Sélection de variables dans le modèle linéaire général en grande dimension : application à des approches "multi-omiques" pour l'étude de la qualité des graines

Marie Perrot-Dockès, Céline Lévy-Leduc, Gwendal Cueff, Loïc Rajjou

SÉLECTION DE VARIABLES DANS LE MODÈLE LINÉAIRE GÉNÉRAL EN GRANDE DIMENSION : APPLICATION À DES APPROCHES "MULTI-OMIQUES" POUR L'ÉTUDE DE LA QUALITÉ DES GRAINES

M. PERROT-DOCKÈS, C. LÉVY-LEDUC, G. CUEFF, AND L. RAJJOU

Résumé. Les données "-omiques" sont caractérisées par la présence d'une forte structure de dépendance qui résulte souvent de processus biologiques. Appliquer des méthodes de sélection de variables qui ne tiennent pas compte de cette structure de dépendance sous-jacente peut conduire à la sélection de variables non pertinentes. Le but de ce chapitre est de proposer une nouvelle méthode de sélection de variables dans le modèle linéaire général qui tienne compte de la dépendance pouvant exister entre les colonnes de la matrice d'observations correspondant ici aux variables. Nous montrons que le fait d'inclure l'estimation de la matrice de covariance des observations dans le critère LASSO améliore nettement les performances de la sélection de variables. Nous avons implémenté notre méthode dans le package R `MultiVarSel` qui est disponible sur le CRAN (Comprehensive R Archive Network) et nous l'avons appliquée à des données "-omiques" pour étudier la qualité des graines.

1. Introduction

La reproduction sexuée des plantes à fleurs conduit à la formation de graines qui sont le vecteur principal de dispersion et de propagation des espèces végétales. Les graines sont des organes de réserve, riches en protéines, en lipides ou en carbohydrates qui sont nécessaires pour la germination et l'émergence des plantules. Elles sont très largement exploitées par l'homme pour des utilisations alimentaires ou non. Les conditions environnementales auxquelles sont soumises les plantes mères impactent fortement les rendements et la qualité des graines produites. Cette qualité s'élabore au champ, s'affine à la récolte (tri, traitement), se maintient au stockage et s'exprime au moment du semis. La qualité peut se traduire par différents caractères tels que le taux de remplissage, la morphologie, la capacité de stockage ou encore le potentiel de germination (Blödner et al. (2007), Burghardt et al. (2016)). Ce dernier point est crucial pour les agriculteurs. Un lot de graines de bonne qualité en production agricole se doit de présenter une bonne vigueur germinative, c'est à dire des graines qui vont germer de manière rapide et homogène et par conséquent avec une très faible dormance. Cette germination homogène permet de synchroniser tout le cycle de développement de la culture, la levée au champ, la croissance des plantes, la floraison et les dates de récolte. Chez la plante modèle, Arabidopsis thaliana, la température de production des graines affecte leur potentiel de germination (Springthorpe and Penfield, 2015). Les températures basses au cours de la maturation des graines favorisent la dormance. Aussi, dans le travail réalisé ici, des graines d'Arabidopsis ont été produites à une température basse (14-16°C), à une température intermédiaire (18-22°C) et à une température élevé (25-28°C), offrant un matériel biologique de qualité variable. Diagnostiquer la qualité des graines est un véritable défi pour l'industrie.

Date: 2 octobre 2023.

Les analyses des produits d'expression des gènes (transcrits, protéines) et de la composition biochimique (métabolites) par des approches "omiques" à haut débit (transcriptomique, protéomique, métabolomique) offrent un fort potentiel pour caractériser des biomarqueurs de qualité des graines. Des expérimentations ont montré une forte discordance entre l'accumulation de transcrits et l'abondance de protéines correspondantes dans les graines (Galland et al., 2014). Ceci est corroboré par des observations antérieures indiquant que l'abondance du transcrit ne reflète pas nécessairement sa traduction en protéine, en particulier dans des conditions de stress (Bailey-Serres et al., 2009). Nous avons focalisé ici notre recherche de biomarqueurs sur des analyses de protéomique et de métabolomique des graines matures sèches fraîchement récoltées.

Afin de comprendre l'effet de la température sur la réponse d'un métabolite ou d'une protéine donné observé sur n échantillons indépendants, on peut utiliser le modèle d'analyse de la variance à un facteur (ANOVA à un facteur) suivant :

$$\begin{pmatrix} Y_{1,1} \\ Y_{2,1} \\ \vdots \\ Y_{n,1} \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} B_{1,1} \\ B_{2,1} \\ B_{3,1} \end{pmatrix} + \begin{pmatrix} E_{1,1} \\ E_{2,1} \\ \vdots \\ E_{n,1} \end{pmatrix}$$

où les $E_{i,1}$ sont supposées être des variables aléatoires gaussiennes centrées et indépendantes dans le but d'estimer les coefficients $B_{j,1}$. Dans le cas où $\widehat{B}_{j,1} > 0$ (resp. $\widehat{B}_{j,1} < 0$), cela signifiera que le métabolite 1 ou la protéine 1 a tendance à être sur-accumulé (resp. sous-accumulé) dans la modalité j de la variable température. Dans une expérience de métabolomique ou de protéomique, on a en général accès respectivement aux réponses de q métabolites ou protéines. On est donc amené dans une expérience de métabolomique par exemple à considérer le modèle de MANOVA (Multivariate ANOVA) suivant :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E} \tag{1}$$

où

$$\underset{n\times q}{\boldsymbol{Y}} = \begin{pmatrix} \text{Metabolite 1} & \text{Metabolite 2} & \cdots & \text{Metabolite q} \\ Y_{1,1} & Y_{1,2} & \cdots & Y_{1,q} \\ Y_{2,1} & Y_{2,2} & \cdots & Y_{2,q} \\ \vdots & \vdots & & \vdots \\ Y_{n,1} & Y_{n,2} & \cdots & Y_{n,q} \end{pmatrix}, \underset{n\times p}{\boldsymbol{X}} = \begin{pmatrix} 1 & 0 & 0 \\ \vdots & \vdots & \vdots \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ \vdots & \vdots & \vdots \\ 0 & 0 & 1 \end{pmatrix},$$

$$\underset{p\times q}{\boldsymbol{B}} = \begin{pmatrix} B_{1,1} & B_{1,2} & \cdots & B_{1,q} \\ B_{2,1} & B_{2,2} & \cdots & B_{2,q} \\ B_{p,1} & B_{p,2} & \cdots & B_{p,q} \end{pmatrix} \text{ et } \underset{n\times q}{\boldsymbol{E}} = \begin{pmatrix} E_{1,1} & E_{1,2} & \cdots & E_{1,q} \\ E_{2,1} & E_{2,2} & \cdots & E_{2,q} \\ \vdots & \vdots & & \vdots \\ E_{n,1} & E_{n,2} & \cdots & E_{n,q} \end{pmatrix}.$$

Pour une présentation plus détaillée du modèle MANOVA et de son inférence, nous renvoyons le lecteur à Mardia et al. (1979) et Muller and Stewart (2006). Ce modèle fait partie des modèles linéaires généraux qui ne doivent pas être confondus avec les modèles linéaires généralisés. Dans ce cas, $(Y_{1,k}, \ldots, Y_{n,k})'$ représente la réponse du métabolite k sur n échantillons et en estimant les $B_{j,k}$, on pourra savoir si la modalité j de la variable température a un effet positif, négatif ou nul sur le métabolite k.

Différentes approches d'apprentissage statistique ont été proposées pour analyser des données "omiques", elles sont décrites dans Saccenti et al. (2013); Ren et al. (2015); Boccard and Rudaz (2016); Zhang et al. (2017). Parmi elles, on peut notamment citer la PLS-DA et la sPLS-DA proposées dans Lê Cao et al. (2011); Durif et al. (2017); el Bouhaddani et al. (2018).

En supposant que chaque colonne de la matrice d'observations $\boldsymbol{Y}$ a une moyenne empirique nulle, la question de savoir si une modalité de la variable température a une influence nulle ou pas sur un métabolite ou une protéine revient à un problème de sélection de variables dans le modèle linéaire général (1). Plusieurs approches peuvent être utilisées pour faire de la sélection de variables dans ce type de modèle. On peut utiliser des tests statistiques classiques dans des modèles d'ANOVA univariés tels que ceux décrits dans Mardia et al. (1979) ou Faraway (2004) pour analyser séparément chaque colonne de $\boldsymbol{Y}$. On peut également utiliser sur chaque colonne de $\boldsymbol{Y}$ des approches de type Lasso telles que celle décrites dans Tibshirani (1996). Cependant, ces méthodes ne tiennent pas compte de la dépendance potentielle qui peut exister entre les colonnes de $\boldsymbol{Y}$.

Dans ce chapitre, nous proposons une méthode qui modélise la dépendance pouvant exister entre les colonnes de $\boldsymbol{Y}$ et qui l'utilise dans la sélection de variables. Plus précisément, nous supposerons que les lignes de $\boldsymbol{E}$ sont indépendantes et que pour chaque ligne i, le vecteur $\boldsymbol{E}_i$ est un vecteur gaussien d'espérance nulle et de matrice de covariance $\boldsymbol{\Sigma}_q$:

$$\boldsymbol{E}_i = (E_{i,1}, \ldots, E_{i,q}) \sim \mathcal{N}(0, \boldsymbol{\Sigma}_q). \tag{2}$$

Estimer convenablement $\boldsymbol{\Sigma}_q$ est en général impossible dans le cas où $n << q$ sans hypothèse supplémentaire. Nous supposerons ici que chaque $\boldsymbol{E}_i$ peut être modélisé comme une réalisation d'un processus stationnaire et nous montrerons comment estimer $\boldsymbol{\Sigma}_q$ dans ce cadre. Il est par ailleurs à noter que notre méthode retire la dépendance potentielle entre les colonnes et utilise ensuite une approche Lasso associée à une étape de "stability selection" proposée par Meinshausen and Buhlmann (2010) pour être sûr de ne garder que les variables les plus stables. Notre méthode est implémentée dans le package R `MultiVarSel` qui est disponible sur le CRAN (Comprehensive R Archive Network).

Le reste du chapitre est organisé comme suit. Notre méthodologie est décrite dans la partie 2. Nous proposons de la valider à partir de simulations numériques dans la partie 3 et nous l'appliquons à des données de métabolomique et de protéomique dans la partie 4.

2. Méthodologie

La méthode que nous proposons peut être résumée comme suit.

— 1ère étape : En supposant que chaque colonne de la matrice $\boldsymbol{Y}$ suit un modèle d'ANOVA à un facteur nous obtenons une estimation $\widehat{\boldsymbol{E}}$ de la matrice d'erreurs $\boldsymbol{E}$.

— 2ème étape : On estime la matrice $\mathbf{\Sigma}_q$ en utilisant les méthodes décrites dans les paragraphes 2.1.1 et 2.1.2. Ensuite, on choisit l'estimateur $\widehat{\mathbf{\Sigma}}_q$ de $\mathbf{\Sigma}_q$ le plus adapté grâce au test statistique décrit dans le paragraphe 2.1.3.
— 3ème étape : A l'aide de $\widehat{\mathbf{\Sigma}}_q$, nous transformons les données afin de retirer la dépendance présente entre les colonnes de la matrice $\boldsymbol{Y}$.
— 4ème étape : On applique aux données transformées la méthode Lasso décrite au paragraphe 2.2.

La première étape fournit un estimateur préliminaire $\widetilde{\boldsymbol{B}}$ de $\boldsymbol{B}$. On définit alors un estimateur $\widehat{\boldsymbol{E}}$ de $\boldsymbol{E}$ par :

$$\widehat{\boldsymbol{E}} = \boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{B}}. \tag{3}$$

Dans la suite, nous nous focaliserons sur la description des trois autres étapes.

2.1. **Estimation de la matrice de covariance $\mathbf{\Sigma}_q$.** Dans la suite, afin d'estimer $\mathbf{\Sigma}_q$, nous proposons de modéliser chaque ligne de $\boldsymbol{E}$ comme la réalisation d'un processus stationnaire, nous utiliserons donc plusieurs types de processus stationnaires tels que les processus ARMA par exemple. Pour plus de détails sur ce sujet, le lecteur pourra consulter le livre de Brockwell and Davis (1991).

Nous considérerons dans ce paragraphe différents modèles de processus stationnaires et dans chaque cas nous porterons une attention particulière à l'estimation de $\mathbf{\Sigma}_q^{-1/2}$ puisque nous utiliserons la transformation suivante :

$$\boldsymbol{Y}\ \mathbf{\Sigma}_q^{-1/2} = \boldsymbol{X}\boldsymbol{B}\ \mathbf{\Sigma}_q^{-1/2} + \boldsymbol{E}\ \mathbf{\Sigma}_q^{-1/2} \tag{4}$$

afin de retirer la dépendance existant entre les colonnes de $\boldsymbol{Y}$. En effet, la matrice de corrélation de chaque ligne de $\boldsymbol{E}\mathbf{\Sigma}_q^{-1/2}$ est égale à la matrice identité. Une telle procédure sera appelée "blanchiment" dans la suite du chapitre.

2.1.1. *Processus ARMA.* L'un des processus ARMA les plus simples est le processus autorégressif d'ordre 1 (AR(1)). Plus précisément, cela revient à supposer que pour chaque i dans $\{1, \dots, n\}$, $E_{i,t}$ vérifie l'équation suivante :

$$E_{i,t} - \phi_1 E_{i,t-1} = W_{i,t}, \forall t \in \mathbb{Z}, \tag{5}$$

où $|\phi_1| < 1$ et $W_{i,t} \sim BB(0, \sigma^2)$. La notation $BB(0, \sigma^2)$ désigne un bruit blanc d'espérance nulle et de σ^2, défini comme suit :

$$Z_t \ \sim\ BB(0,\sigma^2) \text{ si } \begin{cases} \mathbb{E}(Z_t) = 0, \\ \mathbb{E}(Z_t Z_{t'}) = 0 \text{ si } t \neq t', \\ \mathbb{E}(Z_t^2) = \sigma^2. \end{cases} \tag{6}$$

Il est à noter que plus le paramètre ϕ_1 est proche de 1 plus la dépendance entre les $E_{i,t}$ est importante à i fixé.

Dans le cas où $\sigma^2 = 1$, l'inverse de la racine de $\mathbf{\Sigma}_q$ a la forme explicite simple suivante :

$$\mathbf{\Sigma}_q^{-1/2} = \begin{pmatrix} \sqrt{1-\phi_1^2} & -\phi_1 & 0 & \cdots & 0 \\ 0 & 1 & -\phi_1 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\phi_1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}. \tag{7}$$

Il est à noter que dans le cas où σ^2 est différent de 1, la matrice de covariance de chaque ligne de $\boldsymbol{E}\boldsymbol{\Sigma}_q^{-1/2}$ est égale à $\sigma^2\mathrm{Id}$ et que la matrice de corrélation est égale à la matrice identité.

Ainsi, pour obtenir un estimateur $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$ de $\boldsymbol{\Sigma}_q^{-1/2}$, il suffit de savoir estimer le paramètre ϕ_1 et de le remplacer par son estimateur $\widehat{\phi}_1$ dans (7). Pour cela nous utilisons $\widehat{\boldsymbol{E}}$ défini par (3) et nous définissons $\widehat{\phi}_1$ par :

$$\widehat{\phi}_1 = \frac{1}{n}\sum_{i=1}^{n} \widehat{\phi}_{1,i},$$

où $\widehat{\phi}_{1,i}$ désigne l'estimateur de ϕ_1 obtenu par les équations de Yule-Walker à partir de $(\widehat{E}_{i,1}, \ldots, \widehat{E}_{i,q})$, cf. Brockwell and Davis (1991) pour plus de détails sur cette méthode.

Plus généralement, il est aussi possible d'avoir accès à $\boldsymbol{\Sigma}_q^{-1/2}$ pour des processus ARMA(p,q) définis comme suit : pour chaque i dans $\{1,\ldots,n\}$,

$$E_{i,t} - \phi_1 E_{i,t-1} - \cdots - \phi_p E_{i,t-p} = W_{i,t} + \theta_1 W_{i,t-1} + \ldots \theta_q W_{i,t-q}, \tag{8}$$

où $W_{i,t} \sim BB(0,\sigma^2)$, les ϕ_k et les θ_k sont des paramètres réels.

2.1.2. *Processus stationnaires généraux.* Dans le cas où la modélisation par un processus ARMA(p,q) n'est pas appropriée, on peut modéliser chaque ligne de $\boldsymbol{E}$ comme un processus faiblement stationnaire général et estimer $\boldsymbol{\Sigma}_q$ comme suit :

$$\widehat{\boldsymbol{\Sigma}}_q = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \cdots & \widehat{\gamma}(q-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \cdots & \widehat{\gamma}(q-2) \\ \vdots & & & \\ \widehat{\gamma}(q-1) & \widehat{\gamma}(q-2) & \cdots & \widehat{\gamma}(0) \end{pmatrix}, \tag{9}$$

où

$$\widehat{\gamma}(h) = \frac{1}{n}\sum_{i=1}^{n} \widehat{\gamma}_i(h),$$

et $\widehat{\gamma}_i(h)$ est l'estimateur classique de $\gamma_i(h) = \mathbb{E}(E_{i,t}E_{i,t+h})$ pour tout t. Plus précisément, $\widehat{\gamma}_i(h)$ est l'autocovariance empirique de $\widehat{E}_{i,t}$ en h *i.e.* la covariance empirique entre $(\widehat{E}_{i,1}, \ldots, \widehat{E}_{i,n-h})$ et $(\widehat{E}_{i,h+1}, \ldots, \widehat{E}_{i,n})$). Pour plus de détails, voir le chapitre 7 de Brockwell and Davis (1991). La matrice $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$ est alors obtenue en inversant le facteur de Cholesky de $\widehat{\boldsymbol{\Sigma}}_q$.

2.1.3. *Choix de la meilleure structure de dépendance.* Dans le but de savoir quelle modélisation de la dépendance est la plus adaptée nous proposons d'utiliser le test statistique défini ci-dessous. Si la structure de dépendance est bien choisie chaque ligne de $\widetilde{\boldsymbol{E}} = \widehat{\boldsymbol{E}}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$ doit être un bruit blanc défini dans (6), où $\widehat{\boldsymbol{E}}$ est donné par (3).

Pour tester si un processus aléatoire est un bruit blanc une des méthodes les plus classiquement utilisées est le test de Portmanteau qui est fondé sur le théorème de Bartlett (Théorème 7.2.2 dans Brockwell and Davis (1991)). D'après ce théorème, nous avons que sous l'hypothèse nulle (H_0) : "Pour chaque i dans $\{1,\ldots,n\}$, $(\widetilde{E}_{i,1}, \ldots, \widetilde{E}_{i,q})$ est un bruit blanc",

$$q\sum_{h=1}^{H} \widehat{\rho}_i(h)^2 \approx \chi^2(H), \text{ lorsque } q \to \infty, \tag{10}$$

pour chaque i dans $\{1,\ldots,n\}$, où $\widehat{\rho}_i(h)$ désigne l'auto-corrélation empirique de $(\widetilde{E}_{i,1}, \ldots, \widetilde{E}_{i,q})$ en h et $\chi^2(H)$ désigne la loi du khi-deux à H degrés de liberté. Ainsi d'après (10), nous

disposons d'une p-valeur pour chaque i dans $\{1, \ldots, n\}$. Dans le but d'avoir une seule p-valeur au lieu de n, nous considérerons l'approximation suivante

$$q \sum_{i=1}^{n} \sum_{h=1}^{H} \widehat{\rho}_i(h)^2 \approx \chi^2(nH), \text{ lorsque } q \to \infty, \tag{11}$$

où l'approximation vient du fait que les lignes $\widetilde{\boldsymbol{E}}$ sont supposées être indépendantes. L'équation (11) fournit ainsi une p-valeur : Pval. On en conclut donc que si Pval $\leq \alpha$, on rejette (H_0) au niveau α, où α est en général égal à 5%. Si au contraire Pval $> \alpha$, cela signifie que la structure de dépendance de $\boldsymbol{E}$ est bien choisie. Si parmi les différentes structures de dépendance testées aucune n'est bien choisie, on pourra tester une structure de dépendance par blocs telle que celle proposée dans Perrot-Dockès et al. (2019).

2.2. **Estimation de $\boldsymbol{B}$.**

2.2.1. *Approche Lasso.* Nous rappelons dans un premier temps le contexte dans lequel l'approche Lasso est habituellement utilisée. Considérons le modèle linéaire univarié suivant :

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{12}$$

où $\mathcal{Y}$, $\mathcal{B}$ et $\mathcal{E}$ sont des vecteurs. En général, dans les modèles linéaires en grande dimension, la matrice $\mathcal{X}$ a plus de colonnes que de lignes ce qui signifie que le nombre de variables est plus grand que le nombre d'observations et $\mathcal{B}$ est en général un vecteur parcimonieux ce qui signifie qu'il a beaucoup de composantes nulles.

Dans de tels modèles, la méthode LASSO (Least Absolute Shrinkage and Selection Operator) initialement proposée par Tibshirani (1996) est très utilisée. Elle est définie comme suit pour $\lambda > 0$:

$$\widehat{\mathcal{B}}(\lambda) = \text{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}, \tag{13}$$

où, pour $u = (u_1, \ldots, u_n)$, $\|u\|_2^2 = \sum_{i=1}^n u_i^2$ et $\|u\|_1 = \sum_{i=1}^n |u_i|$ qui correspond à la norme ℓ_1 du vecteur u. Dans (13), le premier terme correspond au critère des moindres carrés et $\lambda\|\mathcal{B}\|_1$ peut être vu comme une pénalité. L'intérêt de l'approche LASSO est de fournir un estimateur parcimonieux $\widehat{\mathcal{B}}$ de $\mathcal{B}$. Le nombre de composantes non nulles de $\widehat{\mathcal{B}}$ est d'autant plus petit que λ est grand.

Cette méthodologie ne peut pas être directement appliquée à notre modèle puisque les observations dont on dispose ne sont pas des vecteurs mais des matrices. Cependant, nous allons voir que le modèle (1) peut se réécrire comme (12) où $\mathcal{Y}$, $\mathcal{B}$ et $\mathcal{E}$ sont des vecteurs de tailles respectives nq, pq et nq. En effet, si on note $vec(\boldsymbol{A})$ le vecteur obtenu à partir de la matrice $\boldsymbol{A}$ en mettant les colonnes de $\boldsymbol{A}$ les unes en dessous des autres et que l'on applique l'opérateur vec au modèle (1) alors on obtient :

$$vec(\boldsymbol{Y}) = vec(\boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}) = vec(\boldsymbol{X}\boldsymbol{B}) + vec(\boldsymbol{E}).$$

Si on pose $\mathcal{Y} = vec(\boldsymbol{Y})$, $\mathcal{B} = vec(\boldsymbol{B})$ et $\mathcal{E} = vec(\boldsymbol{E})$, on obtient :

$$\mathcal{Y} = vec(\boldsymbol{X}\boldsymbol{B}) + \mathcal{E} = (\mathrm{I}_q \otimes \boldsymbol{X})\mathcal{B} + \mathcal{E},$$

où on a utilisé l'identité suivante :

$$vec(AXB) = (B' \otimes A)vec(X),$$

d'après l'annexe A.2.5 de Mardia et al. (1979). Dans cette équation, B' désigne la transposée de la matrice B. Ainsi,

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

où $\mathcal{X} = \mathrm{I}_q \otimes \boldsymbol{X}$ et $\mathcal{Y}$, $\mathcal{B}$ et $\mathcal{E}$ sont des vecteurs de tailles respectives nq, pq et nq.

Estimer les positions des composantes non nulles de $\mathcal{B}$ constitue donc une première approche pour sélectionner les variables les plus pertinentes. Cependant, cette méthode ne tient pas compte de la dépendance pouvant potentiellement exister entre les colonnes de $\boldsymbol{Y}$. Dans la suite, nous proposons une version modifiée de la méthode LASSO standard prenant en compte cette dépendance potentielle.

Comme nous l'avons expliqué précédemment, notre méthode consiste tout d'abord à "blanchir" les observations c'est-à-dire à supprimer la dépendance pouvant exister entre les colonnes de la matrice d'observations en multipliant l'équation (1) à droite par $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$, voir l'équation (4) où $\boldsymbol{\Sigma}_q^{-1/2}$ est remplacé par $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$. En utilisant la même astuce de vectorisation que celle utilisée ci-dessus pour transformer le modèle (1) en (12), le critère LASSO peut être appliqué à la version vectorisée du modèle (4) où $\boldsymbol{\Sigma}_q^{-1/2}$ est remplacée par $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$. On peut obtenir les expressions de $\mathcal{Y}$, $\mathcal{X}$, $\mathcal{B}$ et $\mathcal{E}$ comme suit.

En appliquant l'opérateur *vec* operator au modèle (4) dans lequel $\boldsymbol{\Sigma}_q^{-1/2}$ est remplacé par $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$, on obtient

$$vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) = vec(\boldsymbol{X}\boldsymbol{B}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}) = ((\widehat{\boldsymbol{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X})vec(\boldsymbol{B}) + vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2}).$$

Ainsi,

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{14}$$

où $\mathcal{Y} = vec(\boldsymbol{Y}\widehat{\boldsymbol{\Sigma}}_q^{-1/2})$, $\mathcal{X} = (\widehat{\boldsymbol{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X}$ et $\mathcal{E} = vec(\boldsymbol{E}\widehat{\boldsymbol{\Sigma}}_q^{-1/2})$.

Il est à noter que Rothman et al. (2010) ont également eu une idée de "blanchiment" similaire qu'ils ont implémentée dans le package R `MRCE`. Cependant, nous avons observé à l'aide de simulations numériques que pour les valeurs de n et q que nous souhaitions utiliser le temps de calcul de leur approche était tellement important pour une valeur donnée du paramètre λ qu'il était impossible d'utiliser leur méthode dans notre contexte. En conséquence, nous ne nous comparerons pas à cette approche.

2.2.2. *Choix du paramètre de régularisation.* L'estimateur défini dans (13) dépend du paramètre λ qui permet de calibrer le niveau de parcimonie de $\widehat{\mathcal{B}}$ c'est à dire la proportion de composantes nulles de cet estimateur. Pour régler le niveau de parcimonie de $\widehat{\mathcal{B}}$, nous proposons d'utiliser deux approches classiques : la validation croisée et la méthode de la "stability selection" qui garantit la robustesse des variables sélectionnées.

Plus précisément, nous commençons par appliquer notre méthode à $\mathcal{Y}$ défini dans (14) et nous obtenons λ_{CV}. Nous choisissons alors aléatoirement un sous-échantillon de taille $nq/2$ de $\mathcal{Y}$ et nous appliquons notre méthode avec $\lambda = \lambda_{\mathrm{CV}}$ et nous enregistrons les indices i des composantes non nulles de $\widehat{\mathcal{B}}(\lambda)$. Nous répétons ces étapes de sous-échantillonnage aléatoire et d'application du critère LASSO N fois. A l'issue de ces différentes étapes, nous avons accès au nombre de fois N_i où chaque composante $\widehat{\mathcal{B}}_i$ de $\widehat{\mathcal{B}}$ a été estimée comme étant non nulle. Nous conservons les composantes i dont la fréquence N_i/N est plus grande qu'un seuil donné. L'influence du choix de N et du seuil est étudié dans la partie 3.

Il est par ailleurs à noter que nous avons établi dans Perrot-Dockès et al. (2018) des résultats théoriques de consistance en signe de l'estimateur $\widehat{\mathcal{B}}$ pour valider notre approche.

3. Expériences numériques

Le but de cette partie est d'évaluer les performances statistiques et numériques de notre méthodologie que nous avons implémentée dans le package R `MultiVarSel` et de la comparer aux méthodes existantes. Pour cela nous avons généré des observations $\boldsymbol{Y}$ satisfaisant (1) avec $q = 1000$, $p = 3$, $n = 30$ ($n_1 = 9$, $n_2 = 8$, $n_3 = 13$, n_k correspondant au nombre de répétitions de la modalité k de la variable qualitative, c'est-à-dire au nombre de 1 dans la colonne k de la matrice $\boldsymbol{X}$) et différentes structures de dépendance c'est-à-dire différentes matrices $\boldsymbol{\Sigma}_q$ associées au modèle AR(1) décrit dans (5) avec $\sigma = 1$, $\phi_1 = 0.7$ ou 0.9. Il est à noter que les valeurs des paramètres p, q et n ont été choisies car elles correspondent à des ordres de grandeur usuels en métabolomique et en protéomique.

Dans cette partie nous étudierons également l'effet de la parcimonie et du rapport signal sur bruit (SNR) sur les performances statistiques des méthodes. Le niveau de parcimonie s correspond à la proportion d'éléments non nuls dans $\mathcal{B}$ et différents SNR sont obtenus en multipliant $\boldsymbol{B}$ défini dans (1) par un coefficient κ.

3.1. **Performances statistiques.**

Le but de cette partie est de comparer les performances de nos différentes méthodes de blanchiment aux méthodes existantes.

3.1.1. *Performances en termes de sélection de variables.* Nous comparons notre approche à la méthode d'ANOVA (notée ANOVA), à la méthode LASSO standard (notée Lasso) c'est-à-dire le critère LASSO sans étape de blanchiment et à la méthode sPLSDA qui est très utilisée dans le domaine de la métabolomique et décrite dans Lê Cao et al. (2011). Il est de plus à noter que cette méthode est implémentée dans le package R `mixOmics` et est disponible sur le site web MetaboAnalyst.

La méthode ANOVA consiste à modéliser chaque colonne de la matrice d'observations $\boldsymbol{Y}$ comme une ANOVA à un facteur en faisant comme si les colonnes de $\boldsymbol{Y}$ étaient indépendantes. Nos différentes méthodes de blanchiment décrites dans les paragraphes 2.1.1 et 2.1.2 sont notées AR1 et Nonparam. Elles sont comparées à la méthode Oracle où $\boldsymbol{\Sigma}_q$ est connue ce qui n'est jamais le cas en pratique.

Pour comparer ces différentes méthodes nous utilisons des courbes ROC qui représentent le taux de vrais positifs (TPR) en fonction du taux de faux positifs (FPR). Puisque les variables sélectionnées par la méthode sPLSDA ne sont pas affectées à une modalité donnée de la variable qualitative, nous considérerons qu'une variable est un vrai positif dès qu'elle est sélectionnée ce qui donne un avantage à sPLSDA.

Nous observons dans la figure 1 que dans le cas d'une dépendance de type AR(1) le fait de prendre en compte la dépendance produit de meilleurs résultats que les méthodes qui considèrent les colonnes de la matrice $\boldsymbol{E}$ comme indépendantes. Par ailleurs, plus le niveau de parcimonie est élevé plus les différences entre les méthodes sont faibles et plus le rapport signal sur bruit est élevé meilleures sont les performances des différentes approches.

3.1.2. *Performances en termes de choix du modèle sélectionné.* Nous étudions dans cette partie les performances de la méthode de "stability selection" décrite dans le paragraphe 2.2.2. La figure 2 représente le taux de vrais positifs (TPR) et de faux positifs (FPR) pour différentes valeurs de N et différents seuils. Nous observons que lorsque N est plus grand que 1000 et lorsque le seuil est égal à 0.95 le taux de faux positifs est très faible et le taux de vrais positifs est très élevé et ce quel que soit le scénario étudié.

Les ronds ('•') dans la figure 3 montrent les positions des variables sélectionnées par notre méthode pour deux valeurs de seuils : 0.95 et 1 et $N = 1000$ réplications. Les positions des

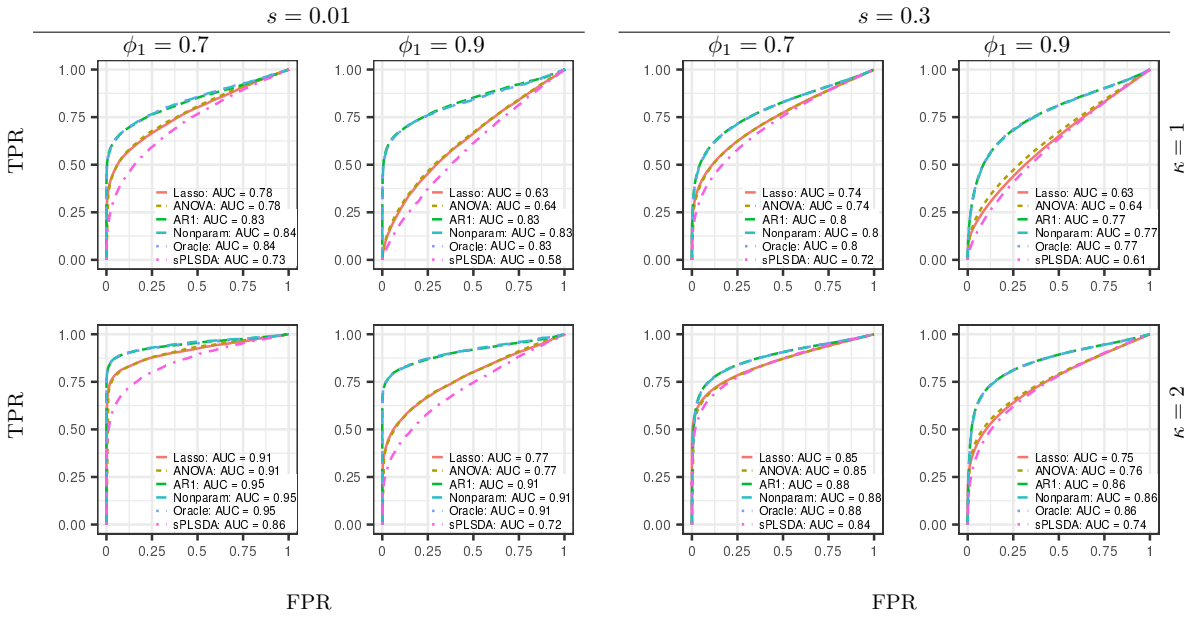

FIGURE 1. Moyennes empiriques des courbes ROC obtenues à partir de 200 simulations pour les différentes méthodologies pour un modèle AR(1). Sur la 1ère ligne, $\kappa = 1$, sur la seconde ligne : $\kappa = 2$, ϕ_1 est le paramètre de l'AR(1) et s est le degré de parcimonie.

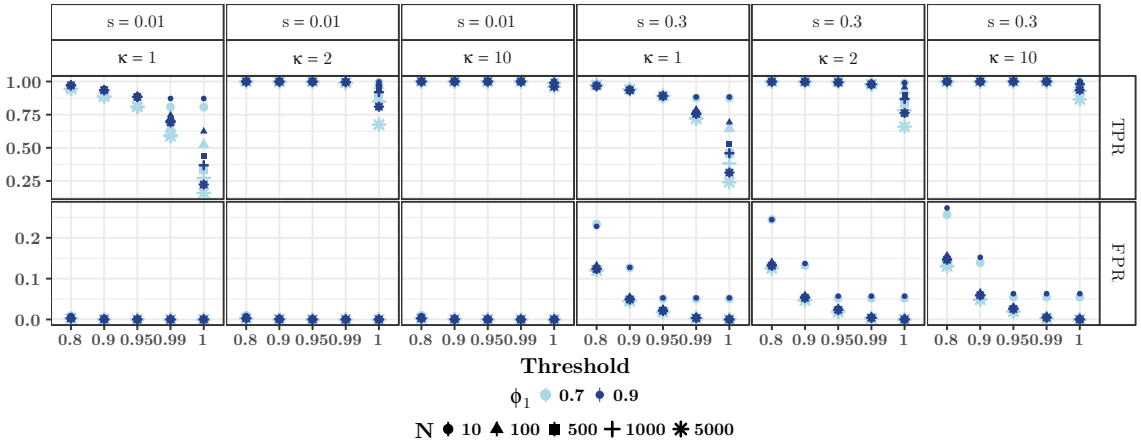

FIGURE 2. Influence du nombre N de réplications, du seuil de la "stability selection" et du paramètre ϕ_1 du processus AR(1).

coefficients non nuls de $\boldsymbol{B}$ sont représentées avec des croix ('+'). Dans ce paragraphe, les observations $\boldsymbol{Y}$ sont générées avec les paramètres donnés au début de la partie 3 dans le cas d'un processus AR(1) avec $\phi_1 = 0.9$ et $\kappa = 1$. Nous observons dans cette figure que les positions des coefficients non nuls sont retrouvées beaucoup plus fréquemment pour le seuil 0.95 que pour le seuil 1 et que les faux positifs restent rares même pour le seuil 0.95.

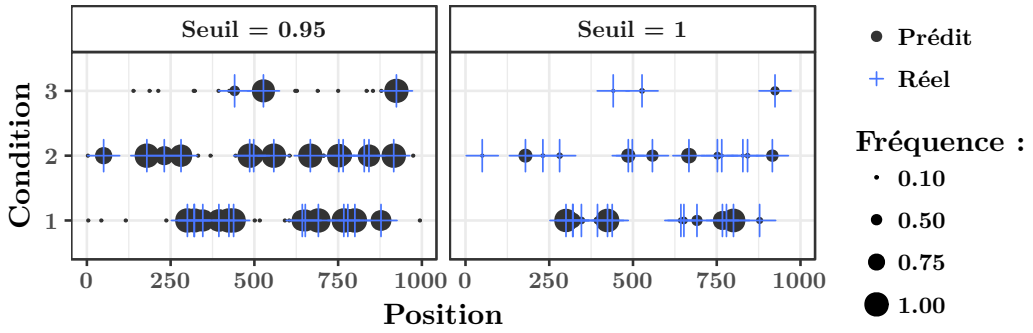

FIGURE 3. Positions des variables sélectionnées par notre approche ('•') lorsque $\kappa = 1$. Les valeurs sur l'axe des ordonnées correspondent aux trois conditions. Les résultats obtenus lorsque le seuil est égal à 0.95 sont dans la figure de gauche et ceux obtenus lorsque le seuil est égal à 1 sont dans la figure de droite. La taille des points est d'autant plus grande que la fréquence de sélection est élevée.

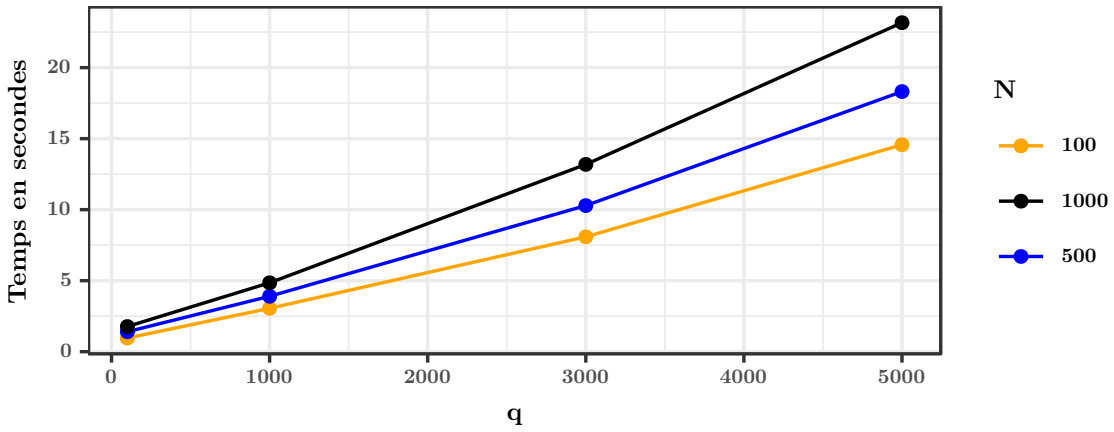

FIGURE 4. Temps d'exécution en secondes de `MultiVarSel` en fonction du nombre de colonnes q de la matrice d'observations $\mathbf{Y}$. Le nombre de réplications correspond au nombre N de sous-échantillonnages dans l'étape de la stability selection.

3.2. **Performances numériques.** Afin d'étudier la complexité algorithmique de notre méthode, nous avons généré des matrices $\boldsymbol{Y}$ satisfaisant le modèle (1) avec $n = 30$, un niveau de parcimonie de $\boldsymbol{B}$ égal à 0.01 et $q = 100, 1000, 3000, \ldots, 5000$. Nous avons de plus généré les lignes de $\boldsymbol{E}$ comme des réalisations d'un processus AR(1). La figure 4 représente les temps d'exécution de `MultiVarSel` pour différents nombres de réplications dans l'étape de "stability selection". Ces temps d'exécution ont été obtenus pour un ordinateur ayant 16 Go de RAM et 8 coeurs de type Intel Core i7 (3.66GHz). D'après cette figure, `MultiVarSel` prend seulement quelques secondes pour analyser une matrice ayant 5000 colonnes.

4. Application à l'étude de la qualité des graines

Dans cette partie, nous donnons les résultats obtenus en appliquant le package R `MultiVarSel` aux données "-omiques" afin de comprendre l'influence de la température de production des graines d'Arabidopsis sur leur composition en protéines et en métabolites. Les commandes R à utiliser pour étudier les données métabolomiques et protéomiques sont décrites respectivement dans les annexes A et B.

4.1. **Données de métabolomique.** Nous avons représenté dans la figure 5 les estimations des coefficients $B_{i,j}$ de la matrice $\mathbf{B}$ obtenues à l'aide de notre méthode avec un seuil égal à 0.93 et les boxplots des abondances des métabolites sélectionnés dans la figure 6. Il est à noter que ce seuil a été choisi afin d'une part de limiter le nombre de métabolites sélectionnés et d'autre part de garder les métabolites intéressants d'un point de vue biologique. Ainsi, 19 métabolites ont été sélectionnés comprenant 2 glucosinolates, le X6MTH (6-methylthiohexyl glucosinolate) et le X4MTB (4-methylthiobutyl glucosinolate) qui sont plus abondants dans les graines matures sèches lorsque la température de production est élevée. A l'inverse, 2 produits du catabolisme des glucosinolates suivent un profil d'accumulation opposé, en étant caractéristiques des graines produites à basse température. Les glucosinolates sont des métabolites spécialisés riches en soufre comportant une molécule de glucose et un groupe aglycone variable. Ils sont impliqués dans la protection des plantes face aux ravageurs et peuvent présenter des propriétés antifongiques et antioxydantes (Sønderby et al., 2010). Ainsi la température de production des graines modifie le métabolisme des glucosinolates chez Arabidopsis et par conséquent leur qualité biochimique et physiologique.

4.2. **Données de protéomique.** Nous avons représenté dans la figure 7 les estimations des coefficients $B_{i,j}$ de la matrice $\mathbf{B}$ obtenues à l'aide de notre méthode avec un seuil égal à 0.95 et les boxplots des abondances des protéines sélectionnées dans la figure 8. Les résultats ont permis de mettre en évidence sept protéines caractérisant les graines produites à basse température. Elles sont impliquées dans des fonctions biologiques et moléculaires assez différentes. Parmi elles, la protéine codée par le gène At1g07985. Ce gène n'a quasiment pas été étudié chez les plantes. Il code pour une petite protéine de 16.4 kDa. Une analyse de séquence via PROSITE (`https://prosite.expasy.org/prosite.html`) révèle la présence d'un signal bipartite (en deux parties) de localisation nucléaire (NLS_BP, PS50079) pour cette protéine. Il a été proposé que le gène At1g07985 serait sous le contrôle d'un promoteur bidirectionnel également impliqué dans la régulation du gène At1g07980 codant pour le facteur de transcription NF-YC10 (nuclear factor Y, subunit C10) (Kourmpetli et al., 2013). Cette région promotrice contient plusieurs éléments de type G-box impliqués dans le développement des graines et la réponse à l'acide abscissique (ABA) (Keddie et al., 1994). L'ABA est une phytohormone centrale dans le remplissage, la tolérance à la dessiccation et la mise en place de la dormance des graines.

Les graines produites à une température élevée sont également caractérisables par sept protéines. Deux d'entre-elles sont impliquées dans la machinerie traductionnelle, la protéine ribosomale L14p/L23e (At1g04480) et la threonyl-tRNA synthetase (At5g26830). Il a été démontré que la synthèse de nouvelles protéines est essentielle pour permettre la germination et le développement de la jeune plante. Des protéines spécifiques et indispensables à la germination sont issues de la traduction sélective et séquentielle d'ARNm stockés dans la graine et d'ARNm néotranscrits au cours de l'imbibition pour permettre le succès de la germination et la croissance de la future plantule. Au cours de l'imbibition, les gènes At1g04480 et At5g26830

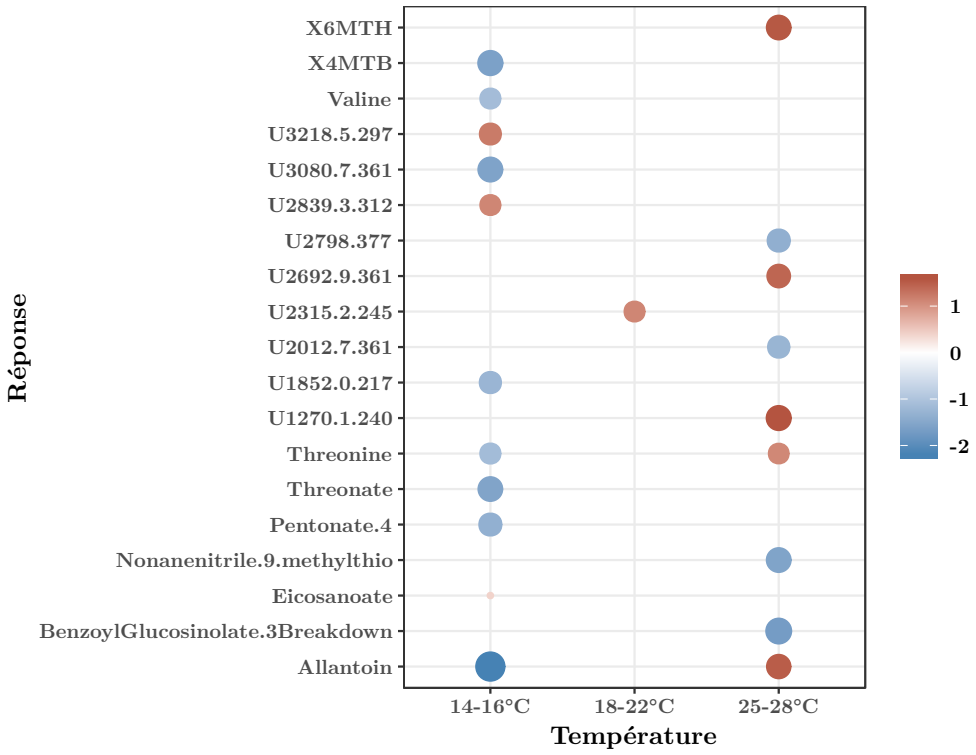

FIGURE 5. Estimation des coefficients $B_{i,j}$ pour les métabolites sélectionnés avec un seuil égal à 0.93. Il est à noter que la surface des ronds est d'autant plus grande que les coefficients sont élevés en valeur absolue.

sont positivement régulés lors du programme de germination stricto sensu. L'accumulation des protéines codées par ces gènes dans les graines matures sèches apparaît être un élément indicateur du potentiel de germination.

5. CONCLUSION

Dans ce chapitre, nous proposons une nouvelle méthode de sélection de variables dans le modèle linéaire général prenant en compte la dépendance pouvant exister entre les colonnes de la matrice d'observations. Notre approche est implémentée dans le package R `MultiVarSel` qui est disponible sur le CRAN (Comprehensive R Archive Network). Nous avons montré que notre méthode avait de très bonnes performances statistiques mais également numériques ce qui la rend tout à fait adaptée pour analyser des données "-omiques" de grande dimension.

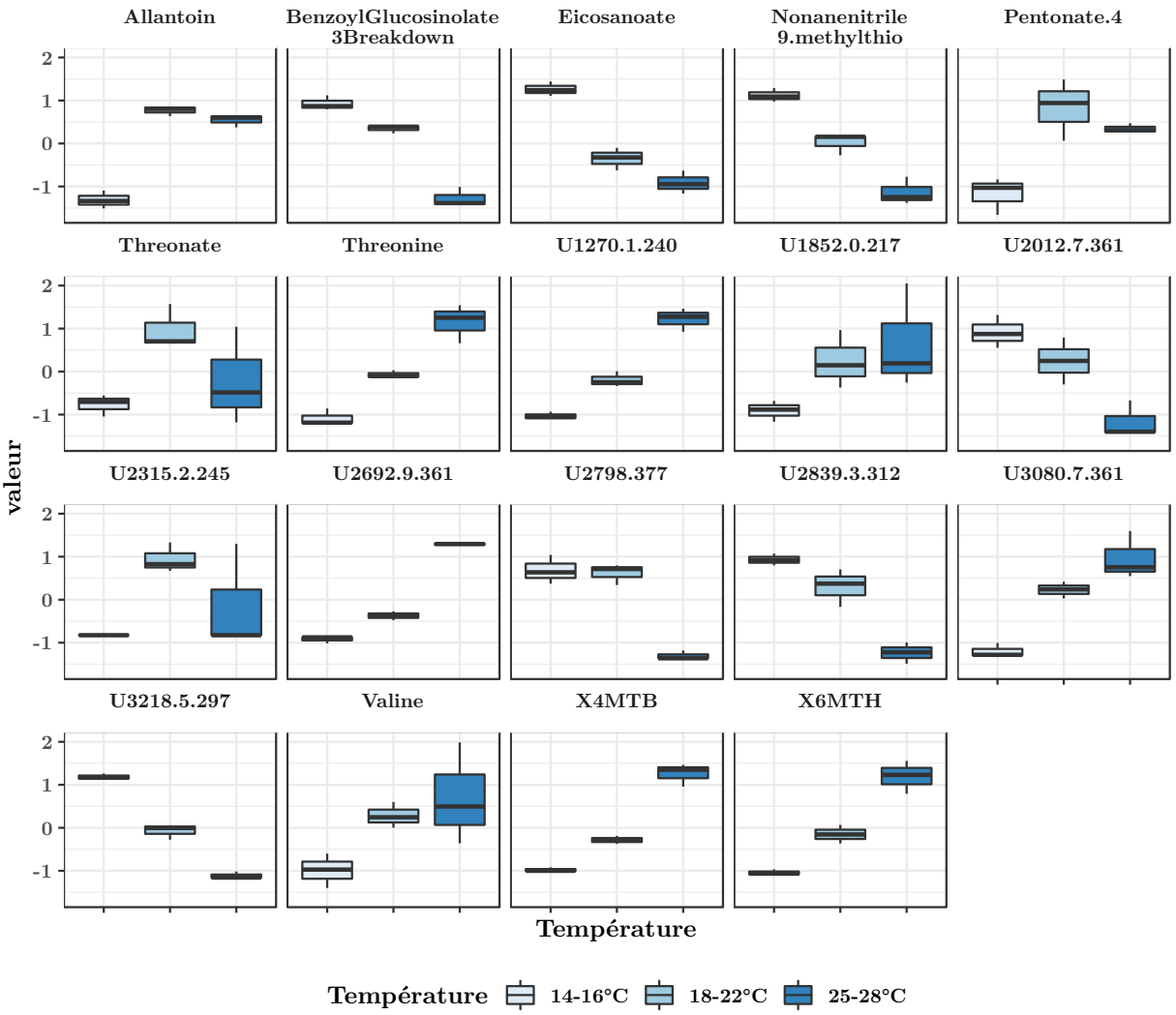

FIGURE 6. Boxplots des abondances des métabolites sélectionnés avec un seuil égal à 0.93.

ANNEXE A. EXEMPLE D'UTILISATION DU PACKAGE MULTIVARSELPOUR L'ANALYSE DES DONNÉES DE MÉTABOLOMIQUES

```
require(MultiVarSel) # Chargement du package MultiVarSel

data("metabolomAth") # permet de charger la table de données metab dans R

# Definition des matrices X et Y :
temperature <- metab$temperature
Y <- as.matrix(metab[, - 1])
X <- model.matrix(lm(Y ~ temperature + 0, data = metab))
```

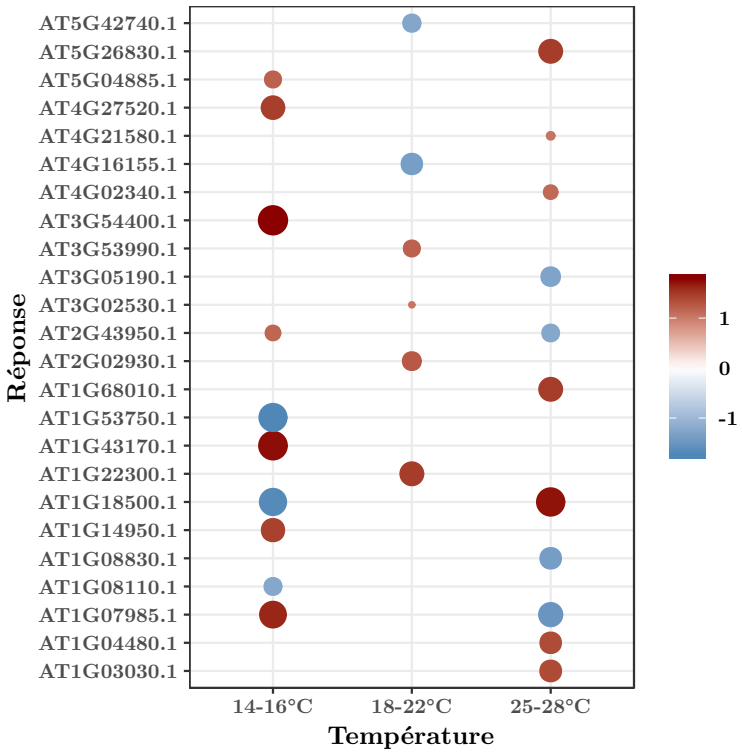

FIGURE 7. Estimation des coefficients $B_{i,j}$ pour les protéines sélectionnées avec un seuil égal à 0.95. Il est à noter que la surface des ronds est d'autant plus grande que les coefficients sont élevés en valeur absolue.

```
p <- ncol(X)
n <- nrow(X)
q <- dim(Y)[2]
Y <- scale(Y) # Renormalisation de la matrice Y
```

Ici on renormalise la matrice Y pour forcer la moyenne empirique des colonnes à être nulle et la variance empirique des colonnes à être égale à 1

```
residus <- lm(as.matrix(Y) ~ X - 1)$residuals # Définition des résidus

pvalue <- whitening_test(residus) # Test de blanchiment
print(pvalue)
## [1] 0.03038582
```

Ce test de blanchiment permet de savoir si de la dépendance est présente dans les données. La p-valeur est inférieure à 0.05 donc au niveau 5% on a mis en évidence de la dépendance dans les données qu'il faut retirer.

On peut tester plusieurs types de dépendances : ici AR(1) et Toeplitz symétrique (cas le plus général noté ici "nonparam")

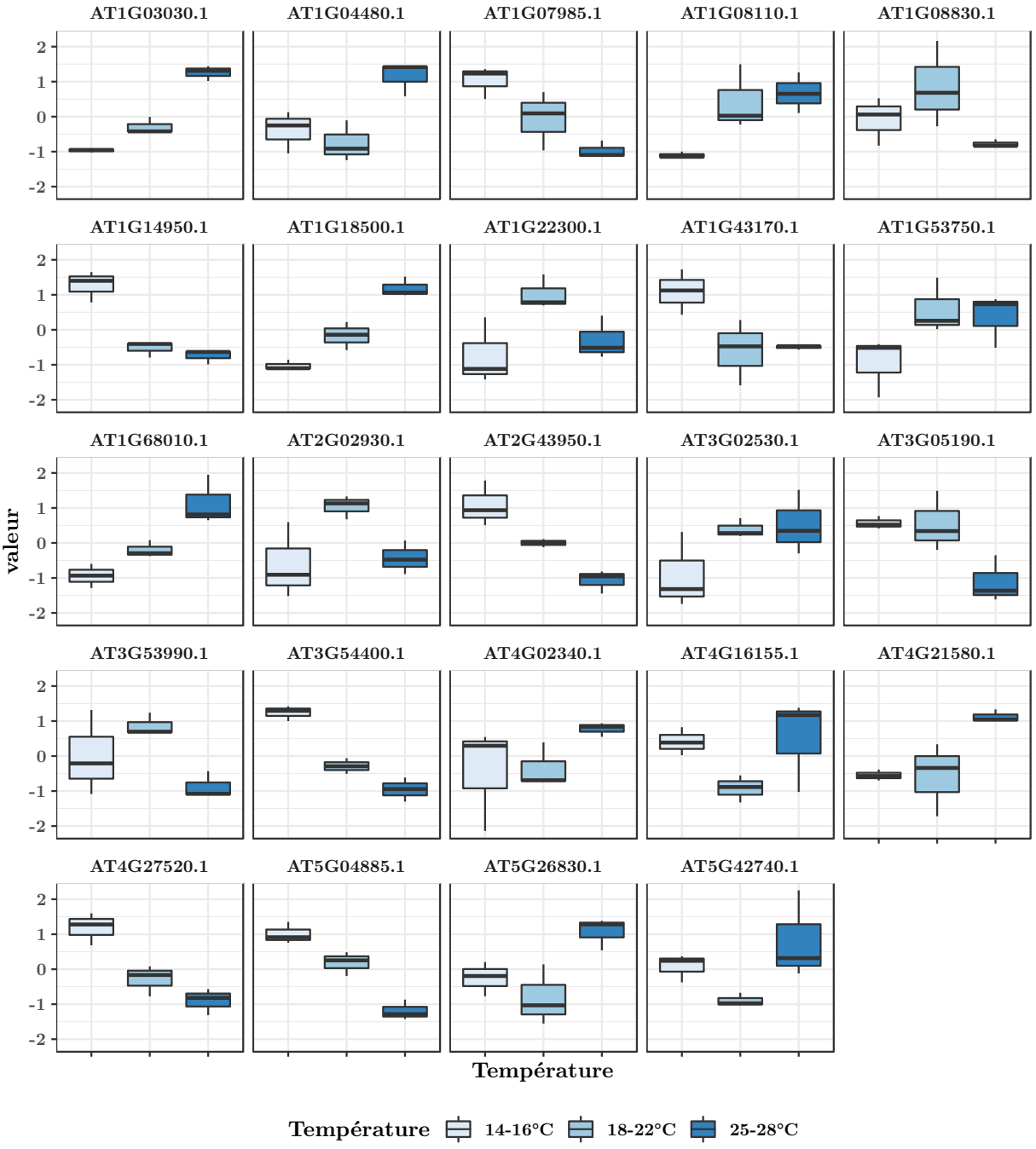

FIGURE 8. Boxplots des abondances des protéines sélectionnées avec un seuil égal à 0.95.

```
whitening_choice(residus, typeDeps = c("AR1", "nonparam"), pAR = 1, qMA = 0)
```

```
##          Pvalue       Decision
```

```
## AR1       0.035 NO WHITE NOISE
## nonparam  0.741    WHITE NOISE
```

On choisit donc la modélisation qui a la p-valeur la plus élevée. Ici il s'agit de la modélisation "nonparam" *i.e.* celle correspondant à la modélisation de la dépendance par une matrice de Toeplitz symétrique.

```
## Calcul de l'estimateur de la racine carrée de l'inverse de Sigma
square_root_inv_hat_Sigma <- whitening(residus, "nonparam", pAR  = 1, qMA = 0)

## Calcul de la fréquence de sélection de chaque variable à l'aide du Lasso
# et de la "stability selection"
require(doMC)  # pour paralléliser le calcul (ne fonctionne pas sous windows)
registerDoMC(cores = 4) # 4 correspond au nombre de coeurs utilisés

Freqs <- variable_selection(Y, X, square_root_inv_hat_Sigma, nb_repli = 5000,
                            parallel = TRUE, nb.cores = 4)
# 5000 correspond au nombre de réplications utilisées

seuil <- 0.93 # Choix du seuil pour les fréquences des variables sélectionnées
indices <- which(Freqs$frequency >= seuil)

## Estimation des coefficients sélectionnés dans B :
Yvec <- as.numeric(Y %*% square_root_inv_hat_Sigma)
Xvec <- kronecker(t(square_root_inv_hat_Sigma), X)
Xvec_sel <- Xvec[, indices]
B_sel_hat <- solve(t(Xvec_sel) %*% Xvec_sel, t(Xvec_sel) %*% Yvec)
Freqs$estim <- rep(0, p * q)
Freqs$estim[indices] <- as.vector(B_sel_hat)

## Graphe des métabolites sélectionnés avec l'estimation de leur
# coefficient B associé
ggplot(data =  Freqs[Freqs$frequency >= seuil, ],
             aes(y = Names_of_Y, x = Names_of_X,
                 color = estim, size = abs(estim))) +
  geom_point() +
  scale_size_continuous(name = "", breaks = c(0, 0.5, 1, 2, 3),
                        guide = FALSE)+
  scale_color_gradient2(low = "steelblue", mid = "white", high = "darkred") +
  labs(y = "Réponse", x = "Température", color = "")
```

Ce graphe est représenté dans la figure 5.

On donne à présent les commandes pour obtenir des boxplots similaires à ceux obtenus dans la figure 6.

```
require(reshape2) # package utilisé pour transformer les données
```

```
## Boxplots des métabolites sélectionnés
# On adapte la base de données pour le graphe
metab_sel <- as.character(Freqs[Freqs$frequency >= seuil, "Names_of_Y"])
temperature <- metab$temperature
metab_scale <- cbind.data.frame(temperature, Y[, metab_sel])
mm_sel <- melt(metab_scale, id.vars = "temperature")

# Graphe
ggplot(data = mm_sel, aes(x = temperature, y = value, fill = temperature)) +
  geom_boxplot() +
  facet_wrap(~variable) +
  theme(axis.text.x = element_blank())  +
  labs(y = "valeur", x = "Température", fill = "Température")+
  scale_fill_brewer(palette="Blues")
```

Annexe B. Exemple d'utilisation du package MultiVarSel pour l'analyse des données protéomiques

```
require(MultiVarSel) # Chargement du package MultiVarSel

data("proteomAth") # permet de charger la table de données prot dans R

# Definition des matrices X et Y :
temperature <-prot$temperature
Y <- as.matrix(prot[, - 1])
X <- model.matrix(lm(Y ~ temperature + 0, data = prot))
p <- ncol(X)
n <- nrow(X)
q <- dim(Y)[2]
Y <- scale(Y) # Renormalisation de la matrice Y
```

Ici on renormalise la matrice Y pour forcer la moyenne empirique des colonnes à être nulle et la variance empirique des colonnes à être égale à 1.

```
residus <- lm(as.matrix(Y) ~ X - 1)$residuals # Définition des résidus

pvalue <- whitening_test(residus) # Test de blanchiment
print(pvalue)
## [1] 0.06240354
```

Ce test de blanchiment permet de savoir si de la dépendance est présente dans les données. La p-valeur est inférieure à 0.05 donc au niveau 5% on a mis en évidence de la dépendance dans les données qu'il faut retirer.

On peut tester plusieurs types de dépendances : ici AR(1) et Toeplitz symétrique (cas le plus général noté ici "nonparam")

```
whitening_choice(residus, typeDeps = c("AR1", "nonparam"), pAR = 1, qMA = 0)
##          Pvalue    Decision
## AR1       0.187 WHITE NOISE
## nonparam      1 WHITE NOISE
```

On choisit donc la modélisation qui a la p-valeur la plus élevée. Ici, il s'agit de la modélisation "nonparam" *i.e.* celle correspondant à la modélisation de la dépendance par une matrice de Toeplitz symétrique.

```
## Calcul de l'estimateur de la racine carrée de l'inverse de Sigma
square_root_inv_hat_Sigma <- whitening(residus, "nonparam", pAR  = 1, qMA = 0)

## Calcul de la fréquence de sélection de chaque variable à l'aide du Lasso
# et de la "stability selection"
require(doMC)  # pour paralléliser le calcul (ne fonctionne pas sous windows)
registerDoMC(cores = 4) # 4 correspond au nombre de coeurs utilisés


Freqs <- variable_selection(Y, X, square_root_inv_hat_Sigma, nb_repli = 5000,
                            parallel = TRUE, nb.cores = 4)
# 5000 correspond au nombre de réplications utilisées
seuil <- 0.95 # Choix du seuil pour les fréquences des variables sélectionnées
indices <- which(Freqs$frequency >= seuil)

## Estimation des coefficients sélectionnés dans B :
Yvec <- as.numeric(Y %*% square_root_inv_hat_Sigma)
Xvec <- kronecker(t(square_root_inv_hat_Sigma), X)
Xvec_sel <- Xvec[, indices]
B_sel_hat <- solve(t(Xvec_sel) %*% Xvec_sel, t(Xvec_sel) %*% Yvec)
Freqs$estim <- rep(0, p * q)
Freqs$estim[indices] <- as.vector(B_sel_hat)

## Graphe des protéines sélectionnées avec l'estimation de leur
# coefficient B associé
ggplot(data =  Freqs[Freqs$frequency >= seuil, ],
             aes(y = Names_of_Y, x = Names_of_X,
                 color = estim, size = abs(estim))) +
  geom_point() +
  scale_size_continuous(name = "", breaks = c(0, 0.5, 1, 2, 3),
                        guide = FALSE)+
  scale_color_gradient2(low = "steelblue", mid = "white", high = "darkred") +
  labs(y = "Réponse", x = "Température", color = "")
```

Ce graphe est représenté dans la figure 7.

On donne à présent les commandes pour obtenir des boxplots similaires à ceux obtenus dans la figure 8.

```
require(reshape2) # package utilisé pour transformer les données

## Boxplots des protéines sélectionnées
# On adapte la base de données pour le graphe
prot_sel <- as.character(Freqs[Freqs$frequency >= seuil, "Names_of_Y"])
temperature <- prot$temperature
prot_scale <- cbind.data.frame(temperature, Y[, prot_sel])
mp_sel <- melt(prot_scale, id.vars = "temperature")

# Graphe
ggplot(data = mp_sel, aes(x = temperature, y = value, fill = temperature)) +
  geom_boxplot() +
  facet_wrap(~variable) +
  theme(axis.text.x = element_blank())  +
  labs(y = "valeur", x = "Température", fill = "Température")+
  scale_fill_brewer(palette="Blues")
```

Nous proposons ici une autre façon de créer les figures 7 et figures 8. en utilisant le package R `tidyverse`.

```
require(tidyverse)
Yvec <- as.numeric(Y %*% square_root_inv_hat_Sigma)
Xvec <- kronecker(t(square_root_inv_hat_Sigma), X)
colnames(Xvec) <-  paste(rep(colnames(Y), each = ncol(X)),
                         rep(colnames(X), ncol(Y)), sep = "_")

 # récupération des associations protéines températures selectionnées :
sel  <- Freqs[Freqs$frequency >= seuil, c("Names_of_Y","Names_of_X")] %>%
  unite("sel") %>% pull(sel)

# récupération des protéines selectionnées:
sel_prot <- Freqs[Freqs$frequency >= seuil, "Names_of_Y"] %>%
  as.character()

# Estimation de B
Xvec_sel <- as.matrix(Xvec)[,sel]
B_sel_hat <- solve(t(Xvec_sel) %*% Xvec_sel, t(Xvec_sel) %*% Yvec)

# Graphe représentant les valeurs non nulles de B
p <- B_sel_hat %>% as.data.frame() %>%
  rownames_to_column() %>%
  mutate(rowname = str_remove_all(rowname,"sel")) %>%
  separate(rowname, into = c("Response", "Temperature" ), sep="_") %>%
  mutate(Temperature= str_remove( Temperature,"temperature")) %>%
  ggplot(aes(y = Response, x = Temperature, color = V1, size = abs(V1))) +
  geom_point() +
  scale_size_continuous(name = "", breaks = c(0, 0.5, 1, 2, 3))+
```

```
  scale_color_gradient2(low = "steelblue",mid = "white", high = "darkred") +
  labs(y = "Réponse", x = "Température", color = "")

# Graphe représentant les boxplots des protéines selectionnées
p <- Y %>% as.data.frame() %>%
  select(sel_prot) %>%
  gather(key = "Response", value = value, - temperature) %>%
  ggplot(aes(x = temperature, y = value, fill = temperature)) +
  geom_boxplot() +
  facet_wrap( ~ Response) +
  theme(axis.text.x = element_blank(), legend.position = "bottom")  +
  labs(y = "valeur", x = "Température", fill = "Température") +
  scale_fill_brewer(palette = "Blues")
```

Remerciements

Nous remercions l'ensemble du consortium du projet européen EcoSeed (FP7 Environment, Grant/Award Number : 311840 "EcoSeed"). L'IJPB bénéficie du soutien du LABEX Saclay Plant Sciences-SPS (ANR-10-LABX-0040-SPS). Nous remercions l'ensemble des personnes ayant contribué à la production du matériel biologique et des données de protéomique et de métabolomique. En particulier, nous souhaitons remercier l'Université de Warwick (UWAR, Finch-Savage WE and Awan S) pour la production des graines, la plateforme de Biochimie de l'Observatoire du Végétal de l'IJPB (OV-Biochimie, Bailly M) pour la gestion et la préparation des échantillons pour la protéomique et la métabolomique, la Plateforme d'Analyse Protéomique de Paris Sud-Ouest (PAPPSO, Balliau T, Zivy M) pour les analyses en spectrométrie de masse des protéomes et la plateforme de Chimie-Métabolisme de l'Observatoire du Végétal de l'IJPB (OV-Chimie-Métabolisme, Clément G) pour l'analyse des métabolomes en GC/MS.

Références

Bailey-Serres, J., R. Sorenson, and P. Juntawong (2009). Getting the message across : cytoplasmic ribonucleoprotein complexes. *Trends in plant science 14*(8), 443–453.

Blödner, C., C. Goebel, I. Feussner, C. Gatz, and A. Polle (2007). Warm and cold parental reproductive environments affect seed properties, fitness, and cold responsiveness in arabidopsis thaliana progenies. *Plant, cell & environment 30*(2), 165–175.

Boccard, J. and S. Rudaz (2016). Exploring omics data from designed experiments using analysis of variance multiblock orthogonal partial least squares. *Analytica Chimica Acta 920*, 18 – 28.

Brockwell, P. and R. Davis (1991). *Time Series : Theory and Methods*. Springer Series in Statistics. Springer-Verlag New York.

Burghardt, L. T., B. R. Edwards, and K. Donohue (2016). Multiple paths to similar germination behavior in arabidopsis thaliana. *New Phytologist 209*(3), 1301–1312.

Durif, G., L. Modolo, J. Michaelsson, J. E. Mold, S. Lambert-Lacroix, and F. Picard (2017, 09). High dimensional classification with combined adaptive sparse PLS and logistic regression. *Bioinformatics 34*(3), 485–493.

el Bouhaddani, S., H.-W. Uh, C. Hayward, G. Jongbloed, and J. Houwing-Duistermaat (2018). Probabilistic partial least squares model : Identifiability, estimation and application. *Journal of Multivariate Analysis 167*, 331 – 346.

Faraway, J. J. (2004). *Linear Models with R*. Chapman & Hall/CRC.

Galland, M., R. Huguet, E. Arc, G. Cueff, D. Job, and L. Rajjou (2014). Dynamic proteomics emphasizes the importance of selective mrna translation and protein turnover during arabidopsis seed germination. *Molecular & Cellular Proteomics 13*(1), 252–268.

Keddie, J. S., M. Tsiantis, P. Piffanelli, R. Cella, P. Hatzopoulos, and D. J. Murphy (1994). A seed-specific brassica napus oleosin promoter interacts with a g-box-specific protein and may be bi-directional. *Plant molecular biology 24*(2), 327–340.

Kourmpetli, S., K. Lee, R. Hemsley, P. Rossignol, T. Papageorgiou, and S. Drea (2013). Bidirectional promoters in seed development and related hormone/stress responses. *BMC plant biology 13*(1), 187.

Lê Cao, K.-A., S. Boitard, and P. Besse (2011). Sparse pls discriminant analysis : biologically relevant feature selection and graphical displays for multiclass problems. *BMC Bioinformatics 12*(1), 253.

Mardia, K., J. Kent, and J. Bibby (1979). *Multivariate analysis*. Probability and mathematical statistics. Academic Press.

Meinshausen, N. and P. Buhlmann (2010). Stability selection. *Journal of the Royal Statistical Society 72*(4), 417–473.

Muller, K. E. and P. W. Stewart (2006). *Linear Model Theory : Univariate, Multivariate, and Mixed Models*. John Wiley & Sons.

Perrot-Dockès, M., C. Lévy-Leduc, and L. Rajjou (2019). Estimation of large block structured covariance matrices : Application to "multi-omic" approaches to study seed quality. arXiv :1806.10093v2.

Perrot-Dockès, M., C. Lévy-Leduc, L. Sansonnet, and J. Chiquet (2018). Variable selection in multivariate linear models with high-dimensional covariance matrix estimation. *Journal of Multivariate Analysis 166*, 78–97.

Ren, S., A. A. Hinzman, E. L. Kang, R. D. Szczesniak, and L. J. Lu (2015). Computational and statistical analysis of metabolomics data. *Metabolomics 11*(6), 1492–1513.

Rothman, A. J., E. Levina, and J. Zhu (2010). Sparse multivariate regression with covariance estimation. *Journal of Computational and Graphical Statistics 19*(4), 947–962.

Saccenti, E., H. C. J. Hoefsloot, A. K. Smilde, J. A. Westerhuis, and M. M. W. B. Hendriks (2013). Reflections on univariate and multivariate analysis of metabolomics data. *Metabolomics 10*(3), 361–374.

Sønderby, I. E., F. Geu-Flores, and B. A. Halkier (2010). Biosynthesis of glucosinolates–gene discovery and beyond. *Trends in plant science 15*(5), 283–290.

Springthorpe, V. and S. Penfield (2015). Flowering time and seed dormancy control use external coincidence to generate life history strategy. *Elife 4*, e05557.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. Royal. Statist. Soc B. 58*(1), 267–288.

Zhang, H., Y. Zheng, G. Yoon, Z. Zhang, T. Gao, B. Joyce, W. Zhang, J. Schwartz, P. Vokonas, E. Colicino, A. Baccarelli, L. Hou, and L. Liu (2017, Jul). Regularized estimation in sparse high-dimensional multivariate regression, with application to a DNA methylation study. *Stat Appl Genet Mol Biol 16*(3), 159–171.

UMR MIA-Paris, INRA, AgroParisTech, Université Paris-Saclay, 75005, Paris, France
Email address: marie.perrot-dockes@agroparistech.fr

UMR MIA-Paris, INRA, AgroParisTech, Université Paris-Saclay, 75005, Paris, France
Email address: celine.levy-leduc@agroparistech.fr

Institut Jean-Pierre Bourgin, INRA, AgroParisTech, Université Paris-Saclay, 78026, Versailles, France
Email address: gwendal.cueff@inra.fr

Institut Jean-Pierre Bourgin, INRA, AgroParisTech, Université Paris-Saclay, 78026, Versailles, France
Email address: loic.rajjou@agroparistech.fr