Présentation de projet : Méthodes de statistiques 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 (***F****reshly* ***H****arvested*).

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 données[1] :

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

[1] Les 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 statistiques en grande dimension

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 blanchiment permettrait d'améliorer cela.

Principe du blanchiment

Le principe du blanchiment est de supprimer les corrélations existant entre les colonnes.

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

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

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

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

avec

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

Il faut donc estimer $\boldsymbol{\Sigma}^{-1/2}$ avec un estimateur $\widehat{\boldsymbol{\Sigma}}^{-1/2}$.

Pour cela le package R `MultiVarSel` (Perrot-Dockès, Lévy-Leduc, and Chiquet (2019)) implémente 3 méthodes pour blanchir les données en utilisant 3 structures de dépendance :

- $AR(1)$
- $ARMA(p, q)$
- Non paramétrique[2]

[2] Suppose uniquement que le processus est stationnaire.

Blanchiement 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

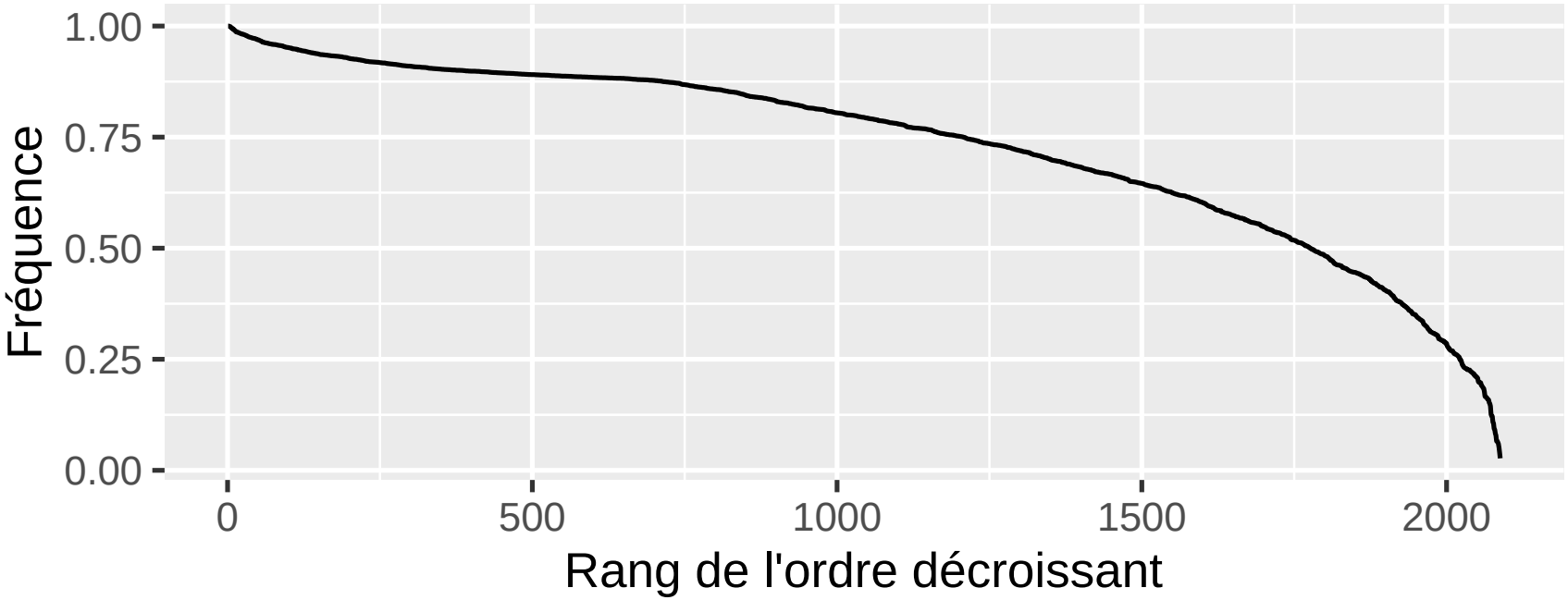
[3]Ce qui est empiriquement régulièrement le cas.

Sélection de variable

La *stability selection*

Le principe est

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



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

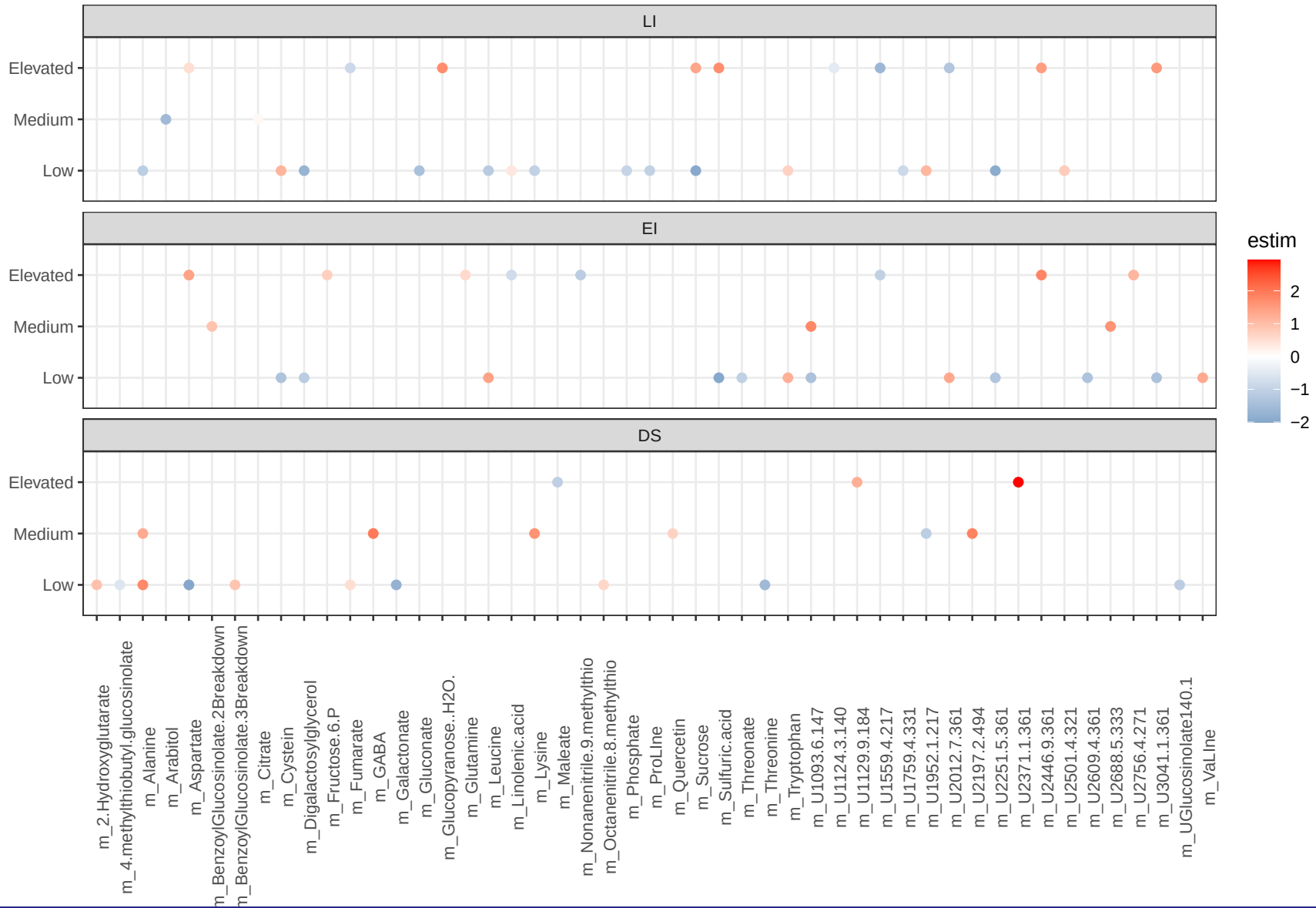
Sur le graphique, on observe une cassure de la fréquence aux alentours de la 750e fréquence par ordre décroissant.

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.


LI
EI
DS
Elevated
Medium
Low
estim
2
1
0
−1
−2
m_2.Hydroxyglutarate
m_4.methylthiobutyl.glucosinolate
m_Alanine
m_Arabitol
m_Aspartate
m_BenzoylGlucosinolate.2Breakdown
m_BenzoylGlucosinolate.3Breakdown
m_Citrate
m_Cystein
m_Digalactosylglycerol
m_Fructose.6.P
m_Fumarate
m_GABA
m_Galactonate
m_Gluconate
m_Glucopyranose..H2O.
m_Glutamine
m_Leucine
m_Linolenic.acid
m_Lysine
m_Maleate
m_Nonanenitrile.9.methylthio
m_Octanenitrile.8.methylthio
m_Phosphate
m_ProLIne
m_Quercetin
m_Sucrose
m_Sulfuric.acid
m_Threonate
m_Threonine
m_Tryptophan
m_U1093.6.147
m_U1124.3.140
m_U1129.9.184
m_U1559.4.217
m_U1759.4.331
m_U1952.1.217
m_U2012.7.361
m_U2197.2.494
m_U2251.5.361
m_U2371.1.361
m_U2446.9.361
m_U2501.4.321
m_U2609.4.361
m_U2688.5.333
m_U2756.4.271
m_U3041.1.361
m_UGlucosinolate140.1
m_ValIne

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. 2019. "Introduction to MultiVarSel," March.

Annexes