Introduction à la sparse PLS

Hadrien Lorenzo
hadrienlorenzo.netlify.com *hadrien.lorenzo@u-bordeaux.fr*

22 janvier 2018
STA301, Master 2 Biostatistique, ISPED

Ce cours est très largement inspiré de documents rédigés par ***Boris Hejblum*** *et* ***Robin Genuer***

Modèle de régression linéaire multiple

Soit la régression linéaire suivante :

$$Y = X\beta + \varepsilon$$

avec :

- n observations
- Y, la variable à expliquer (vecteur de dimension n)
- $X_{n \times p}$, la matrice des p variables explicatives
- β, les coefficients de régression (vecteur de dimension p)
- ε, les erreurs (vecteur de dimension n)

Estimateur des Moindres Carrés Ordinaires (MCO) :

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

Modèle de régression linéaire multiple

L'estimateur des MCO est trouvé grâce à la résolution du problème

$$\min_{\beta} ||Y - X\beta||_2^2$$

En détaillant

$$\begin{aligned} f(\beta) &= ||Y - X\beta||_2^2 = (Y - X\beta)^T(Y - X\beta) \\ &= Y^T Y + \beta^T X^T X\beta - 2Y^T X\beta \\ &= \beta^T X^T X\beta - 2Y^T X\beta + \text{cste} = ||X\beta||_2^2 - 2 < Y, X\beta > + \text{cste} \end{aligned},$$

on peut alors réécrire le problème de MCO

$$\max_{\beta} < Y, X\beta > -\frac{1}{2}||X\beta||_2^2$$

Modèle de régression linéaire multiple

Trouver une combinaison linéaire des covariables telle que la variable ainsi créée tende à positionner les individus comme la variable réponse, sans pour autant donner trop d'importance à X.

En notant $g(\beta) = \beta^T X^T Y - \frac{1}{2}||X\beta||_2^2$, fonction deux fois continument dérivable en β, on peut écrire

$$g'(\beta) = X^T Y - X^T X\beta,$$

au point de minimum, $\widehat{\beta}$, on obtient $g'(\widehat{\beta}) = 0$ et alors $X^T Y = X^T X\widehat{\beta}$ et par inversion il vient directement $\widehat{\beta} = (X^T X)^{-1} X^T Y$. Une analyse rapide des dérivées secondes montre que ce point est bien un maximum.

Limites de la régression linéaire en grande dimension

$$\widehat{\beta} = (X^T X)^{-1} X^T Y$$

1. $n < p \Rightarrow X^T X$ non inversible
2. colinéarité $\Rightarrow X^T X$ non inversible

Très souvent le cas pour les données "omiques".

Idée de la régression Partial Least Squares

Trouver successivement des variables latentes (ou composantes), combinaisons linéaires des colonnes de X, orthogonales deux à deux, expliquant au mieux Y.

"Expliquant au mieux" = maximisation de la covariance entre la variable latente et Y. Soit donc

$$\max_u Y^T X u,$$

u est appelé le vecteur de *poids* ou *weight* en anglais, que l'on contraint à être de norme 1 : $||u||_2^2 = 1$.

Centrage et réduction

- Les colonnes de X sont nécessairement centrées afin de ne pas donner plus de poids à certaines variables

- Également recommandé (et usuel) de réduire les colonnes de X:
  - ⊕ homogénéise les variables explicatives
  - ⊕ avantage les variables explicatives à forte variabilité (lors de la sélection de leurs scores pour les variables latentes – cf. *sparse*)
  - ⊖ donne de la variabilité aux variables qui n'en avait que très peu
    - ↪ amplification du bruit sur ces variables : il est important de **pré-sélectionner des variables explicatives variant suffisamment!**

Notation lagrangienne

Le problème PLS s'écrit :

$$\max_{u|u^T u=1} Y^T X u$$

Afin de résoudre ce problème on adopte la notation **lagrangienne**

Introduire la contrainte dans le problème à maximiser grâce à un coefficient ($\lambda > 0$). La fonction est notée $\mathcal{L}$

Soit $\mathcal{L}(u, \lambda) = u^T X^T Y - \frac{\lambda}{2}(u^T u - 1)$ et ainsi, pour un point critique, les dérivées y sont nulles, noté (u_1, λ_1)

$$\begin{cases} \partial_u \mathcal{L}(u_1, \lambda_1) & = X^T Y - \lambda_1 u_1 = 0 \\ \partial_\lambda \mathcal{L}(u_1, \lambda_1) & = u_1^T u_1 - 1 = 0 \end{cases}$$

$$\begin{cases} \lambda_1 & = ||X^T Y||_2 \\ u_1 & = \dfrac{X^T Y}{||X^T Y||_2} \end{cases}$$

Interprétation

Le *poids* de la première composante PLS est en fait la matrice de covariance $X^T Y$ normalisée : c'est la proportion de chaque variable de X qui permet de reconstruire un maximum d'information à la fois de X et de Y.

Remarque

Le problème de PLS permet de construire une *composante*.
L'objectif initial était de construire des composantes **successives** et **différentes** permettant de décrire l'information commune à X et Y.
Il faut *retirer* l'information de la composante courante pour créer la nouvelle composante.
C'est la **déflation**.

La déflation

On note $t_1 = Xu_1$, retirer l'information de cette variable dans X peut être réalisé en retirant l'information de X projetée sur cette composante à X. On note alors le **projecteur** $\frac{t_1 t_1^T}{t_1^T t_1}$ (fonction qui vérifie qu'appliquée deux fois elle renvoie le même résultat) et on définit X_2 tel que :

$$X_2 = X - \frac{t_1 t_1^T}{t_1^T t_1} X$$

Il suffit maintenant de résoudre de nouveau le problème de PLS afin de trouver la seconde composante.

Remarque

Cette opération est à réitérer jusqu'à avoir construit r composantes, fixé par l'utilisateur.

Algorithme PLS1 – détail

On l'appelle PLS**1**, car Y est univarié. On note $X_1 = X$ et l'algorithme général devient :

- Pour $h \in \{1, \cdots, r\}$:
  - $u_h = \dfrac{X_h^T Y}{||X_h^T Y||_2}$
  - $t_h = X_h u_h$
  - $X_{h+1} = X_h - \dfrac{t_h t_h^T}{\|t_h\|^2} X_h$ (déflation)

Régression PLS

Finalement on régresse Y sur les r variables latentes construites:

$$Y = T\Gamma + \varepsilon \qquad \text{où} \begin{cases} T = \begin{pmatrix} t_1 & \dots & t_r \end{pmatrix} \\ \Gamma = \begin{pmatrix} \gamma_1 \\ \vdots \\ \gamma_r \end{pmatrix} \end{cases}$$

Il n'est pas nécessaire de connaître la méthode permettant d'expliciter Γ.

Algorithme PLS2

On suppose désormais que Y est multivarié : on a $q > 1$ variables à expliquer.

$\Rightarrow$ on cherche les variables latentes $(t_1, \ldots, t_r)$ de X et les variables latentes $(s_1, \ldots, s_r)$ de Y, *i.e.* qui maximisent la covariance entre X et Y :

$$(u_h, v_h) = \underset{(u,v)\,:\,\|u\|=1,\,\|v\|=1}{\operatorname{argmax}} (X_h u)^T Y_h v$$

Différentes approches pour la déflation de Y

- *Regression*

$$\hookrightarrow \qquad Y_{h+1} \leftarrow Y_h - \frac{t_h t_h^T}{\| t_h \|^2} Y_h$$

On déflate Y_h en retranchant la partie expliquée par t_h

- *Canonic*

$$\hookrightarrow \qquad Y_{h+1} \leftarrow Y_h - \frac{s_h s_h^T}{\| s_h \|^2} Y_h$$

On déflate Y_h pour travailler à l'étape suivante orthogonalement à s_h

Algorithme PLS général

- $X_1 = X$ et $Y_1 = Y$
- Pour $h = 1 \dots r$
  - (a) résoudre : $(u_h, v_h) = \underset{\|u\|=1,\ \|v\|=1}{\arg\min} - cov\,(X_h u, Y_h v)$
  - (b) $t_h = X_h u_h$
    $s_h = Y_h v_h$
  - (c) $X_{h+1} = X_h - \dfrac{t_h t_h^T}{\| t_h \|^2} X_h$
  - (d) $Y_{h+1} = \begin{cases} Y_h - \dfrac{t_h t_h^T}{\| t_h \|^2} Y_h & regression \\ Y_h - \dfrac{s_h s_h^T}{\| s_h \|^2} Y_h & canonic \end{cases}$

sparse PLS

$sPLS = PLS + LASSO$

Algorithme sPLS

- $X_1 = X$ et $Y_1 = Y$
- Pour $h = 1 \ldots H$
  - (a) résoudre :
    $$(u_h, v_h) = \underset{\|u\|=1,\ \|v\|=1}{\arg\min} \; - cov\,(Xu, Yv) + p_{\lambda_1}(u) + p_{\lambda_2}(v)$$
  - (b) $t_h = X_h u_h$
    $s_h = Y_h v_h$
  - (c) $$X_{h+1} = X_h - \frac{t_h t_h^T}{\| t_h \|^2} X_h$$
  - (d) $$Y_{h+1} = \begin{cases} Y_h - \dfrac{t_h t_h^T}{\| t_h \|^2} Y_h & regression \\ Y_h - \dfrac{s_h s_h^T}{\| s_h \|^2} Y_h & canonic \end{cases}$$

La pénalisation

On choisit la pénalité suivante :

$$p_\lambda(u) = 2\lambda \sum_{j=1}^{p} |u_j|$$

qui permet de récupérer toutes les bonnes propriétés du LASSO :

- pénalise les loadings avec une norme L1 trop élevée
- met à zéro des coordonnées $\Rightarrow$ sélection de variables intervenant dans les variables latentes

sparse Partial Least Squares - Discriminant Analysis

Problèmes de discrimination (*i.e.* classification supervisée)

Extension de la (s)PLS où la réponse, qui est au départ un vecteur qualitatif est recodé dans une matrice de 0 et de 1 où chaque colonne correspond à la variable indicatrice de chaque catégorie.

$$\underline{\text{Ex :}} \qquad y = \begin{pmatrix} A \\ B \\ A \\ A \\ C \\ B \\ \vdots \end{pmatrix} \qquad \Rightarrow \qquad Y = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \\ \vdots & \vdots & \vdots \end{pmatrix}$$

(s)PLS est ensuite utilisée comme si Y était continue.

Package R `mixOmics`

González I., Lé Cao K.-A. and Déjean S. *mixOmics : Integrate Omics data project*, 2011.

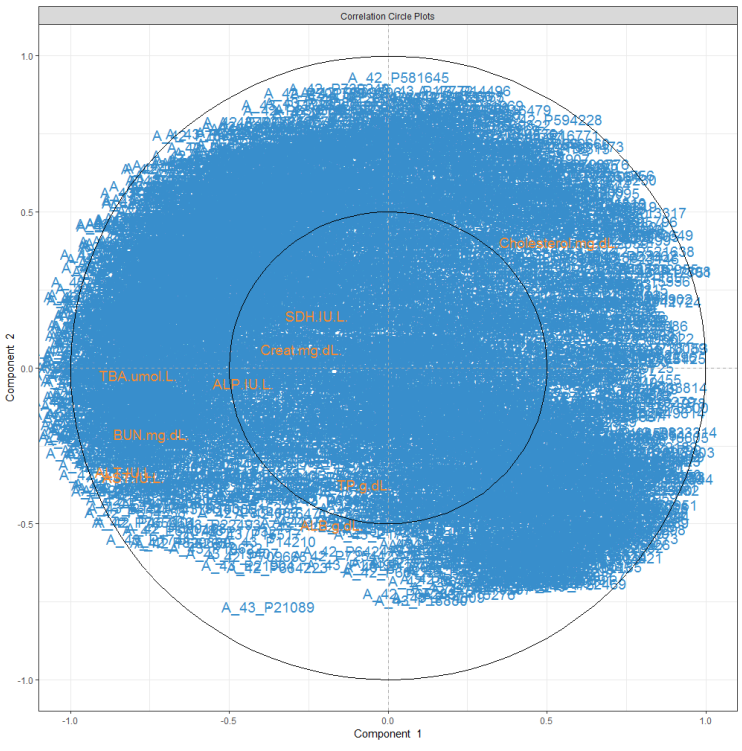
- PLS
- sPLS
- (s)PLS-DA
- graphiques

Une analyse de régression via PLS

On étudie le dataset `liver.toxicity`, voir mixOmics. Dans ce dataset on peut extraire 10 variables Y continues. Il y a 64 individus décrites au travers de 3116 variables.
On construit 2 composantes. On obtient donc :

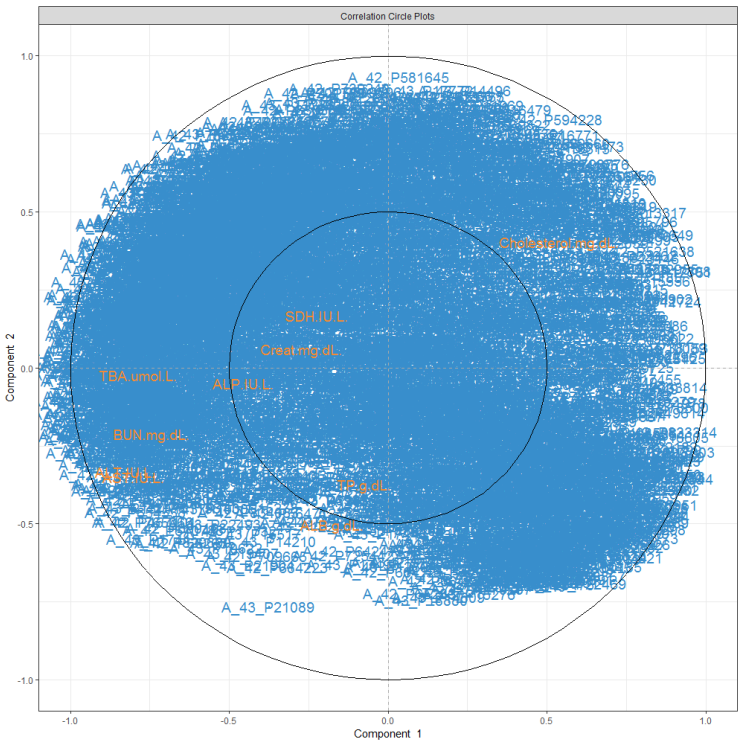
- Les poids u_1 et u_2 pour X et v_1 et v_2 pour Y : Montrent l'importance de chaque variable. Chaque coefficient est inf à 1 en valeur absolue.
- Les composantes t_1 et t_2 pour X et s_1 et s_2 pour Y : Montrent les positions des individus pour chaque composante. Permet d'interpréter quels individus guident la composante à prendre cette forme.

Une analyse de régression via PLS



Un problème ?

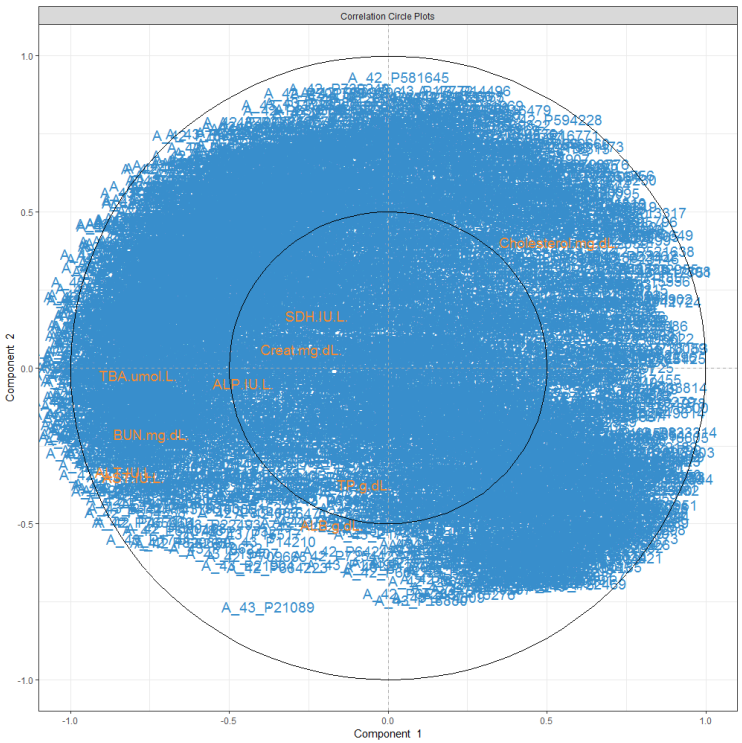
Une analyse de régression via PLS



Un problème ?

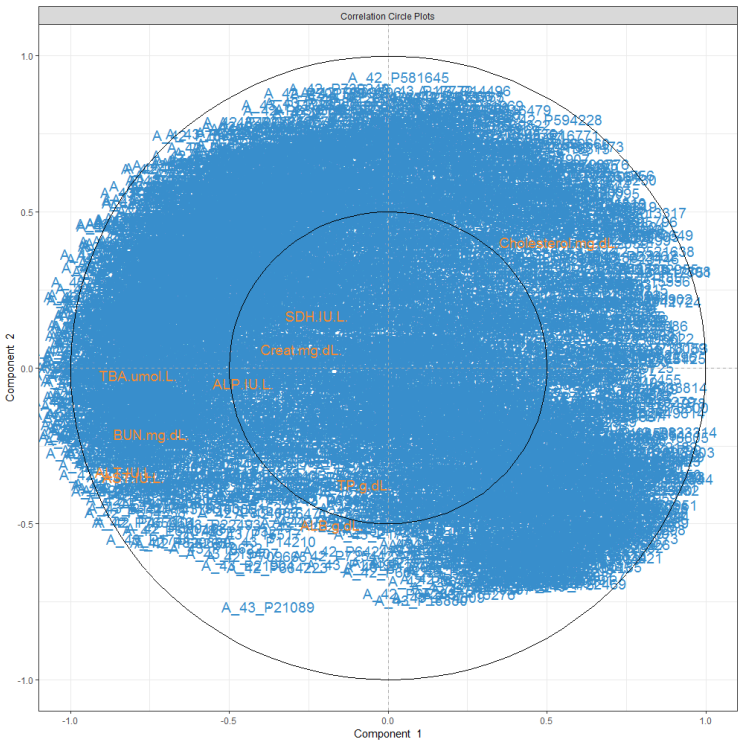
On ne voit pas quelles variables de X sont intéressantes

Une analyse de régression via PLS



Une solution ?

Une analyse de régression via PLS

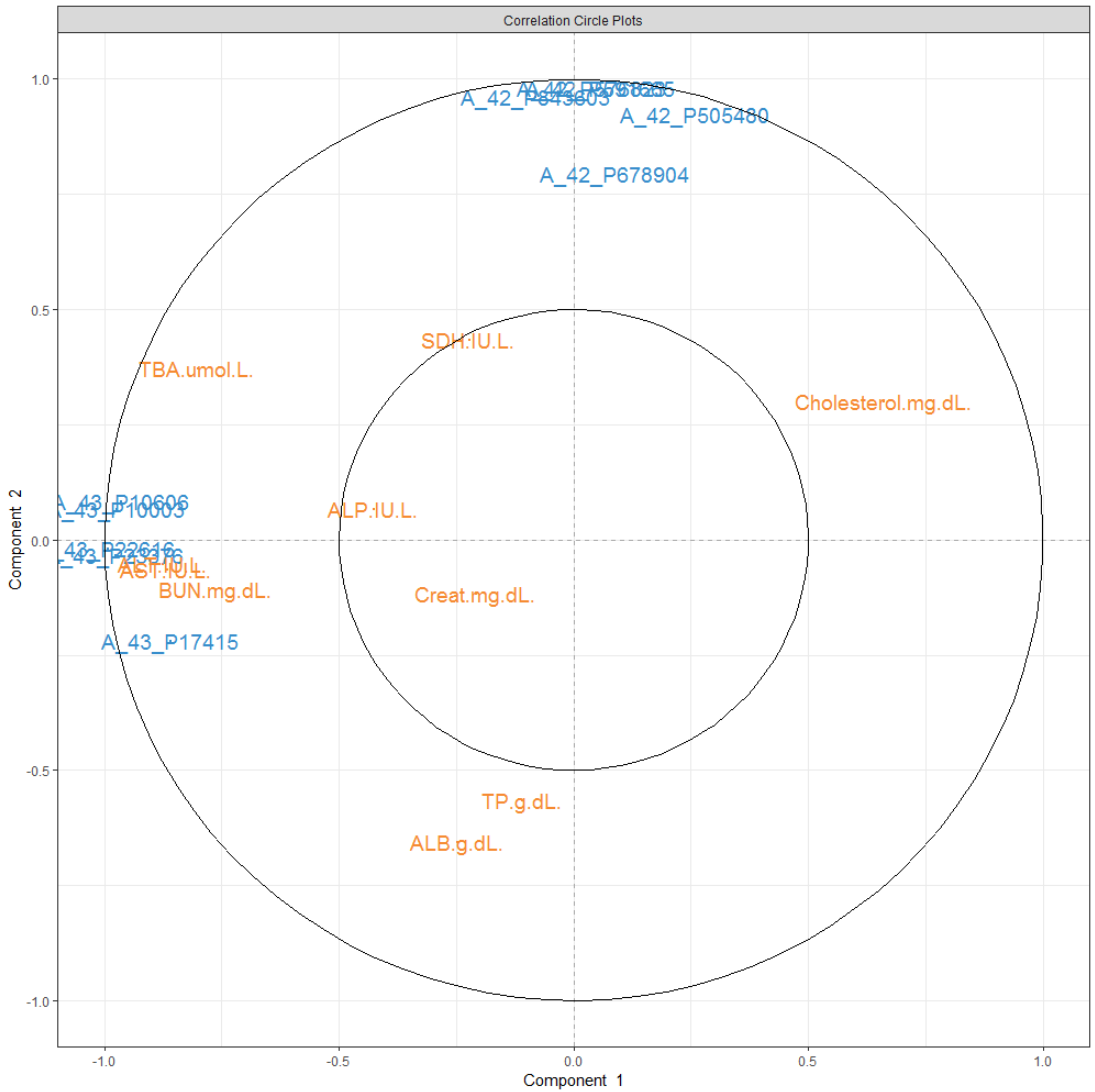


Une solution ?

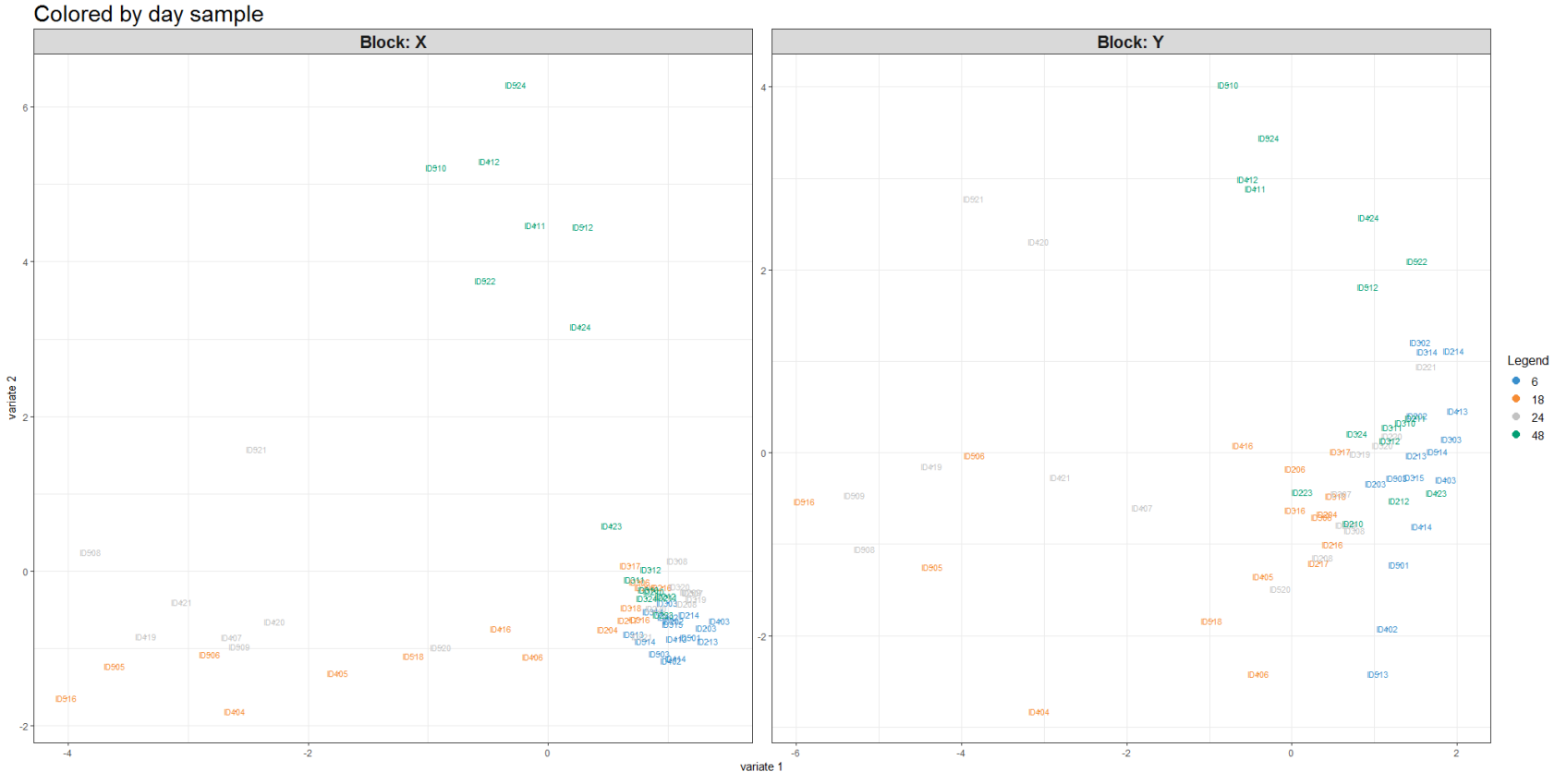
Passer à un modèle **parcimonieux** → sPLS

Recours à la sPLS - plot variables

On conserve 5 variables par composante

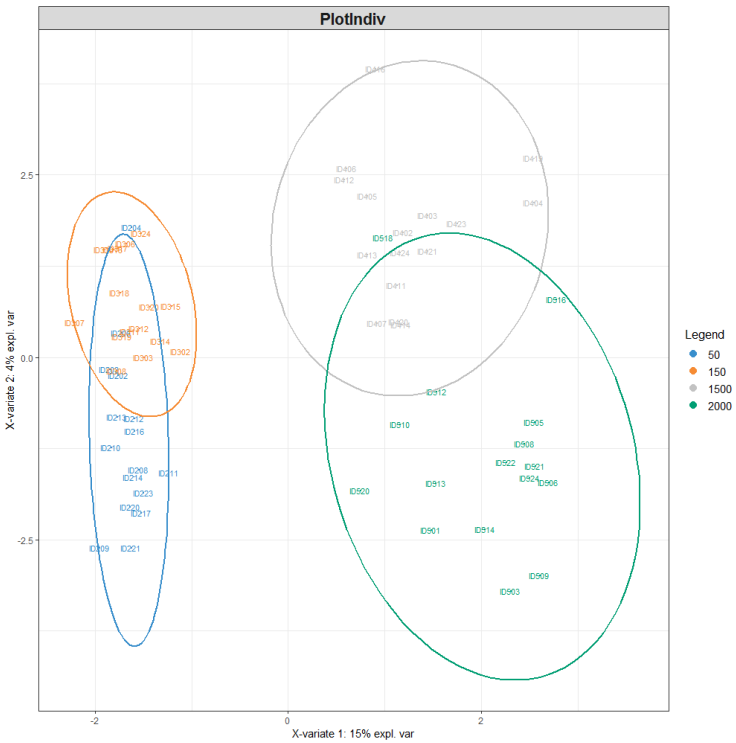


Recours à la sPLS - plot individus



sPLS-DA

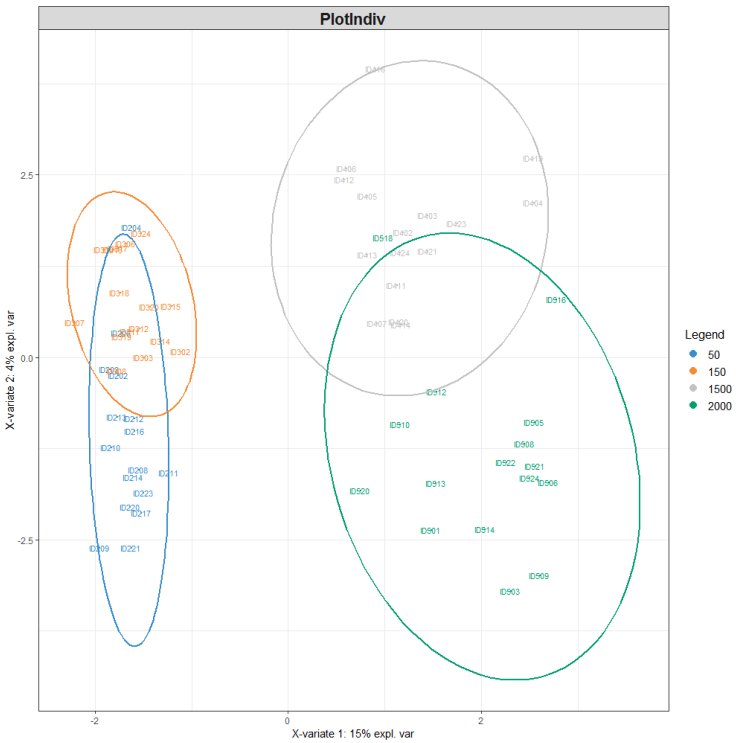
Toujours 5 variables par composante. On explique les groupes de vaccination.



Que constatez-vous ?

sPLS-DA

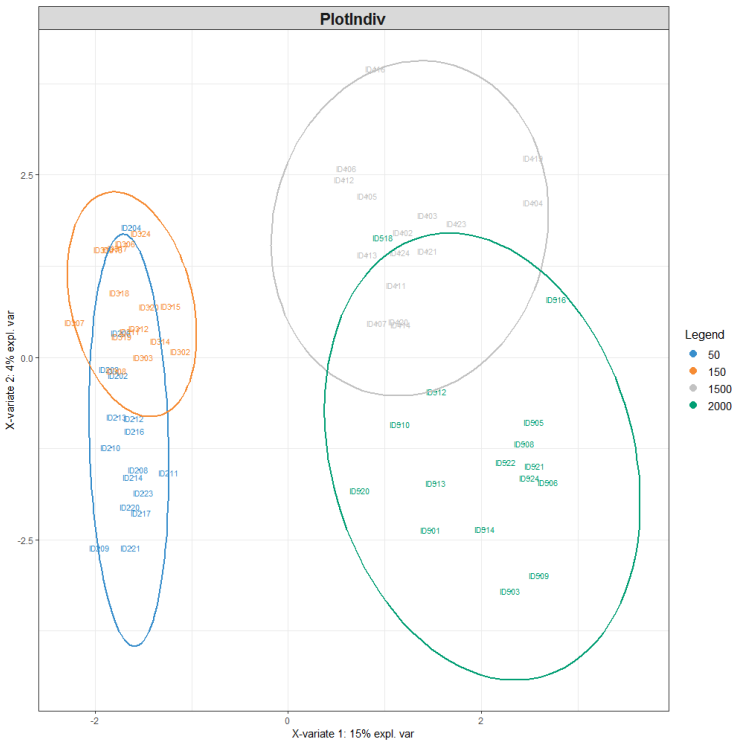
Toujours 5 variables par composante. On explique les groupes de vaccination.



Que constatez-vous ?
Chaque axe permet de discriminer deux ensembles de groupes par rapport à deux autres.

sPLS-DA

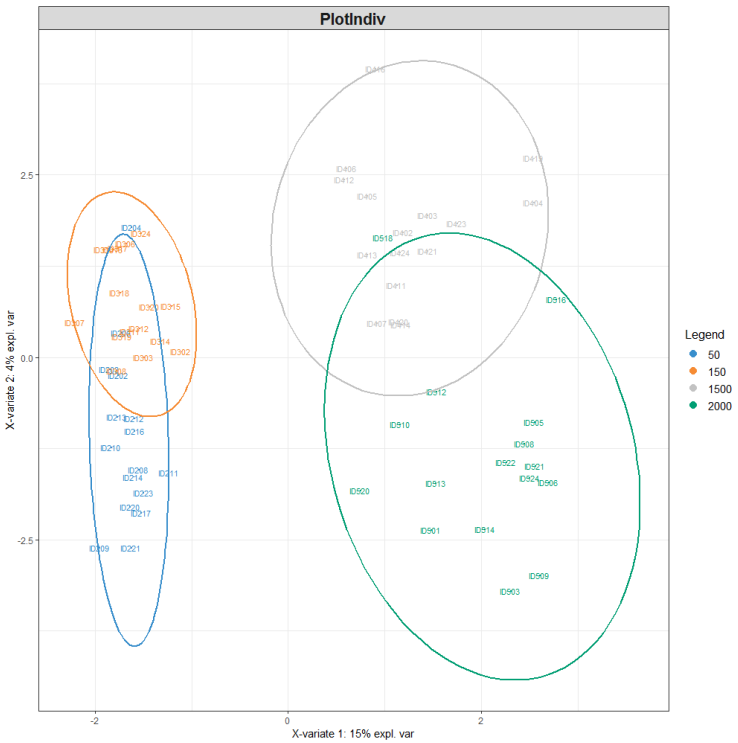
Toujours 5 variables par composante. On explique les groupes de vaccination.



Mais encore ?

sPLS-DA

Toujours 5 variables par composante. On explique les groupes de vaccination.

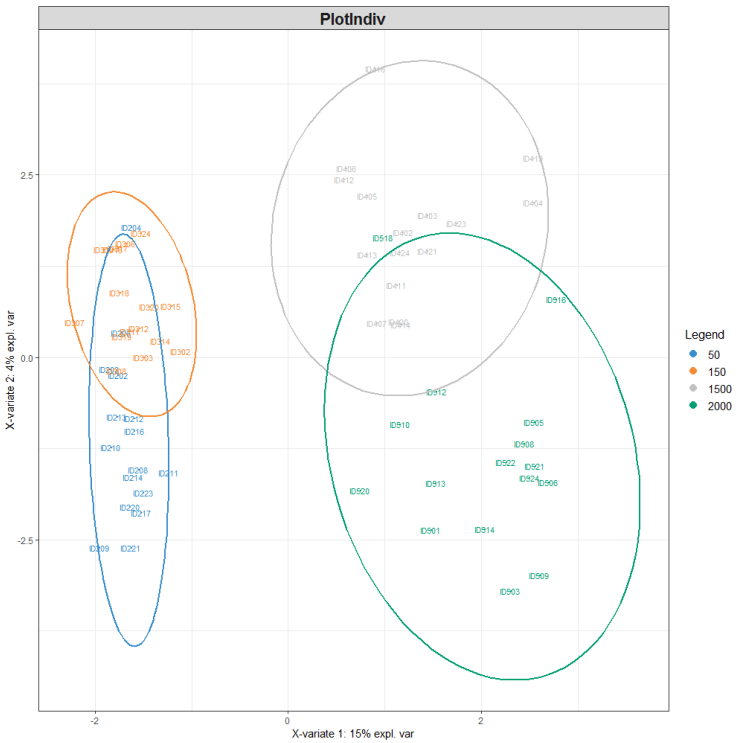


Mais encore ?

Le première axe discrimine parfaitement mais le suivant moins bien.

sPLS-DA

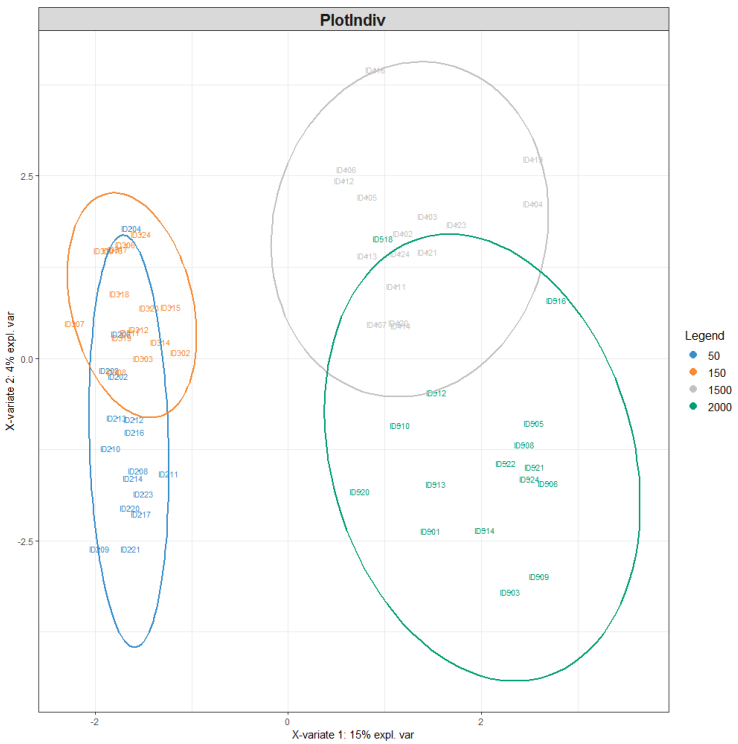
Toujours 5 variables par composante. On explique les groupes de vaccination.



Comment résoudre ce problème ?

sPLS-DA

Toujours 5 variables par composante. On explique les groupes de vaccination.

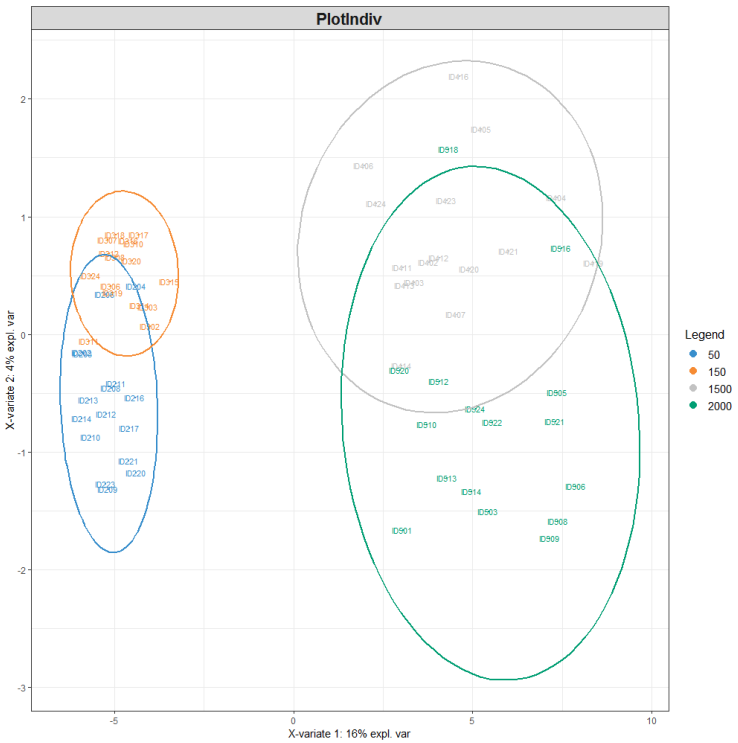


Comment résoudre ce problème ?

Modifier le nombre de gènes à conserver sur chaque composante.

sPLS-DA, validation croisée

On cherche le nombre de variables sur chaque axe permettant de minimiser l'erreur en validation croisée



Ceci pour $keepX_1 = 50$ et $keepX_2 = 2$

Remarques concernant la sPLS

Avantages :

- prend en compte un Y multivarié
- sélection de variables
- réduction de la dimension : permet de faire des graphiques faciles à lire (on projète sur 2 ou 3 axes)

Inconvénients :

- "beaucoup" de paramètres à régler (nombre de variables latentes (r), nombres de variables intervenant dans chacune de ces variables)
- méthode très "linéaire"

Références

- V. Esposito Vinzi *et al.*
  Handbook of Partial Least Squares, Springer, 2010
- R. Tibshirani
  Regression shrinkage and selection via the lasso,
  Journal of the Royal Statistical Society: Series B (Methodological), 58:267–288, 1996.
- J. Friedman *et al.*,
  Pathwise coordinate optimization
  The Annals of Applied Statistics, 1: 302–332, 2007.
- K.-A. Lé Cao *et al.*
  A Sparse PLS for Variable Selection when Integrating Omics Data,
  Statistical Applications in Genetics and Molecular Biology, 7(1):35, 2008.
- H. Shen and J. Z. Huang
  Sparse Principal Component Analysis via Regularized Low Rank Matrix Approximation,
  Journal of Multivariate Analysis, 99:1015-1034, 2008.
- K.-A. Lé Cao *et al.*
  Sparse Canonical Methods for Biological Data Integration: application to a cross-platform study,
  BMC Bioinformatics, 10:34, 2009.
- González I. *et al.*
  mixOmics : Integrate Omics data project,
  2011, URL: `http://www.math.univ-toulouse.fr/~biostat/mixOmics`.