Modèle linéaire multivarié parcimonieux avec estimation de covariance : une application à des données de métabolomique

Marie Perrot-Dockès [1,2] & Céline Lévy-Leduc[1,3] & Julien Chiquet[1,4] & Laure Sansonnet[1,5]

[1] *AgroParisTech/INRA UMR518 MIA-Paris, 16 rue Claude Bernard 75005 Paris*
[2] *marie.perrot-dockes@agroparistech.fr ;* [3] *celine.levy-leduc@agroparistech.fr*
[4] *julien.chiquet@agroparistech.fr ;* [5] *laure.sansonnet@agroparistech.fr*

Résumé : Les données 'omiques' sont caractérisées par une forte structure de dépendance qui peut être due à l'acquisition des données ou à un phénomène biologique sous-jacent. En métabolomique, par exemple, il est intéressant de trouver quelles variables permettent de caractériser un phénotype donné. Ne pas tenir compte de la structure de dépendance présente dans les données de métabolomique peut conduire à la sélection de variables non pertinentes. Dans cet article, nous proposons une nouvelle méthode utilisant le critère Lasso adapté aux modèles multivariés en grande dimension et prenant en compte la structure de dépendance en utilisant différentes modélisations de la matrice de covariance des résidus. Les résultats des simulations numériques que nous avons menées ont montré que la prise en compte de la structure de dépendance sous-jacente améliore de façon significative la sélection de variables. Nous présentons également une application de notre méthode à des données de métabolomique analysant des échantillons de résine d'arbres.

Mots-clés. Modèle linéaire multivarié, dépendance, Lasso, grande dimension, sélection de variable.

Abstract. 'Omic' data are characterized by the presence of strong dependence structures that result either from data acquisition or from some underlying biological processes. In metabolomics, for instance, an important goal is to select metabolites characterizing a phenotype of interest associated with the samples. However, not taking into account the dependence pattern in the variable selection step may result in the selection of spurious variables. In this paper we propose a novel Lasso-based approach in the multivariate framework of the general linear model taking into account the dependence structure by using various modelings of the covariance matrix of the residuals. Our numerical experiments show that including the estimation of the covariance matrix of the residuals in the Lasso criterion dramatically improves the variable selection performance. Our approach is also successfully applied to a data set made of African copals samples for which it is able to provide a small list of metabolites without altering the phenotype discrimination.

Keywords. Linear multivariate model, dependence, Lasso, high-dimension, variable selection.

1 Introduction

Dans une expérience de métabolomique standard où n échantillons sont analysés, les résultats se présentent sous la forme d'une matrice de taille $n \times q$ où q correspond au nombre de métabolites ("petites molécules"). Les métabolites y sont ordonnés par ordre croissant de valeur de masse sur charge (m/z). Lorsque les n échantillons ont été obtenus dans diverses conditions, on cherche à comprendre l'effet de chaque condition sur chaque métabolite. Dans le cas où C conditions expérimentales sont comparées, on note n_c le nombre d'observations dans la condition c. Notons $Y_{c,r}^{(j)}$ la réponse centrée du j-ième métabolite pour la r-ième observation dans la condition c. La méthode la plus classique pour analyser l'effet d'une variable qualitative sur une variable quantitative est le modèle d'ANOVA à un facteur qui s'écrit comme suit :

$$Y_{c,r}^{(j)} = \mu_c^{(j)} + E_{c,r}^{(j)}, \tag{1}$$

où $\mu_c^{(j)}$ est l'effet de la condition c de la variable qualitative sur le métabolite j et où les $E_{c,r}^{(j)}$ sont supposées être des variables iid gaussiennes et centrées. Le but d'une telle modélisation est de comprendre quels sont parmi les $\mu_1^{(j)}, \mu_2^{(j)}, \ldots, \mu_C^{(j)}$ ceux qui sont les plus significatifs sur le métabolite j. Pour simplifier les notations nous remplacerons dans la suite les indices c, r par un indice unique i.

Lorsque l'on considère la matrice entière de taille $n \times q$ à la place de la colonne j, le modèle peut être réécrit comme suit :

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{B} + \boldsymbol{E}, \tag{2}$$

où $\boldsymbol{Y} = (Y_{i,j})_{1\leq i\leq n,\, 1\leq j\leq q}$ est la matrice des observations de taille $n\times q$, $\boldsymbol{X}$ est la matrice de design de taille $n\times p$, $\boldsymbol{B}$ est la matrice des coefficients de taille $p\times q$ et $\boldsymbol{E} = (E_{i,j})_{1\leq i\leq n,\, 1\leq j\leq q}$ est la matrice de l'erreur résiduelle de taille $n\times q$. Afin de prendre en compte la dépendance potentielle qui existe entre les colonnes de $\boldsymbol{Y}$ nous supposerons que

$$(E_{i,1}, \ldots, E_{i,q}) \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{\Sigma}_q), \text{ où } i \in \{1, \ldots n\}. \tag{3}$$

Trouver les paramètres qui sont les plus significatifs parmi les $(\mu_c^{(j)})_{1\leq c\leq C, 1\leq j\leq q}$ dans le modèle (1) revient à chercher les coefficients non nuls de $\boldsymbol{B}$ dans le modèle (2) et donc à faire de la sélection de variables dans le modèle linéaire multivarié (2). Pour cela, nous proposons d'étendre le critère Lasso proposé par Tibshirani (1996) afin de prendre en compte la dépendance présente dans $\boldsymbol{Y}$. Nous illustrerons notre méthode sur des données simulées et réelles issues d'une application en métabolomique.

2 Description de la méthode

Notre méthode s'articule en trois étapes :

1re étape – estimation des erreurs : la matrice des erreurs $\boldsymbol{E}$ est estimée par les résidus $\widehat{\boldsymbol{E}}$ obtenus en modélisant indépendamment chacune des colonnes de $\boldsymbol{Y}$ à l'aide d'un modèle linéaire univarié.

2e étape – estimation de la dépendance : la matrice de covariance $\boldsymbol{\Sigma}_q$ est estimée par différentes méthodes à l'aide des résidus $\widehat{\boldsymbol{E}}$. Le meilleur estimateur $\widehat{\boldsymbol{\Sigma}}_q$ est sélectionné à l'aide d'un test d'indépendance des résidus « blanchis » définis comme suit : $\widehat{\boldsymbol{E}}\ \widehat{\boldsymbol{\Sigma}}_q^{-1/2}$.

3e étape – sélection de variables : les données sont blanchies selon la même opération afin de retirer la dépendance entre les colonnes de $\boldsymbol{Y}$. Une procédure de type Lasso est ensuite mise en place afin d'effectuer de la sélection de variables sur les données blanchies, accompagnée d'une étape de sous-échantillonnage pour assurer la stabilité des variables sélectionnées.

Nous décrivons par la suite plus en détails les étapes 2 et 3.

2.1 Estimation de la structure de dépendance

Il est à noter que les différents métabolites sont caractérisés et ordonnés dans la matrice $\boldsymbol{Y}$ en fonction de leur rapport masse sur charge (m/z). Nous proposons d'estimer la structure de dépendance de $\boldsymbol{E}$ en modélisant chaque ligne de la matrice $\boldsymbol{E}$ comme la réalisation d'un processus stationnaire pour lequel nous considèrerons des modélisations paramétriques ou non-paramétriques de séries temporelles développées dans Brockwell and Davis (1991).

2.1.1 Modélisation paramétrique de la dépendance

La modélisation paramétrique la plus simple est le modèle auto-régressif d'ordre 1 noté AR(1). Plus précisément, pour chaque $i \in \{1, \dots, n\}$, ceci revient à supposer que $E_{i,t}$ satisfait l'équation

$$E_{i,t} - \phi_1 E_{i,t-1} = W_{i,t}, \text{ avec } W_{i,t} \sim BB(0, \sigma^2), \tag{4}$$

où $|\phi_1| < 1$ et $BB(0, \sigma^2)$ correspond à un bruit blanc d'espérance nulle et de variance σ^2. Dans ce cas particulier $\boldsymbol{\Sigma}_q^{-1/2}$ a une forme explicite qui ne dépend que de ϕ_1. Ainsi, pour obtenir l'expression de $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$, il suffit d'estimer le paramètre ϕ_1 ce qui peut être fait en utilisant les équations de Yule-Walker (voir Brockwell and Davis (1991)). Plus généralement il est aussi possible d'accéder à $\widehat{\boldsymbol{\Sigma}}_q^{-1/2}$ pour des modélisations paramétriques plus compliquées, telles que les ARMA(p, q).

2.1.2 Modélisation non-paramétrique de la dépendance

Dans le cas où un modèle paramétrique n'est pas adapté, $\mathbf{\Sigma}_q$ peut être estimée par

$$\widehat{\mathbf{\Sigma}}_q = \begin{pmatrix} \widehat{\gamma}(0) & \widehat{\gamma}(1) & \cdots & \widehat{\gamma}(q-1) \\ \widehat{\gamma}(1) & \widehat{\gamma}(0) & \cdots & \widehat{\gamma}(q-2) \\ \vdots & & & \\ \widehat{\gamma}(q-1) & \widehat{\gamma}(q-2) & \cdots & \widehat{\gamma}(0) \end{pmatrix}, \tag{5}$$

où $\widehat{\gamma}(h)$ est un estimateur de la fonction d'auto-covariance des processus stationnaires $(\widehat{E_{i,t}})$ en h. La matrice $\widehat{\mathbf{\Sigma}}_q^{-1/2}$ est ensuite obtenue en inversant le facteur de Cholesky de $\widehat{\mathbf{\Sigma}}_q$.

2.1.3 Choix de l'estimateur de $\mathbf{\Sigma}_q$

On choisit parmi les estimateurs précédents celui qui garantit que les résidus blanchis $\widehat{\boldsymbol{E}}\widehat{\mathbf{\Sigma}}_q^{-1/2}$ sont un bruit blanc. Ceci est réalisé à l'aide d'un test de type Portmanteau qui vise à tester l'hypothèse nulle que les résidus blanchis sont un bruit blanc, pour plus de détails voir Brockwell and Davis (1991). On sélectionne ensuite l'estimateur $(\widehat{\mathbf{\Sigma}}_q^{-1/2})$ pour lequel la p-valeur de ce test est la plus élevée.

2.2 Sélection de variables

Pour sélectionner les variables les plus pertinentes après prise en compte de la dépendance des colonnes de $\boldsymbol{Y}$, nous proposons d'utiliser l'estimateur Lasso introduit par Tibshirani (1996). La sélection des variables est appliquée sur les données blanchies, obtenues selon la transformation suivante :

$$\boldsymbol{Y}\ \widehat{\mathbf{\Sigma}}_q^{-1/2} = \boldsymbol{X}\boldsymbol{B}\ \widehat{\mathbf{\Sigma}}_q^{-1/2} + \boldsymbol{E}\ \widehat{\mathbf{\Sigma}}_q^{-1/2}, \tag{6}$$

où $\widehat{\mathbf{\Sigma}}_q^{-1/2}$ est l'estimateur sélectionné à l'aide du test du Portmanteau.

2.2.1 Approche fondée sur le lasso

La méthodologie Lasso ne pouvant pas être directement appliquée au modèle linéaire général, nous vectorisons l'équation (4) afin de se ramener à la forme

$$\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}, \tag{7}$$

où $\mathcal{Y} = vec(\boldsymbol{Y}\widehat{\mathbf{\Sigma}}_q^{-1/2})$, $\mathcal{X} = (\widehat{\mathbf{\Sigma}}_q^{-1/2})' \otimes \boldsymbol{X}$ and $\mathcal{E} = vec(\boldsymbol{E}\widehat{\mathbf{\Sigma}}_q^{-1/2})$ sont de tailles respectives $nq \times 1$, $nq \times pq$ et $nq \times 1$. Ainsi, retrouver les positions non nulles de $\mathcal{B}$ revient à trouver

les variables pertinentes en ayant pris en compte la structure de dépendance sous-jacente. Le critère lasso adapté à notre contexte s'écrit pour $\lambda > 0$,

$$\widehat{\mathcal{B}}(\lambda) = \text{Argmin}_{\mathcal{B}} \left\{ \|\mathcal{Y} - \mathcal{X}\mathcal{B}\|_2^2 + \lambda \|\mathcal{B}\|_1 \right\}, \tag{8}$$

où, pour $u = (u_1, \ldots, u_n) \in \mathbb{R}^n$, $\|u\|_2^2 = \sum_{i=1}^n u_i^2$ et $\|u\|_1 = \sum_{i=1}^n |u_i|$.

2.2.2 Choix de modèle et stabilité

Afin de calibrer le niveau de parcimonie du vecteur $\widehat{\mathcal{B}}$, nous proposons la stratégie suivante : *i*) sélectionner une valeur de λ_{cv} à l'aide de la validation croisée ; *ii*) appliquer la méthode de rééchantillonnage *Stability Selection* de Meinshausen and Buhlmann (2010) pour cette valeur de λ_{cv} afin de garantir la stabilité des variables sélectionnées. Les variables retenues sont celles qui sont sélectionnées à chaque étape de rééchantillonnage.

3 Simulations numériques

Le but de cette partie est d'illustrer les performances statistiques de notre méthode sur des données simulées. Pour cela, nous avons généré des observations $\boldsymbol{Y}$ selon le modèle (2) où $q = 1000$, $p = 3$, $n = 30$, $\boldsymbol{X}$ est la matrice de design d'une ANOVA à un facteur et $\boldsymbol{\Sigma}_q$ est la matrice de covariance d'un AR(1) avec $\phi_1 = 0.9$. Les résultats des simulations numériques sont donnés dans la figure 1. Le graphe de gauche compare à l'aide de courbes ROC les performances *(i)* d'une méthode Lasso sans blanchiment ('Lasso'), *(ii)* d'une méthode de sélection de variables par ANOVA sans prise en compte de dépendance ('ANOVA'), *(iii)* de notre approche en supposant Σ_q connue ('Oracle') ou en l'estimant ('AR1', 'Nonparam'). On observe que le blanchiment améliore considérablement la sélection de variables. Le graphe de droite de la figure 1 montre que notre méthode de calibration de λ nous assure aucun faux positif dans les variables sélectionnées. Par contre, certains variables peuvent être omises.

4 Application

Dans cette section, nous appliquons notre méthode à des données de métabolomique correspondant à l'analyse de 30 échantillons de résines d'arbres. Le but de cette analyse est de comprendre comment l'origine des arbres influe sur leur métabolisme. Ces données peuvent être modélisées par (2) où $\boldsymbol{X}$ est la matrice de design d'une ANOVA à un facteur, $q = 1000$, $p = 3$ et $n = 30$. Nous avons comparé notre approche à une méthode classiquement utilisée en métabolomique : sPLS-DA développée par Lê Cao et al. (2011) et implémentée dans le package R `MixOmics`. Les résultats sont donnés dans la figure 2.

Dans le graphe de gauche de la figure 2, les métabolites sélectionnés par les deux approches sont représentés. "Comp1" et "Comp2" sont les deux premières composantes

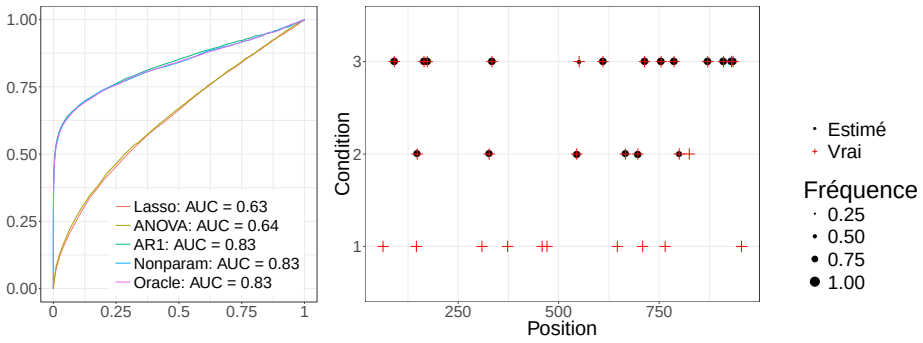

FIGURE 1 – À gauche : moyenne de 200 réplications de courbes ROC. À droite : fréquences des variables sélectionnées par notre méthode ('•').

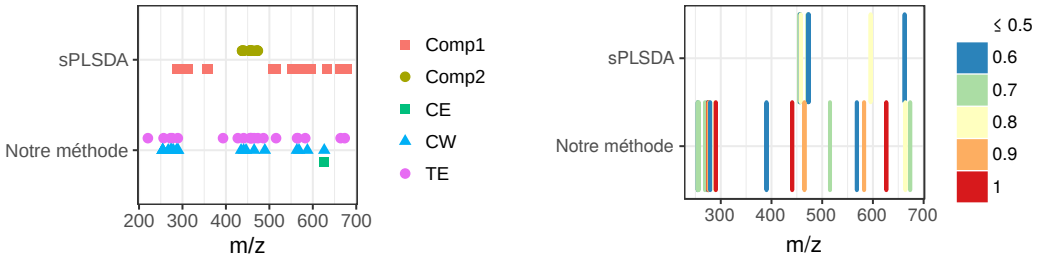

FIGURE 2 – Comparaison des positions (à gauche) et des fréquences (à droite) des métabolites sélectionnés par notre approche et par sPLS-DA.

de la sPLS-DA, "CE", "CW" et "TE" représentent les différentes modalités de la variable qualitative (origine des arbres). Dans le graphe de droite de la figure 2, on a représenté la fréquence des variables sélectionnées par les deux méthodologies après ré-échantillonnage afin d'étudier la stabilité des variables sélectionnées. Notre approche fournit des variables plus stables que sPLS-DA.

Références

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

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.

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

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