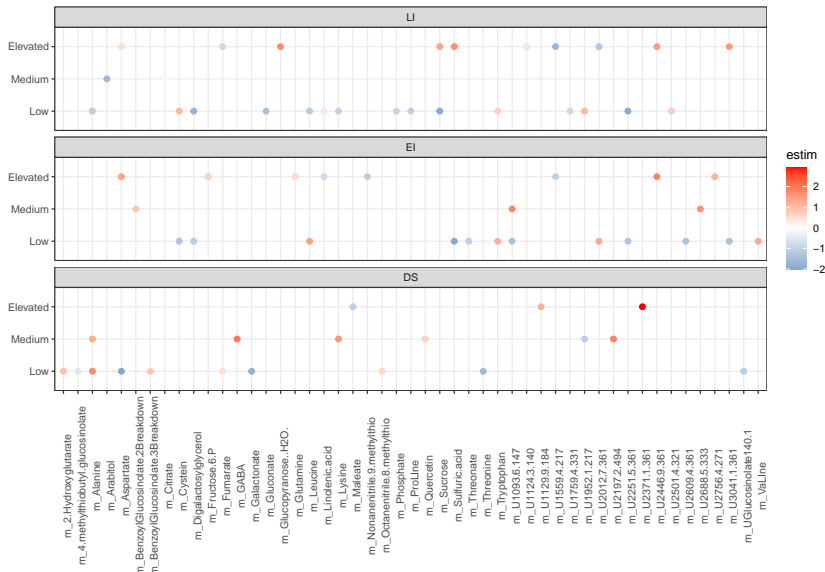
Afin de pouvoir interpréter nos résultats plus facilement, notamment du point de vue biologique nous allons nous limiter à un seuil de

Ré-estimation des paramètres

Pourquoi ré-estimer ?

Dans le cours (Lévy-Leduc (2024)), nous avons vu que les Théorèmes 1 et 2 garantissent la consistance en signe des estimateurs des $\tilde{\mathcal{B}}$.

Cependant, l'estimation de la valeur tend à être biaisée, cette étape nous permet donc de ré-estimer les valeurs des \mathcal{B} qui ont été estimés non nuls.



Bibliographie

Bibliographie

Lévy-Leduc, Céline. 2024. “Notes pour le cours : ‘Méthodes de statistique en grande dimension pour l’analyse de données de biologie moléculaire.’”

Perrot-Dockès, Marie, Céline Lévy-Leduc, and Julien Chiquet. n.d. “Introduction to MultiVarSel.”

Annexes