Statistique exploratoire de données hétérogènes : la régression PLS (Partial Least Squares)

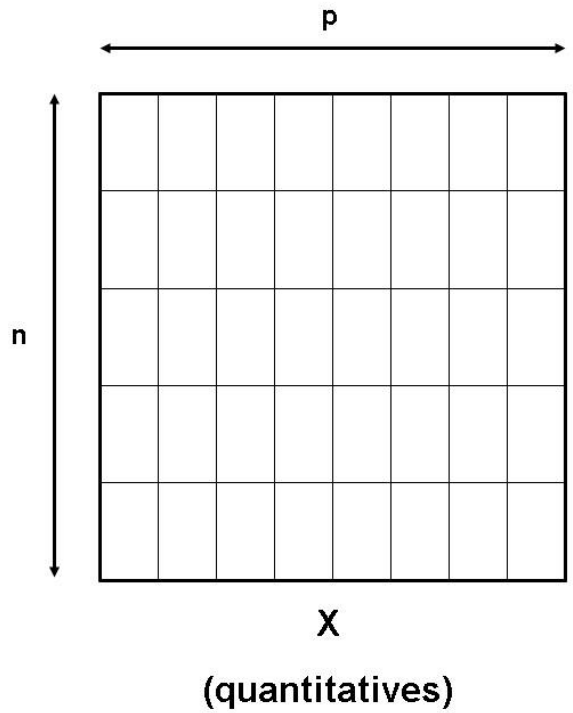
Simon Boitard[†], Kim-Anh Lê Cao[*]

[†] INRA, Laboratoire de Génétique Cellulaire, Toulouse.
[*] Queensland Facility for Advanced Bioinformatics, University of Queensland, Australia.

Formation Interbio, 7 mars 2012

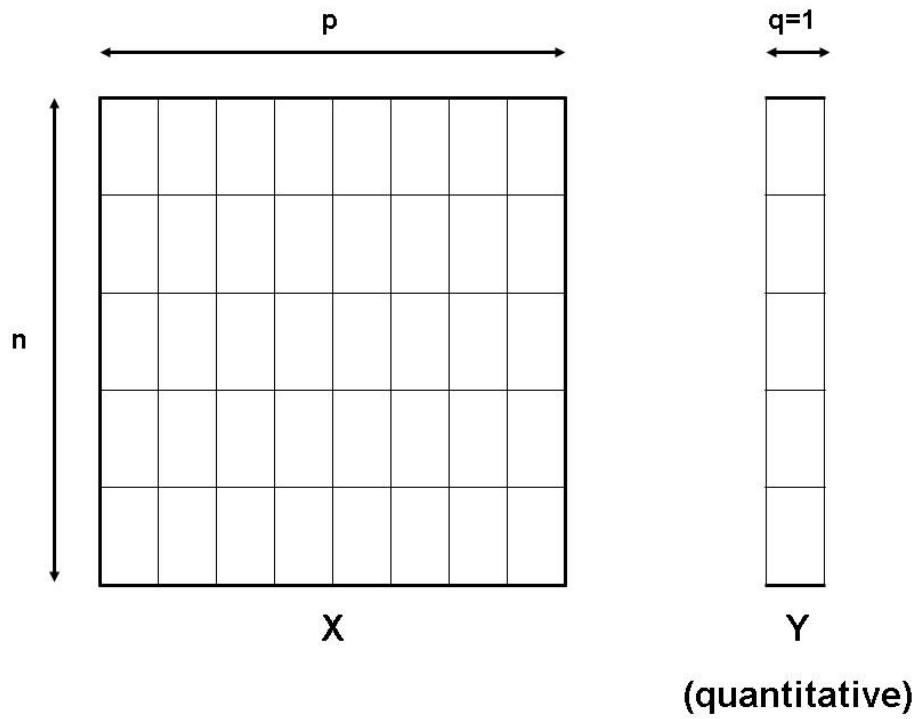
Plan de l'exposé

Previously on Interbio . . .



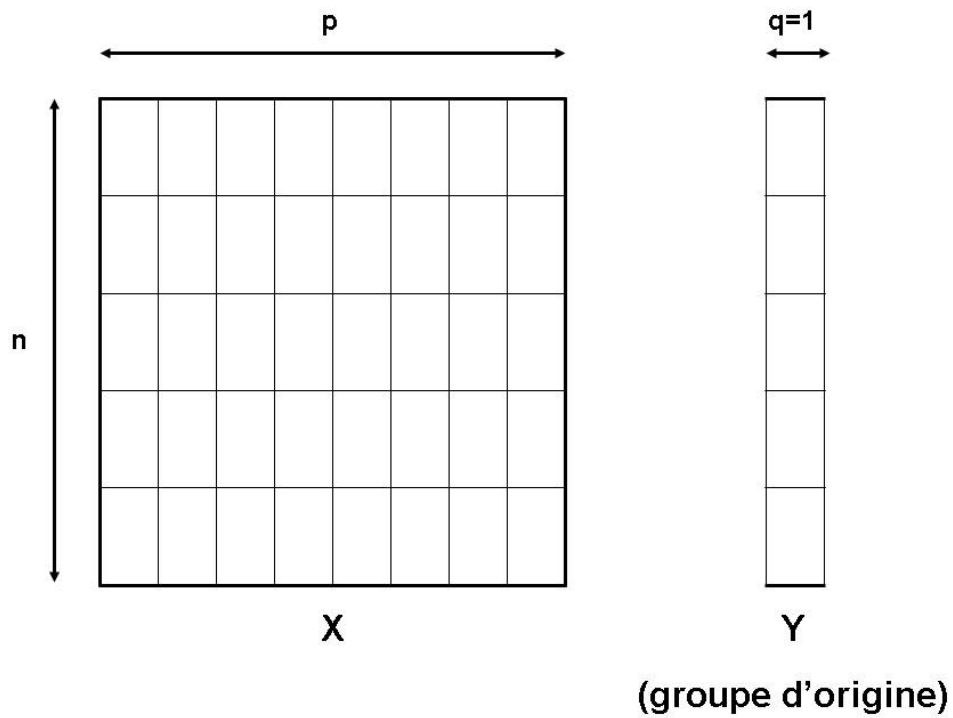
Description / Simplification / Projection d'un tableau de p variables quantitatives : **ACP**.

Previously on Interbio . . .



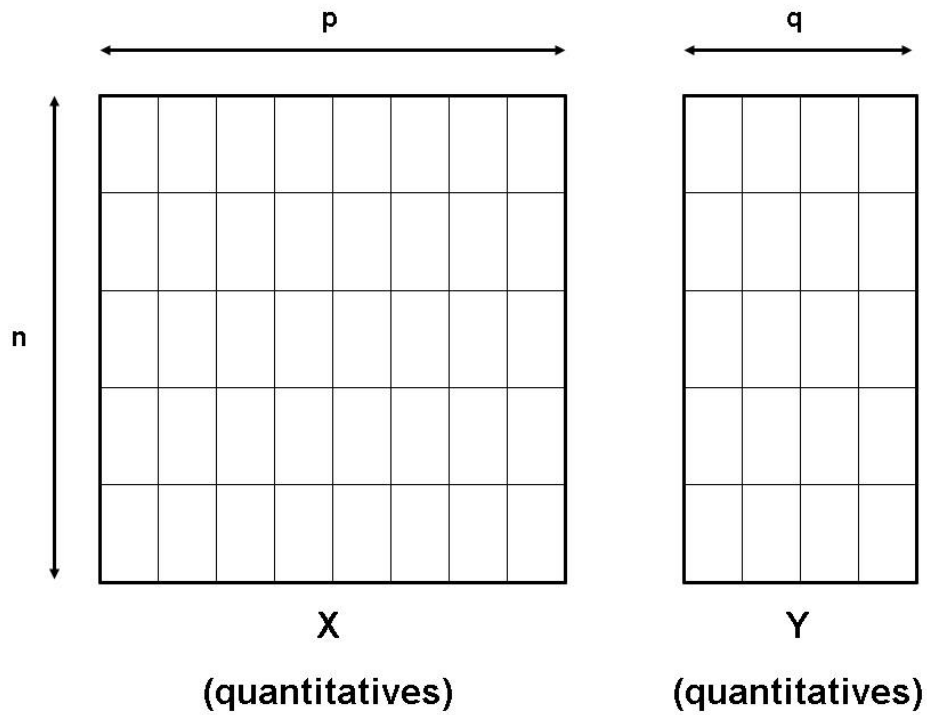
Explication / Prédiction d'une variable quantitative à l'aide de p variables (quantitatives ou qualitatives) : **modèle linéaire, modèle linéaire mixte**.

Previously on Interbio . . .



Explication / Prédiction d'une variable qualitative représentant un groupe d'origine à l'aide de p variables : **classification supervisée**.

Previously on Interbio . . .



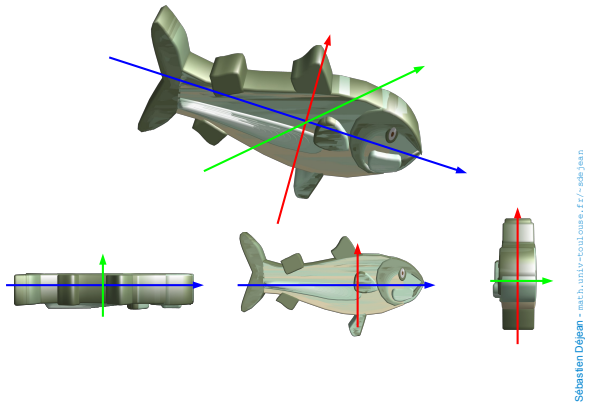
Description / Simplification / Projection de 2 tableaux de p et q variables quantitatives hétérogènes : **analyse canonique, régression PLS**

Biologie intégrative

- Données de nature différente (génome, transcriptome, protéome, métabolome, phénome ...) prélevées chez les mêmes individus.
  $\rightarrow$ Besoin de méthodes statistiques pour mettre en relation ces groupes de variables.
- Beaucoup plus de variables que d'individus : $n << p + q$.

Rappel sur l'ACP

- Tableau X de taille $n \times p$ = représentation des individus dans un espace de dimension p.
- ACP = projection des n individus dans un espace de dimension $H < p$, de manière à simplifier la visualisation et l'interprétation des résultats.
- Pour perdre le moins possible d'information, on projette sur le sous-espace qui maximise la variance des observations.



Rappel sur l'ACP

Soit c_h la h-éme composante principale.

- c_h est un vecteur de taille n qui contient les projections (les coordonnées) des individus sur le nouvel axe h.
- c_h peut être vue comme une nouvelle variable (artificielle), qui est une combinaison linéaire des variables initiales :

$$\underset{(n \times 1)}{c_h} = \underset{(n \times p)}{X} \underset{(p \times 1)}{u_h} = \sum_{j=1}^{p} (u_h)_j X_{.j}$$

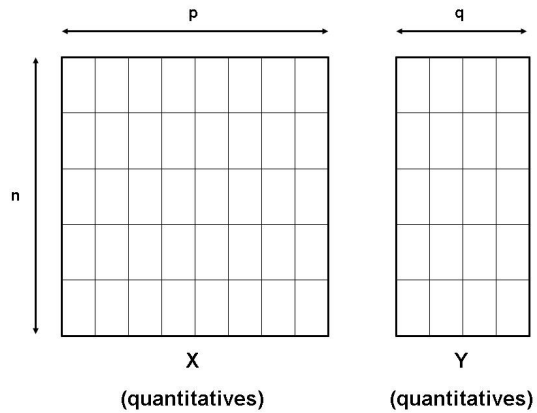
où u_h donne les contributions des variables initiales sur le nouvel axe h (loading vector).

- u_h est le vecteur qui maximise la fonction

$$var(Xu_h) = var(c_h)$$

sous les contraintes $||u_h||_2 = 1$ et $u_h \perp u_1 \dots u_{h-1}$.

Régression PLS



- Comme pour l'ACP, On résume X et Y à l'aide de H variables synthétiques ($H < p, q$), combinaisons linéaires des variables initiales. On note ces variables ξ_h pour X et ω_h pour Y ($h = 1 \ldots H$).
- **Contrairement à l'ACP**, on ne cherche pas à maximiser $var(\xi_h)$ ou $var(\omega_h)$, mais

$$Cov^2(\xi_h, \omega_h)$$

pour mettre en évidence les liens entre X et Y.

Et l'analyse canonique alors?

- Comme pour la PLS, on cherche des combinaisons linéaires ξ_h et ω_h des variables initiales X et Y, mais qui maximisent cette fois:

$$Cor^2(\xi_h, \omega_h)$$

- Résolution facile d'un point de vue théorique, mais nécessite l'inversion de $X'X$ et $Y'Y$
  $\rightarrow$ impossible si les variables sont colinéaires, ce qui arrive forcément si $n < p, q$.
- Pour résoudre ces problèmes, des techniques de régularisation existent (rCCA).
- Interprétation des coefficients moins claire qu'avec la PLS.

Un peu d'histoire

- Méthode introduite par Herman Wold en 1975, algorithme de résoluton appellé NIPALS.
- Popularisée dans les années 1980 par Svante Wold, chercheur en chimiométrie (données spectrales de grande dimension).
- Egalement utilisée dans de nombreux autres domaines : économie, psychologie, médecine ...
- Philosophie : existence d'un petit nombre de variables latentes qui expliquent toutes les variables de X et Y.

Principe de l'algorithme

On part de matrices X et Y centrées.

1. On choisit $\xi_1 = Xu_1$ et $\omega_1 = Yv_1$ qui maximisent $cov^2(Xu_1, Yv_1)$ sous la contrainte $||u_1||_2 = ||v_1||_2 = 1$.
2. On essaie d'expliquer chaque variable initiale du tableau X par ξ_1 (régression linéaire simple), ce qui conduit à l'estimation

$$\underset{(n \times p)}{\hat{X}} = \underset{(n \times 1)}{\xi_1} \underset{(1 \times p)}{c_1}$$

   où c_1 contient les p coefficents de regression.
   On calcule de même $\hat{Y} = \omega_1 e_1$.
3. Déflation : on calcule les résidus $X_1 = X - \hat{X}$ et $Y_1 = Y - \hat{Y}$

On recommence ensuite à partir de X_1 et Y_1.

Propriétés

- Comme pour l'ACP, u_h $(p \times 1)$ et v_h $(q \times 1)$ représentent les loadings : $(u_h)_j$ poids de la j-éme colonne de X dans ξ_h, $(v_h)_j$ poids de la j-éme colonne de Y dans ω_h
- $(u_h)_j \propto cov((X_{h-1})_{.j}, \omega_h)$, $(v_h)_j \propto cov((Y_{h-1})_{.j}, \xi_h)$.
- u_h et v_h sont les vecteurs singuliers associés à la plus grande valeur singulière de $X'_{h-1} Y_{h-1}$.

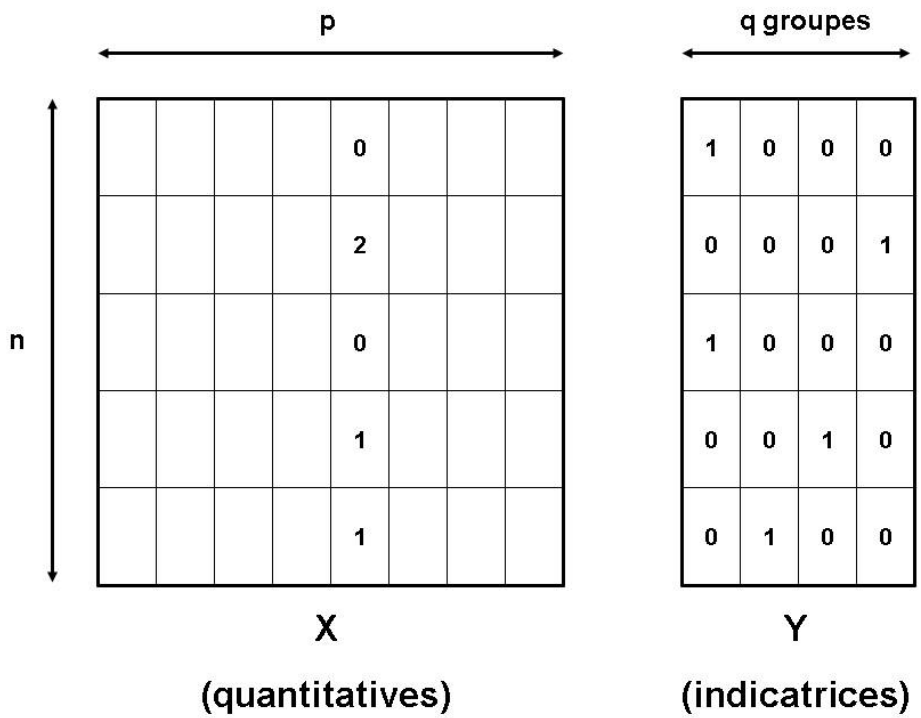
Plusieurs variantes de la PLS

- La relation entre X et Y peut être :
  - asymétrique, s'il y a une relation de causalité logique entre les variables de X et celles de Y. On applique alors la PLS en **mode régression**.
    ex: X = transcriptome, Y = phénotypes.
  - symétrique, s'il n'y a pas de relation de causalité évidente entre X et Y. On applique alors la PLS en **mode canonique**.
    ex: X = transcritome par hybridation, Y = transcritome par RNA seq.
- En mode régression, on régresse Y sur ξ_1 ($\hat{Y} = \xi_1 d_1$) et non sur ω_1 (étape 3).
- La plupart des méthodes PLS s'intéressent au cas où Y est multidimensionnel (PLS2, SIMPLS), mais la PLS en mode régression peut aussi être utile pour Y unidimensionnel (PLS1)

Plusieurs variantes de la PLS

- Différents algorithmes PLS existent (NIPALS, Kernel)
- Ils diffèrent notamment dans la manière d'identifier les loadings u_h et v_h (étape 1) le type de déflation (étape 4).
- Toutes les variantes de la PLS donnent des résultats similaires pour la dimension 1.
- Toutes cherchent des variables synthétiques maximisant la covariance entre X et Y.

PLS et classification

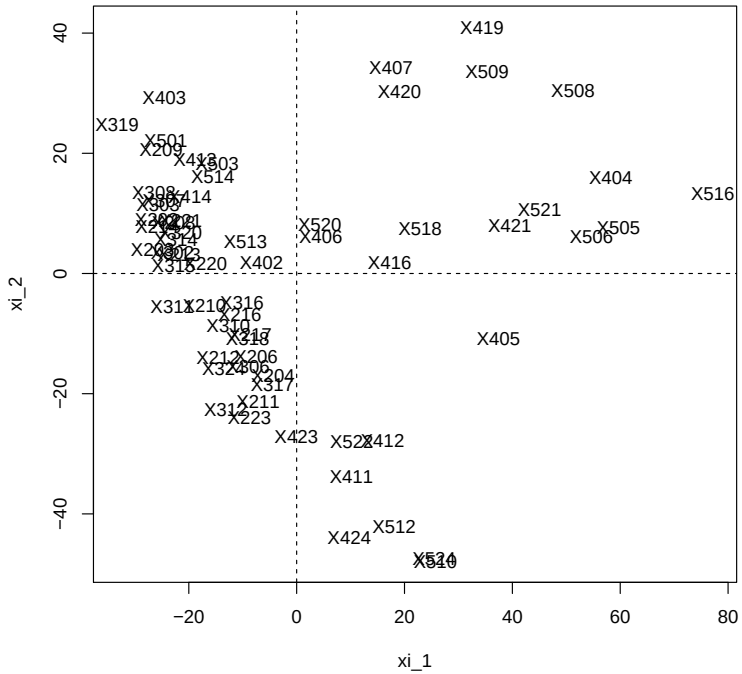


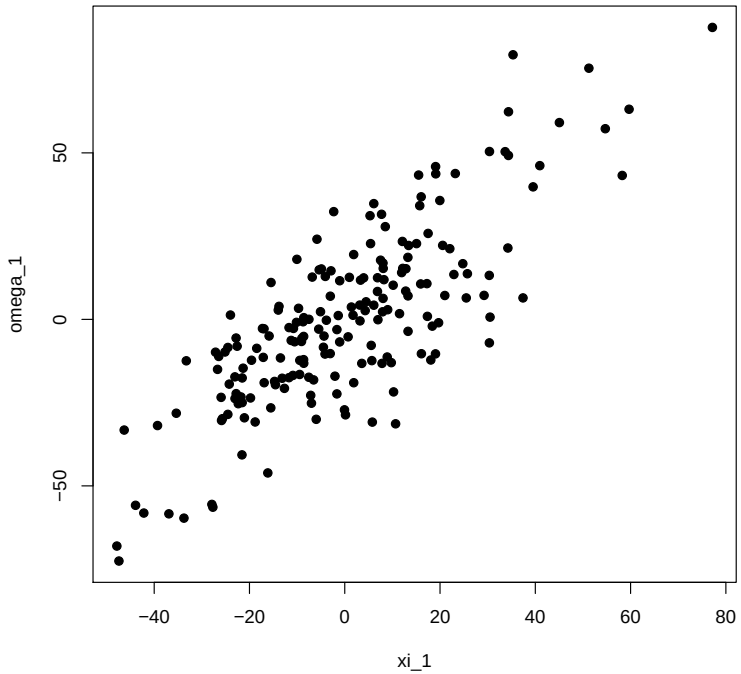
Si Y décrit la répartition des n individus en q groupes, la PLS en mode régression fournit un modèle prédictif de classification.

Jeu de données "Liver toxicity" (Bushel et al., 2007)

- $n = 64$ rats, exposés à diverses doses d'acetaminophen (paracetamol)
- X : expression dans le foie de $p = 3116$ gènes.
- Y : $q = 10$ variables cliniques mesurées dans le serum, marqueurs des dommages causés au foie.
- Dose injectée : 50 mg/kg (non toxique), 150 mg/kg (modérément toxique), 1500 mg/kg (très toxique).
- instant d'extraction des ARN (6h, 18h, 24h, 48h après injection).

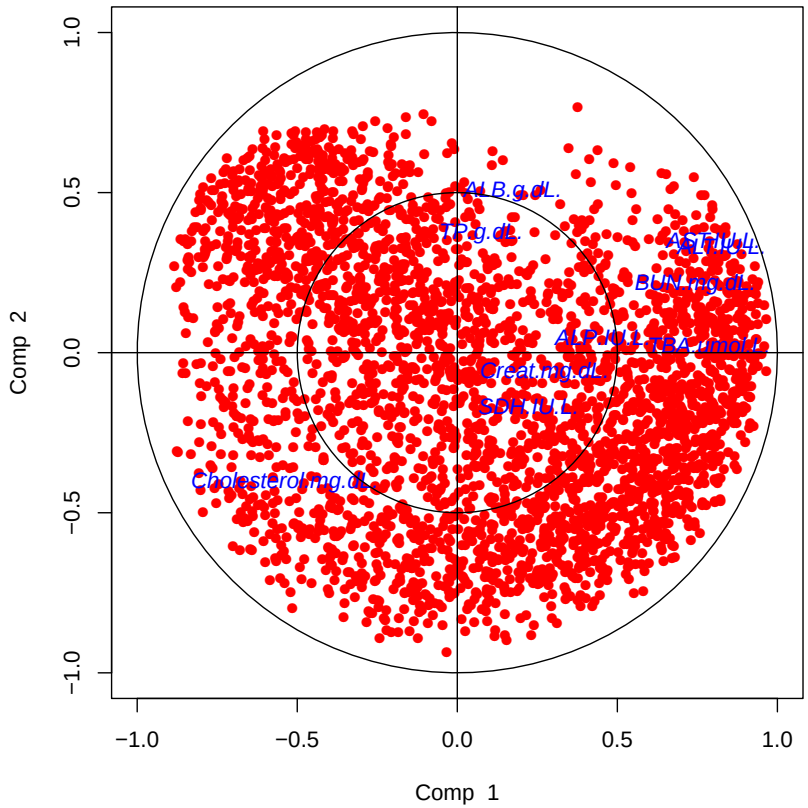
Représentation des individus





On projette les individus sur les composantes de X, mais on obtiendrait une image similaire en projetant sur les composantes de Y (par construction).

Représentation des variables



Comme pour l'ACP, c'est ce graphe qui nous renseigne sur la signification des axes.

Choix de la dimension H

- Mode régression : on cherche à minimiser l'erreur de prévision, que l'on peut évaluer par **validation croisée** :
  - les n individus sont divisés en M groupes.
  - $M-1$ groupes utilisés pour estimer le modèle (apprentissage), le groupe restant utilisé pour évaluer l'erreur de prévision (validation).
  - On fait varier le groupe de validation et on moyenne.
- Mode canonique : pas d'objectif de prédiction donc c'est surtout l'interprétation biologique qui prime.

Critères d'évaluation

- PRediction Error Sum of Squares pour la variable k et la dimension h.

$$PRESS_{k,h} = \sum_{i=1}^{M} \sum_{j \in g_i} (y_j^k - \hat{y}_{h,j}^k)^2$$

où $\hat{y}_{h,j}^k$ est estimé sans les individus du groupe i.

- Root Mean Squared Error Prediction pour la variable k et la dimension h.

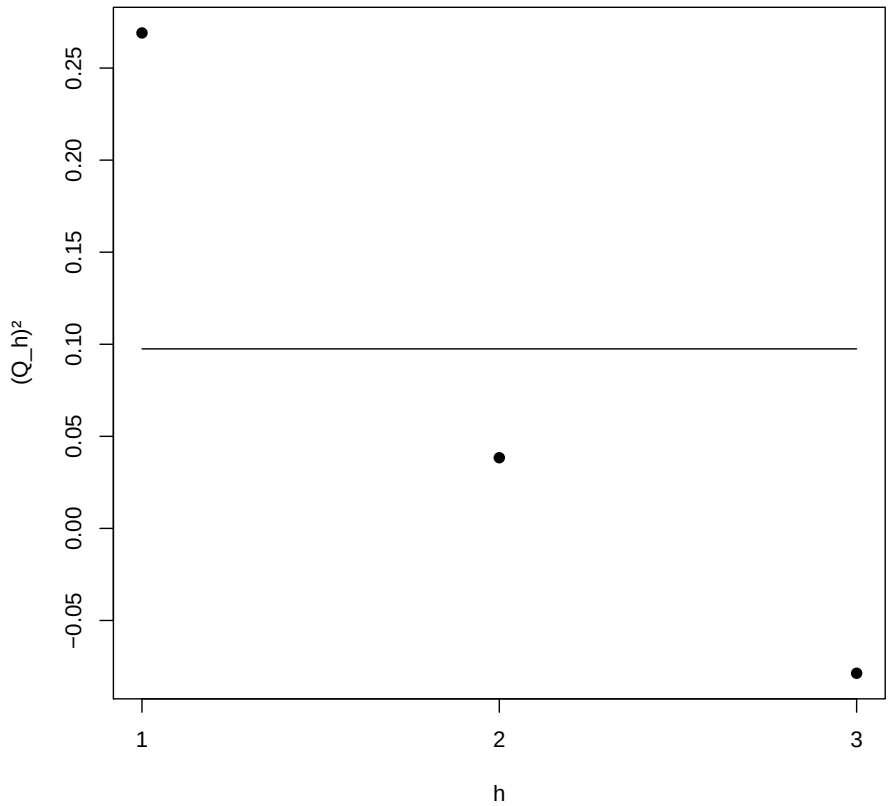
$$RMSEP_{k,h} = \sqrt{PRESS_{k,h}}$$

- pour tenir compte de toutes les variables :

$$Q_h^2 = 1 - \frac{\sum_{k=1}^{q} PRESS_{k,h}}{\sum_{k=1}^{q} RSS_{k,h-1}}$$

où $RSS_{k,h}$ mesure la variance résiduelle non expliquée par les h premières composantes.

Choix de variable avec Q_h^2

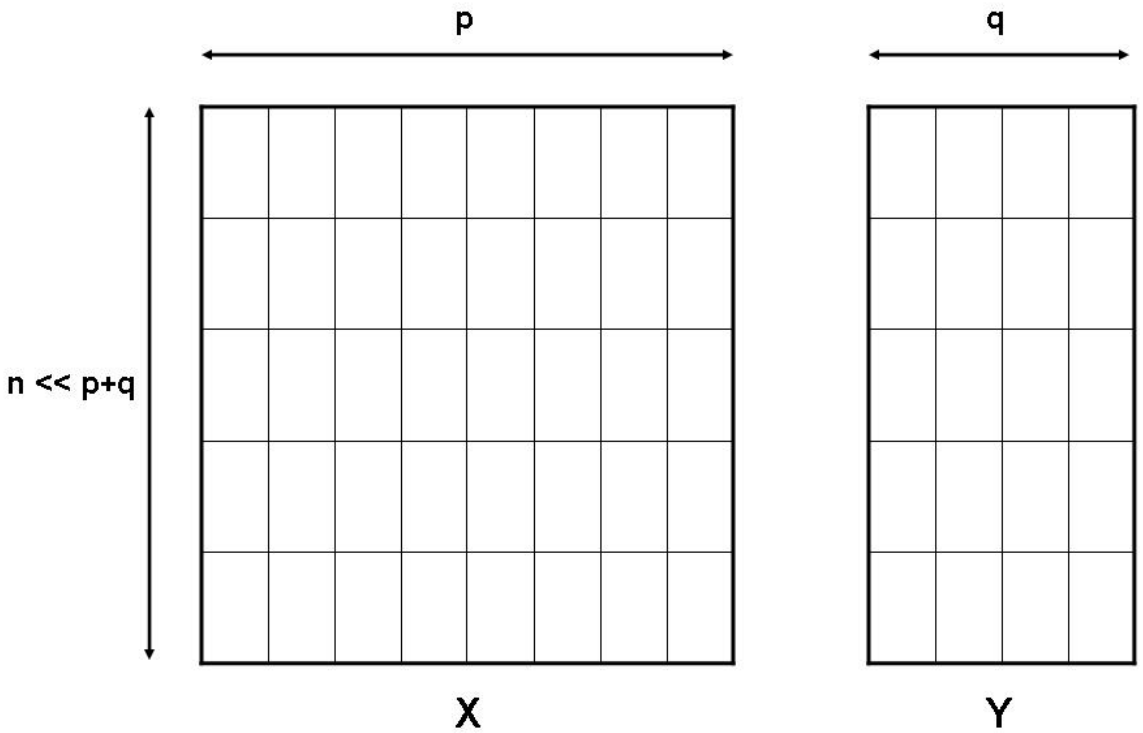


Seules les dimensions telles que $Q_h^2 > 0.0975$ sont considérées importantes.

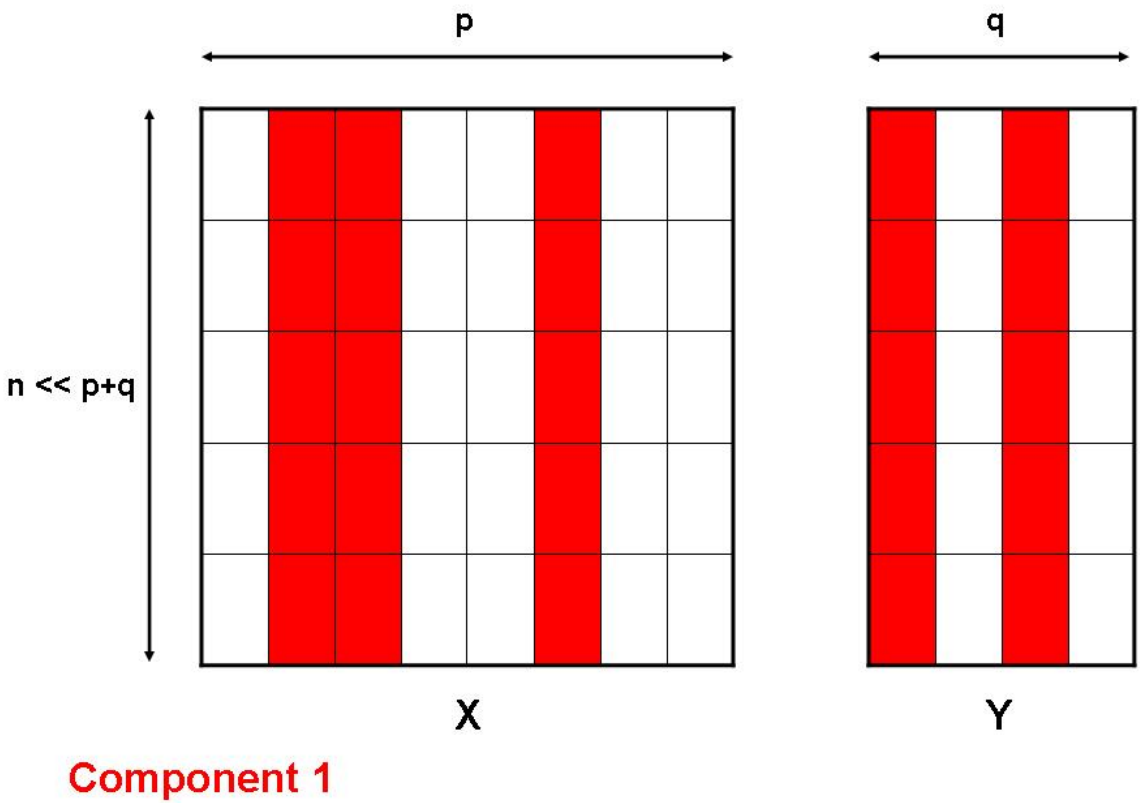
Principe

- Presque tout fonctionne comme pour le PLS :
  - Recherche de variables synthétiques de covariance maximale.
  - Mode régression ou canonique.
  - Représentation des variables et des individus.
  - Choix de la dimension h.
- Une différence fondamentale, la **sélection de variables : chaque variable synthétique n'utilise qu'un petit nombre de variables initiales**
  - Statistique : réduction de l'erreur de prévision.
  - Biologie : interprétation plus facile.
  - Représentation plus claire.

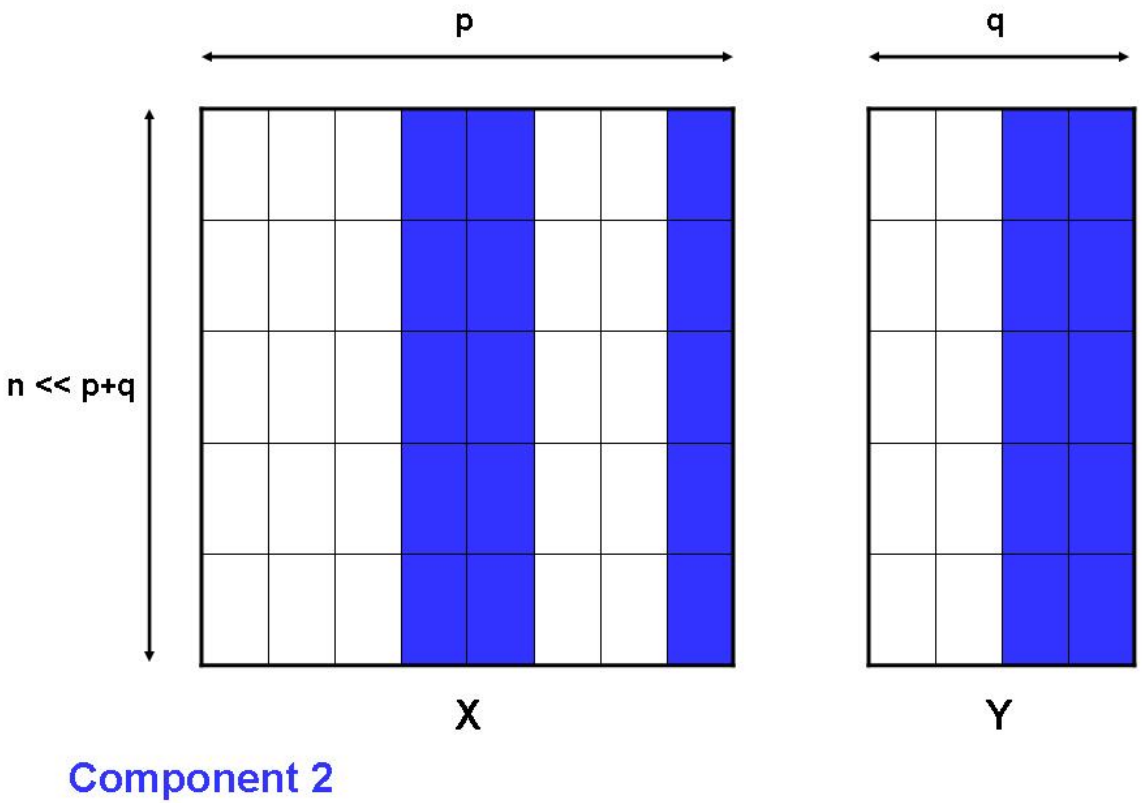
Principe



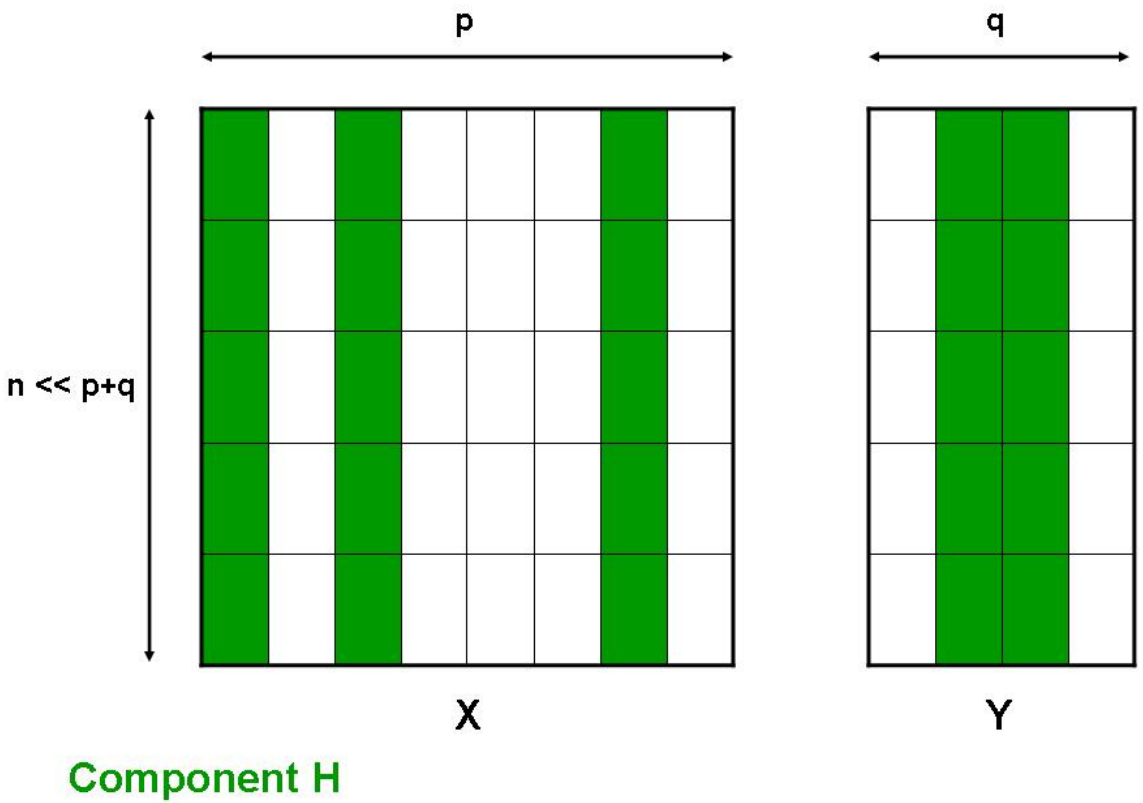
Principe



Principe



Principe



Rappel : la sélection de variables en statistique

- Erreur de prévision = biais^2 + variance
- Ajouter des variables explicatives diminue le biais mais augmente la variance.
- 2 stratégies permettent de trouver un compromis :
  - Sélection de variables : exploration des modèles possibles (algorithme forward, backward, stepwise, Furnival et Wilson) et comparaison (test de Fisher, C_p de Mallows ...)
  - Régularisation : pénaliser la vraisemblance en ajoutant un terme $\lambda||\theta||_1$ (ridge) ou $\lambda||\theta||_2$ (lasso) qui augmente avec la dimension du paramètre θ.

Grande dimension

- Quand on dispose d'un grand nombre de variables, les algos de sélection de variables sont trop longs ou pas optimaux.
- La régularisation ridge donne des résultats difficiles à interpréter.
- La régularisation lasso est un bon compromis et sélectionne **automatiquement** les variables.
- Sparse PLS = PLS2 + pénalisation lasso.

Comment ça marche?

- Pour la Composante h de la PLS, on maximise $cov^2(X_{h-1}u_h, Y_{h-1}v_h)$ sous la contrainte $||u_h||_2 = ||v_h||_2 = 1$.
- Une formulation alternative consiste à dire qu'on minimise :

$$||M_{h-1} - u_h v_h'||_F = \sum_{i=1}^{n}\sum_{j=1}^{n}(M_{h-1} - u_h v_h')_{i,j}^2$$

sous la contrainte $||u_h||_2 = 1$, avec $M_h = X_h' Y_h$.

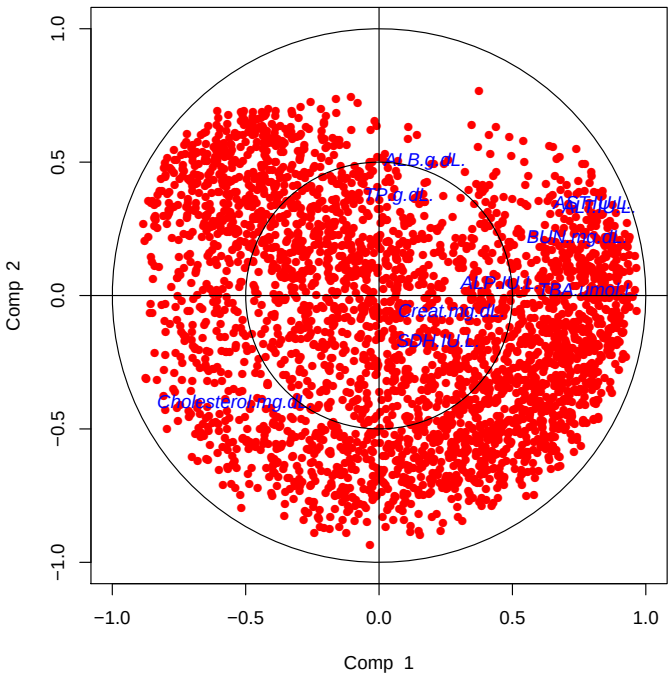
- Pour la Sparse PLS, on minimise

$$||M_{h-1} - u_h v_h'||_F + \lambda_1 \sum_{j=1}^{p} |(u_h)_j| + \lambda_2 \sum_{j=1}^{q} |(v_h)_j|$$

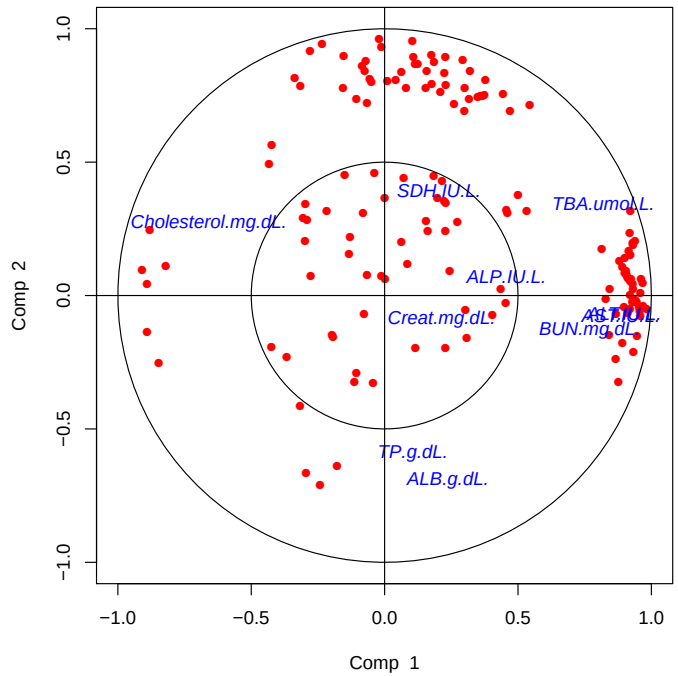
L'ajout des pénalités lasso $\lambda_1||u_h||_1$ et $\lambda_2||v_h||_1$ contraint de manière automatique les vecteurs u_h et v_h à avoir beaucoup de zéros.

- Remarque : Cette technique s'inspire de la Sparse PCA, qui minimise $||X_{h-1} - u_h v_h'||_F$ avec une contrainte lasso.

Représentation des variables



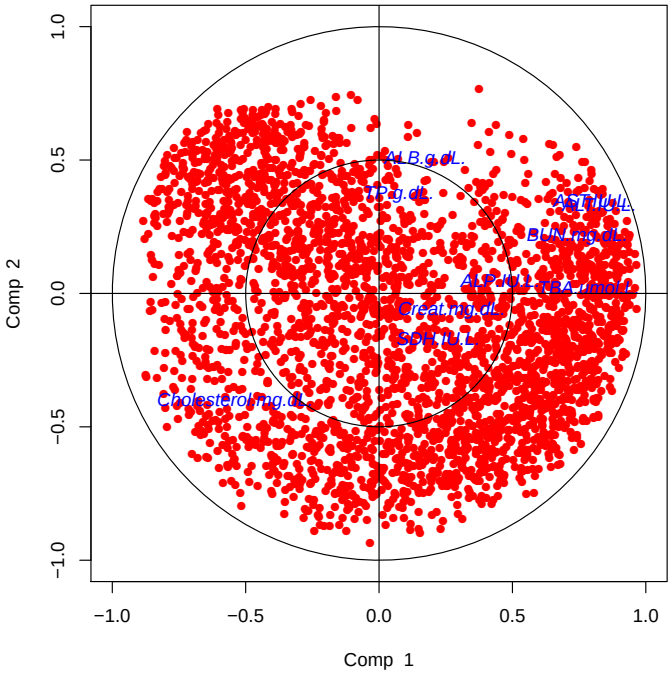
PLS



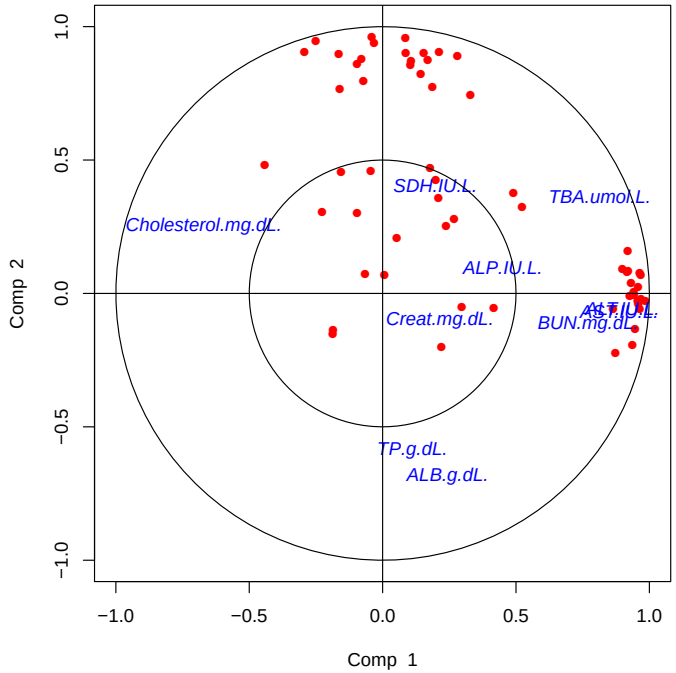
sPLS (keepX=50,keepY=10)

Ici, on a utilisé le package R mixOmics. Il permet de choisir directement le nombre de variables non nulles par composante; plutôt que λ_1 ou λ_2.

Représentation des variables



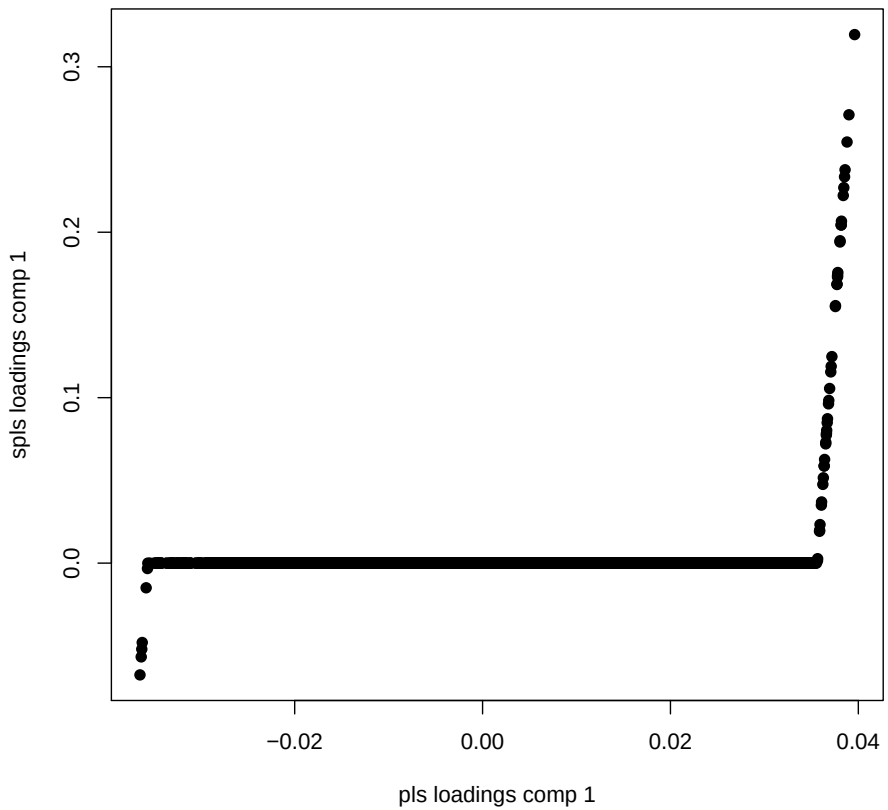
PLS



sPLS (keepX=20,keepY=10)

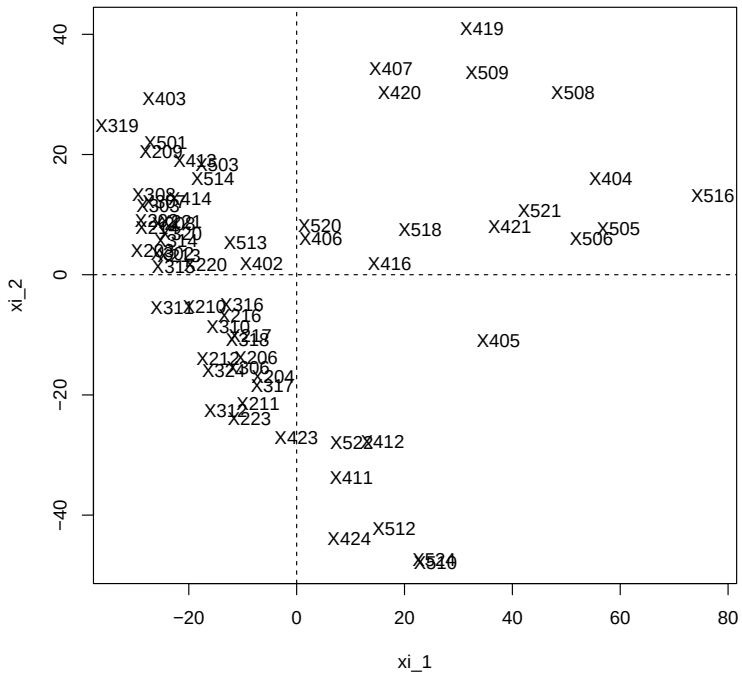
Moins de variables non nulles pour X = λ_1 plus élevé. $\lambda_2 = 0$.

Effet de la sPLS

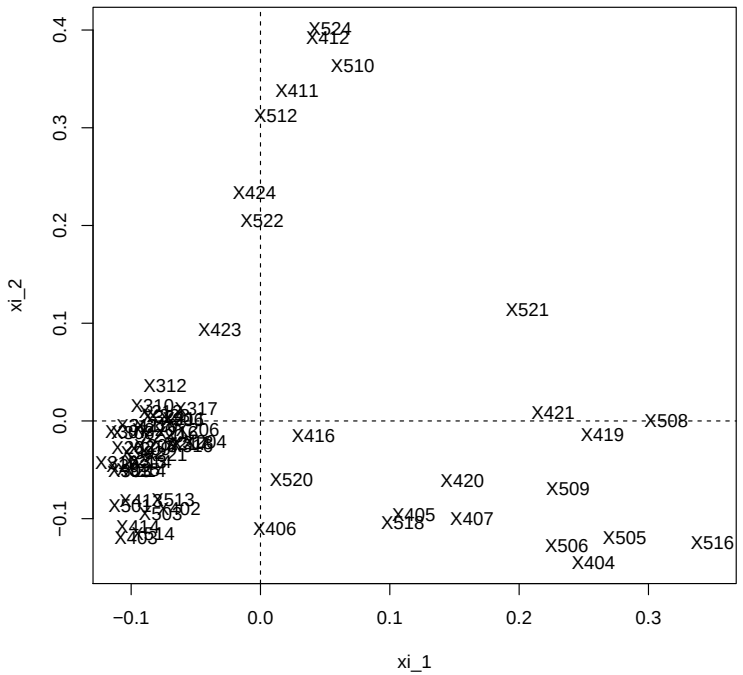


Les petits loadings de la PLS tendent à être éliminés par la sPLS.

Représentation des individus



PLS



sPLS (keepX=50,keepY=10)

Les deux méthodes donnent des résultats assez proches.

Conclusions

- La PLS permet de mettre en évidence les liens entre deux tableaux de données hétérogènes mesurées sur les mêmes échantillons.
- Contrairement à d'autres méthodes, elle est adaptée au cas $n << p + q$.
- Elle permet de répondre à différentes questions : régression, analyse canonique, classification.
- La Sparse PLS permet en plus une sélection des variables sur chaque composante, facilitant l'interprétation biologique.

Références

- PLS
  - Tenenhaus M. (1998) La régression PLS: théorie et pratique. Paris: Editions Technic.
  - Wegelin, J. (2000). A survey of Partial Least Squares (PLS) methods, with emphasis on the two-block case. Technical Report 371, Department of Statistics, University of Washington, Seattle.
- Sparse PLS
  - Lê Cao K.-A., Martin P.G.P., Robert-Granié C. and Besse P. (2009) Sparse canonical methods for biological data integration: application to a cross-platform study. BMC Bioinformatics 10(34).
  - Lê Cao K.-A., Rossouw D., Robert-Granié C. and Besse P. (2008) A sparse PLS for variable selection when integrating Omics data. Statistical Applications in Genetics and Molecular Biology 7, article 35.
  - Lê Cao K.-A., Boitard S. and Besse P. (2011) Sparse PLS Discriminant Analysis: biologically relevant feature selection and graphical displays for multiclass problems. BMC Bioinformatics, 22:253.
- package R : http://www.math.univ-toulouse.fr/ biostat/mixOmics/