

Méthode de sélection de variables dans le modèle linéaire multivarié avec prise en compte de la dépendance

Marie Perrot-Dockès

Céline Lévy-Leduc, Julien Chiquet, Laure Sansonnet

AgroParisTech/UMR INRA MIA-Paris (Équipe Stat & Génome)



Introduction

- **Description des données :**
 - \mathbf{X} matrice de design de taille $n \times p$
 - \mathbf{Y} matrice réponse de taille $n \times q$ ($n \approx 30$, $q \approx 1000$)
- **Question :** Quelles sont les variables qui influencent les réponses ?
- **Approche :**
 - Sélection de variables dans le modèle

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

où

- \mathbf{B} matrice **parcimonieuse** des coefficients de taille $p \times q$
- \mathbf{E} matrice des erreurs de taille $n \times q$ avec

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

- Prise en compte de la dépendance en estimant Σ_q



Méthode en 3 étapes

① Estimation des erreurs E

Résidus des modèles indépendants sur chaque colonne de Y
 $\Rightarrow \widehat{E}$.

② Estimation de la covariance des erreurs Σ_q

- Différentes modélisations de la dépendance \Rightarrow différentes estimations de Σ_q (AR(1), ARMA, Toeplitz, ...).
- Choix de la meilleure modélisation de la dépendance à l'aide d'un test de bruit blanc (test de Portmanteau).

③ “Blanchiment” et sélection de variables :

- “Blanchiment” : $Y \widehat{\Sigma}_q^{-1/2} = XB \widehat{\Sigma}_q^{-1/2} + E \widehat{\Sigma}_q^{-1/2}$
- Sélection de variables par procédure Lasso
- Stability selection*

Estimation de $\Sigma_q^{-1/2}$ dans le cas d'un AR(1)

$\forall i \in \{1, \dots, n\}, \forall t \in \mathbb{Z}, E_{i,t} - \phi_1 E_{i,t-1} = W_{i,t},$
 avec $(W_{i,t})_t \sim BB(0, \sigma^2), |\phi_1| < 1.$

Dans ce cas :

$$\Sigma_q^{-1/2} = \begin{pmatrix} \sqrt{1 - \phi_1^2} & -\phi_1 & 0 & \cdots & 0 \\ 0 & 1 & -\phi_1 & \cdots & 0 \\ 0 & 0 & \ddots & \ddots & \vdots \\ \vdots & \vdots & \ddots & \ddots & -\phi_1 \\ 0 & 0 & \cdots & 0 & 1 \end{pmatrix}.$$

Estimateur de ϕ_1 : $\widehat{\phi}_1 = \frac{1}{n} \sum_{i=1}^n \widehat{\phi}_1^{(i)}$

où $\widehat{\phi}_1^{(i)}$: estimateur de Yule-Walker de ϕ_1 obtenu à partir de la i ème ligne de \widehat{E} .

Estimation de $\Sigma_q^{-1/2}$ dans le cas “Toeplitz”

$\forall i \in \{1, \dots, n\}$, on modélise $(E_{i,1}, \dots, E_{i,q})$ comme un processus stationnaire.

Dans ce cas :

$$\Sigma_q = \begin{pmatrix} \gamma(0) & \gamma(1) & \cdots & \gamma(q-1) \\ \gamma(1) & \gamma(0) & \cdots & \gamma(q-2) \\ \vdots & & & \\ \gamma(q-1) & \gamma(q-2) & \cdots & \gamma(0) \end{pmatrix},$$

Estimateur de $\gamma(h)$: $\hat{\gamma}(h) = \frac{1}{n} \sum_{i=1}^n \hat{\gamma}_i(h),$

où $\hat{\gamma}_i(h)$ est un estimateur de la fonction d'auto-covariance du processus stationnaire $(E_{i,t})_t$ en h .

Estimation de B et sélection de variables

Approche de type LASSO

- Dans le cas d'un modèle univarié : $\mathcal{Y} = \mathcal{X}\mathcal{B} + \mathcal{E}$ le critère LASSO pour estimer \mathcal{B} est

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

- On vectorise le modèle blanchi :
- $$\mathbf{Y}\widehat{\Sigma}_q^{-1/2} = \mathbf{X}\mathcal{B}\widehat{\Sigma}_q^{-1/2} + \mathbf{E}\widehat{\Sigma}_q^{-1/2}$$

$$\begin{aligned} \mathcal{Y} &= \operatorname{vec}(\mathbf{Y}\widehat{\Sigma}_q^{-1/2}) = \operatorname{vec}(\mathbf{X}\mathcal{B}\widehat{\Sigma}_q^{-1/2}) + \operatorname{vec}(\mathbf{E}\widehat{\Sigma}_q^{-1/2}) \\ &= ((\widehat{\Sigma}_q^{-1/2})' \otimes \mathbf{X}) \operatorname{vec}(\mathcal{B}) + \operatorname{vec}(\mathbf{E}\widehat{\Sigma}_q^{-1/2}) \\ &= \mathcal{X}\mathcal{B} + \mathcal{E}. \end{aligned}$$

Résultat Théorique

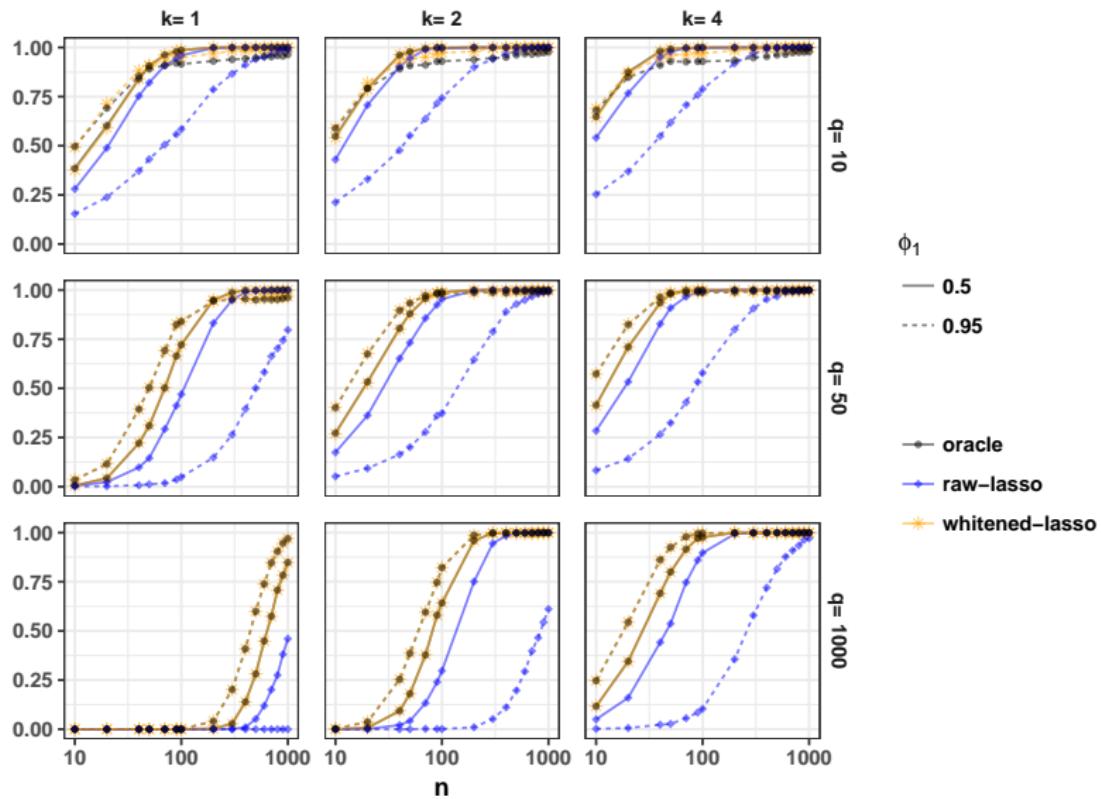
Théorème : Si les conditions suivantes sont respectées

- ① $|J| = O(q^{c_1})$
- ② $q^{c_2} \min_{j \in J} |\mathcal{B}_j| \geq M_3$.
- ③ $\|\Sigma^{-1} - \widehat{\Sigma}^{-1}\|_\infty = O_P((nq)^{-1/2})$
- ④ $\rho(\Sigma - \widehat{\Sigma}) = O_P((nq)^{-1/2})$
- ⑤ $q = q_n = o\left(n^{\frac{1}{2(c_1+c_2)}}\right) = o(n^k)$ où $c_1 + c_2 = \frac{1}{2k}$,
 $\frac{\lambda}{\sqrt{n}} \rightarrow \infty$ and $\frac{\lambda}{n} = o\left(q^{-(c_1+c_2)}\right)$, as $n \rightarrow \infty$,

alors,

$$\mathbb{P}\left(\text{sign}(\widehat{\mathcal{B}}(\lambda)) = \text{sign}(\mathcal{B})\right) \rightarrow 1, \text{ as } n \rightarrow \infty.$$

Vrai support : Simulation numérique

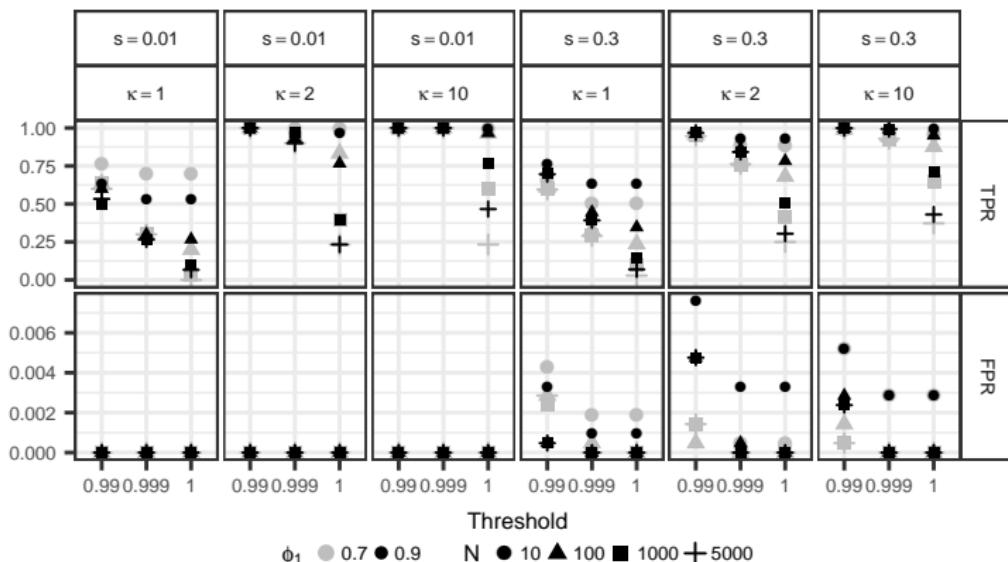


Choix de λ en pratique

- On divise le jeu de donnée en 10 groupes
- Appelons k l'échantillon à qui on retire le k^{iem} groupe. Pour chaque échantillon :
 - Calcul λ_{CV}
 - Stability selection $\rightarrow (N_i^k)_{1 \leq i \leq n}$: le nombre de fois où chaque variable a été sélectionnée dans cet échantillon k
- $N_i = \sum_{i=1}^{10} N_i^k / 10$
- On garde les variables i tel que $N_i >$ seuil

Implémenté dans un package R **MultiVarSel**

Choix du seuil en pratique



Données réelles: cas de l'ANOVA à un facteur

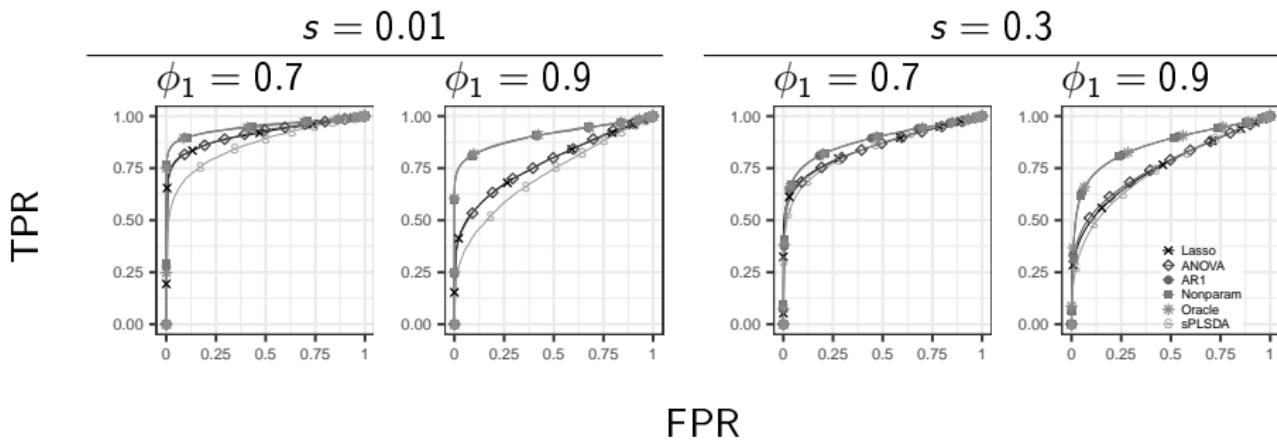
Défi CNRS AREA Enviromics

- Échantillon de résines de 30 arbres de type :
 - CE (*Copaifera East Africa*)
 - TE (*Trachylobium East Africa*)
 - CW (*Copaifera West Africa*)
- Réponses : concentrations de 1000 métabolites (petites molécules)



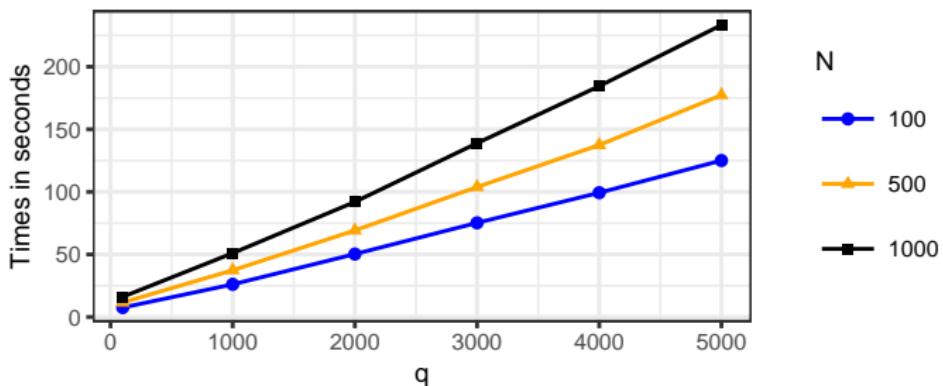
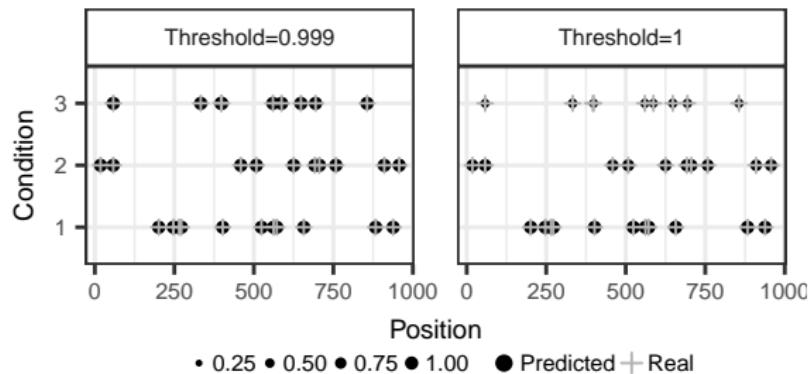
Simulations : comparaisons de courbes ROC

Simulations faites sur 1000 répliques en utilisant notre package R **MultiVarSel**.

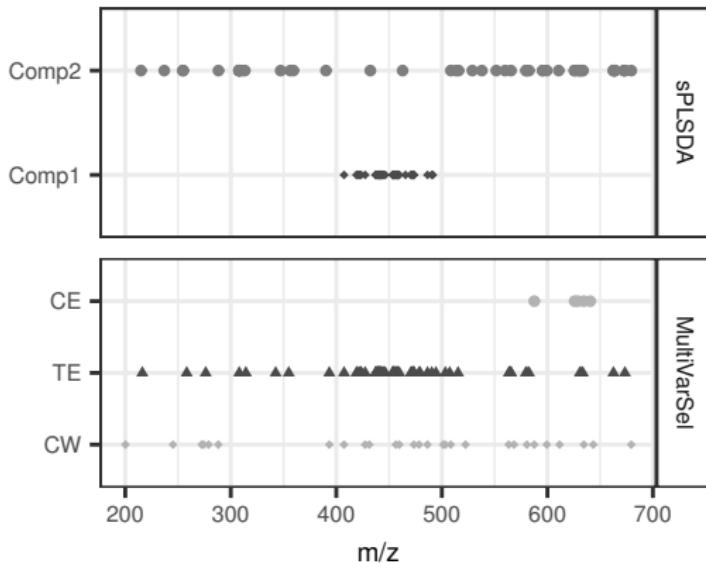


Nous retrouvons mieux le support que la sPLS-DA.

Simulations : performances statistiques et numériques



Comparaison des métabolites sélectionnées par notre méthode et la sPLS-DA



Conclusion et Perspectives

- Bonnes performances d'un point de vue pratique et théorique de notre méthode
- Performances similaires voire meilleures que les méthodes existantes

Conclusion et Perspectives

- Bonnes performances d'un point de vue pratique et théorique de notre méthode
- Performances similaires voire meilleures que les méthodes existantes
- Travaux en cours
 - Utilisation d'autres types de structures de dépendance pour Σ_q
 - En pratique : Généralisation à d'autres matrices de design
 - Collaboration avec l'Institut Curie : dialogue entre cellule

Références

- M. Perrot-Dockès et al. “A multivariate variable selection approach for analyzing LC-MS metabolomics data”, arXiv:1704.00076.
- M. Perrot-Dockès et al. “Variable selection in multivariate linear models with high-dimensional covariance matrix estimation”, arXiv:1707.04145
- Package R **MultiVarSel** sur le CRAN

Merci pour votre attention !