Partial Least Squares, an introduction.

Hadrien Lorenzo

hadrien.lorenzo@inria.fr

DU BDSI, ENSC

23 mars 2022

Rappel sur l'analyse supervisée, cas de la régression

Le modèle linéaire multiple

$$Y = b^t X + \varepsilon,$$

où

- X est une variable aléatoire p-dimensionelle,
- Y est une variable aléatoire réelle (se généralise à q dimensions),
- b est le vecteur des coefficients de régression,
- ε est l'erreur du modèle.

Rappel sur l'analyse supervisée, cas de la régression

Recueil d'un jeu de données $\mathcal{D}_n$

On suppose un jeu de données de n observations indépendantes et identiquement distribuées $(X_i, Y_i)_{i=1..n}$.
On note $\mathbf{X} = (X_1^t, \ldots, X_n^t)^t \in \mathbb{R}^{n \times p}$ et $\mathbf{Y} = (Y_1, \ldots, Y_n)^t \in \mathbb{R}^n$.
On suppose aussi que les colonnes de $\mathbf{X}$ et $\mathbf{Y}$ sont centrées.

Idée

n doit être assez important pour *remonter* au modèle caché et représenté par b.
Notion d'estimateur de b, noté $\hat{b}$ par exemple.

Rappel sur l'analyse supervisée, cas de la régression

Estimateur des moindres carrés ordinaires (MCO, *OLS in English*)

On recherche à estimer b en considérant que l'erreur du modèle, selon une certaine norme $\|.\|_2$ normée norme euclidienne par exemple, doit être minimale. Plus précisément :

$$\forall z \in \mathbb{R}^n,\ \|z\|_2 = \sqrt{\sum_{i=1}^{n} |z_i|^2},$$

et

$$\hat{b} = \arg\min_{b \in \mathbb{R}^p} \|\mathbf{Y} - \mathbf{X}b\|_2^2 .$$

Rappel sur l'analyse supervisée, cas de la régression

Résolution de l'estimateur des moindres carrés ordinaires

Résolution, avec $\mathcal{L}(b) = 1/2 \left\| \mathbf{Y} - b^t \mathbf{X} \right\|_2^2$, doublement continûment différentiable :

$$\nabla \mathcal{L}(b) = \mathbf{X}^t (\mathbf{Y} - \mathbf{X} b),$$

qui s'annule en

$$\hat{b} = (\mathbf{X}^t \mathbf{X})^{-1} \mathbf{X}^t \mathbf{Y},$$

défini uniquement si $\mathbf{X}^t \mathbf{X}$ est inversible...

La grande dimension

Qu'est-ce que c'est ?

On rassemble sous cette appellation l'ensemble des contextes tels que :

- $n < p$ et alors $\mathbf{X}^t\mathbf{X}$ n'est pas de rang plein,
- certaines variables soient colinéaires.

Observation

On note la matrice de covariance **empirique** de X

$$\frac{\mathbf{X}^t\mathbf{X}}{n}.$$

On remarque que les MCO sont accessibles si la matrice de covariance empirique de X est inversible... problème numérique.

La grande dimension

Solutions ?

Ne pas rechercher la solution des MCO mais une solution approchée qui soit accessible.
On **régularise** le problème (Ridge, Lasso, etc...).
On peut aussi choisir de projeter l'information sur des sous-espaces plus ou moins bien choisis, appelés **scores** et symbolisés par t, et ensuite faire la régression sur ces scores.
Historiquement, il y avait la méthode PCR, Principal Component Regression, où les scores sont les R premières composantes de l'ACP.
Problème : Si les R premières composantes ne contiennent pas l'information associée à Y, comment faire ?

PLS!

Idée de la PLS

Construire des scores directement associés avec l'information recherchée, l'information à prédire. On recherche une direction, il faut donc contraindre le problème

Problème d'optimisation sous contrainte associé

$$\max_{w\in\mathbb{R}^p \text{ tel que } w^t w=1} \operatorname{cov}(Y, w^t X),$$

mais $\mathbf{X}$ et $\mathbf{Y}$ sont à colonnes centrées, donc l'estimateur empirique de la covariance s'écrit, sur $\mathcal{D}_n$

$$\max_{w\in\mathbb{R}^p \text{ tel que } w^t w=1} \mathbf{Y}^t\mathbf{X}w,$$

w est appelé **poids/weights** et il y en a un par composante. On a alors

$$t = \mathbf{X}w.$$

Résolution de la PLS

Résolution d'un problème d'optimisation sous contrainte

Soit le lagrangien

$$\mathcal{L}(w, \lambda) = \mathbf{Y}^t\mathbf{X}w - \lambda/2(w^t w - 1),$$

pour $\lambda > 0$ et ainsi par annulation de la fonctionnelle précédente, il vient par annulation du gradient

$$\mathbf{X}^t\mathbf{Y} = \lambda w \text{ et } w^t w = 1 \text{et donc } w = \frac{\mathbf{X}^t\mathbf{Y}}{\|\mathbf{X}^t\mathbf{Y}\|_2} \text{ et } \lambda = \|\mathbf{X}^t\mathbf{Y}\|_2.$$

Interprétations

Le poids/weight w

ATTENTION : Interprétable si les variables ont une variance égale.
C'est l'impact de chaque variable sur la composante courante.
Plus la valeur absolue est importante et plus la variable joue sur la composante.

Le score $t = \mathbf{X}w$

C'est la position de chaque observation sur la composante courante.
Plus la valeur absolue est importante et plus l'observation drive sur la composante.
On peut regarder la dispersion des observations pour savoir si un individu est exotique.

Interprétations 2

Le loading $p = \mathbf{X}^t t/(t^t t)$

ATTENTION : Interprétable si les variables ont une variance unitaire.

C'est la corrélation de chaque variable avec la composante courante.

Plus la valeur absolue est importante et plus la composante est corrélée avec la variable courante.

Observation générale

Les variables de la PLS t et p sont interprétables si les variables de $\mathbf{X}$ sont standardisées, soit

$$\forall j \in [\![1, p]\!], \ \frac{\mathbf{X}_j^t \mathbf{X}_j}{n} = 1$$

Construction des composantes suivantes, la **déflation**

Rendre le jeu de données indépendant de t

La composante créée ne décrit qu'une source de variabilité, il en existe souvent plusieurs dans les problèmes multivariés.
Pour décrire les sources suivantes, il faut **retirer** la variabilité décrite par t du jeu de données.
Pour cela on projette $\mathbf{X}$ sur le sous-espace engendré par t :

$$\mathbf{X}_2 = \mathbf{X} - \frac{tt^t}{t^t t}\mathbf{X} = \mathbf{X} - tp^t.$$

En effet

$$\begin{aligned} t^t\mathbf{X}_2 = t^t(\mathbf{X} - \frac{tt^t}{t^t t}\mathbf{X}) &= t^t\mathbf{X} - \frac{t^t tt^t}{t^t t}\mathbf{X}, \\ &= t^t\mathbf{X} - \frac{t^t t}{t^t t}t^t\mathbf{X} = 0. \end{aligned}$$

Notations mises à jour

On note alors w_1, t_1 et p_1 les variables associées à la première composante ainsi que $\mathbf{X}_1 = \mathbf{X}$.
pour chaque composante $r \in [\![2, R]\!]$, on note

$$
\begin{aligned}
w_r &= \frac{\mathbf{X}_{r-1}^t \mathbf{Y}}{\|\mathbf{X}_{r-1}^t \mathbf{Y}\|_2}, \\
t_r &= \mathbf{X}_{r-1} w_r, \\
p_r &= \mathbf{X}_{r-1}^t t_r / (t_r^t t_r), \\
\mathbf{X}_r &= \mathbf{X}_{r-1} - \frac{t_r t_r^t}{t_r^t t_r} \mathbf{X}_{r-1}.
\end{aligned}
$$

On note alors $\mathbf{W}$, $\mathbf{T}$, $\mathbf{P}$ les matrices à R colonnes issues des concaténations des vecteurs associés.

Oui mais *quid* de la régression ?

Le problème initiale veut prédire Y à partir de X via le vecteur de régression b, pour rappel le modèle de régressiojn linéaire s'écrit

$$Y = b^t X + \varepsilon.$$

Dans le cas de la PLS, on remarque que

$$\begin{aligned} t_r &= \mathbf{X}_{r-1} w_r, \\ &= (\mathbf{X}_{r-2} - \frac{t_{r-1} t_{r-1}^t}{t_{r-1}^t t_{r-1}} \mathbf{X}_{r-2}) w_r, \\ &= \ldots, \\ &= \mathbf{X} w_r^\star. \end{aligned}$$

où l'on peut montrer que $\mathbf{W}^\star = (w_1^{\star t}, \ldots, w_R^{\star t})^t = \mathbf{W}(\mathbf{P}^t\mathbf{W})^{-1}$.

Régression PLS, l'estimateur PLS

Ainsi en régressant $\mathbf{Y}$ sur $\mathbf{T} = \mathbf{XW}^\star$, on régresse indirectement sur $\mathbf{X}$ avec

$$\hat{b} = \mathbf{W}^\star \mathbf{C}^t,$$

où $c_r = \dfrac{\mathbf{Y}^t t_r}{t_r^t t_r}$ et $\mathbf{C} = (c_1, \ldots, c_R)^t$.

La force de la PLS, gérer des Y multivariés!

Si maintenant Y est q-dimensionnel, on fait presque pareil!
On cherche R directions telles que $\mathbf{X}$ projetée sur w et $\mathbf{Y}$ projetée sur v, donnant les scores t et s respectivement, soient de covariance maximale.

Le problème d'optimisation par composante

$$\max_{(w,v)\in\mathbb{R}^p\times\mathbb{R}^q \text{ tel que } w^tw=v^tv=1} \operatorname{cov}(v^tY, w^tX),$$

et l'estimateur empirique de la covariance s'écrit, sur $\mathcal{D}_n$

$$\max_{(w,v)\in\mathbb{R}^p\times\mathbb{R}^q \text{ tel que } w^tw=v^tv=1} v^t\mathbf{Y}^t\mathbf{X}w,$$

Les déflations

La déflation sur $\mathbf{X}$ reste inchangée mais il est maintenant nécessaire de déflater aussi $\mathbf{Y}$. On peut vouloir retirer de $\mathbf{Y}$ l'information

- portée par $\mathbf{X}$ (Régression)

$$\mathbf{Y} \leftarrow \mathbf{Y} - \frac{tt^t}{t^t t}\mathbf{Y}.$$

- Portée par $\mathbf{Y}$ (Canonique)

$$\mathbf{Y} \leftarrow \mathbf{Y} - \frac{ss^t}{s^t s}\mathbf{Y}.$$

Algorithme PLS2

Pour chaque composante $r \in [\![1, R]\!]$, on note $\mathbf{X}_0 = \mathbf{X}$, $\mathbf{Y}_0 = \mathbf{Y}$ et :

$$(w_r, v_r) = \underset{(w,v)\in\mathbb{R}^p\times\mathbb{R}^q \text{ tel que } w^t w = v^t v = 1}{\arg\max} v^t \mathbf{Y}_{r-1}^t \mathbf{X}_{r-1} w,$$

$$t_r = \mathbf{X}_{r-1} w_r,$$

$$s_r = \mathbf{Y}_{r-1} v_r,$$

$$p_r = \mathbf{X}_{r-1}^t t_r / (t_r^t t_r),$$

$$\mathbf{X}_r = \mathbf{X}_{r-1} - \frac{t_r t_r^t}{t_r^t t_r} \mathbf{X}_{r-1}.$$

$$\text{Régression} \begin{cases} c_r = \mathbf{Y}_{r-1}^t t_r / (t_r^t t_r), \\ \mathbf{Y}_r = \mathbf{Y}_{r-1} - \dfrac{t_r t_r^t}{t_r^t t_r} \mathbf{Y}_{r-1} \end{cases}$$

$$\text{Canonique} \begin{cases} c_r = \mathbf{Y}_{r-1}^t s_r / (s_r^t s_r), \\ \mathbf{Y}_r = \mathbf{Y}_{r-1} - \dfrac{s_r s_r^t}{s_r^t s_r} \mathbf{Y}_{r-1} \end{cases}$$

Et l'estimateur de la matrice de régression

On peut réécrire le problème de régression tel que

$$Y = \mathbf{B}^t X + \mathbf{e},$$

où $\mathbf{e}$ est le bruit résiduel et la matrice de régression peut s'écrire

$$\widehat{\mathbf{B}} = \mathbf{W}^\star \mathbf{C}^t,$$

où les grandeurs $\mathbf{W}^\star$ et $\mathbf{C}$ ont déjà été définies.

Choix de R

Pour choisir R, on utilise le principe de généralisabilité d'un modèle. On choisit une valeur pour R et alors :

- construire M jeux de données depuis $\mathcal{D}_n$,
- estimer $\widehat{\mathbf{B}}$ sur chacun des jeux,
- estimer une erreur de prédiction sur des jeux de données indépendant,
- agréger les erreurs.

On compare les erreurs agrégées pour sélectionner le modèle avec la plus faible erreur.

Erreur classique, RMSEP

Root Mean Square Error in Prediction. Soit un jeu d'entraînement de taille n_0 et un jeu de test de taille $n_1 - n_0$

$$RMSEP = \sqrt{\frac{1}{n_1 - n_0} \sum_{i=n_0+1}^{n_1} \|Y_i - \widehat{Y}_i\|^2},$$

où $\widehat{Y}_i = \hat{\mu}_Y + \widehat{B}^t(X_i - \hat{\mu}_X)$, $\hat{\mu}_Y = \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i$ et $\hat{\mu}_X = \frac{1}{n_0}\sum_{i=1}^{n_0} X_i$.

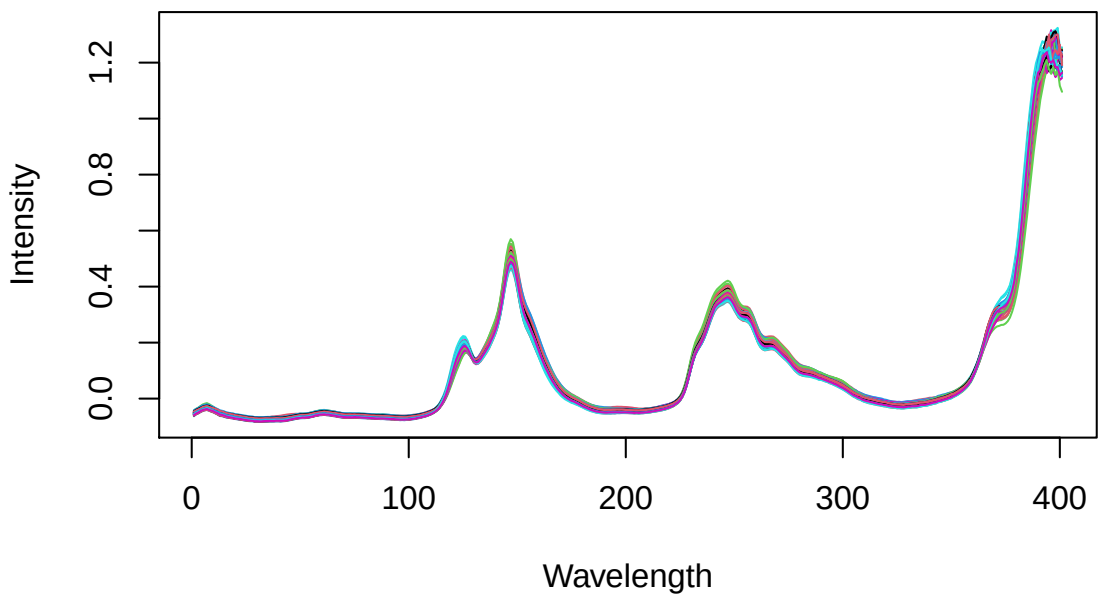
La validation croisée

On divise aléatoirement $\mathcal{D}_n$ en N parties de la même taille.
Chaque partie est le jeu de test associé à son complémentaire dans $\mathcal{D}_n$ qui est le jeu de test.
Il y a donc $N \leqslant n$ erreurs à agréger.
Remarque : Si $N = n$ on parle de *Leave-One-Out* (LOO) et il n'y a pas d'aléa.

Package `pls` Wehrens and Mevik (2007)

Permet de construire des modèles PLS en choisissant *R* par validation croisée. Soit le jeu de données `gasoline` :

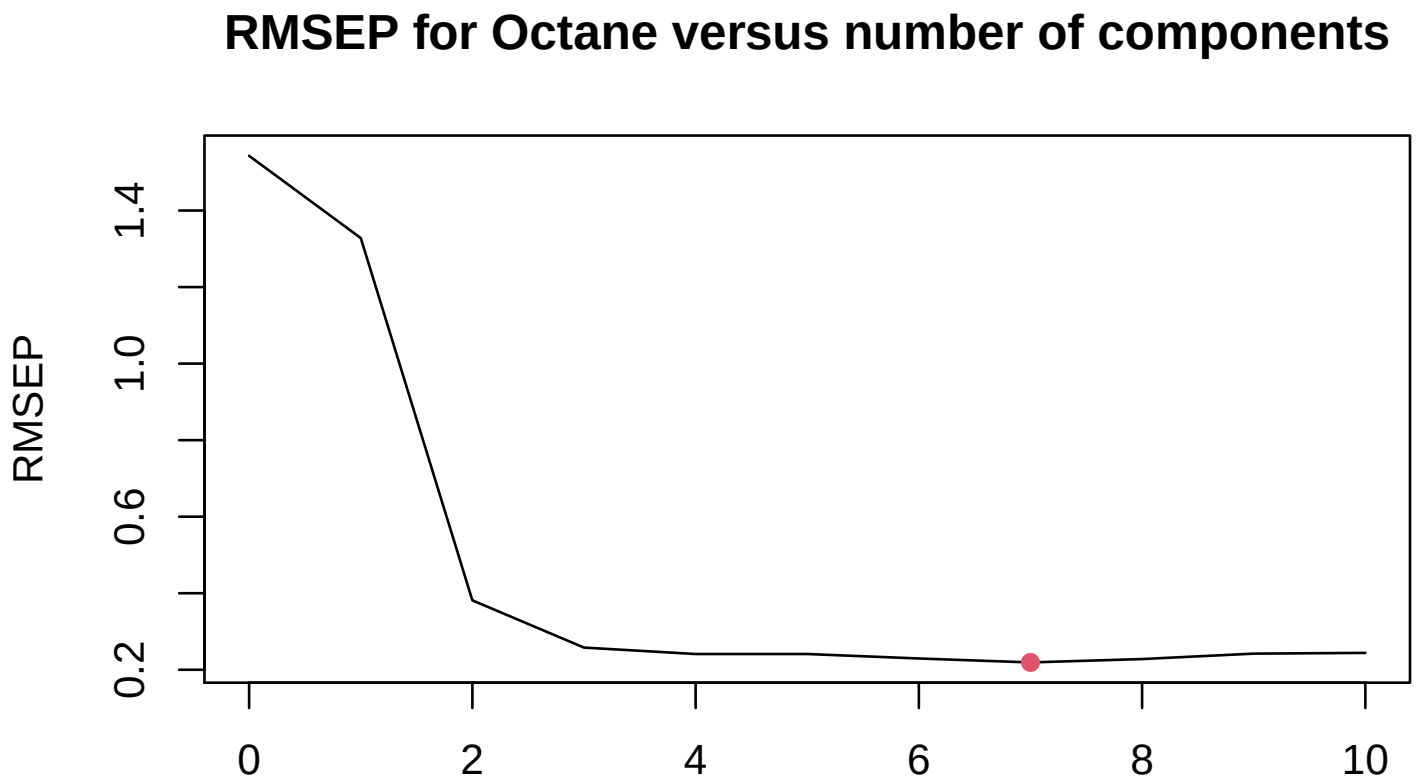
A data set consisting of octane number (octane) and NIR spectra (NIR) of 60 gasoline samples. Each NIR spectrum consists of 401 diffuse reflectance measurements from 900 to 1700 nm.



Analyse

Permet de construire des modèles PLS en choisissant *R* par validation croisée.

```
gas1 <- plsr(octane ~ NIR, ncomp = 10, data = gasoline,
validation = "LOO")
```

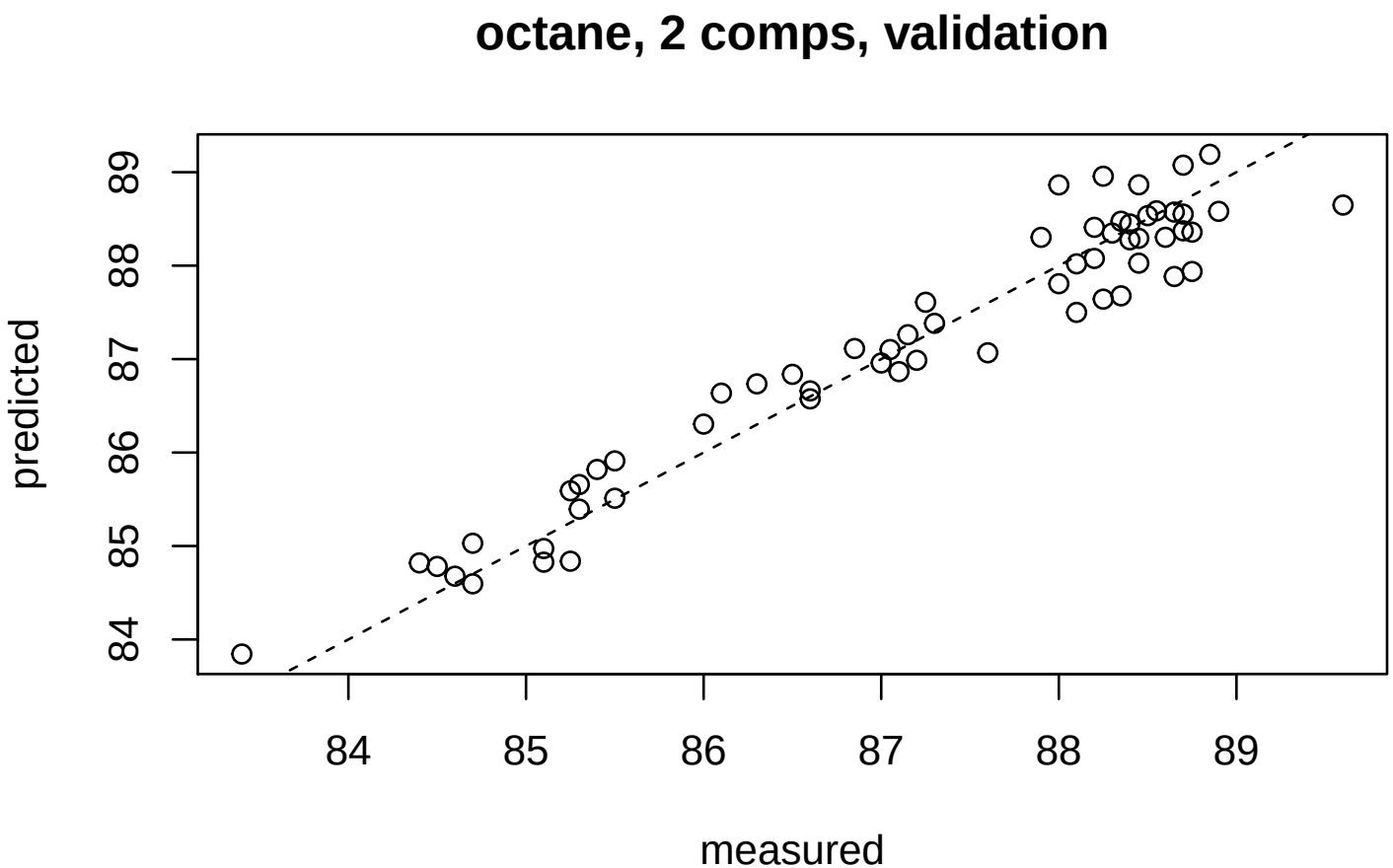


Choix de R

La courbe est minimale en $R = 7$ mais on voit que les estimations ne bougent plus vraiment à partir de $R = 3$. On choisirait $R = 2$ dans la pratique car la troisième composante n'apporte pas grand chose.

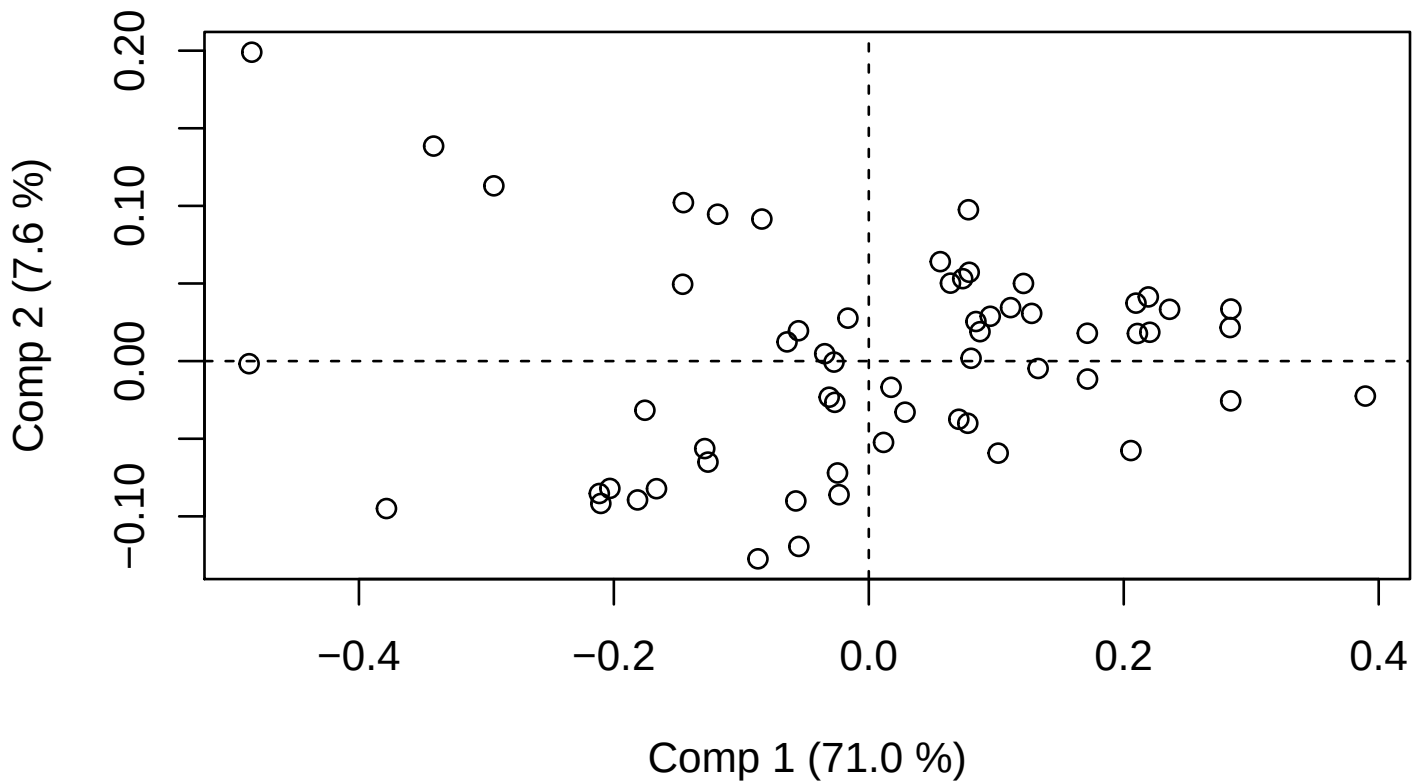
Interprétation des résultats

```
plot(gas1,ncomp=2);abline(a=0,b=1,lty=2)
```



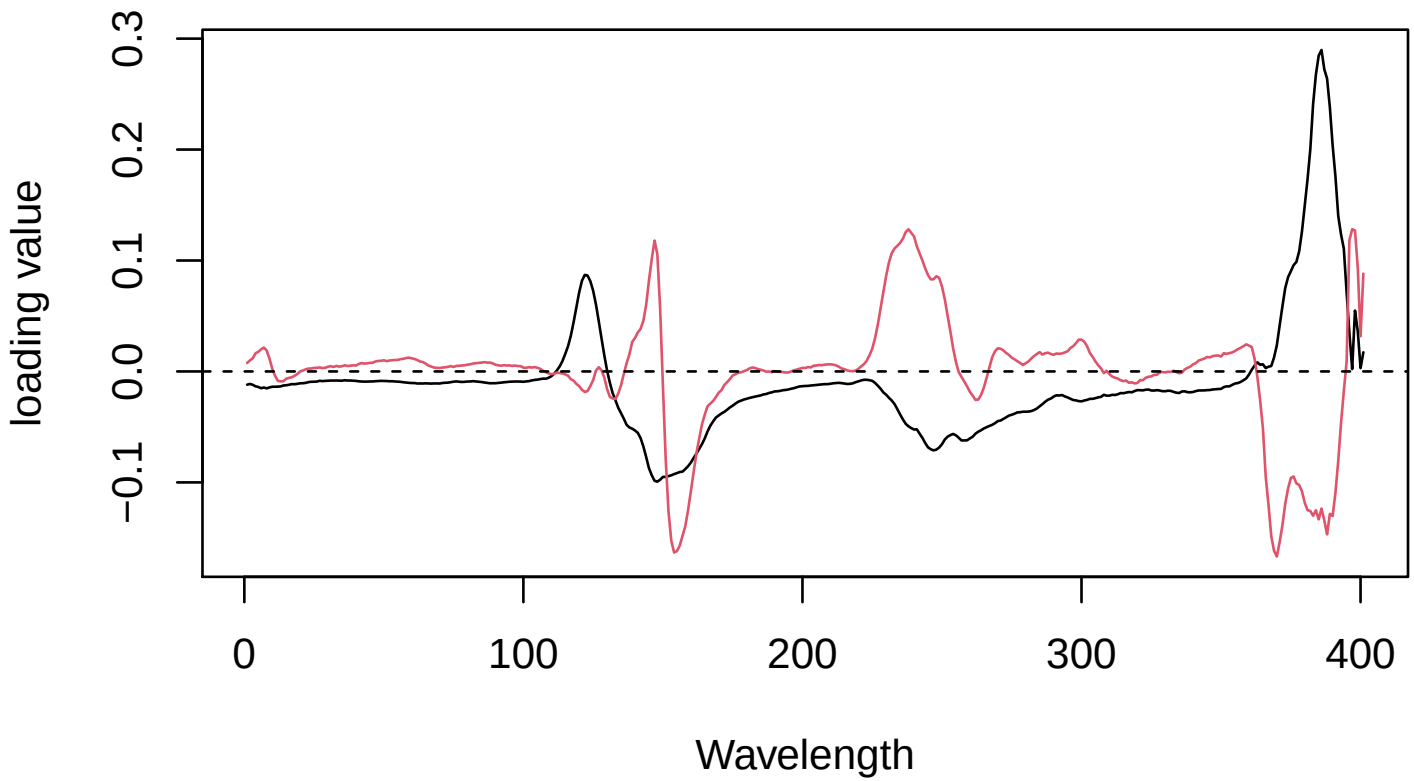
Interprétation des résultats

```
plot(gas1,comps=1:2,plottype="scores")
abline(h=0,v=0,lty=2)
```



Interprétation des résultats

```
plot(gas1,comps=1:2,plottype="loading",lty=1,
xlab="Wavelength")
abline(h=0,lty=2)
```



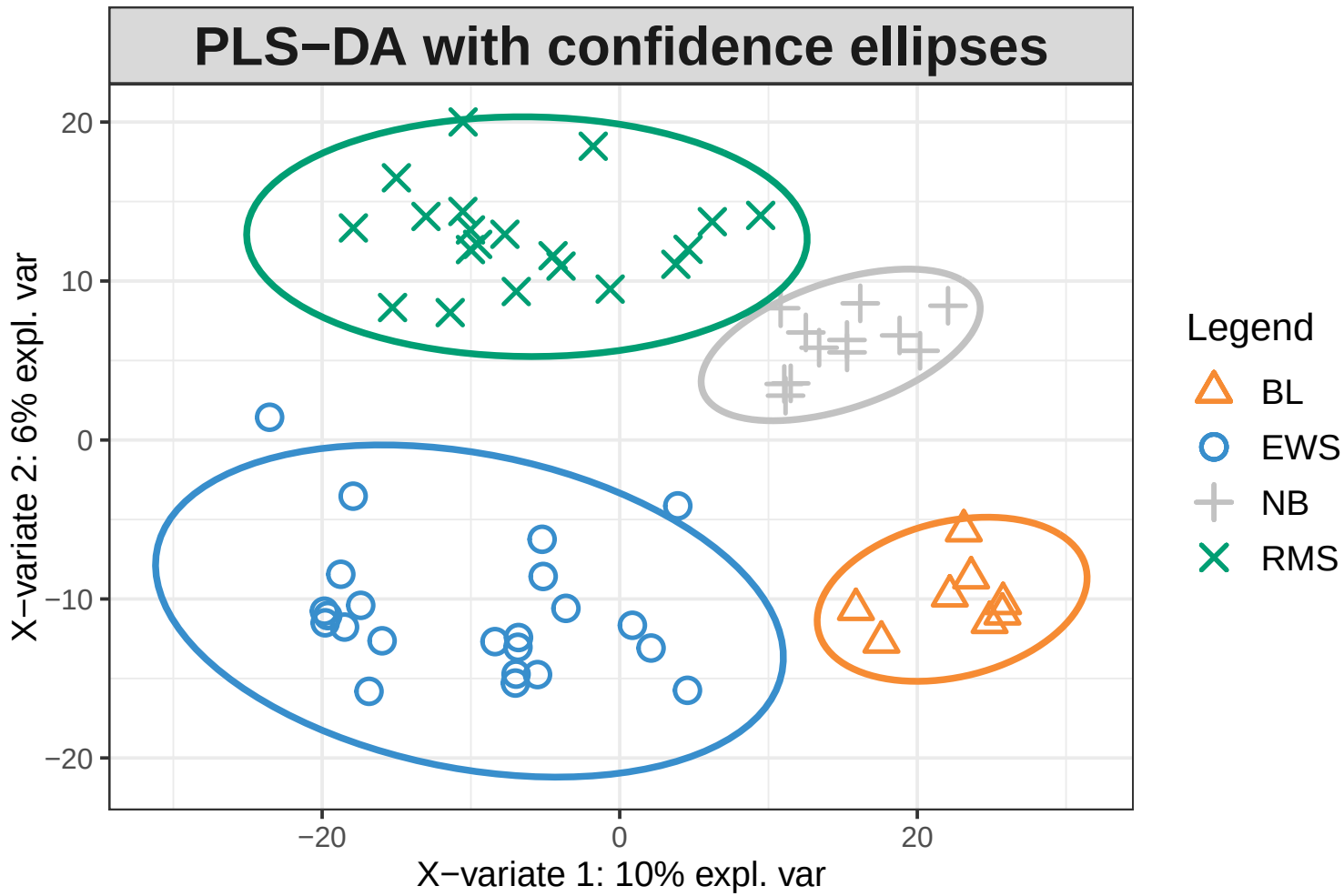
Analyse discriminante

Lorsque Y divisé en K classes non ordonnées.
Recoder Y en **dummies** : une colonne par classe avec des 1 pour les membres de la classe et des 0 ailleurs, voir Pérez-Enciso and Tenenhaus (2003).
Utiliser le package `mixOmics` Rohart et al. (2017).
On choisit souvent $R = K - 1$, mais il faudrait valider...

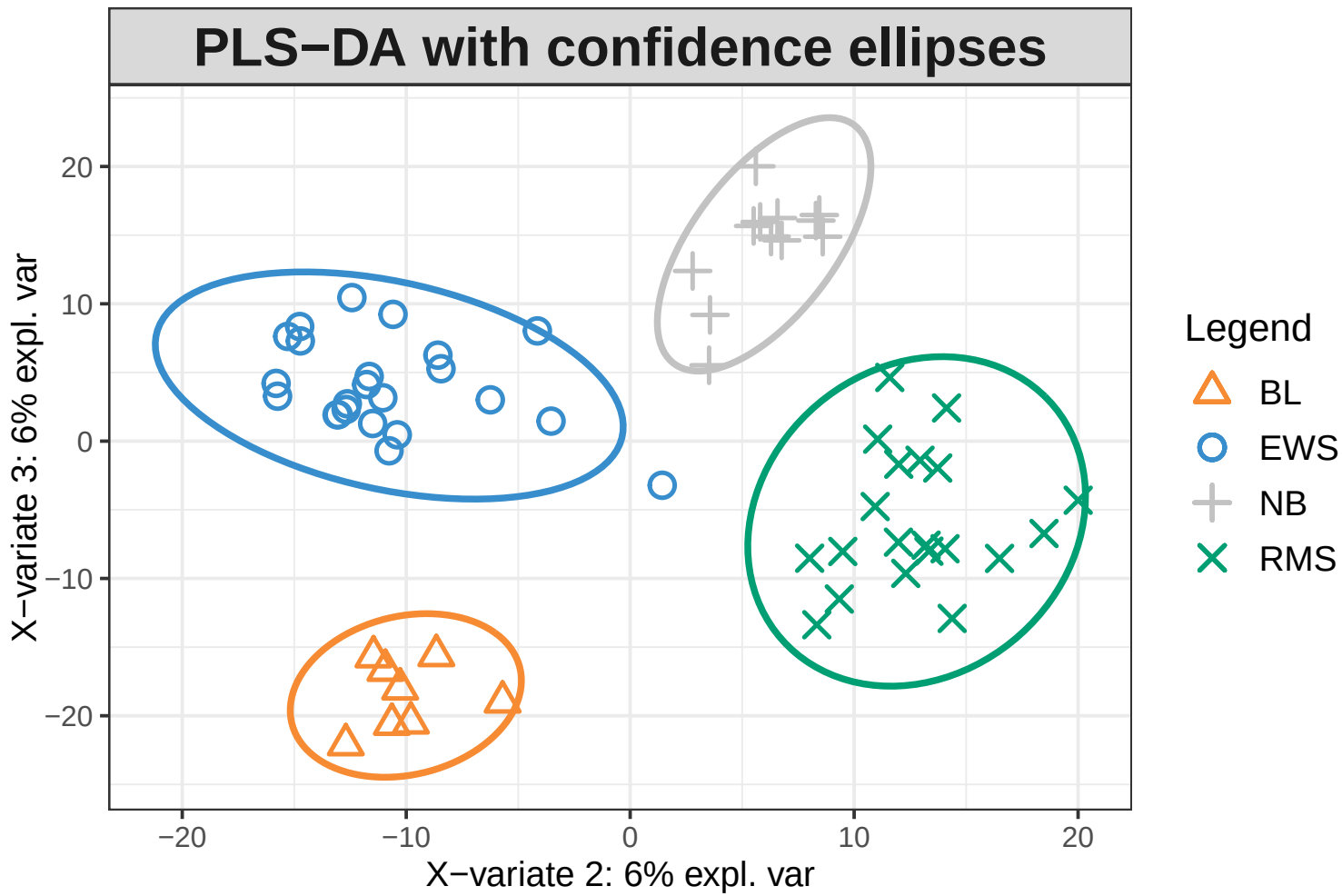
Nouveau jeu de données : Small Round Blue Cell Tumours `srbct`

gene: the expression levels of the 2308 genes across the 63 tested subjects.
class: contains the tumour class of each individual: Burkitt Lymphoma (BL), Ewing Sarcoma (EWS), neuroblastoma (NB) and rhabdomyosarcoma (RMS).

Analyses

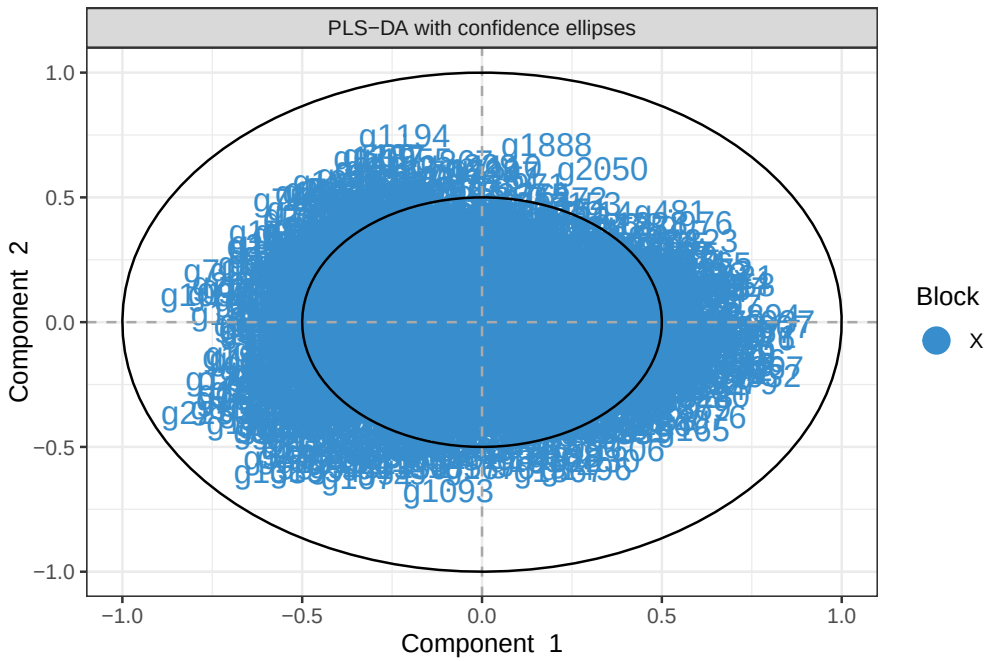

PLS−DA with confidence ellipses
X−variate 2: 6% expl. var
X−variate 1: 10% expl. var
20
10
0
−10
−20
−20
0
20
Legend
BL
EWS
NB
RMS

Analyses


PLS−DA with confidence ellipses
X−variate 3: 6% expl. var
X−variate 2: 6% expl. var
20
10
0
−10
−20
−20
−10
0
10
20
Legend
BL
EWS
NB
RMS

Analyses

Interpretation ?...



...Non

Revue de la littérature

Il existe un moyen de stabiliser + rendre plus interprétable les modèles de prédiction:

La régularisation parcimonieuse (*sparse* in English).

Il en existe plusieurs versions, où $\mathbf{M}_r = \mathbf{Y}_r{}'\mathbf{X}_r/(n-1)$ et $\mathbf{S}^{X,(r)} = \mathbf{X}'\mathbf{Y}_r\mathbf{Y}_r{}'\mathbf{X}$:

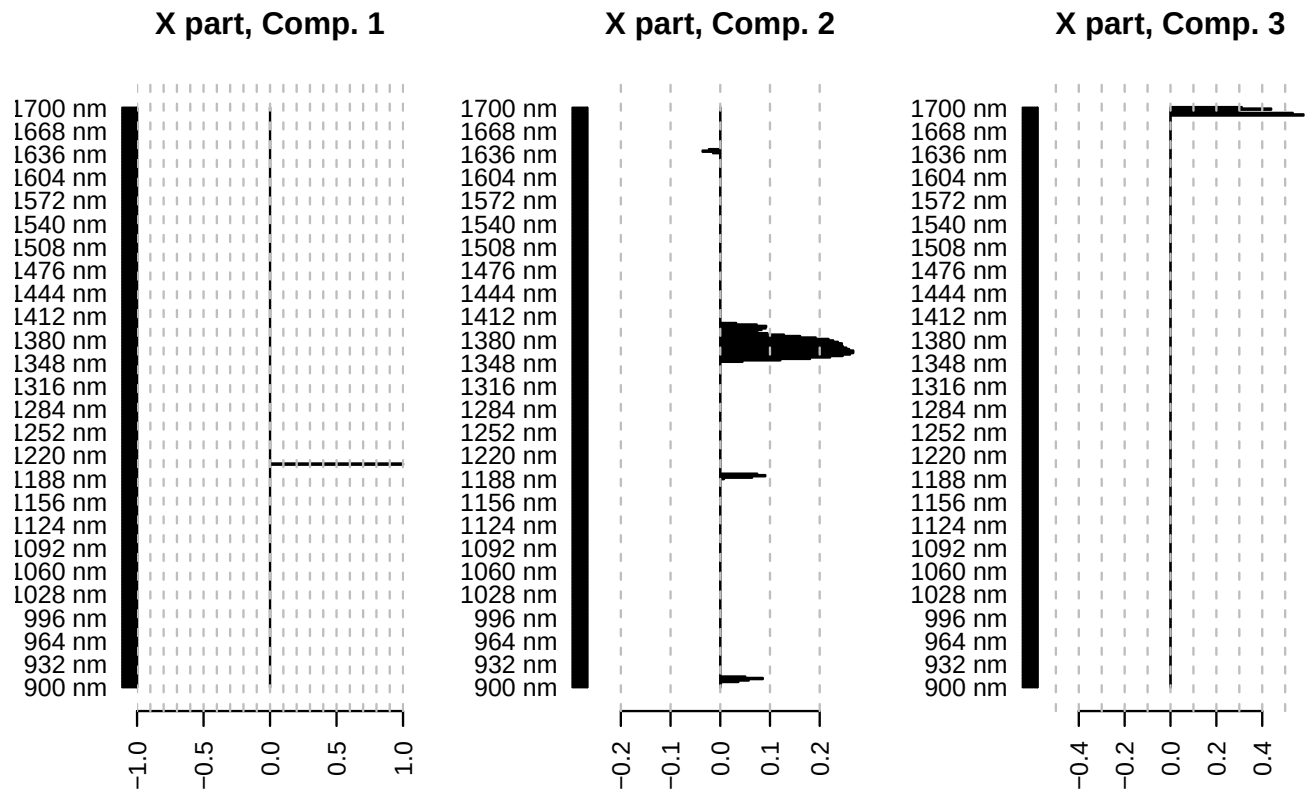
Method	Optimization Problem	Parameters
Wold (1966)	$\mathbf{v}'\mathbf{M}^{(r)}\mathbf{u}$	R
Lê Cao et al. (2008)	$\left\|(n-1)\mathbf{M}^{(r)} - \mathbf{v}\mathbf{u}'\right\|^2 + \lambda_u^{(r)}\|\mathbf{u}\| + \lambda_v^{(r)}\|\mathbf{v}\|$	$R, \left(\lambda_u^{(r)}, \lambda_v^{(r)}\right)_r$
Chun and Keleş (2010)	$-\kappa\mathbf{w}\mathbf{S}^{X,(r)}\mathbf{w} + (1-\kappa)(\mathbf{c}-\mathbf{w})'\mathbf{S}^{X,(r)}(\mathbf{c}-\mathbf{w}) + \lambda_1\|\mathbf{c}\| + \lambda_2\|\mathbf{c}\|$	$R, \kappa, \lambda_1, \lambda_2$
Lorenzo et al. (2022)	$\mathbf{v}' S_{\lambda_r}\left(\mathbf{M}^{(r)\prime}\right)\mathbf{u}$	$R, (\lambda_r)_r$

Inconvénient majeur des méthodes...

Il faut fixer les hyperparamètres.
Idée de **ddsPLS** : les hyperparamètres associés à **X** et **Y** sont liés et leur estimation devrait donc être simultanée.

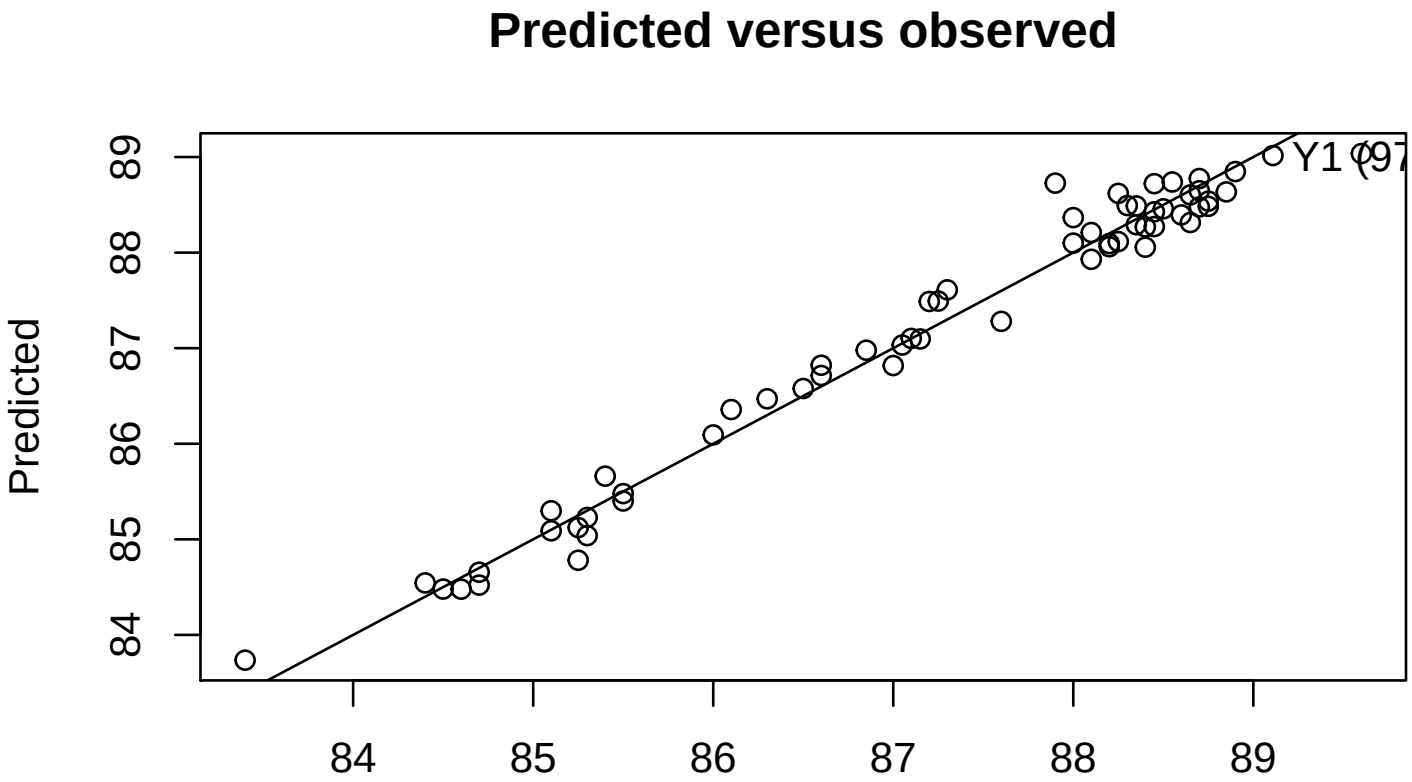
Exemple 1

Le nombre de composantes et les coefficients de régularisation sont fixés automatiquement.
On obtient 3 composantes.
Les **weights** sont



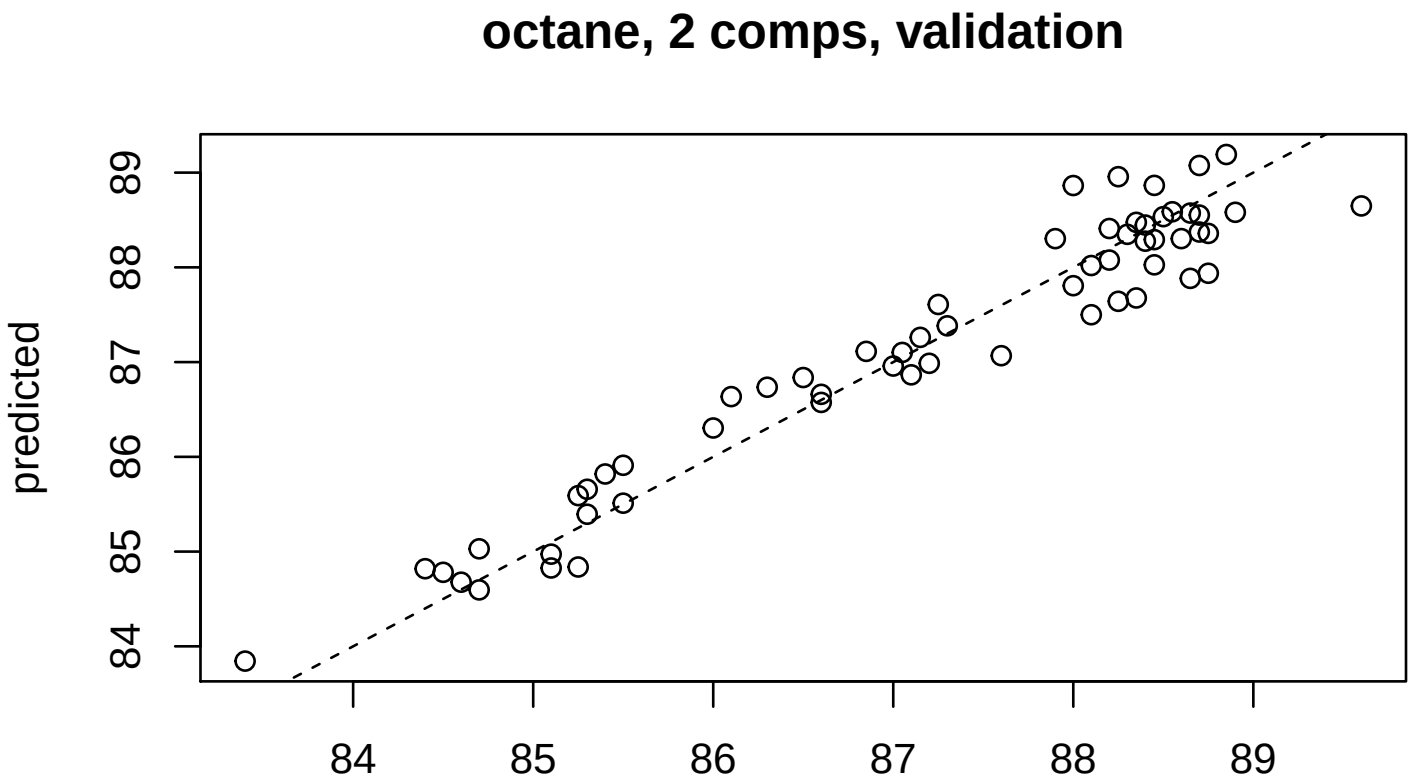
Exemple 1

Le nombre de composantes et les coefficients de régularisation sont fixés automatiquement.
On obtient 3 composantes.
Les prédictions sont



Exemple 1

Le nombre de composantes et les coefficients de régularisation sont fixés automatiquement.
On obtient 3 composantes.
Dans le cas sans sélection de variables



Conclusion

Avantages

- problème de la grande dimension contourné,
- Y multivarié,
- Sélection de variables et interprétation.

Inconvenients

- Beaucoup de paramètres à régler,
- Approche composante par composante,
- Méthode très linéaire (voir kernel PLS).

References I

[1] Hyonho Chun and Sündüz Keleş. “Sparse partial least squares regression for simultaneous dimension reduction and variable selection”. In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.1 (2010), pp. 3–25.

[2] Kim-Anh Lê Cao et al. “A sparse PLS for variable selection when integrating omics data”. In: *Statistical applications in genetics and molecular biology* 7.1 (2008).

[3] Hadrien Lorenzo et al. “Data-driven sparse partial least squares”. In: *Statistical Analysis and Data Mining: The ASA Data Science Journal* 15.2 (2022), pp. 264–282.

[4] Miguel Pérez-Enciso and Michel Tenenhaus. “Prediction of clinical outcome with microarray data: a partial least squares discriminant analysis (PLS-DA) approach”. In: *Human genetics* 112.5 (2003), pp. 581–592.

[5] Florian Rohart et al. “mixOmics: An R package for ‘omics feature selection and multiple data integration”. In: *PLoS computational biology* 13.11 (2017), e1005752.

[6] Ron Wehrens and B-H Mevik. “The pls package: principal component and partial least squares regression in R”. In: (2007).

[7] Herman Wold. “Estimation of principal components and related models by iterative least squares”. In: *Multivariate analysis* (1966), pp. 391–420.