

# 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 chez *Arabidopsis Thaliana* à l'aide de données de métabolomique pour des graines fraîchement récoltées (***F*reshly *H*arvested**).

# Les paramètres considérés

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 7 premières colonnes (7 sur 237) du jeu de données<sup>1</sup> :

temperature	imbibition	m_Alanine	m_Arginine	m_Asparagine	m_Aspartate	m_Cystein
Low	DS	0.7670471	0.0293251	0.4197357	0.1473605	0.0736553
Low	DS	0.7360741	0.0349146	0.4447546	0.1494870	0.0666249
Low	DS	0.8128032	0.0464299	0.4347309	0.1403788	0.1273317
Medium	DS	0.4299879	0.0210281	0.5830220	0.2817400	0.0543873
Medium	DS	0.5130800	0.0119001	0.5458675	0.3087902	0.0331085
Medium	DS	0.4696609	0.0197695	0.5696875	0.3278582	0.0288687

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

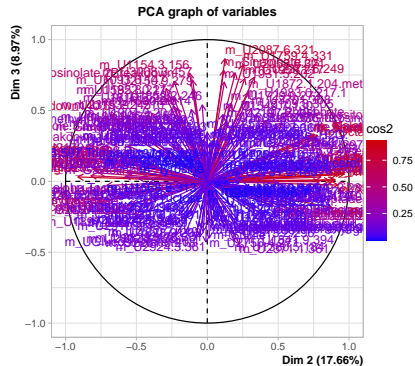
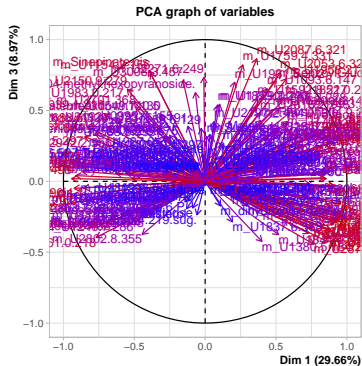
# Statistiques descriptives et classification

# Statistiques descriptives et classification

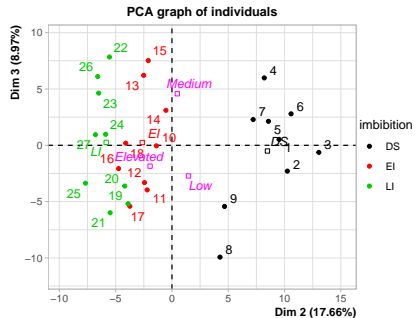
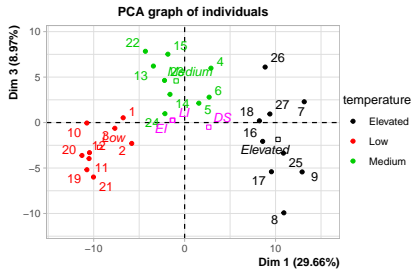
## Vérifications élémentaires

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

## ACP







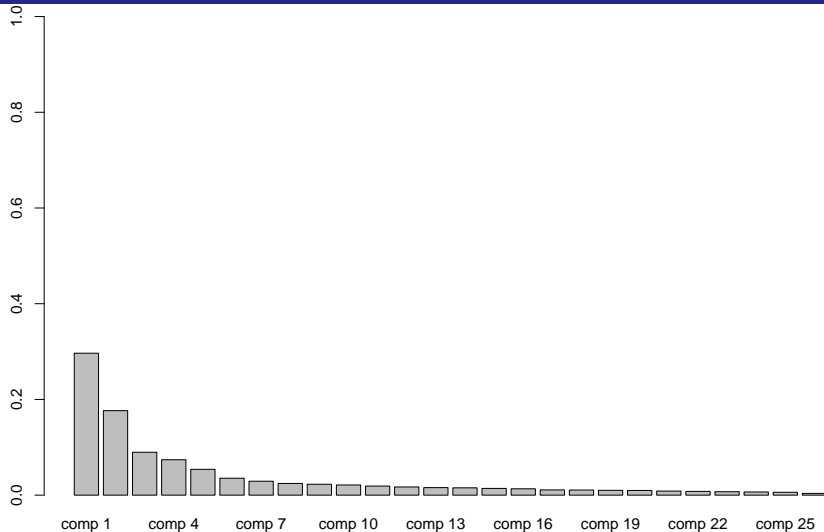
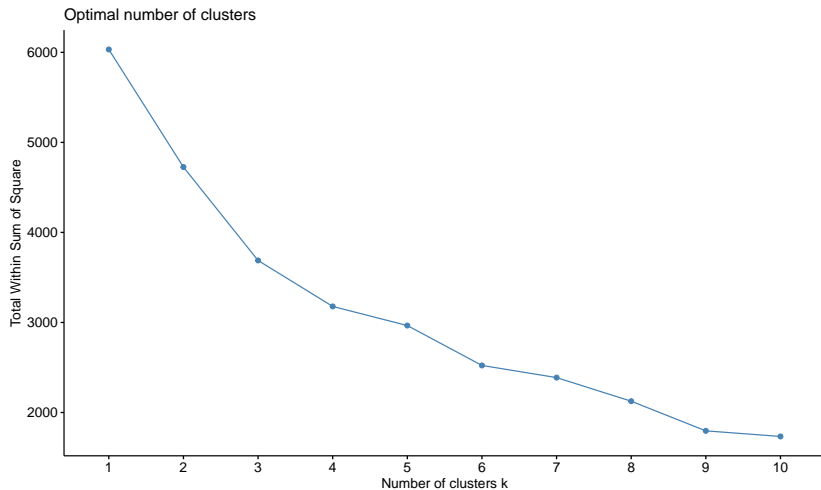
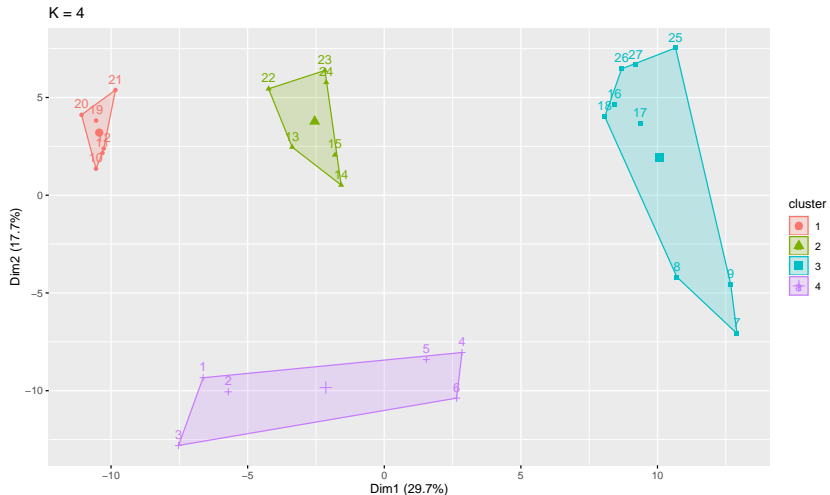


Figure 1: Pourcentage de variance expliqué par les composantes

# Clustering (*K-means*)

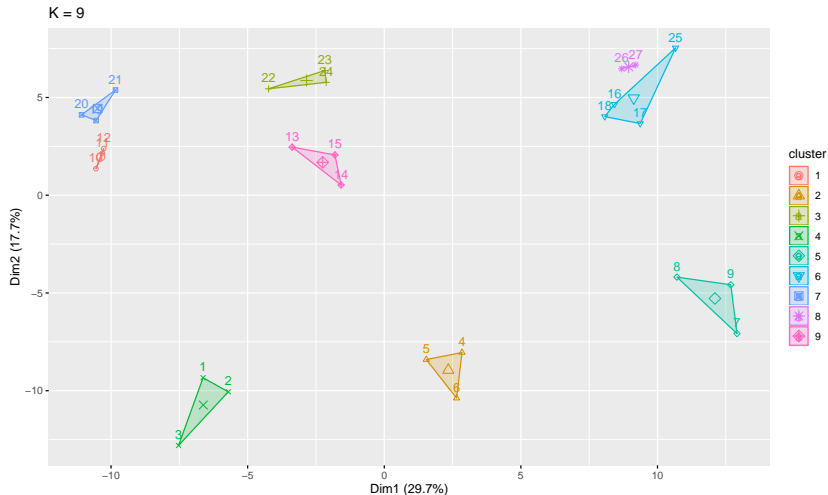


$K = 4$ 

Nous avons tenté de faire des moyennes par groupe sans voir de tendance se dégager clairement.

km\$cluster	temperature	imbibition
4	Low	DS
4	Low	DS
4	Low	DS
4	Medium	DS
4	Medium	DS
4	Medium	DS
3	Elevated	DS
3	Elevated	DS
3	Elevated	DS
1	Low	EI
1	Low	EI
1	Low	EI
2	Medium	EI

km\$cluster	temperature	imbibition
14	2	Medium
15	2	Medium
16	3	Elevated
17	3	Elevated
18	3	Elevated
19	1	Low
20	1	Low
21	1	Low
22	2	Medium
23	2	Medium
24	2	Medium
25	3	Elevated
26	3	Elevated
27	3	Elevated

$K = 9$ 

km9\$cluster	temperature	imbibition
4	Low	DS
4	Low	DS
4	Low	DS
2	Medium	DS
2	Medium	DS
2	Medium	DS
5	Elevated	DS
5	Elevated	DS
5	Elevated	DS
1	Low	EI
1	Low	EI
1	Low	EI
9	Medium	EI

km9\$cluster	temperature	imbibition
14	9	Medium
15	9	Medium
16	6	Elevated
17	6	Elevated
18	6	Elevated
19	7	Low
20	7	Low
21	7	Low
22	3	Medium
23	3	Medium
24	3	Medium
25	6	Elevated
26	8	Elevated
27	8	Elevated

# Méthode



# Principe du modèle

Nous allons poser le modèle suivant (détaillé sur les slides suivantes):

$$\begin{array}{c} \text{matrice des observations} \\ \underbrace{\mathbf{Y}} \end{array} = \begin{array}{c} \underbrace{\mathbf{X}} \\ \text{matrice de design} \end{array} \mathbf{B} + \begin{array}{c} \underbrace{\mathbf{E}} \\ \text{erreur résiduelle} \end{array}$$

Nous supposons que :

$$\forall i \in \{1, \dots, n\}, (E_{i,1}, \dots, E_{i,q}) \sim \mathcal{N}(0, \Sigma)$$

et que  $\forall (i, k) \in \{1, \dots, n\}^2 | i \neq k, (E_{i,1}, \dots, E_{i,q}) \perp\!\!\!\perp (E_{k,1}, \dots, E_{k,q})$

Et nous cherchons à estimer  $\mathbf{B}$ , la matrice des coefficients, par  $\hat{\mathbf{B}}$  en faisant en sorte que l'estimateur soit parcimonieux car  $\mathbf{B}$  est supposée contenir beaucoup de zéros.

# Modèle linéaire général, ANOVA interaction à 2 facteurs

$$\mathbf{Y} = \mathbf{XB} + \mathbf{E}$$

$$\mathbf{X} = \begin{pmatrix} \mathbf{1}_{Ele.:DS} & \cdots & \mathbf{1}_{Med.:LI} \end{pmatrix}_{p=9} \quad n=27$$

$$\mathbf{B} = \begin{pmatrix} B_{1,1} & \cdots & B_{1,q} \\ \vdots & \cdots & \vdots \\ B_{p,1} & \cdots & B_{p,q} \end{pmatrix}_{q=232} \quad p=9$$

Avec  $\mathbf{1}_{A:B,i} = 1$  ou  $0$  si la  $i^e$  observation n'est pas dans les 2 modalités A et B.

Les observations pour les différentes valeurs de métabolites sont alors mise en forme dans la matrice  $\mathbf{Y}$  :

$$\mathbf{Y} = \begin{pmatrix} Y_{1,1} & \cdots & Y_{1,q} \\ \vdots & \cdots & \vdots \\ Y_{n,1} & \cdots & Y_{n,q} \end{pmatrix}_{q=232} \quad n=27 \quad \mathbf{Y} \text{ a été centrée et réduite.}$$

# Calcul des résidus

Ici nous allons ajuster le modèle linéaire en faisant comme si les colonnes de  $\mathbf{Y}$  étaient indépendantes afin d'estimer par  $\hat{\mathbf{E}}$  la matrice  $\mathbf{E}$ , l'erreur résiduelle.

Puis nous allons tester avec le test de Portmanteau (grâce au théorème de Bartlett) si chaque ligne de  $\hat{\mathbf{E}}$  est un bruit blanc.

En calculant les résidus du modèle linéaire on obtient une *p-value* de 0.052 qui est à peine au-dessus du seuil 5%. 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  $\Sigma^{-1/2}$  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 estimateur  $\hat{\Sigma}^{-1/2}$ .

Pour cela le package R `MultiVarSel` (Perrot-Dockès, Lévy-Leduc, and Chiquet (2019)) permet d'utiliser 3 structures de dépendances et implémente les méthodes d'estimation de  $\hat{\Sigma}^{-1/2}$  pour chacun des cas :

- $AR(1)$
- $ARMA(p, q)$
- Non paramétrique<sup>2</sup>

---

<sup>2</sup>Suppose uniquement que le processus est stationnaire.

# Blanchiment des données

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

Table 1: 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

<sup>3</sup>Ce qui est empiriquement régulièrement le cas.

## Test de $\Sigma = Id$

# Principe de l'article de Fisher (2012)

Dans cet article, Fisher développe de nouvelles statistiques de test afin de vérifier si l'on peut rejeter ou non l'hypothèse

$$(H_0) : \Sigma = Id$$

pour les cas où  $(n, q) \rightarrow +\infty$ .

En utilisant les moyennes arithmétiques et leurs estimateurs  $(\hat{a}_i)^4$ , et  $c = q/n$  Fisher démontre que sous  $H_0$  et d'autres conditions:

$$T_1 = \frac{n}{c\sqrt{8}}(\hat{a}_4 - 4\hat{a}_3 + 6\hat{a}_2 - 4\hat{a}_1 + 1) \xrightarrow{D} \mathcal{N}(0, 1)$$

$$T_2 = \frac{n}{\sqrt{8(c^2 + 12c + 8)}}(\hat{a}_4 - 2\hat{a}_2 + 1) \xrightarrow{D} \mathcal{N}(0, 1)$$

---

<sup>4</sup>Voir l'article pour les formules



# Comportement des statistiques $T_1$ et $T_2$ sur données simulées

Nous avons testé en simulant plusieurs jeux de données avec différentes corrélations :

- Sous un AR(1) ( $\phi_1 = 0.5$ ), donc avec des corrélations entre colonnes.
- Les données d'AR(1) mais blanchies par  $\Sigma^{-1/2}$ , donc sans corrélations.
- Des vecteurs gaussiens, donc sans corrélations.

Le tout pour différentes valeurs de  $n$  répétitions indépendantes, avec des vecteurs de longueur  $q^5$  en utilisant le paramètre d'échelle  $c$  qui donne la relation  $q = cn$ .

---

<sup>5</sup>correspondant au  $p$  de l'article

# Résultats de simulations

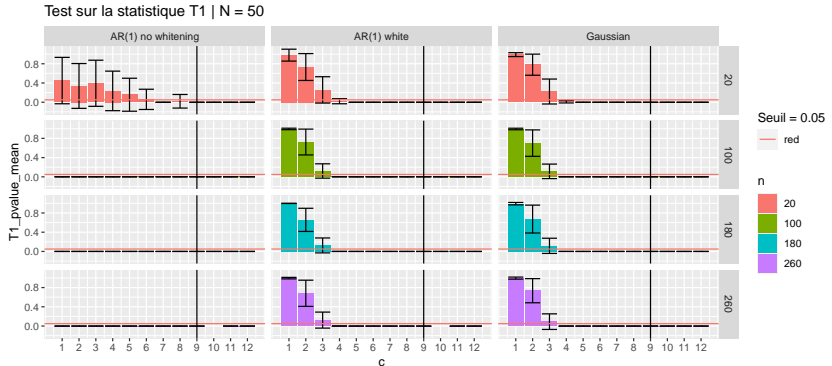


Figure 2:  $p$ -value pour le test basé sur la statistique  $T_1$

La ligne verticale, indique la valeur de  $c$  la plus proche de nos données.

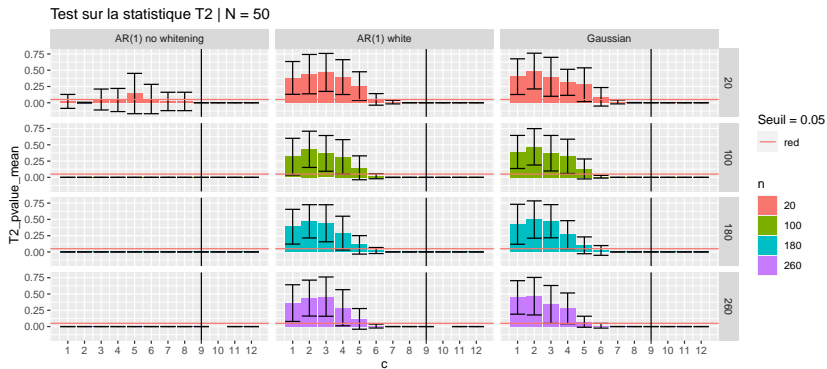


Figure 3:  $p$ -value pour le test basé sur la statistique  $T_2$

La ligne verticale, indique la valeur de  $c$  la plus proche de nos données.

# Notre cas

Nous avons  $n = 27$  et  $q = 232$  soit  $c \simeq 9$

En appliquant les tests à nos données nous avons :

- Pour les données non blanchies,  $p\text{-value}_{T_1}^{\text{avant blanch.}} = 0$  et  $p\text{-value}_{T_2}^{\text{avant blanch.}} = 0$ , ainsi le test indique que nos données sont corrélées, en accord avec le test de Portmanteau (“no\_whitening”).
- Pour les données blanchies,  $p\text{-value}_{T_1}^{\text{après blanch.}} = 0$  et  $p\text{-value}_{T_2}^{\text{après blanch.}} = 0$ , ainsi le test indique que nos données sont corrélées, en opposition au test de Portmanteau et au processus de blanchiment.

Aux vues des simulations, notre  $c$  est sûrement trop grand pour ces tests.

# Résultats

# Sélection de variable

Le critère LASSO consiste à résoudre le problème d'optimisation suivant :

$$\tilde{\mathcal{B}}(\lambda) = \arg \min_{\mathcal{B}} \{ \|\tilde{\mathcal{Y}} - \tilde{\mathcal{X}}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \}$$

## Choix du $\lambda$

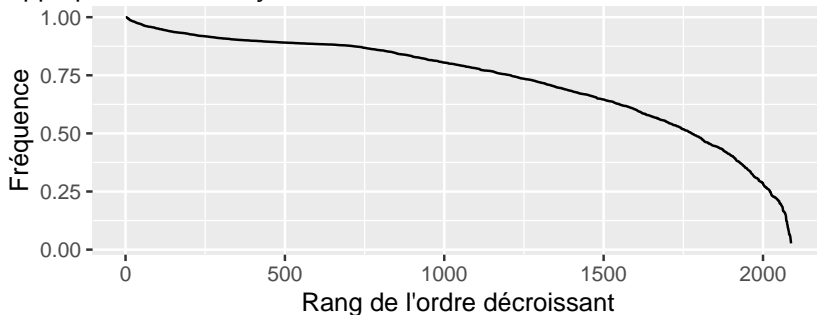
Pour mener à bien la procédure il faut choisir un  $\lambda$ . Le choix fait dans `MultiVarSel` consiste à réaliser une validation croisée et à choisir le  $\lambda_{CV}$ .

## La *stability selection*

La méthode (proposée par Meinshausen and Bühlmann (2010)) consiste à échantillonner  $nq/2$  indices de  $\mathcal{Y}$  et à résoudre le problème d'optimisation un grand nombre de fois en relevant les indices des coefficients non nuls de  $\mathcal{B}$ .

Une fois cela fait, on obtient une fréquence de sélection pour chacun des coefficients.

Voici un graphique des fréquences obtenues par ordre décroissant en appliquant la *stability selection*<sup>6</sup>



<sup>6</sup>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 0.96.

# 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.

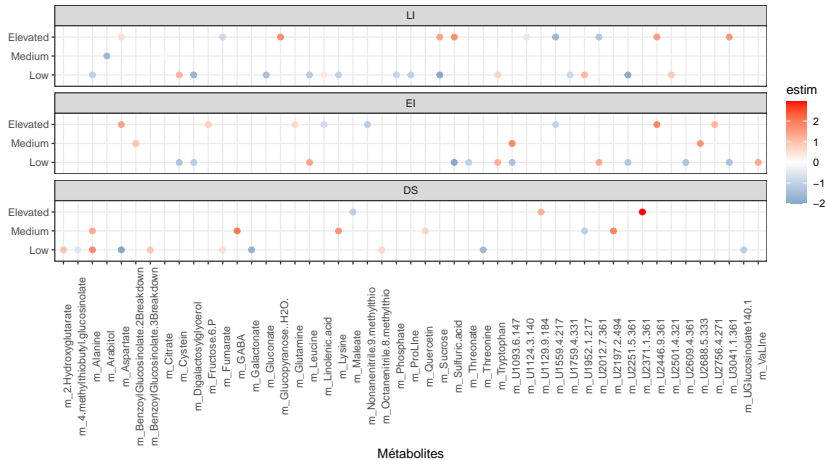


Figure 4: Graphique en boulier des estimations pour les métabolites sélectionnés en fonction de la température et de l'imbibition

## Bibliographie

# Bibliographie I

Fisher, Thomas J. 2012. "On Testing for an Identity Covariance Matrix When the Dimensionality Equals or Exceeds the Sample Size." *Journal of Statistical Planning and Inference* 142 (1): 312–26. <https://doi.org/10.1016/j.jspi.2011.07.019>.

Lévy-Leduc, Céline. 2024. "Notes pour le cours : 'Méthodes de statistique en grande dimension pour l'analyse de données de biologie moléculaire'."

Meinshausen, Nicolai, and Peter Bühlmann. 2010. "Stability Selection." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72 (4): 417–73. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>.

# Bibliographie II

Perrot-Dockès, Marie, Céline Lévy-Leduc, and Julien Chiquet. 2019.  
“Introduction to MultiVarSel,” March.

Merci pour votre attention