Composantes principales et régressions PLS parcimonieuses

Résumé

L'introduction de pénalisations en norme L_1 induit une sélection de variables optimale en régression. Même si, numériquement, ce n'est pas indispensable pour les méthodes de régression ou projection sur composantes orthogonales avec réduction de dimension, le même type de pénalisation est introduit afin de simplifier la construction des composantes et donc leur interprétation lorsque le nombre de variables est important. Cette démarche conduit à la définition de versions parcimonieuses de l'Analyse en Composantes Principales et de la régression PLS pour différents objectifs : exploration, comparaison ou intégration de deux jeux de données en version régression ou canonique, analyse discriminante PLS.

Retour au plan du cours

1 Introduction

1.1 Objectif

L'intérêt principal des méthodes de cette vignette réside dans leur capacité à prendre en compte des données de grande dimension et même de très grande dimension lorsque le nombre de variables p est largement plus grand que le nombre d'individus n : $p >> n$. La sélection de variables devient inefficace et même ingérable par les algorithmes usuels. La construction d'un modèle de régression requiert alors une pénalisation (*ridge, Lasso, elastic net*) ou une réduction de dimension : régression sur composantes principales ou régression PLS.

1.2 Régression sur composantes principales

La régression sur composantes principales ou PCR est simple par son principe et sa mise en œuvre. L'objectif est de résumer l'ensemble des variables $X^1, \ldots, X^p$ par un sous-ensemble de variables $Z^1, \ldots, Z^r$ deux à deux orthogonales et combinaisons linéaires des variables $X^1, \ldots, X^p$. Avec $r = p$ il n'y a pas de réduction de dimension et le même ajustement qu'en régression classique est obtenu : même espace de projection engendré. Les variables $Z^1, \ldots, Z^p$ sont simplement les composantes principales associées des variables $X^1, \ldots, X^p$ obtenues par l'analyse en composantes principales ou encore la décomposition en valeurs singulières de la matrice $\mathbf{X}$. Pour éviter les problèmes d'unité et l'influence d'une hétérogénéité des variances, les variables sont centrées et réduites ; c'est donc l'ACP réduite qui est calculée.

La première composante $Z^1 = \sum_{j=1}^{p} \alpha_j X^j$ est de variance maximale la première valeur propre λ_1 de la matrice des corrélations avec $\sum \alpha_j^2 = 1$. Tandis que Z^m est combinaison linéaire de variance maximale λ_j et orthogonale à $Z^1, \ldots, Z^{m-1}$.

La PCR considère un prédicteur de la forme :

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

avec

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

obtenu par une procédure classique de régression.

Le choix $r = p$ redonne l'estimateur des moindres carrés car le même espace est engendré tandis que $r < p$ élimine les composantes de variances nulles ou très faibles et donc résout par là les problèmes de colinéarité même dans les cas extrêmes où ($p > n$). Le choix de r est optimisé de par validation croisée.

Bien évidemment, l'interprétation des composantes est rendu difficile si p est grand. La PCR est à rapprocher de la régression *ridge* qui seuille les coefficients des composantes principales tandis que la PCR annule ceux d'ordre supérieur à r.

Le principal *Problème* posée par la PCR est que les premières composantes, associées aux plus grandes valeurs propres, ne sont pas nécessairement corrélées avec Y et ne sont donc pas nécessairement les meilleures candidates pour résumer ou modéliser Y.

Cette remarque justifie les développements de la régression PLS ou *partial least square*.

1.3 Régression PLS

La régression PLS (*partial least square*) est une méthode ancienne (Wold, 1966)[10] largement utilisée, notamment en chimiométrie dans l'agro-alimentaire, lors de l'analyse de données spectrales (Near Infra-Red ou HPLC) discrétisées et donc toujours de grande dimension. La régression PLS s'avère concrètement une méthode efficace qui justifie son emploi très répandu mais présente le défaut de ne pas se prêter à une analyse statistique traditionnelle qui exhiberait les lois de ses estimateurs. Elle est ainsi restée un marge des approches traditionnelles de la Statistique mathématique.

Différentes version de régression PLS sont proposées en fonction de l'objectif poursuivi ; consulter Tenenhaus (1998)[8] pour une présentation détaillée :

PLS1 Une variable cible Y quantitative est à expliquer, modéliser, prévoir par p variables explicatives quantitatives X^j.

PLS2 Version canonique. Mettre en relation un ensemble de q variables quantitatives Y^k et un ensemble de p variables quantitatives X^j.

PLS2 Version régression. Chercher à expliquer, modéliser un ensemble de q variables Y^k par un ensemble de p variables explicatives quantitatives X^j.

PLS-DA Version discriminante. Cas particulier du cas précédent. La variable Y qualitative à q classes est remplacée par q variables indicatrices (*dummy variables*) de ces classes.

Une application utile de la PLS2 en version canonique s'opère, par exemple en Biologie à haut débit, dans la comparaison de deux plates-formes ou deux technologies de mesures sur le même échantillon : Affymetrix *vs.* Agilent ou encore entre les résultats obtenus par séquençage (RNA Seq) et biopuces. Toujours en Biologie, la PLS2 en version régression permet d'intégrer des jeux de données observées à des niveaux différents sur le même échantillon : expliquer par exemple un ensemble de métabolites ou de phénotypes par des transcrits. Des applications industrielles se trouvent,par exemple, en suivi de la qualité : plusieurs variables de mesure de la qualité ou de défaillances expliquées par un ensemble de mesures du procédé de fabrication.

Dans un objectif seulement prévisionnel, l'approche PLS s'avère plutôt efficace mais, si l'objectif est aussi la recherche d'une interprétation, c'est-à-dire nécessairement la recherche des variables les plus pertinentes parmi un très grand nombre, les composantes obtenues sont difficilement exploitables. C'est pourquoi il a été proposé (Lê Cao et al. 2008[5], 2009[4], 2011[3]) de coupler les deux approches : pénalisation L_1 de type Lasso pour une sélection des variables utilisées dans la construction des composantes orthogonales. Cette démarche passe par l'utilisation d'un algorithme parcimonieux (Shen et Huang, 2008)[7] de SVD (décomposition en valeur singulière). Celui-ci permet, à la fois, de définir des versions parcimonieuses de l'ACP et aussi de la PLS en remarquant que l'algorithme de la PLS peut être défini comme une succession de premières étapes de SVD.

L'objectif principal est donc la construction de versions parcimonieuses (en anglais *sparse*) des différentes méthodes de régression PLS. Aux résultats numériques, éventuellement de prévision, s'ajoutent des *représentations graphiques* en petite dimension très utiles pour aider à l'interprétation.

2 Régression PLS

Quelques rappels pour introduire cette méthode largement employée pour traiter les situations présentant une forte multicolinéarité et même lorsque le nombre d'observations est inférieur au nombre de variables explicatives.

2.1 Régression PLS1

Une variable cible Y quantitative est à expliquer, modéliser, prévoir par p variables explicatives quantitatives X^j. Comme pour la régression sur composantes principales, le principe est de rechercher un modèle de régression linéaire sur un ensemble de composantes orthogonales construites à partir de combinaisons linéaires des p variables explicatives centrées X^j. Dans le cas de la PLS, la construction des composantes est optimisée pour que celles-ci soient les plus liées à la variable Y à prédire au sens de la covariance empi-

rique, alors que les composantes principales ne visent qu'à extraire une part de variance maximale sans tenir compte d'une variable cible.

Soit $\mathbf{X}(n \times p)$ la matrice des variables explicatives centrées avec n pouvant être inférieur à p. On cherche une matrice $\mathbf{U}$ de coefficients ou pondérations (*loading vectors*) définissant les r composantes Ξ_h (ou variables latentes) par combinaisons linéaires des variables X_j :

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

La matrice $\mathbf{U}$ est solution du problème suivant :

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

La matrice $\mathbf{U}$ est obtenue par la démarche itérative de l'algorithme 0 ; il suffit ensuite de calculer la régression de Y sur les r variables $\boldsymbol{\xi}_h$ centrées, appelées *variables latentes* ainsi construites. Le choix du nombre de composantes r est optimisé par validation croisée.

Algorithm 1 Régression PLS1

$\mathbf{X}$ matrice des variables explicatives centrées,
Calcul de la matrice $\mathbf{U}$ des coefficients.
for $h = 1$ à r **do**
  $\mathbf{u}_h = \frac{\mathbf{X}'\mathbf{Y}}{\|\mathbf{X}'\mathbf{Y}\|}$,
  $\boldsymbol{\xi}_h = \mathbf{X}\mathbf{u}_h$
  Déflation de $\mathbf{X} : \mathbf{X} = \mathbf{X} - \xi_h\xi_h'\mathbf{X}$
end for

Exemple de PLS1 sur les données de cancer de la prostate

La figure 1 donne l'estimation par validation croisée (10-fold) de l'erreur de prévision en fonction de la dimension tandis que la figure 2 (gauche) est une aide à l'interprétation. Les *loadings* sont les coefficients ou importance

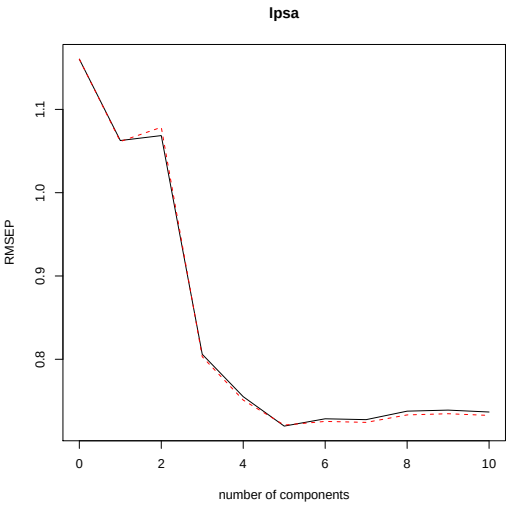

FIGURE 1 – Données cancer : optimisation du nombre de composantes en PLS1

des variables sur la première composante PLS. Le graphe de droite de la figure 2 indique simplement la plus ou moins bonne qualité de l'ajustement avec un choix de 6 composantes PLS.

2.2 Régression PLS2

Définition

L'algorithme précédent de PLS1 se généralise à une variable à expliquer Y multidimensionnelle (PLS2) : Mettre en relation ou chercher à expliquer, modéliser un ensemble de q variables Y^k par un ensemble de p variables explicatives X^j. Le critère à optimiser devient une somme des carrés des covariances entre une composante et chacune des variables réponses. Plusieurs variantes de la régression PLS multidimensionnelle ont été proposées ; le même critère est optimisé mais sous des contraintes différentes. La version *canonique* (par référence à l'analyse canonique de deux ensembles de variables), où les

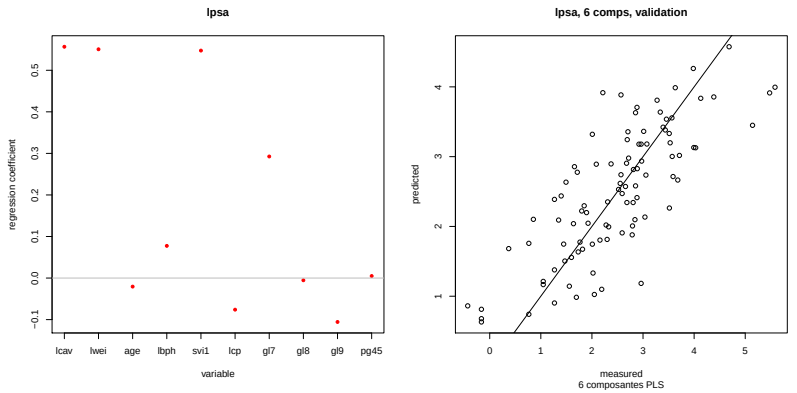

FIGURE 2 – Données cancer : Coefficient (loadings) des variables sur la première composante et qualité de l'ajustement avec 6 composantes.

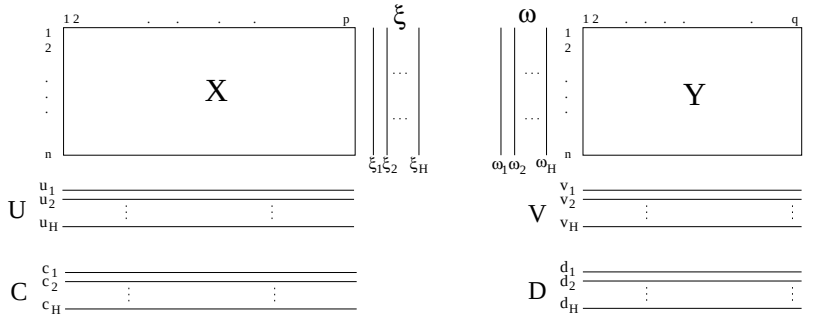

FIGURE 3 – *PLS2 : les matrices* $\mathbf{X}$ *and* $\mathbf{Y}$ *sont successivement décomposées en ensembles de coefficients (loading vectors)* $(\mathbf{u}_1, \ldots, \mathbf{u}_r)$, $(\mathbf{v}_1, \ldots, \mathbf{v}_r)$ *et ensembles de variables latentes* $(\xi_1, \ldots, \xi_r)$, $(\boldsymbol{\omega}_1, \ldots, \boldsymbol{\omega}_r)$, *où* r *est la dimension recherchée ou nombre de composantes.*

deux ensembles de données jouent des rôles symétriques, diffère de la version *régression* (un paquet de variable expliqué par un autre) par l'étape dite de déflation de l'algorithme général de PLS.

Dans les deux cas, la PLS se définit par la recherche (cf. 3) de :

- 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 sous contraintes d'orthogonalité par déflations de $\mathbf{X}$ et $\mathbf{Y}$.
- Les vecteurs de coefficients $(\mathbf{u}_h, \mathbf{v}_h)_{h=1,\ldots,r}$ sont appelés vecteurs *loadings*.

Tous ces vecteurs sont schématisés dans la figure 3.

Deux types de déflation sont considérés, l'un faisant jouer un rôle symétrique entre les variables (mode canonique), tandis que l'autre suppose que les variables X sont expliquées par celles Y. La régression PLS est à rapprocher de l'analyse canonique des corrélations qui s'utilise dans le même contexte de deux variables multidimensionnelles X et Y à mettre en relation. La différence vient du critère optimisé en analyse canonique qui est la corrélation entre les variables latentes plutôt que la covariance :

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

Cette optimisation requiert l'inversion des matrices $\mathbf{X}'\mathbf{X}$ et $\mathbf{Y}'\mathbf{Y}$. Ces inversions sont impossibles en cas de colinéarité des variables et donc évidemment si $n < p$ ou $n < q$. Une version régularisée ou *ridge* de l'analyse canonique rend les calculs possibles (Gonzales et al. 2008) mais les interprétations restent difficiles pour des grandes valeurs de p ou q.

Algorithme

Historiquement, la régression PLS est construite par l'algorithme NIPALS (Non linear Iterative PArtial Least Square algorithm) (cf. 2) dans lequel chaque itération h, $h = 1, \ldots, r$ de l'algorithme décompose $\mathbf{X}$ et $\mathbf{Y}$ en faisant intervenir une étape de déflation spécifique à l'objectif.

Cet algorithme, en itérant des régressions partielles, présente de nombreux avantages. Il n'est pas nécessaire d'inverser une matrice comme en analyse

canonique ; de plus il accepte des données manquantes et même propose, par la PLS, une méthode d'imputation de celles-ci.

Algorithm 2 NIPALS

$\mathbf{X}$ et $\mathbf{Y}$ matrices des données centrées
Initialiser $\boldsymbol{\omega}_1$ par la première colonne de Y
for $h = 1$ à r **do**
  while Convergence pas atteinte **do**
    $\mathbf{u}_h = \mathbf{X}'\boldsymbol{\omega}_h / \boldsymbol{\omega}_h'\boldsymbol{\omega}_h$
    $\mathbf{u}_h = \mathbf{u}_h / \mathbf{u}_h'\mathbf{u}_h$ est le vecteur *loading* associé à $\mathbf{X}$
    $\boldsymbol{\xi}_h = \mathbf{X}\mathbf{u}_h$ est la *variable latente* associée à $\mathbf{X}$
    $\mathbf{v}_h = \mathbf{Y}'\boldsymbol{\xi}_h / (\boldsymbol{\xi}_h'\boldsymbol{\xi}_h)$
    $\mathbf{v}_h = \mathbf{v}_h / \mathbf{v}_h'\mathbf{v}_h$ est le vecteur *loading* associé à $\mathbf{Y}$
    $\boldsymbol{\omega}_h = \mathbf{Y}'\mathbf{v_h}$ est la variable latente associée à $\mathbf{Y}$
  end while
  $\mathbf{c}_h = \mathbf{X}'\boldsymbol{\xi} / \boldsymbol{\xi}'\boldsymbol{\xi}$ régression partielle de $\mathbf{X}$ sur $\boldsymbol{\xi}$
  $\mathbf{d}_h = \mathbf{Y}'\boldsymbol{\omega} / \boldsymbol{\omega}'\boldsymbol{\omega}$ régression partielle de $\mathbf{Y}$ sur $\boldsymbol{\omega}$
  Résidus $\mathbf{X} \leftarrow \mathbf{X} - \boldsymbol{\xi}\mathbf{c}'$ ou *déflation*
  Résidus $\mathbf{Y} \leftarrow \mathbf{Y} - \boldsymbol{\omega}\mathbf{d}'$ ou *déflation*
end for

Le nombre r d'itérations est à fixer ou optimiser par l'utilisateur tandis que la convergence de chaque étape h est analogue à celle, relativement rapide (moins d'une dizaine d'itérations), d'un algorithme de puissance itérée. En effet, à la convergence, les vecteurs vérifient :

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

où $\mathbf{u}, \boldsymbol{\omega}, \mathbf{v}$ et $\boldsymbol{\xi}$ sont donc les vecteurs propres respectifs des matrices $\mathbf{Y}\mathbf{Y}'\mathbf{X}\mathbf{X}'$, $\mathbf{Y}'\mathbf{X}\mathbf{X}'\mathbf{Y}$, $\mathbf{X}\mathbf{X}'\mathbf{Y}\mathbf{Y}'$, $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$ associés à la même plus grande valeur propre λ. L'étape de déflation permet donc de calculer successivement les vecteurs propres associés aux valeurs propres décroissantes.

En résumé,

- La régression PLS2 gère des données incomplètes, bruitées, colinéaires ou de très grande dimension
- calcule les variables latentes $\boldsymbol{\xi}_h$ et $\boldsymbol{\omega}_h$ qui renseignent (graphes) sur les similarités et/ou dissimilarités des observations,
- et les vecteurs *loading* $\mathbf{u}_h$ et $\mathbf{v}_h$ qui renseignent sur l'importance des variables X_j et Y_k,
- trace les Graphes illustrant les covariations des variables.

Variante de l'algorithme

Une autre approche consiste à calculer directement les vecteurs propres de la matrice $\mathbf{X}'\mathbf{Y}\mathbf{Y}'\mathbf{X}$ ou encore et c'est équivalent, les valeurs et vecteurs singuliers de la décomposition en valeurs singulières (SVD) de la matrice $\mathbf{X}'\mathbf{Y}$. Néanmoins, la perspective de gérer les données manquantes ou encore celle de réaliser les calculs sans avoir à stocker des matrices $p \times p$ pour p très grand, rend l'algorithme NIPALS tout à fait pertinent même s'il est numériquement moins performant.

PLS mode Régression vs. canonique

Deux modes de déflation sont proposés selon que les variables jouent un rôle symétrique ou que les variables X sont supposées expliquées par celles Y.

- 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 $\mathbf{Y}_h = \mathbf{Y}_{h-1} - \boldsymbol{\xi}_h\mathbf{v}'_h$

La PLS en mode canonique poursuit donc le même objectif que l'analyse canonique des corrélations en rendant les calculs possibles même si $p > n$ car la PLS ne nécessite pas l'inversion des matrices de corrélation. Toujours avec le même objectif de rendre possible les calculs, des versions régularisées (norme L_2) de l'analyse canonique ont été proposées de façon analogue à la régression ridge. Néanmoins, cette approche conduit à des graphiques et interprétations difficiles lorsque p est grand.

PLS-DA ou discrimination PLS

La régression PLS peut facilement s'adaptée au cas de la classification supervisée, ou analyse discriminante décisionnelle (PLS-*Discriminant Analysis*), dans lequel p variables quantitatives X^j expliquent une variable qualitative Y à m modalités. Il suffit de générer le paquet des m variables indicatrices

ou *dummy variables* Y^k et d'exécuter l'algorithme PLS2 (mode régression) en considérant comme quantitatives ces variables indicatrices. Le choix du nombre de dimensions peut être optimisé en minimisant l'erreur de prévision des classes par validation croisée.

2.3 Représentations graphiques

Les représentations graphiques des individus, comme celles des variables initiales, sont analogues à celles obtenues en analyse canonique.

- Les variables initiales sont représentées par leurs coefficients sur les variables latentes ;
- les individus par leurs valeurs sur les composantes de X (ou de Y) comme en ACP.

3 Méthodes parcimonieuses

3.1 Objectif

La régression PLS est une régression sur composantes orthogonales qui résout efficacement les problèmes de multicolinéarité ou de trop grand nombre de variables en régression comme en analyse canonique. La contre partie, ou prix à payer, est l'accroissement souvent rédhibitoire de la complexité de l'interprétation des résultats. En effet, chaque composante est obtenue par combinaison linéaire d'un nombre pouvant être très important de l'ensemble des p variables.

Pour aider à l'interprétation, l'objectif est donc de limiter, ou contraindre, le nombre de variables participant à chaque combinaison linéaire. La façon simple de procéder est d'intégrer une contrainte de type Lasso dans l'algorithme PLS2. Plusieurs approches on été proposées, celle décrite ci-après s'avère rapide et efficace.

3.2 Sparse SVD

La démarche adoptée est issue d'une construction d'une version parcimonieuse de l'ACP proposée par Shen et Huang (2008)[7]. Considérant que l'ACP admet pour solution la décomposition en valeurs singulières (SVD) de la matrice centrée $\bar{\mathbf{X}}$, la *sparse* PCA (s-PCA) est basée sur un algorithme qui résout le problème :

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

où le vecteur $\mathbf{v}$ contient les paramètres des combinaisons linéaires des variables initiales. Une pénalisation de type L_1 ($\lambda||\mathbf{v}||_1$) conduit à l'annulation des paramètres les plus petits pour ne laisser qu'un ensemble restreint de paramètres non-nuls dont l'effectif dépend directement de la valeur λ de la pénalisation.

Algorithm 3 *sparse* SVD

Décomposer $\mathbf{M} = \mathbf{U}\Delta\mathbf{V}'$
$\mathbf{M}_0 = \mathbf{M}$
for h de 1 à r **do**
  Fixer $v_{old} = \delta_h v_h^\star$
  $u_{old} = u_h^\star$ avec $v_h^\star$ et $v_h^\star$ de norme 1
  while Pas de convergence de u_{new} et v_{new} **do**
    $v_{new} = g_\lambda(\mathbf{M}'_{h-1}u_{old})$
    $u_{new} = \mathbf{M}'_{h-1}v_{new}/||\mathbf{M}_{h-1}v_{new}||$
    $u_{old} = u_{new}$, $v_{old} = v_{new}$
  end while
  $v_{new} = v_{new}/||v_{new}||$
  $\mathbf{M}_h = \mathbf{M}_{h-1} - \delta_h u_{new} v'_{new}$
end for

L'algorithme peut adopter différents types de fonction de pénalisation, celle retenue est une fonction de seuillage "doux" avec

$$g_\lambda(y) = sign(y)(|y| - \lambda)_+.$$

3.3 Sparse PLS

Ayant remarqué qu'un étape h de PLS2 est la première étape de la décomposition en valeur singulière de la matrice $\mathbf{M}_h = \mathbf{X}'_h\mathbf{Y}_h$, la version parcimonieuse de la PLS2 est simplement construite en itérant r fois l'algorithme de *sparse* SVD (s-SVD) qui cherche à résoudre :

$$\min_{\mathbf{u}_h,\mathbf{v}_h} ||\mathbf{M}_h - \mathbf{u}_h\mathbf{v}'_h||_F^2 + P_{\lambda_1}(\mathbf{u}_h) + P_{\lambda_2}(\mathbf{v}_h.$$

Comme pour l'algorithme de *sparse*-SVD, une pénalisation de type L_1 ($\lambda||\mathbf{v}||_1$) conduit à l'annulation des paramètres les plus petits pour ne laisser qu'un ensemble restreint de paramètres non-nuls dont l'effectif dépend directement des valeurs λ_1 et λ_2 de pénalisation.

Plus précisément, l'algorithme adopte pour pénalisation des fonctions de seuillage "doux" composante par composante avec

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

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

Entre deux étapes de s-SVD, les matrices $\mathbf{X}_h$ et $\mathbf{Y}_h$ subissent une déflation (mode régression ou canonique) avant de passer à l'étape suivante.

Cette démarche soulève des questions délicates d'optimisation du nombre r de dimensions et celle des valeurs des paramètres de la fonction de pénalisation. En mode régression (PLS2 ou PLS-DA) il est possible d'optimiser ces choix en minimisant des erreurs de prévision estimées par validation croisée. En mode canonique, le "degré" de parcimonie comme le nombre de dimensions doivent être fixés *a priori* par l'utilisateur. Plus concrètement, ce sont souvent des choix a priori qui sont opérés en fonction de l'objectif de l'utilisateur : recherche de peu de variables assimilées, par exemple, à des biomarqueurs ou de "beaucoup" de variables dans le cadre d'une tentative de compréhension globale de la structure des données. De même, le nombre de composantes r est choisi avec une valeur réduite afin de construire des représentations graphiques $r \leq 3$ plus élémentaire pour aider à l'interprétation.

En résumé, ce sont donc les capacités d'interprétation d'un problème qui guident concrètement le choix à moins qu'un objectif de construction d'un meilleur modèle de prévision conduisent à une optimisation par validation croisée.

Dans le cas particulier de PLS-DA, la sélection de variables s'opère sur le seul ensemble des variables X et donc un seul paramètre λ est à régler.

Attention, les variables latentes successivement calculées perdent leur propriété de stricte orthogonalité du fait de la pénalisation. Cela ne s'est pas avéré gênant sur les quelques premières dimensions et donc composantes calculées en pratique.

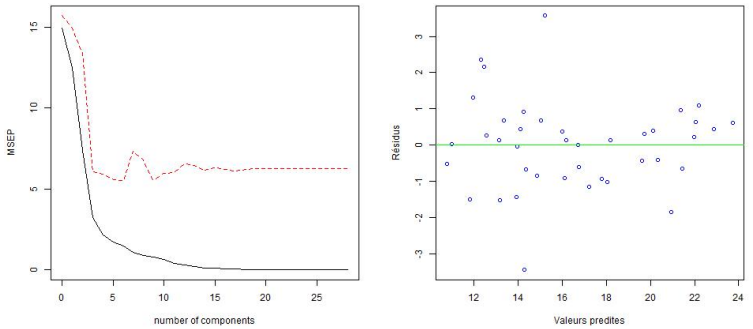

FIGURE 4 – Cookies : Optimisation du nombre de composante en régression PLS par validation croisée et graphe des résidus calculés sur l'échantillon test.

4 Exemples

4.1 PLS1 de données de spectrométrie NIR

Les données (*cookies*) sont celles étudiées par régression pénalisée. Comme pour les autres techniques, le paramètre de complexité, ici le nombre de composantes, est optimisé par validation croisée. Le graphe de la figure 4 montre l'évolution de l'erreur quadratique (ou risque) d'apprentissage (en noir) et de celle estimée par validation croisée (en rouge).

Une fois la dimension optimale déterminée, les prévisions des taux de sucre sont calculés pour l'échantillon test afin d'obtenir le graphe des résidus.

4.2 sPLS de données simulées

Le modèle de simulation est celui proposé par (Chun et Keles, 2010)[2]. Les données générées permettent de voir le rôle de la pénalisation dans la sélection des variables en PLS mode canonique. Elles sont constituées de

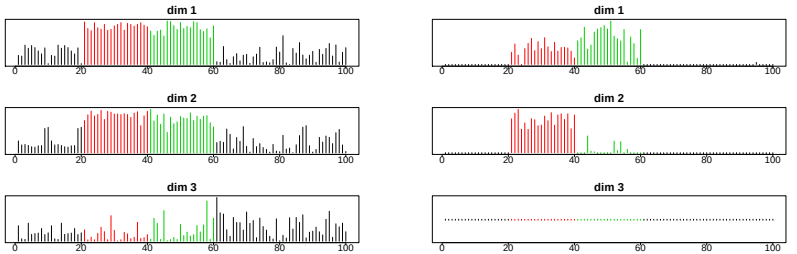

FIGURE 5 – Effet de la pénalisation sur les vecteurs “loading” associés à la matrice $\mathbf{X}$; PLS à gauche et sPLS à droite.

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

TABLE 1 – Effectifs des répartitions des échantillons des lignées cellulaires en 8 types de cancer et 3 types de cellules : épithéliales, mésenchymales, mélanomes

- $n = 40$, $p = 5000$ (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

4.3 Analyse canonique par sPLS2

Les données (NCI) concernent 60 lignées cellulaires de tumeurs. L’objectif est de comparer deux plate-formes. Sur la première (*cDNA chip*) ont été observées les expressions de $p = 1375$ gènes tandis que sur la 2ème (*Affymetrix*) ce sont $q = 1517$ gènes qui sont concernés. Une grande majorité des gènes, sont communs aux deux tableaux $\mathbf{X}(60 \times 1375)$ et $\mathbf{Y}(60 \times 1517)$.

Les deux technologies de mesure d’expression des gènes conduisent-elles à des résultats globalement comparables pour l’étude de ces lignées cellulaires cancéreuses ?

4.4 Recherche de bio-marqueurs par sPLS-DA

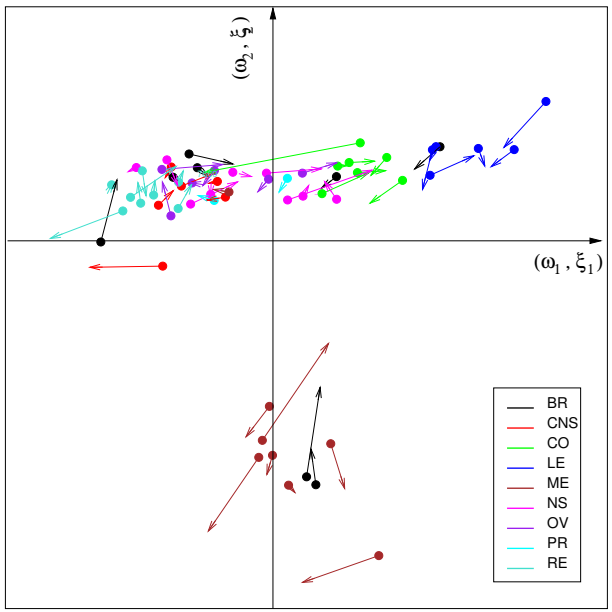

FIGURE 6 – Les “individus “lignées cellulaires” sont représentées dans les deux espaces : $(\boldsymbol{\xi}_1, \boldsymbol{\omega}_1)$ vs. $(\boldsymbol{\xi}_2, \boldsymbol{\omega}_2)$. La longueur de chaque vecteur souligne l’impact de la technologie utilisée sur chaque type de cellule.

Les données

Les qualités prédictives de la sPLS-DA peuvent-être comparées avec celles des autres méthodes abordées dans ce cours, notamment celles d'agrégation de modèles. Lê Cao et al. (2011)[3] ont mené cette comparaison systématique sur un ensemble de jeux de données publiques dont des données relatives à la discrimination de 5 types de cancer du cerveau. Les expressions de $p = 6144$ gènes sont observés sur $n = 90$ individus.

La question est donc de savoir si la mesure des expressions d'une sélection de gènes est pertinente pour aider à diagnostiquer ces différents types de cancer. Les gènes les plus discriminants au sens de la régressions sPLS-DA sont représentés par leur coefficients (*loadings*) dans la base des deux premières composantes (figure 7). Le réseau (figure 8) est celui des gènes connus dans la littérature pour intervenir sur ces pathologies.

Parmi les gènes sélectionnés par la sPLS-DA (figure 7), une couleur particulière est attribuée à ceux déjà connus et présents dans le réseau.

5 Robustesse d'une sélection

5.1 Principe

Le grand nombre de variables au regard de la taille de l'échantillon soulève quelques doutes quand à la robustesse ou la stabilité d'une sélection de variables au sein d'un modèle ; n'est-elle pas finalement qu'un artefact lié à l'échantillon observé ?

Bach (2008)[1] d'une part, Meinshausen et Bülhmann (2008)[6] d'autre part ont proposé des stratégies pour évaluer cette stabilité, éventuellement l'optimiser en les utilisant pour régler le paramètre de pénalisation. Ils partent de la même idée : étudier les occurrences ou non des sélections des variables dans un modèle pour une pénalisation donnée sur des échantillons bootstrap (Bach, 2008)[1] ou sur des sous-échantillons aléatoires (Meinshausen et Bülhmann, 2010)[6]. Ils étudient ces stratégies dans le cas du modèle linéaire avec pénalisation Lasso et montrent dans ce cas des propriétés asymptotiques de convergence vers la bonne sélection. Bach (2008)[1] s'intéresse à la sélection obtenue par intersection de toutes les sélections sur chacun des échantillons bootstrap tandis que Meinshausen et Bülhmann (2010)[6] compte le nombre de fois où

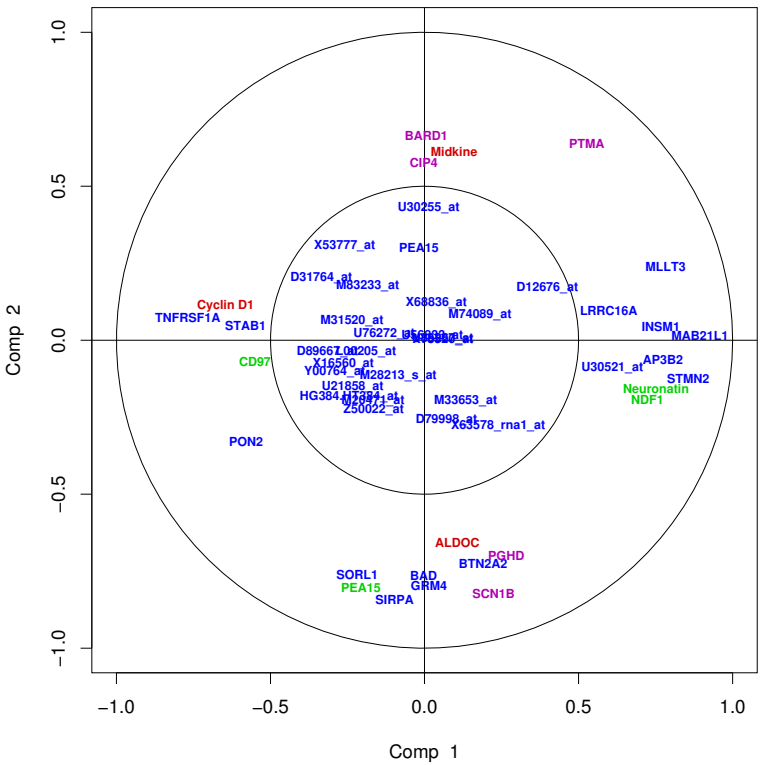

FIGURE 7 – Représentation des individus dans les deux premières composantes

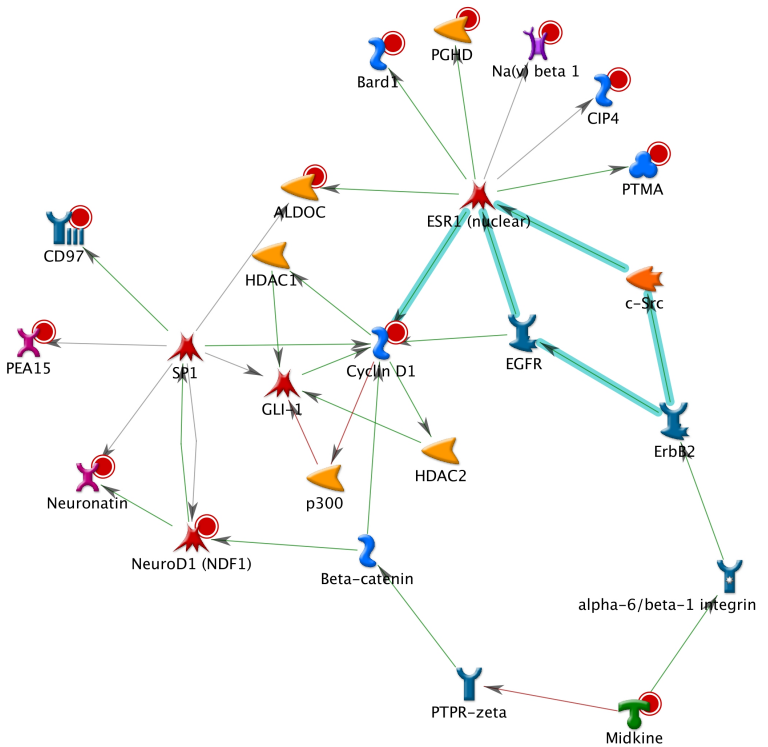

FIGURE 8 – Représentation (Gene Go software) en réseau des gènes déjà identifiés comme liés à ces pathologies de tumeurs cérébrales.

une variables est sélectionnée pour une valeur donnée de la pénalisation.

5.2 Exemple

Le graphique de la figure 9 est obtenu en synthétisant les stratégies précédentes. Sur chacun des 50 échantillons bootstrap, une sPLS-DA est calculée pour différentes valeurs de la pénalisation. On ne s'intéresse ici qu'à la première composante ($h = 1$). Dans ce cas de seuillage doux, la pénalisation revient à fixer le nombre de variables intervenant dans la construction de la première variable latente. La probabilité d'occurrence d'une variable ou gène est tout simplement estimée par le ratio du nombre de fois où elle a été sélectionnée. Quelques variables ou gènes apparaissent assez systématiquement sélectionnés,principalement 4 d'entre eux. Il apparaît que les données observées ne peuvent garantir la sélection que d'un nombre restreint de gènes. Ce constat serait à rapprocher du résultat théorique de Verzelen (2012)[9] dans le cas du modèle gaussien. Celui-ci met en évidence qu'un problème de ultra-haute dimension se manifeste si

$$\frac{2k\log(p/k)}{n} > \frac{1}{2}.$$

Avec les effectifs (n=90, p=6144) de l'exemple présenté, cette contrainte, dans le cas gaussien, signifierait qu'il est illusoire de vouloir sélectionner plus de 6 gènes. Pour un tout autre modèle, c'est aussi ce que nous signifie le graphique. Seule la considération d'un petit nombre de gènes dont la sélection est relativement stable sur les différents échantillons bootstrap est raisonnable sur ces données compte tenu de la faible taille de l'échantillon.

Références

[1] F. Bach, *Bolasso : model consistent Lasso estimation through the bootstrap*, Proceedings of the Twenty-fifth International Conference on Machine Learning (ICML) (2008), 33–40.

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

[3] K. A. Lê Cao, S. Boistard et P. Besse, *Sparse PLS Discriminant Analysis : biologically relevant feature selection and graphical displays for*

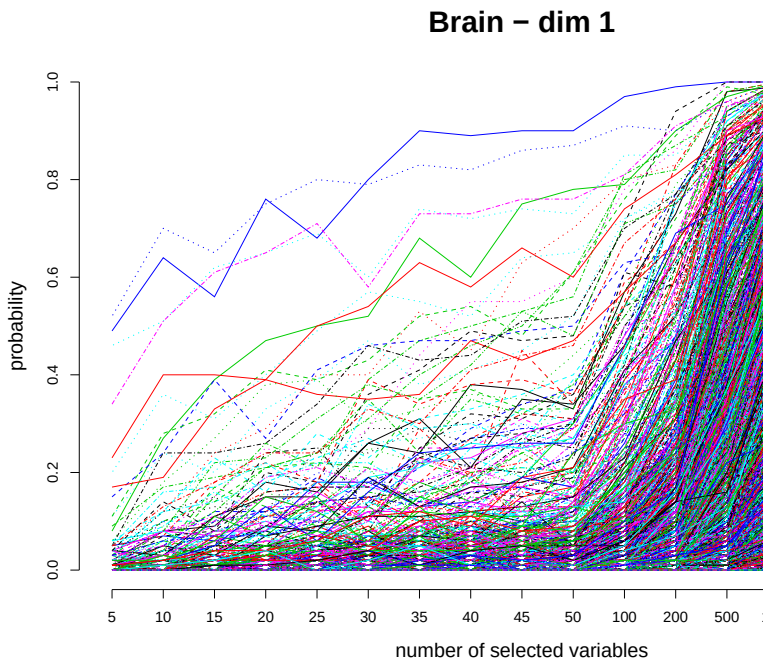

FIGURE 9 – Probabilités de sélection des différentes variables (gènes) sur la première composante en fonction de la valeur de la pénalisation en sPLS-DA.

multiclass problems, BMC Bioinformatics **12** (2011), n° 253.

[4] K. A. Lê Cao, P.G.P Martin, C. Robert-Granié et P. Besse, *Sparse Canonical Methods for Biological Data Integration : application to a cross-platform study*, BMC Bioinformatics **10** (2009), n° 34.

[5] K. A. Lê Cao, D. Rossouw, C. Robert-Granié et P. Besse, *A sparse PLS for variable selection when integrating Omics data*, Statistical Applications in Genetics and Molecular Biology **7** (2008), n° 35.

[6] N. Meinshausen et P. Bühlmann, *Stability selection*, Journal of the Royal Statistical Society : Series B **72** (2008), 417–473.

[7] H. Shen et J.Z. Huang, *Sparse principal component analysis via regularized low rank matrix approximation*, Journal of Multivariate Analysis **99** (2008), 1015–1034.

[8] M. Tenenhaus, *La régression PLS : théorie et applications*, Technip, 1998.

[9] Nicolas Verzelen, *Minimax risks for sparse regressions : Ultra-high-dimensional phenomenons*, Electron. J. Statistics **6** (2012), 38–90, http://arxiv.org/pdf/1008.0526.pdf.

[10] H. Wold, *Multivariate analysis*, Academic Press,, 1966.