Apprentissage Machine / Statistique

PCA et PLS parcimonieuses

PHILIPPE BESSE

INSA de Toulouse
Institut de Mathématiques

Données de grande dimension

- Contexte de grande dimension ($p >> n$)
- Régression par sélection ou pénalisée (ridge, lasso)
- Régression sur composantes principales ou PCR
- Régression PLS (Wold 1966) développée en Chimiométrie (logiciel SIMCA-P)

Régression sur composantes principales (PCR)

- $Z^1, \ldots, Z^p$: composantes principales associées des variables $X^1, \ldots, X^p$:
  - $Z^1 = \sum_{j=1}^{p} \alpha_j X^j$ de variance maximale avec $\sum \alpha_j^2 = 1$
  - Z^m combinaison linéaire de variance maximale et orthogonale à $Z^1, \ldots, Z^{m-1}$.
- La PCR considére un prédicteur de la forme :

$$\hat{Y}^{PCR} = \sum_{m=1}^{r} \hat{\theta}_m Z^m$$

avec

$$\hat{\theta}_m = \frac{\langle Z^m, Y \rangle}{\|Z^m\|^2}$$

Propriétés de la PCR

- $r = p$ redonne l'estimateur des moindres carrés
- $r < p$ pour réduire la variance lors de variables colinéaires ($p > n$)
- Optimisation du choix de r par validation croisée
- Interprétation des composantes difficile si p est grand
- Régression *ridge* seuille les coefficients des composantes principales, PCR annule ceux d'ordre $> r$
- Problème : premières composantes ne sont pas nécessairement corrélées avec Y
- D'où, l'intérêt de la régression PLS

Régressions PLS

- PLS1 : Y quantitative expliquée par p variables X^j,
- PLS2 : (canonique) p variables X^j et q variables Y^k,
- PLS2 : (régression) q variables Y^k par p variables X^j,
- PLS-DA : Y qualitative expliquée par p variables X^j.
- Pas de propriétés statistiques de la PLS

Principe et objectif de parcimonie

- Exploration et intégration de données
  - *e.g.* données biologiques à haut débit $n << p$
  - phénotypes, métabolites... fonctions de transcrits
- Interprétation
  - Version parcimonieuse de la régression PLS
  - Construite sur un algorithme de *Sparse-SVD*
  - Donc d'ACP parcimonieuse

Définition de la PLS1 (partial least square)

Chercher les r composantes Ξ_h combinaisons linéaires des X_j :

$$\boldsymbol{\Xi} = \mathbf{XU}$$

fortement liées avec Y
$\mathbf{U}$ est solution du problème suivant :

$$\begin{aligned} \textbf{Pour} \quad h = 1, \ldots, r, \quad \mathbf{u}_h &= \arg\max_{\mathbf{u}} \mathrm{Cov}(Y, \Xi_h)^2 \\ &= \arg\max_{\mathbf{u}} \mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{u} \\ \textbf{Avec} \quad \mathbf{u}_h'\mathbf{u}_h &= 1 \\ \textbf{et} \quad \boldsymbol{\xi}_h'\boldsymbol{\xi}_h = \mathbf{u}'\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{u} &= 0, \quad \textbf{pour} \quad \ell = 1 \ldots, h-1. \end{aligned}$$

Les variables X_j sont préalablement centrées et réduites

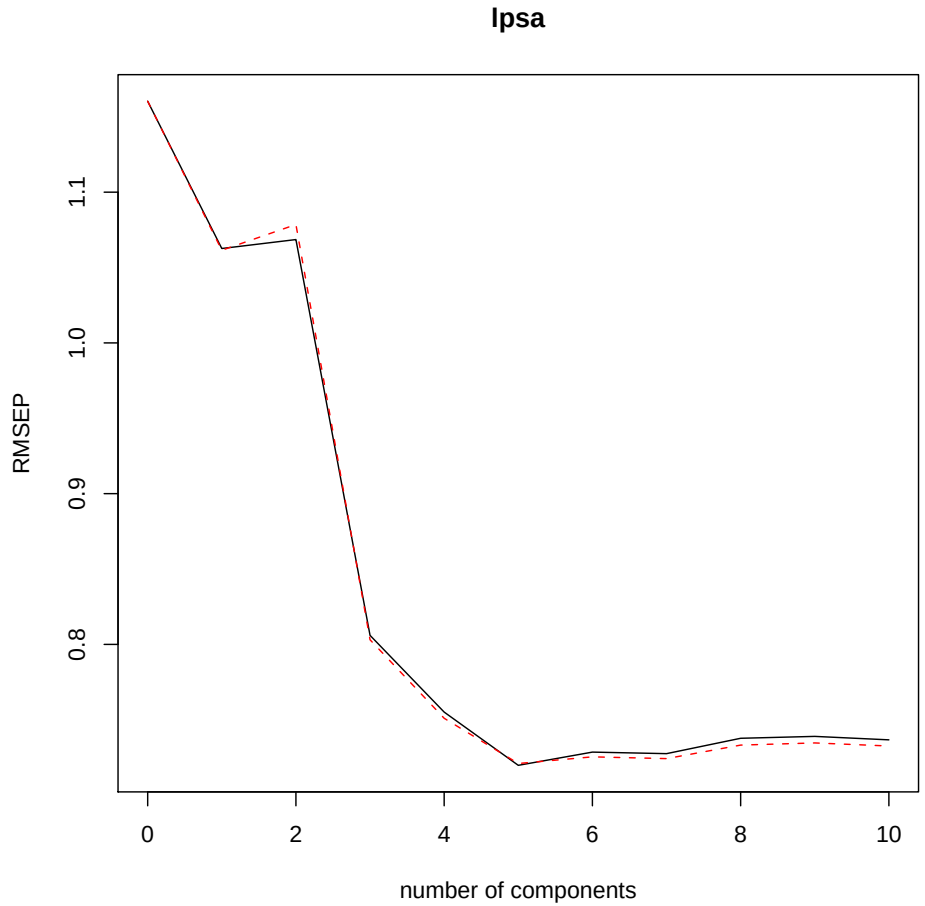
Algorithme de PLS1

- $\mathbf{X}$ matrice des variables explicatives centrées réduites
- Calcul de la matrice $\mathbf{U}$ des coefficients
- **Pour** $h = 1$ à r **Faire**
  1. $\mathbf{u}_h = \frac{\mathbf{X}'\mathbf{Y}}{\|\mathbf{X}'\mathbf{Y}\|}$
  2. $\boldsymbol{\xi}_h = \mathbf{X}\mathbf{u}_h$
  3. Déflation de $\mathbf{X}$: $\mathbf{X} = \mathbf{X} - \xi_h \xi_h' \mathbf{X}$

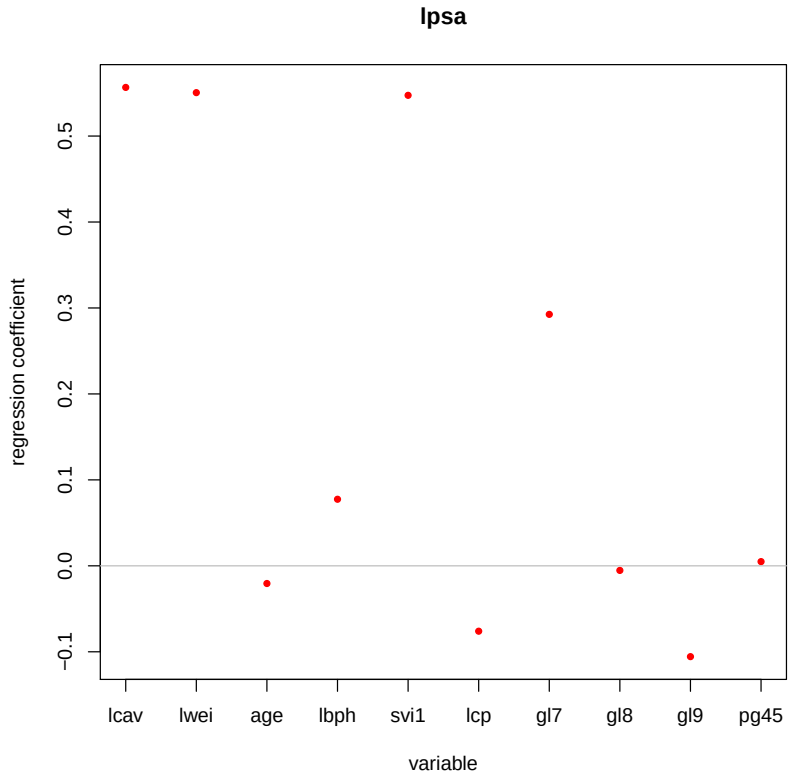
Puis régression de Y sur les r *variables latentes* $\boldsymbol{\xi}_h$

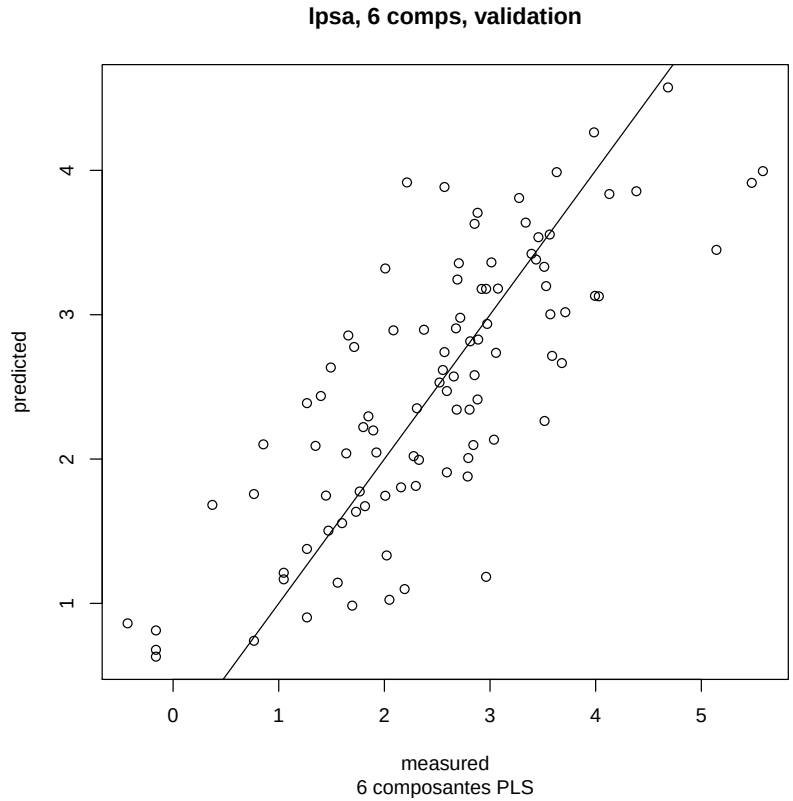
Propriétés de la PLS1

- Réduction de dimension comme avec la PCR
- Régression sur des composantes décorrélées (orthogonales)
- Optimisation de r par validation croisée
- En général : solution de la PLS plus parcimonieuse que celle de la PCR
- Problème d'interprétation : version *sparse*-PLS



Erreur par validation croisée





Loadings et qualité d'ajustement

Définition de la PLS2

- Variables latentes $\boldsymbol{\xi}_h$ et $\boldsymbol{\omega}_h$, $(h = 1, \ldots, r)$

$$\boldsymbol{\xi}_1 = \mathbf{X}\mathbf{u}_1 \text{ et } \boldsymbol{\omega}_1 = \mathbf{Y}\mathbf{v}_1$$

Solutions de

$$\max_{||u||=||v||=1} \text{cov}(\mathbf{X}\mathbf{u}, \mathbf{Y}\mathbf{v})$$

- Puis itérations avec déflations de $\mathbf{X}$ et $\mathbf{Y}$
- $(\mathbf{u}_h, \mathbf{v}_h)_{h=1,\ldots,r}$ sont appelés vecteurs *loading*

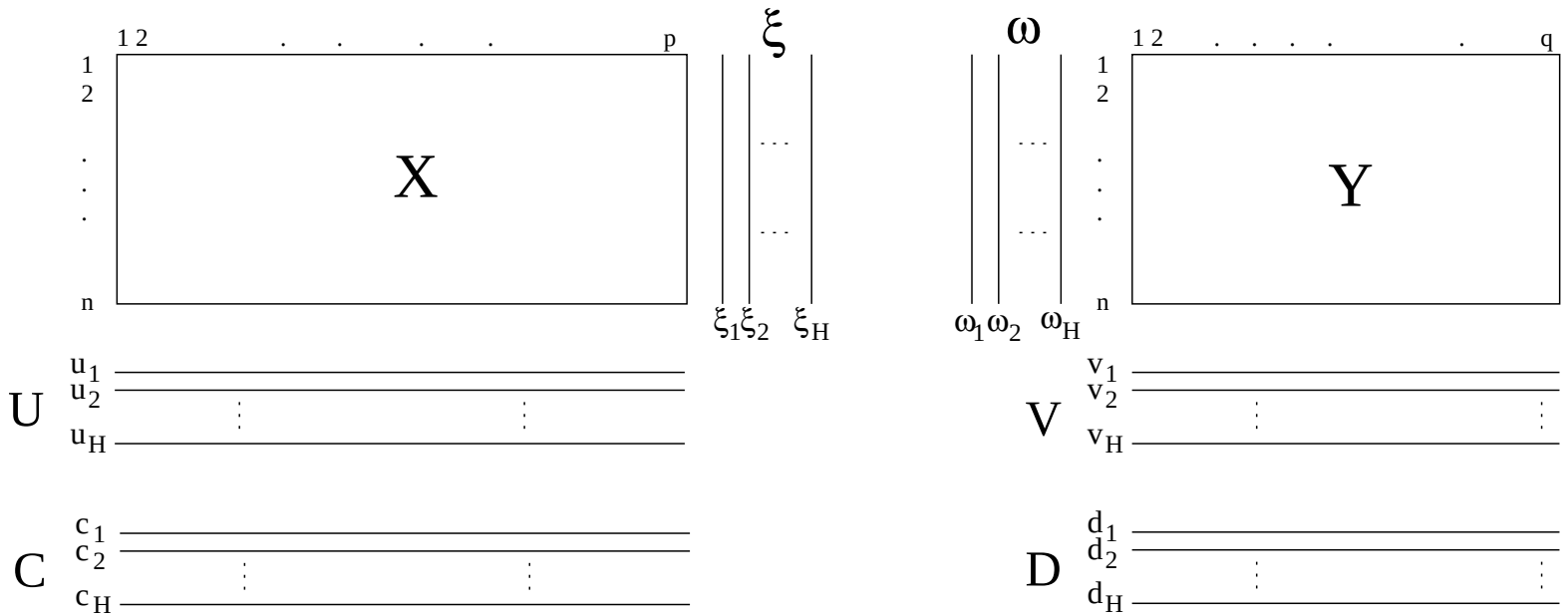


Schéma de la PLS2 : $\mathbf{X}$ and $\mathbf{Y}$ sont décomposées en *loading vectors* $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$, $(\mathbf{v}_1, \ldots, \mathbf{v}_r)$ et variables latentes $(\xi_1, \ldots, \xi_r)$, $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r)$

Algorithme NIPALS de PLS2

- $\mathbf{X}$ et $\mathbf{Y}$ matrices des données centrées
- Initialiser ω_1 par la première colonne de Y
- **Pour** $h = 1$ à r **Faire**
  1. Jusqu'à convergence
     1. $\mathbf{u}_h = \mathbf{X}'\omega_h / \omega_h'\omega_h$
     2. $\mathbf{u}_h = \mathbf{u}_h / \mathbf{u}_h'\mathbf{u}_h$ est le vecteur *loading* associé à $\mathbf{X}$
     3. $\xi_h = \mathbf{X}\mathbf{u}_h$ est la *variable latente* associée à $\mathbf{X}$
     4. $\mathbf{v}_h = \mathbf{Y}'\xi_h / (\xi_h'\xi_h)$
     5. $\mathbf{v}_h = \mathbf{v}_h / \mathbf{v}_h'\mathbf{v}_h$ est le vecteur *loading* associé à $\mathbf{Y}$
     6. $\omega_h = \mathbf{Y}'\mathbf{v_h}$ est la variable latente associée à $\mathbf{Y}$
  2. $\mathbf{c}_h = \mathbf{X}'\xi / \xi'\xi$ régression partielle de $\mathbf{X}$ sur ξ
  3. $\mathbf{d}_h = \mathbf{Y}'\omega / \omega'\omega$ régression partielle de $\mathbf{Y}$ sur ω
  4. Résidus $\mathbf{X} \leftarrow \mathbf{X} - \xi\mathbf{c}'$ ou *déflation*
  5. Résidus $\mathbf{Y} \leftarrow \mathbf{Y} - \omega\mathbf{d}'$ ou *déflation*

Propriétés de NIPALS

- Nombre r d'itérations à fixer ou optimiser
- Algorithme de puissance itérée

$$\begin{aligned} \mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{u} &= \lambda\mathbf{u} \\ \mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}\boldsymbol{\omega} &= \lambda\boldsymbol{\omega} \\ \mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{v} &= \lambda\mathbf{v} \\ \mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}\boldsymbol{\xi} &= \lambda\boldsymbol{\xi} \end{aligned}$$

- Données de grande dimension, colinéaires ou incomplètes
- Graphes de co-variation des variables
- Graphes des individus comme en ACP

PLS par NIPALS ou SVD

- Vecteurs et valeurs propres de $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$
- ou SVD de $\mathbf{X}'\mathbf{Y}$:
- Première étape de la SDV est celle de la PLS
- Plus rapide mais
  - Stocker des matrices $p \times p$
  - Imputation des données manquantes
  - Rendent NIPALS utile

PLS régression *v.s.* canonique

Modes de déflation :

- Mode canonique : $\mathbf{X}_h = \mathbf{X}_{h-1} - \boldsymbol{\xi}_h \mathbf{c}'_h$ et $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\omega}_h \mathbf{d}'_h$
- Mode régression : $\mathbf{X}_h = \mathbf{X}_{h-1} - \boldsymbol{\xi}_h \mathbf{c}'_h$ et $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\xi}_h \mathbf{v}'_h$

PLS-DA

- Y qualitatives à m modalités
- Remplacée par m indicatrices

Dimension et interprétation

- $n << p$ donc p très grand
- PLS et réduction de dimension pour colinéarité
- Composantes ou variables latentes ininterprétables
- Objectif : limité le nombre de coefficients non nuls des variables latentes
- Version parcimonieuse ou *sparse*

Algorithme de Shen et Huang (2008) de *sparse SVD*

Résoudre : $\min_{\mathbf{u},\mathbf{v}} ||\mathbf{M} - \mathbf{u}\mathbf{v}'||_F^2 + P_\lambda(\mathbf{v})$

- Décomposer $\mathbf{M} = \mathbf{U}\Delta\mathbf{V}'$
- $\mathbf{M}_0 = \mathbf{M}$
- **Pour** h de 1 à r **Faire**
  1. Fixer $v_{old} = \delta_h v_h^\star$
  2. $u_{old} = u_h^\star$ avec $v_h^\star$ et $v_h^\star$ de norme 1
  3. **Jusqu'à convergence** de u_{new} et v_{new} **Faire**
     1. $v_{new} = g_\lambda(\mathbf{M}'_{h-1} u_{old})$
     2. $u_{new} = \mathbf{M}'_{h-1} v_{new} / ||\mathbf{M}_{h-1} v_{new}||$
     3. $u_{old} = u_{new}$, $v_{old} = v_{new}$
  4. $v_{new} = v_{new} / ||v_{new}||$
  5. $\mathbf{M}_h = \mathbf{M}_{h-1} - \delta_h u_{new} v'_{new}$

Seuillage doux : $g_\lambda(y) = sign(y)(|y| - \lambda)_+$

Définition sparse-PLS

- Pour résoudre :
  $\min_{\mathbf{u}_h,\mathbf{v}_h} ||\mathbf{M}_h - \mathbf{u}_h\mathbf{v}_h'||_F^2 + P_{\lambda_1}(\mathbf{u}_h) + P_{\lambda_2}(\mathbf{v}_h)$
- Itérer r fois la première étape de sparse-SVD
- Seuillage doux composante par composante :

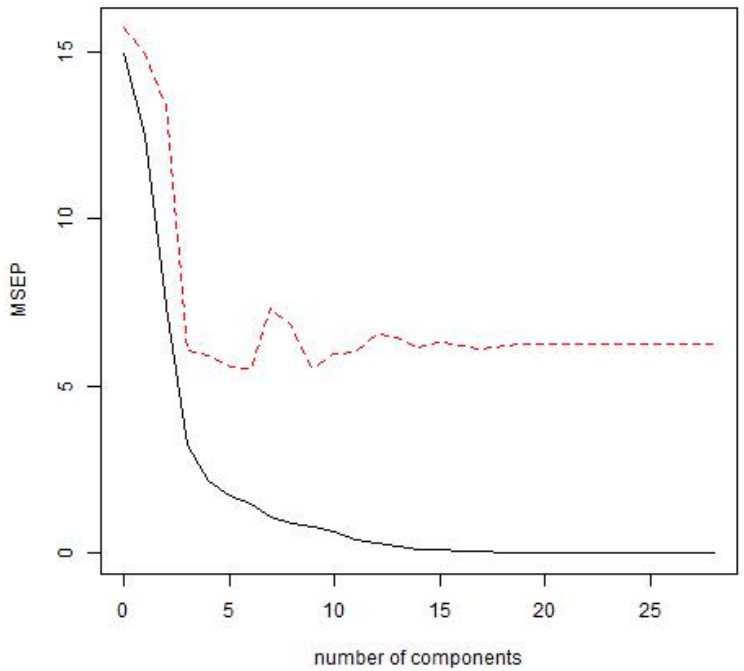
$$P_{\lambda_1}(\mathbf{u}_h) = \sum_{j=1}^{p} sign(\mathbf{u}_{hj})(|\mathbf{u}_{hj}| - \lambda_1)_+$$

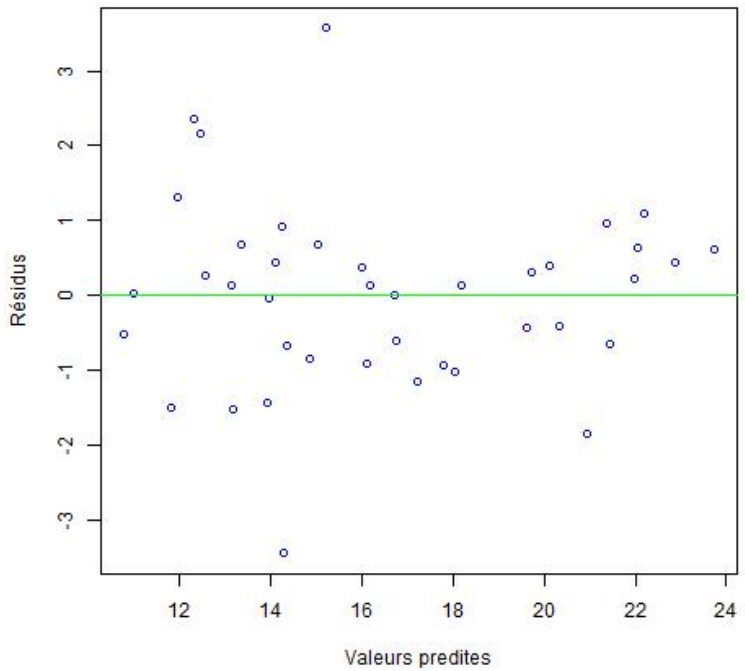
$$P_{\lambda_2}(\mathbf{v}_h) = \sum_{j=1}^{q} sign(\mathbf{v}_{hj})(|\mathbf{v}_{hj}| - \lambda_2)_+$$

- Déflation entre deux SVD et problème d'orthogonalité

Optimisation des paramètres

- Pénalisations Lasso λ_1^h, λ_2^h, $(h = 1, \ldots, r)$:
  - mode régression : erreur de prévision par validation croisée
  - mode canonique : degré de parcimonie, stabilité (bootstrap)
  - sPLS-DA : erreur de prévision
- Dimension r de la PLS : $r \leq 3$ pour l'interprétation



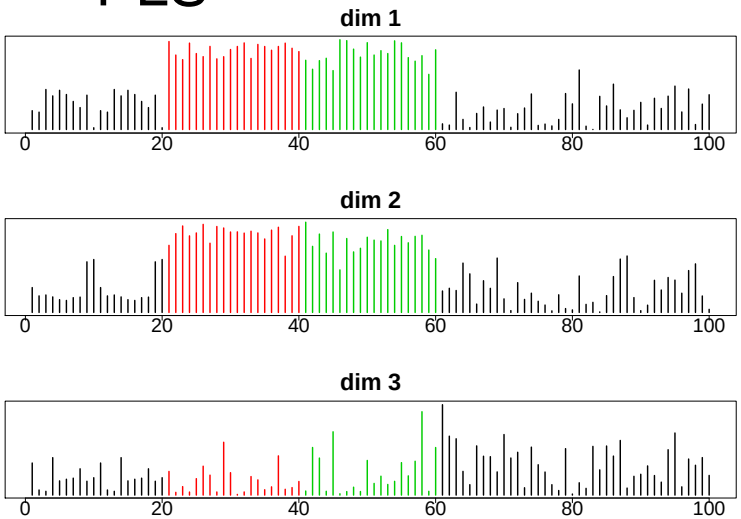


Cookies PLS1 : optimisation de r et résidus

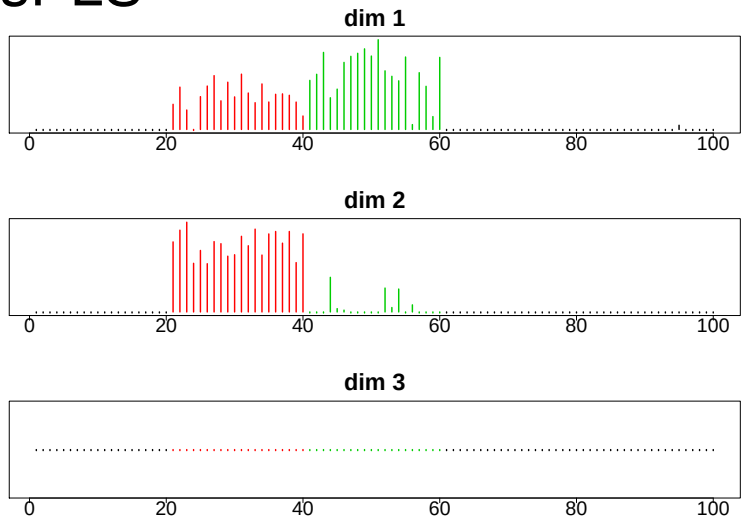
Mode régression : données simulées de Chun et Keles (2010)

- $n = 40$, $p = 5000$ (X var.), $q = 50$ (Y var.)
- 20 variables X et 10 variables Y d'effet μ_1
- 20 variables X et 20 variables Y d'effet μ_2

PLS



sPLS



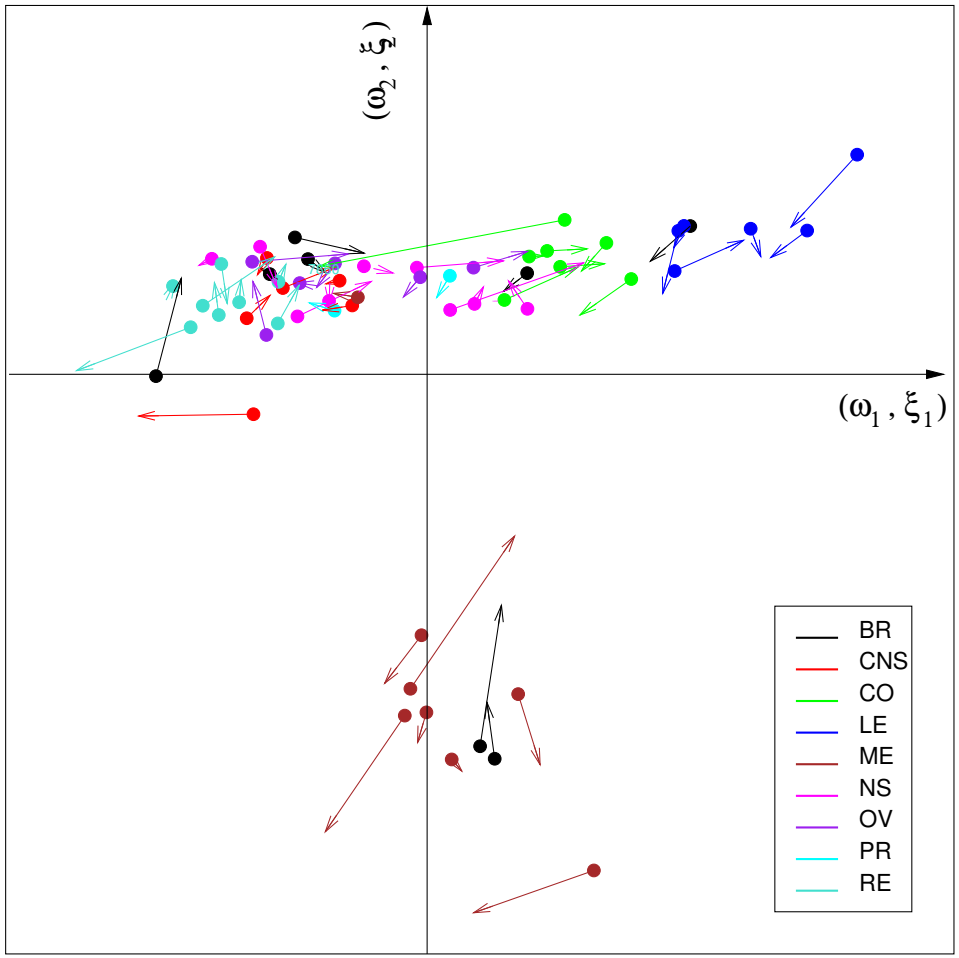
Vecteurs *loading* associés à la matrice $\mathbf{X}$

PLS version analyse canonique (Lê Cao et al. 2009)

NCI : 60 lignées cellulaires de tumeurs

CO	RE	OV	BR	PR	CNS	LEU	ME
7	8	6	8	2	9	6	8

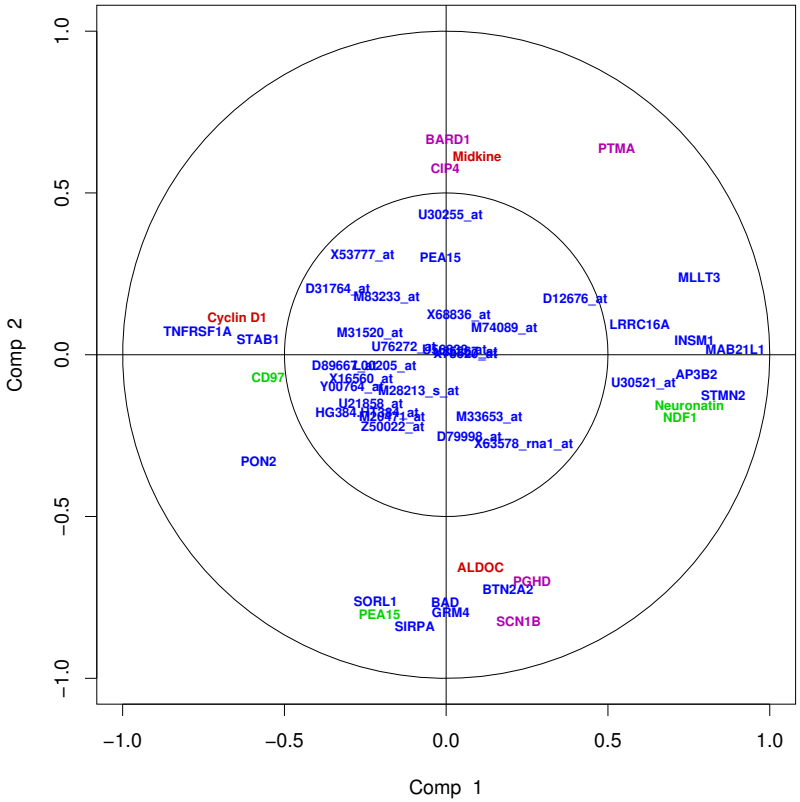
- Épithéliales, Mésenchymales, Mélanomes
- Deux plateformes :
  X = cDNA chip data, $p = 1375$
  Y = Affymetrix chip, $q = 1517$
- Données symétriques
- Recouvrement des gènes exprimés et des compléments

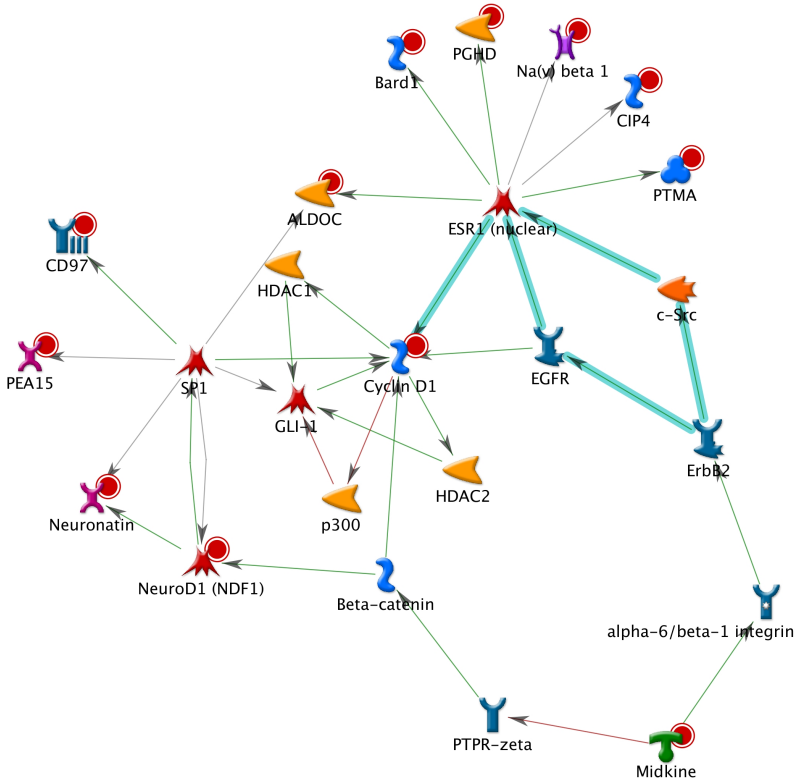


NCI : Representation simultanée

Discrimination par s-PLS-DA : données de tumeur du cerveau

- $n = 90$
- $p = 6144$ expressions de gènes ou variables X
- Variables qualitatives Y à 5 modalités (type de tumeur)
- Objectif : diagnostiquer le type de tumeur à partir de l'expression des gènes
- Lê cao et al. (2011) présente une comparaison détaillée
- Plusieurs jeux de données et plusieurs approches de classification supervisée





Variables (gènes) *Gene Go software*

Brain data set : PLS et réseaux biologiques

Bach (2008)

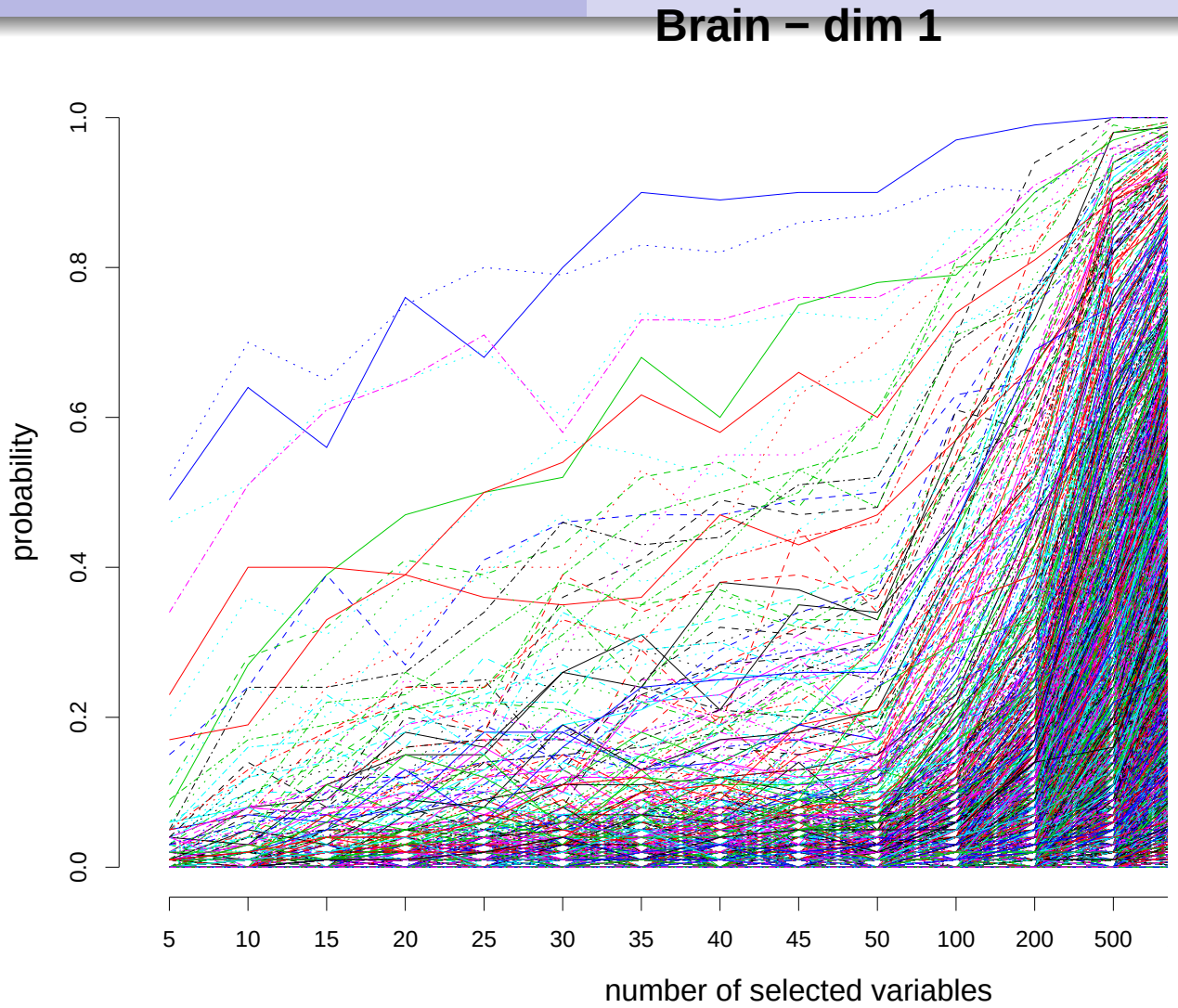
- Modèle linéaire et Lasso, échantillons bootstrap
- Intersection des sélections

Meinshausen et Bülhmann (2010)

- Modèle linéaire, modèles graphiques, Lasso, random lasso
- Échantillons aléatoires taille $n/2$ sans remise
- Graphe : probabilité de sélection fonction de la pénalité

Verzelen (2010)

- Rappel dans le cas gaussien : $\frac{2k\log(p/k)}{n} > \frac{1}{2}$
- $n = 90$ et $p = 6144$ supposent $k < 6$



Brain data set : Stabilité de la sélection

Méthodes avec pénalisation : bibliographie

- Tibshirani (1996) : Modèle linéaire et Lasso
- Zou et Hastie (2005) : Modèle linéaire et Elastic Net
- Jolliffe et al. (2003), Zou et al. (2006), Shen et Huang (2008) : sparse ACP
- González et al. (2009) : Analyse canonique ridge (Vinod, 1976)
- Chun et Keles (2007) : PLS mode régression et Elastic Net
- Waaijenborg et al. (2008), Parkhomenko et al. (2009), Witten et al. (2009) : PLS mode canonique et Elastic Net
- Lê Cao et al. (2008), Chun et Keles (2010) : sparse PLS mode régression
- Lê Cao et al. (2009) : sparse PLS mode canonique
- Ahdesmäki and Strimmer (2009) : sparse LDA
- Chung et Keles (2010), Lê Cao et al. (2010) : sparse PLS-DA

Références

Ahdesmäki, M. and Strimmer, K. (2009). Feature selection in omics prediction problems using cat scores and false non-discovery rate control. Ann. Appl. Stat.

Antoniadis, A., Lambert-Lacroix, S., and Leblanc, F. (2003). Effective dimension reduction methods for tumor classification using gene expression data. *Bioinformatics*, 19(5) :563-570.

Bach, F. (2008). Bolasso : model consistent Lasso estimation through the bootstrap. *Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML)*.

Boulesteix, A. (2004). PLS Dimension Reduction for Classification with Microarray Data. *Statistical Applications in Genetics and Molecular Biology*, 3(1) :1075.

Breiman, L. (2001). Random forests. *Machine learning*, 45(1) :5-32.

Chun, H. and Keles, S. (2010). Sparse partial least squares regression for simultaneous dimension reduction and variable selection. *Journal of the Royal Statistical Society* : Series B, 72(1) :3-25.

Chung, D. and Keles, S. (2010). Sparse Partial Least Squares Classification for High Dimensional Data. *Statistical Applications in Genetics and Molecular Biology*, 9(1) :17.

Dai, J., Lieu, L., and Rocke, D. (2006). Dimension reduction for classification with gene expression microarray data. *Statistical Applications in Genetics and Molecular Biology*, 5(1) :1147.

Ding, B. and Gentleman, R. (2005). Classification using generalized partial least squares. *Journal of Computational and Graphical Statistics*, 14(2) :280-298.

Fort, G. and Lambert-Lacroix, S. (2005). Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(7) :1104.

Gadat, S. and Younes, L. (2007). A stochastic algorithm for feature selection in pattern recognition. *The Journal of Machine Learning Research*, 8 :547.

González I., Déjean S., Martin P.G.P., Goncalves O., Besse P. and Baccini A. (2009) Highlighting Relationships Between Heteregeneous Biological Data Through Graphical Displays Based On Regularized Canonical Correlation Analysis, *Journal of Biological Systems* 17(2), pp 173-199.

Guyon, I., Elisseefi, A., and Kaelbling, L. (2003). An Introduction to Variable and Feature Selection. *Journal of Machine Learning Research*, 3(7-8) :1157-1182.

Guyon, I., Weston, J., Barnhill, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1) :389-422.

...

Huang, X., Pan, W., Park, S., Han, X., Miller, L., and Hall, J. (2004). Modeling the relationship between LVAD support time and gene expression changes in the human heart by penalized partial least squares. *Bioinformatics*, 4991.
Jolliffe, I., Trendafilov, N., and Uddin, M. (2003). A Modified Principal Component Technique Based on the LASSO. *Journal of Computational & Graphical Statistics*, 12(3) :531-547.
Lê Cao K.-A., Boitard, S. and Besse, P. (submitted) Multiclass classification with sPLS-DA, graphical interpretation and comparison with wrapper approaches.
Lê Cao, K.-A., Bonnet, A., and Gadat, S. (2009a). Multiclass classification and gene selection with a stochastic algorithm. *Computational Statistics and Data Analysis*, 53 :3601-3615.
Lê Cao, K.-A., Goncalves, O., Besse, P., and Gadat, S. (2007). Selection of biologically relevant genes with a wrapper stochastic algorithm. *Statistical Applications in Genetics and Molecular Biology*, 6(1) :29.
Lê Cao K.-A., González, I. and Déjean, S. (2009) integrOmics/mixOmics : an R package to unravel relationships between two omics data sets *Bioinformatics*, 25(21) :2855-2856.
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 :Iss. 1, Article 35.
Meinshausen, N. and Bühlmann, P. (2008). Stability selection. *Journal of the Royal Statistical Society* : Series B, 72, 417-473.
Nguyen, D. and Rocke, D. (2002a). Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18(9) :1216.
Nguyen, D. and Rocke, D. (2002b). Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1) :39.
Parkhomenko, E., Tritchler, D., and Beyene, J. (2009). Sparse canonical correlation analysis with application to genomic data integration. *Statistical Applications in Genetics and Molecular Biology*, 8(1) :1.

Shen, H. and Huang, J. Z. (2008). Sparse principal component analysis via regularized low rank matrix

approximation. *Journal of Multivariate Analysis*, 99 :1015-1034.

...

Tan, Y., Shi, L., Tong, W., Gene Hwang, G., and Wang, C. (2004). Multi-class tumor classification by discriminant partial least squares using microarray gene expression data and assessment of classification models. *Computational Biology and Chemistry*, 28(3) :235-243.
Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, Series B, 58(1) :267-288.
Tibshirani, R., Hastie, T., Narasimhan, B., and Chu, G. (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10) :6567.
Waaijenborg, S., de Witt Hamer, V., Philip, C., and Zwinderman, A. (2008). Quantifying the Association between Gene Expressions and DNA-Markers by Penalized Canonical Correlation Analysis. *Statistical Applications in Genetics and Molecu- lar Biology*, 7(3).
Witten, D., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3) :515.
Wold, H. (1966). *Multivariate Analysis*. Academic Press, New York
Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical*, Society Series B, 67(2) :301-320.

Zou, H., Hastie, T. and Tibshirani, R. (2006). Sparse principal component analysis. *Journal of Computational and*

Graphical Statistics, 15(2) :265-286.