Sélection de modèle PLS par rééchantillonnage bootstrap

Abdelaziz Faraj, Hicham Noçairi, Michel Constant

Institut Français du Pétrole
1&4 Av. de Bois-Préau
92500 Rueil Malmaison

{ abdelaziz.faraj, hicham.nocairi, michel.constant }@ifp.fr

Résumé. Le problème de la sélection de modèle en régression PLS est primordial pour la modélisation de phénomènes physiques même si le nombre des variables, pouvant être supérieur à celui des individus, paraît au premier abord peu important pour la mise en œuvre de la méthode. Les techniques de sélection consistent à retenir, parmi les modèles ayant un bon pouvoir de prédiction, ceux qui font intervenir le minimum de variables explicatives. La méthode que nous présentons dans ce papier est basée sur l'utilisation du bootstrap. Elle permet de calculer la distribution empirique des coefficients du modèle et de n'en conserver que les plus significatifs grâce à des tests statistiques. Elle mesure, par ailleurs, le pouvoir prédictif des modèles de régression construits aussi bien pour chaque individu que globalement. Nous illustrons cette approche en l'appliquant à un jeu de données.

Mots-Clés. Régression PLS, bootstrap, validation croisée, sélection de variables, sélection de modèles.

1 Introduction

Dans l'industrie pétrolière, la plupart des processus se présentent sous la forme d'un système à entrées-sorties (données de forage, simulations de gisement pétrolier, chimiométrie, données de procédés, etc.). Il est souvent nécessaire de faire recours à des modèles pour expliciter les relations pouvant exister entre les variables d'entrée et les réponses qui leur sont associées. De tels modèles doivent être explicatifs, c'est-à-dire éclairer les mécanismes des phénomènes physiques qu'ils décrivent. Ils doivent être prédictifs, c'est-à-dire donner, pour des valeurs des variables explicatives fixées, des sorties aussi proches que possible de celles obtenues par le processus expérimental. Et enfin – point primordial pour le praticien – ils doivent être opérationnels, c'est-à-dire pouvoir participer à l'amélioration du processus physique auquel ils sont rattachés (orienter le forage, diminuer le risque d'éboulement, augmenter la production, prédire les propriétés chimiques d'un composé, améliorer une méthode, etc.). Par ailleurs, pour des raisons de coût, le nombre des expériences qui servent à la construction de ces modèles est souvent faible voire inférieur à celui des variables en entrée

du système. La régression PLS apparaît comme la méthode la plus appropriée pour répondre à tous ces critères. Non seulement elle est bien adaptée quand les variables explicatives présentent des fortes colinéarités ou quand leur nombre dépasse celui des individus, elle est aussi une méthode factorielle qui a l'avantage d'apporter un point de vue exploratoire sur les données.

Mais il existe une autre contrainte : le praticien souhaite ne garder que les variables spécifiques de son processus expérimental. Cela aura pour avantage de diminuer le coût des expérimentations tout en focalisant l'étude sur les seules variables d'intérêt. Par ailleurs, les modèles construits, avec un nombre élevé de variables, sont souvent sur-ajustés et/ou d'interprétation confuse. C'est pour ces raisons que nous faisons appel aux méthodes de sélection de variables.

Nous présentons, dans ce papier, une méthode de sélection de variables en régression PLS basée sur le ré-échantillonnage par bootstrap. Notre but n'est pas de comparer les méthodes existantes les unes avec les autres (cela a fait l'objet de multiples travaux que nous détaillons ci-après) mais d'en proposer une qui puisse répondre aux objectifs suivants :

- élimination des variables explicatives non significatives (pour ne garder que les plus pertinentes) avec le souci d'améliorer la qualité de prédiction du modèle ;
- calcul des critères qui mesurent la qualité du modèle ;
- évaluation de son pouvoir de généralisation ;
- calcul des intervalles de confiance des prédictions et des coefficients du modèle ;
- détection des individus ou groupes d'individus atypiques (erreurs de mesures, valeurs extrêmes, classes particulières, …).

Notre méthode est basée sur un processus itératif de sélection de variables ; version modifiée de la PLS-bootstrap proposée par Lazraq et al. (2003). Des échantillons aléatoires sont tirés avec remise dans l'ensemble des points disponibles, et servent à la construction de plusieurs modèles. Des intervalles de confiance des coefficients de ces modèles sont déterminés. On calcule ensuite les prédictions de ces modèles sur les individus n'ayant pas été tirés (individus de test). Des indices statistiques (erreur quadratique de prédiction, biais, variance, …) sont calculés à chaque étape de la sélection afin d'évaluer les performances des modèles construits. Calculés pour chaque point, ces indices permettent de détecter les individus et/ou groupes particuliers d'individus grâce aux distributions empiriques obtenues.

Après une présentation détaillée des méthodes de sélection de variables, nous nous limiterons à la PLS-bootstrap de Lazraq et al. dont nous proposerons une version adaptée à notre problématique. La méthode sera ensuite appliquée à l'analyse et la modélisation d'un jeu de données.

2 Sélection de modèle en régression PLS

De nombreux auteurs ont travaillé sur la sélection de modèles de régression PLS. On peut notamment trouver une description bibliographique détaillée d'un certain nombre de ces travaux par Abrahamsson (2003), Gauchy et al. (2001) et Lazraq et al. (2003). Dans l'article de Lazraq, les auteurs classent les méthodes de sélection de variables - itératives pour la plupart - en 2 catégories : (i) **celles basées sur la réduction de la dimension** qui consistent à retenir (ou à éliminer) les variables les plus (ou les moins) significatives à chaque itération ; et (ii) **celles basées sur le modèle** qui consistent à construire un modèle, à chaque itération, puis à

appliquer une méthode de sélection (ou d'élimination) des variables les plus (ou les moins) significatives dans le modèle.

Gauchy et Chagnon (2001) présentent 20 méthodes de sélection de variables en régression PLS qu'ils appliquent, afin de les comparer, à 5 jeux de données. La méthode qu'ils proposent - appelée BQ (pour Backward-Q_{cum}^2) - est celle qui donne les meilleurs résultats, selon eux. C'est une méthode de sélection descendante qui consiste à éliminer à chaque pas la variable dont le coefficient de régression est le plus faible en valeur absolue. Le critère de sélection du modèle optimal est basé sur Q_{cum}^2 - indicateur de bonne prédiction du modèle - obtenu par validation croisée pour chaque modèle PLS. Le modèle correspondant à la valeur la plus élevée de Q_{cum}^2 est retenue.

Westad (1999) et Martens (2001) proposent une méthode de sélection de variables basée sur le jack-knife. Leur méthode remédie, nous citons, *au déficit d'une base mathématique qui rendrait possibles les tests statistiques pour les modèles de régression PLS*. La méthode du jack-knife est très proche de la PLS-bootstrap. Elle consiste, une fois le modèle PLS construit, à calculer, par validation croisée (leave-one-out), un intervalle de confiance des coefficients de régression des variables explicatives. Des tests, basés sur la statistique de Student, sont alors effectués pour éliminer les variables les moins significatives à un seuil fixé.

Lindgren et al. (1994 et 1995) proposent une méthode appelée IVS-PLS (pour Interactive Variable Selection for PLS). Leur algorithme est basé sur les valeurs des w_j où $\mathbf{t} = \mathbf{X}\boldsymbol{w}/\boldsymbol{w}^t\boldsymbol{w}$ est la composante PLS. Pour un seuil α, entre 0 et 1, les variables $\mathbf{x}^j$ dont les poids vérifient $|w_j| < \alpha$ sont éliminées du modèle (leurs poids w_j est remis à 0). Les poids w_j des variables restantes sont réajustés de sorte que la norme de $\boldsymbol{w}$ reste égale à 1. Ces étapes sont répétées pour des valeurs de α incrémentées par 0.01 entre 0 et 1. Pour chaque valeur de α, des indicateurs de qualité sont estimés par validation croisée. Le modèle optimal retenu est celui pour lequel un critère donné (exemple la valeur du PRESS par validation croisée) est le meilleur. Lindgren et al. proposent 2 variantes de cette méthode qu'ils appellent Inside-Out (celle présentée ci-dessus) et Outside-In.

Höskuldsson (2001) propose une méthode de sélection d'intervalles de variables plutôt que des variables elles-mêmes. Sa méthode est adaptée à la modélisation de données chimiométriques où les variables sont des longueurs d'onde de signaux (raison pour laquelle il parle d'intervalle de variable). Dans son cas, le nombre de variables peut dépasser le millier, voire même atteindre, d'après l'auteur, 8000 variables (longueurs d'onde de signaux mesurés en proche infrarouge).

Forina et al. (1999) proposent une méthode itérative où, à chaque étape, les variables sont pondérées par les coefficients de régression obtenus lors de l'étape précédente.

D'autres méthodes de sélection de variables ont été développées sur la base des algorithmes génétiques (voir Abrahamsson et al., 2003, Broadhurst et a1., 1997 et Kubinyi et al., 1996). Calqués sur le principe de mutation génétique, ces algorithmes consistent à créer, suivant un processus itératif mi-stochastique mi-déterministe, un grand nombre de modèles parmi les meilleurs possibles selon des critères prédéfinis. Ce processus, dans lequel le savoir-faire de l'utilisateur est souvent indispensable pour le succès de la méthode, devrait "raisonnablement" converger vers des modèles optimaux (Kubinyi et al., 1996). Abrahamsson et al. (2003) comparent 3 méthodes de sélection itératives avec une méthode basée sur les algorithmes génétiques. Bien que cette dernière permette d'améliorer significativement les résul-

tats, c'est l'IVS-PLS (dont ils proposent une version améliorée) qui l'emporte dans leur série de tests.

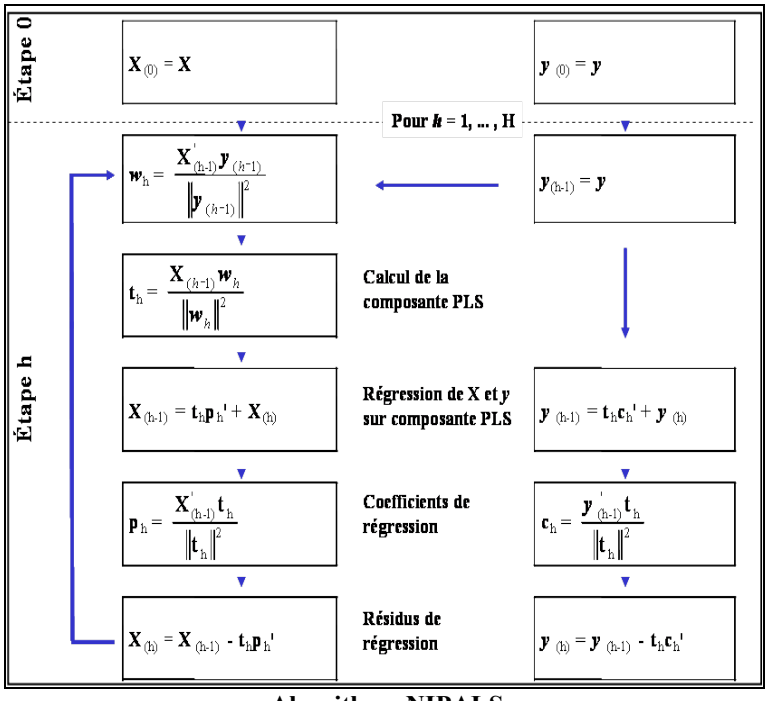
Nous allons présenter une version modifiée de l'algorithme PLS-boostrap dans laquelle nous incluons des calculs d'indices statistiques pour évaluer le pouvoir de prédiction des modèles construits.

3 Algorithme PLS-bootstrap

Soit $\mathbf{X} = (\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^J)$ la matrice NxJ des J variables explicatives ($\mathbf{x}^j$ est le vecteur de dimension N représentant la $j^{ème}$ colonne de $\mathbf{X}$) et $\mathbf{Y}$ la matrice de la (ou des) variable(s) à expliquer. On supposera dans ce qui suit qu'on a une seule variable $\mathbf{y}$ à expliquer.

On notera $\mathbf{Z} = \{(\mathbf{x}_i, y_i),\ i = 1, \ldots, N\}$ l'ensemble des individus où le vecteur $\mathbf{x}_i = (x_i^1, x_i^2, \ldots, x_i^J)^T$ représente la $i^{ème}$ ligne de $\mathbf{X}$ et le scalaire y_i le $i^{ème}$ élément de $\mathbf{y}$.

La régression PLS consiste à calculer, par itérations successives, des combinaisons linéaires $\mathbf{t}=\mathbf{X}\boldsymbol{w}$ des variables explicatives de façon à optimiser la prédiction de la variable à expliquer $\mathbf{y}$ par la maximisation du carré de la covariance $\mathrm{cov}^2(\mathbf{y},\mathbf{t})$. L'algorithme NIPALS (cf. ci-dessous) est l'un des plus connus pour la mise en œuvre de la PLS (voir, par exemple, Tenenahaus, 1998) : à l'itération h de l'algorithme, les coefficients de $\boldsymbol{w}_h$ sont proportionnels aux covariances $\mathrm{cov}(\mathbf{x}^j_{(h-1)}, \mathbf{y}_{(h-1)})$.



Algorithme NIPALS

Les variables $\mathbf{x}^j$ et $\mathbf{y}$ sont d'abord projetées sur la première composante PLS et l'algorithme est appliqué, lors des itérations suivantes, aux résidus des projections successives par une succession de régressions simples, aussi bien des variables $\mathbf{x}^j$ que de $\mathbf{y}$, sur les composantes PLS. Le nombre de celles-ci est en général déterminé par validation croisée. On obtient un modèle linéaire $\mathbf{y} = \mathbf{Xb}$, avec $\mathbf{b} = (b_1, b_2, \ldots, b_J)^T$; les variables $\mathbf{x}^j$ et $\mathbf{y}$ étant centrées-réduites.

Le bootstrap (Efron et Tibshirani, 1993) est une technique de ré-échantillonnage basée sur des tirages aléatoires avec remise dans les données. Son but est de substituer à une distribution inconnue F, dont sont issues les données, la distribution empirique $\hat{F}$ calculée à partir d'échantillons aléatoires. Ces échantillons aléatoires sont déterminés de la manière suivante. Soit $z = \{z_1, z_2, \ldots, z_N\}$ l'échantillon de taille N, représentant les données dont on dispose, issu d'une population de distribution inconnue F. A partir de cet échantillon, on construit un échantillon $z^* = \{z_1^*, z_2^*, \ldots, z_N^*\}$ de même taille N, qu'on appellera *échantillon bootstrap*, par N tirages aléatoires avec remise parmi les N observations de l'échantillon de départ. Dans l'échantillon bootstrap z^*, une observation z_i de z peut apparaître une ou plusieurs fois ou ne pas apparaître du tout. L'astérisque indique que z^* n'est pas identique à z mais en est une duplication aléatoire.

La PLS-bootstrap est une méthode de sélection de variables pour la régression PLS qui a été développée par Lazraq et al. (2003). Les auteurs proposent l'algorithme suivant :

Étape 1 : Répéter pour $\ell = 1, 2, \ldots, L$

1.1. Construire l'échantillon aléatoire $\mathbf{Z}^{*\ell}$ de taille N tiré avec remise dans $\mathbf{Z}$.

1.2. Construire le modèle $\hat{\mathbf{y}}^{*\ell} = \mathbf{X}^{*\ell}\,\mathbf{b}^{*\ell}$ de régression PLS à partir de $\mathbf{Z}^{*\ell}$ où $\mathbf{b}^{*\ell} = (b_1^{*\ell}, b_2^{*\ell}, \ldots, b_J^{*\ell})^T$ est le vecteur colonne des coefficients et $\mathbf{X}^{*\ell}$ la matrice des données associée aux individus tirés.

Étape 2 : Répéter pour j = 1, …, J

2.1. Calculer $\bar{b}_j^* = \frac{1}{L}\sum_{\ell} b_j^{*\ell}$

2.2. Calculer $s_j^{*2} = \frac{1}{L-1}\sum_{\ell}(b_j^{*\ell} - \bar{b}_j^*)^2$

2.3. Déterminer un intervalle de confiance I_j^* pour b_j

2.4. Si $0 \in I_j^*$, éliminer la variable $\mathbf{x}^j$

Étape 3 : Pour j = 1, …, J : Conserver les variables qui n'ont pas été éliminées.

L'intervalle de confiance de l'étape 2.3 est défini par $\bar{b}_j^* \pm c.s_j^*$ où c est un réel fixé par l'utilisateur selon le niveau de confiance souhaité. Dans les exemples présentés par les auteurs, ceux-ci choissent de faire varier c entre 1.4 et 5.2.

Nous proposons la version modifiée suivante (le nombre L de bootsrap et un seuil α - dépendant du niveau de confiance souhaité - sont préalablement fixés par l'utilisateur) :

Étape 1 : Répéter pour ℓ = 1, 2, …, L

1.3. Construire l'échantillon aléatoire $\mathbf{Z}^{*\ell}$ de taille N tiré avec remise dans $\mathbf{Z}$: $\mathbf{Z}^{*\ell} = \{(\mathbf{x}_i, y_i),\ i \in C^{*\ell}\}$ est l'échantillon bootstrap ℓ qui servira en tant qu'ensemble d'apprentissage (i.e. pour la construction du modèle $\hat{\mathbf{y}}^{*\ell}$). $C^{*\ell}$ est l'ensemble des indices des individus ayant été tirés (certains peuvent être dupliqués plusieurs fois ; on a $\left|C^{*\ell}\right| = N$).

1.4. Construire l'échantillon des individus non tirés : $\overline{\mathbf{Z}}^{*\ell} = \{(\mathbf{x}_i, y_i),\ i \in \overline{C}^{*\ell}\}$ désigné par le terme anglais *out of bag* (oob). $\overline{C}^{*\ell}$ est l'ensemble des indices des individus n'ayant pas été tirés dans l'échantillon bootstrap ℓ. $\overline{\mathbf{Z}}^{*\ell}$ servira comme ensemble de test pour le modèle $\hat{\mathbf{y}}^{*\ell}$.

1.5. Construire le modèle $\hat{\mathbf{y}}^{*\ell} = \mathbf{X}^{*\ell}\,\mathbf{b}^{*\ell}$ de régression PLS dont $\mathbf{b}^{*\ell} = (b_1^{*\ell}, b_2^{*\ell}, \ldots, b_J^{*\ell})^T$ est le vecteur colonne des coefficients et $\mathbf{X}^{*\ell}$ la matrice des données qui servent à la construction, dont les lignes sont les individus $\mathbf{x}_i$ où $i \in C^{*\ell}$.

1.6. Calculer les prédictions du modèle pour les individus i non tirés (échantillon de test) : $\hat{y}_i^{*\ell} = (\mathbf{b}^{*\ell})^T \mathbf{x}_i^{*\ell}$, $\mathbf{x}_i^{*\ell} = (x_i^{\mathbf{1}*\ell}, x_i^{\mathbf{2}*\ell}, \ldots, x_i^{\mathbf{J}*\ell})^T$ où $i \in \overline{C}^{*\ell}$.

1.7. Calculer, selon la formule **(1)** donnée au paragraphe suivant, l'erreur quadratique moyenne de prédiction $EQMT^{*\ell}$ à partir des individus de l'échantillon de test.

1.8. Calculer, selon la formule **(2)** donnée au paragraphe suivant, le coefficient de validation $Q^{*\ell\,2}$ à partir des individus de l'échantillon de test.

Étape 2 : Pour i = 1, …, N

2.5. Définir les ensembles $\Lambda^i = \{\ell,\ i \in C^{*\ell}\}$ des indices ℓ des échantillons bootstrap contenant i et $\Lambda^{-i} = \{\ell,\ i \in \overline{C}^{*\ell}\}$ des indices ℓ des échantillons bootstrap ne contenant pas i dont les effectifs respectifs sont notés $\left|\Lambda^i\right|$ et $\left|\Lambda^{-i}\right|$.

2.6. Calculer, selon la formule **(3)** donnée au paragraphe suivant, l'erreur de prédiction $e_{(-i)}^{*\ell}$ pour chaque bootstrap $\ell \in \Lambda^{-i}$

2.7. Calculer, selon la formule **(4)** donnée au paragraphe suivant, le biais $B_{(-i)}^{*}$ de prédiction à partir des $\left|\Lambda^{-i}\right|$ modèles booststrap.

2.8. Calculer, selon la formule **(5)** donnée au paragraphe suivant, la variance de prédiction $\sigma_{(-i)}^{*\,2}$ à partir des $\left|\Lambda^{-i}\right|$ modèles booststrap.

Les notations "(-i)" des indices désignent le fait que $e_{(-i)}^{*\ell}$, $\sigma_{(-i)}^{*\,2}$, et $B_{(-i)}^{*}$ sont calculés par des modèles que les individus i n'ont pas servi à construire.

Étape 3 : Répéter pour j = 1, 2, …, J

3.1. Calculer l'intervalle de confiance $I^*_j(\alpha)$, au seuil α, pour le coefficient b_j de la variable x^j à partir de l'échantillon bootstrap $\{ b_j^{*\ell}, \ \ell = 1, L\}$

3.2. Éliminer les variables $\mathbf{x}^j$ pour lesquelles $0 \in \mathbf{I}^*_j(\alpha)$.

Répéter les étapes 1 à 3 avec les variables X^j retenues, jusqu'à ce qu'aucune variable ne soit éliminée (convergence de l'algorithme).

Les bornes inférieure et supérieure $b_j^{*(inf)}(\alpha)$ et $b_j^{*(sup)}(\alpha)$ de l'intervalle $\mathbf{I}^*_j(\alpha)$ du coefficient b_j à 100.(1-α) % sont définies par :

$b_j^{*(inf)}(\alpha) = 100.\frac{\alpha}{2}$ ème percentile

et

$b_j^{*(sup)}(\alpha) = 100.(1-\frac{\alpha}{2})$ème percentile

obtenues à partir de la distribution $\{ b_j^{*\ell}, \ \ell = 1, L\}$ (Efron et Tibshirani, 1993).

On retient le modèle associé au couple (L,α) qui réalise le maximum de la médiane de la distribution $Q^{*\ell 2}$ et le minimum de sa variance.

En général, selon la nature des données analysées, les résultats de la sélection peuvent beaucoup varier en fonction des valeurs de L et de α. Ces deux paramètres peuvent être optimisés selon une procédure empirique qui permet de présélectionner un couple (L_0,α_0) optimal – celui pour lequel les valeurs des $Q^{*\ell 2}$ se stabilisent autour d'une valeur maximale – avant la mise en oeuvre de l'algorithme (Nocairi et Faraj, 2005). Une méthode de recherche du nombre de bootstrap optimal, hors sélection, est proposée dans (Aji et al., 2003).

4 Qualité d'un modèle

L'algorithme présenté ci-dessus converge au bout de quelques itérations (dont le nombre dépasse rarement 5) en fonction des données analysées. Or, selon la nature de ces données, les modèles sont susceptibles de dégradation au fur et à mesure des itérations. Il est alors nécessaire d'examiner individuellement chacune de ces itérations pour sélectionner celle qui donne le meilleur modèle. Pour ce faire nous nous basons sur les indicateurs calculés lors des étapes 1 et 2 de l'algorithme. Ces indicateurs peuvent être rangés en 2 catégories. D'abord ceux portant sur la qualité d'un modèle, erreur quadratique moyenne $EQMT^{*\ell}$ et coefficient de validation $Q^{*\ell 2}$, respectivement définis par :

$$EQMT^{*\ell} = \frac{1}{\left|\overline{C}^{*\ell}\right|} \sum_{i \in \overline{C}^{*\ell}} (\hat{y}^{*\ell}_{(-i)} - y_i)^2 \qquad \textbf{(1)}$$

$$Q^{*\ell 2} = \mathrm{cor}^2(\hat{\mathbf{y}}^{*\ell}, \mathbf{y}) \qquad \textbf{(2)}$$

avec $\hat{\mathbf{y}}^{*\ell} = (\hat{y}_i^{*\ell})^T$ et $\mathbf{y} = (\mathbf{y}_i)^T$ où i∈ $\overline{C}^{*\ell}$ ($\hat{y}_i^{*\ell}$ est l'estimation de **y** au point i par le modèle construit avec les points excluant i dans le bootstrap ℓ).

L'ensemble $\{EQMT^{*\ell}, \ell = 1, L\}$ représente la distribution empirique de l'erreur quadratique moyenne de test. Calculée pour les individus n'ayant pas servi dans la construction du modèle, $EQMT^{*\ell}$ est un indicateur de l'erreur de généralisation du modèle. Ses valeurs sont d'autant plus proches de 0 que le modèle a un bon pouvoir de généralisation.

L'ensemble $\{Q^{*\ell 2}, \ell = 1, L\}$ représente la distribution empirique du coefficient de validation. Les valeurs de $Q^{*\ell 2}$ sont d'autant plus élevées (proches de 1) que le modèle a un bon pouvoir de prédiction.

La deuxième catégorie d'indicateurs est constituée de ceux portant sur la qualité de prédiction pour un individu, erreur de prédiction $e_{(-i)}^{*\ell}$, biais de prédiction $B_{(-i)}^{*}$ et variance de prédiction $\sigma_{(-i)}^{*\,2}$, respectivement définis par :

$$e_{(-i)}^{*\ell} = \hat{y}_{(-i)}^{*\ell} - y_i \text{ pour } \ell \in \Lambda^{-i} \qquad \textbf{(3)}$$

$$B_{(-i)}^{*} = \bar{\hat{y}}_{(-i)}^{*} - y_i \qquad \textbf{(4)}$$

$$\sigma_{(-i)}^{*\,2} = \frac{1}{\left|\Lambda^{-i}\right|} \sum_{\ell \in \Lambda^{-i}} (\hat{y}_{(-i)}^{*\ell} - \bar{\hat{y}}_{(-i)}^{*})^2 \qquad \textbf{(5)}$$

$\bar{\hat{y}}_{(-i)}^{*} = \frac{1}{\left|\Lambda^{-i}\right|} \sum_{\ell \in \Lambda^{-i}} \hat{y}_i^{*\ell}$ est la moyenne des prédictions des $\left|\Lambda^{-i}\right|$ modèles bootstrap au point i.

L'ensemble $\{e_{(-i)}^{*\ell}, \ell = 1, L\}$ représente la distribution empirique de l'erreur de prédiction du modèle en chaque point i.

$\sigma_{(-i)}^{*\,2}$ est une estimation de la variance de prédiction du modèle de régression PLS pour l'individu i. C'est un indicateur de la dispersion de la distribution empirique $\{\hat{y}_{(-i)}^{*\ell}, \ell = 1, L\}$ de la prédiction du modèle au point i. Il n'est pas nécessaire de connaître la valeur de **y** en i pour le calcul de $\sigma_{(-i)}^{*\,2}$. De ce fait on peut estimer la variance de prédiction en tout point du domaine défini par les J variables explicatives.

$B_{(-i)}^{*}$ mesure le biais du modèle - écart entre la valeur observée y_i et la prédiction moyenne $\bar{\hat{y}}_{(-i)}^{*}$ du modèle - au point i.

Comme nous l'avons précisé ci-dessus, $EQMT^{*\ell}$, $Q^{*\ell 2}$, $e_{(-i)}^{*\ell}$, $B_{(-i)}^{*}$ et $\sigma_{(-i)}^{*\,2}$ ne servent pas dans le processus de sélection des variables. Leur intérêt est de rendre compte, *a*

posteriori, de la qualité des modèles construits au fur et à mesure des itérations de l'algorithme. Ils permettent, de cette façon, d'identifier la (ou les) itération(s) correspondant aux meilleurs ensembles de variables sélectionnées (i.e. celles associées aux modèles dont les qualités de prédiction sont les meilleures). Ils renseignent, à terme, sur la nature (linéaire ou non linéaire) des relations qui existent entre les variables explicatives et la réponse.

Remarque : Les formules **(1)** à **(5)** ci-dessus peuvent s'appliquer aux individus de l'ensemble d'apprentissage pour calculer les critères que nous noterons respectivement $EQMA^{*\ell}$, $R^{*\ell\,2}$, $e_{\mathrm{i}}^{*\ell}$, B_{i}^{*} et $\sigma_{\mathrm{i}}^{*\,2}$ (qui désignent respectivement l'erreur quadratique moyenne de prédiction, le coefficient de détermination, le biais et la variance du modèle sur les individus d'apprentissage). Ces critères ont généralement tendance à surestimer (par sur-ajustement) les quantités qu'elles estiment, alors que $EQMT^{*\ell}$, $Q^{*\ell\,2}$, $e_{(-\mathrm{i})}^{*\ell}$, $B_{(-\mathrm{i})}^{*}$ et $\sigma_{(-\mathrm{i})}^{*\,2}$ auraient plutôt tendance à les sous-estimer (cf. figures 3 et 4 du paragraphe suivant). Pour pallier à cela, Efron propose une moyenne des 2 critères pondérée par les poids 0.632 et 0.368 de sorte que :

$$EQM_{0.632}^{*\ell} = 0.368.\,EQMA^{*\ell} + 0.632.\,EQMT^{*\ell}$$

$$Q_{0.632}^{*\ell\,2} = 0.368.\,R^{*\ell\,2} + 0.632.\,Q^{*\ell\,2}$$

$$B_{\mathrm{i}\,0.632}^{*} = 0.368.\,B_{\mathrm{i}}^{*} + 0.632.\,B_{(-\mathrm{i})}^{*}$$

$$\sigma_{\mathrm{i}\,0.632}^{*\,2} = 0.368.\,\sigma_{\mathrm{i}}^{*\,2} + 0.632.\,\sigma_{(-\mathrm{i})}^{*\,2}$$

0.632 est la probabilité qu'un individu soit tiré par rééchantillonnage bootsrap.

5 Application de la PLS-bootstrap

L'exemple que nous allons traiter est tiré du livre de Cornel (1990) et a été également analysé par Tenenahaus et al. (1995 et 1998). Il concerne la modélisation de l'indice d'octane moteur (variable réponse *y*) à partir de sept composants du carburant : 12 mélanges ont été réalisés suivant un plan d'expériences D-optimal (tableau 1). Les données représentent des proportions ; la somme des variables *x*1 à *x*7 étant égale à 1.

N°	*x1*	*x2*	*x3*	*x4*	*x5*	*x6*	*x7*	*y*
1	0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.7
2	0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.8
3	0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.6
4	0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.0
5	0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.6
6	0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.2
7	0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.9
8	0.17	0.19	0.01	0.38	0.02	0.06	0.08	83.1
9	0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.4
10	0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.2
11	0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.4
12	0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.1

x1 *:* Distillation directe

x2 *:* Réformat

x3 *:* Naphta de craquage thermique

x4 *:* Naphta de craquage catalytique

x5 *:* Polymère

x6 *:* Alkylat

x7 *:* Essence naturelle

y : Indice d'octane Moteur

TAB. 1 – Données de cornell (1990).

Ces données sont intéressantes à plusieurs titres. Tenenhaus et al. les ont pris comme exemple pour montrer les limites de la régression des moindres carrés ordinaires (MCO) et la nécessité, dans certains cas, de lui substituer la régression PLS. Le fait que ces variables soient liées entre elles rend impossible leur prise en compte toutes à la fois dans un modèle de régression MCO. Les auteurs ont choisi les 6 premières variables pour modéliser la réponse *y*. Dans le modèle de régression MCO qu'ils obtiennent, aucun des coefficients n'est significatif. Par ailleurs les signes de ces coefficients ne sont pas en accord avec le sens des corrélations des variables *x*j et de la réponse *y* (cf. la matrice des corrélations du tableau 2).

	x2	*x3*	*x4*	*x5*	*x6*	*x7*	*y*
x1	0.10	0.99	0.37	-0.55	-0.80	0.60	-0.84
x2		0.10	-0.54	-0.29	-0.19	-0.59	-0.07
x3			0.37	-0.55	-0.80	0.61	-0.84
x4				-0.21	-0.64	0.92	-0.71
x5					0.46	-0.27	0.49
x6						-0.66	0.98
x7							-0.74

TAB. 2 – Matrice des corrélations.

Une méthode de sélection pas à pas descendante pour régression MCO, utilisée par les auteurs, leur a permis de retenir les variables $x1$, $x2$, $x4$ et $x5$ comme significatives, sans résoudre pour autant le problème d'accord des signes des coefficients avec les corrélations. Le modèle MCO construit ne donne pas par ailleurs une entière satisfaction aux chimistes : la variable $x6$ est éliminée alors qu'elle est la plus corrélée avec la réponse y.

Les auteurs ont ensuite appliqué la régression PLS aux données pour illustrer la mise en œuvre de celle-ci. Ils obtiennent un modèle cohérent avec 7 variables explicatives (les signes des coefficients ne contredisent pas le sens des corrélations des variables x avec la réponse y). Ils n'ont pas abordé, dans leur article, le problème de sélection de variables en régression PLS.

Dans ce qui suit nous appliquons l'algorithme PLS-bootstrap aux mêmes données. Celui-ci converge au bout de 3 itérations, qui ont permis de retenir un modèle avec 3 variables explicatives :

- A la 1ère itération, toutes les variables (de $x1$ à $x7$) sont utilisées. Les résultats obtenus par ce 1er modèle sont donnés à la figure 1. Les variables $x1$, $x3$, $x4$ et $x6$ sont retenues ;
- A la 2ème itération, on utilise les variables $x1$, $x3$, $x4$ et $x6$. Les variables $x1$, $x4$ et $x6$ sont retenues ;
- A la 3ème itération, on utilise les variables $x1$, $x4$ et $x6$. Aucune variable n'est éliminée (figure 2).

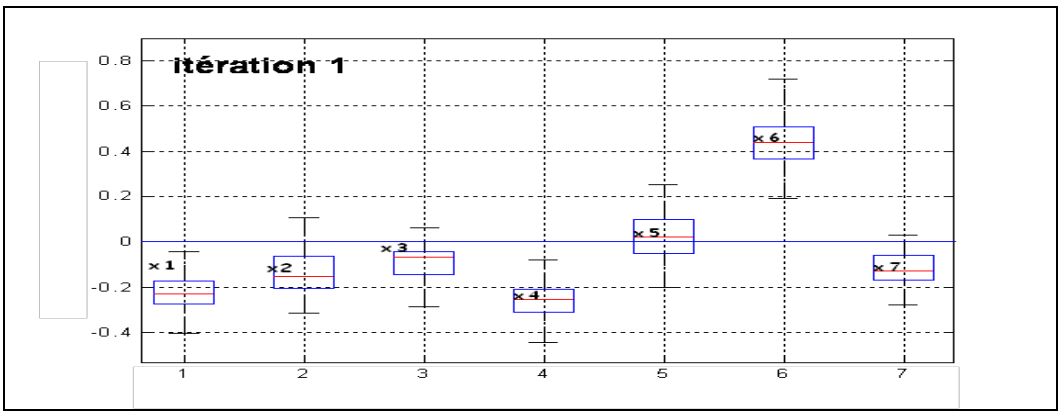


FIG. 1 – Distributions empiriques des coefficients du modèle de régression calculées lors de la première itération de l'algorithme (7 variables explicatives x1 à x7). Les variables x2, x5 et x7 sont éliminées à la fin de cette itération.

Les variables $x1$, $x4$ et $x6$ sont retenues par l'algorithme. Le modèle obtenu avec ces variables est cohérent aussi bien avec le choix des chimistes (la variable $x6$ est sélectionnée) qu'avec les corrélations des xj avec la réponse y. Les signes des coefficients du modèle sont en accord avec ces corrélations : négatifs pour $x1$ et $x4$ et positif pour $x6$ (figure 2). Les prédictions sont améliorées : l'erreur quadratique moyenne est diminuée (figures 3 et 4) de même que les variances de prédiction pour une majorité des points (figure 5).

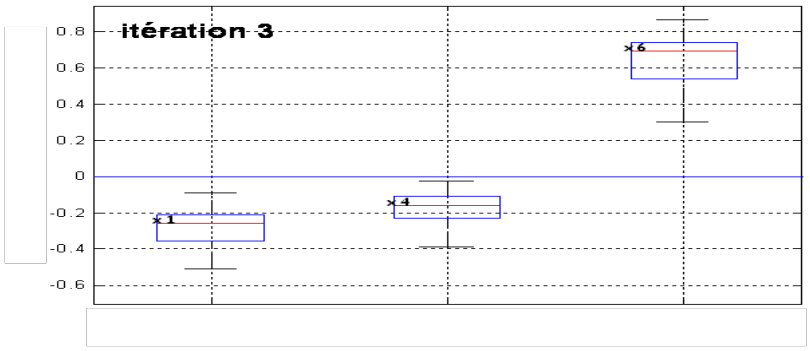


FIG. 2 – Distributions empiriques des coefficients du modèle de régression calculées lors de la dernière itération de l'algorithme. Les variables x1, x4 et x6 sont sélectionnées.

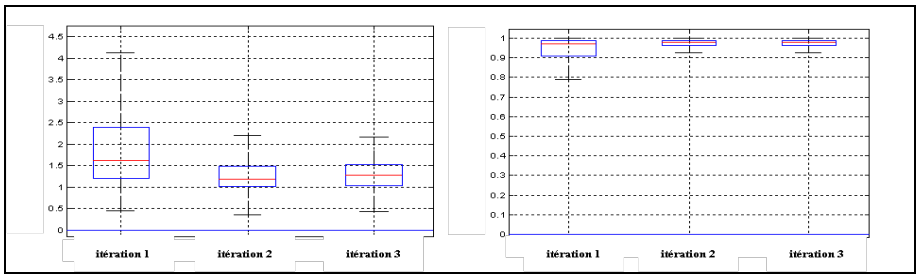


FIG. 3 – Distributions de l'erreur quadratique moyenne de prédiction EQMT (graphique de gauche) et du coefficient de validation Q2 (graphique de droite) calculés sur les individus de test pour les trois itérations.

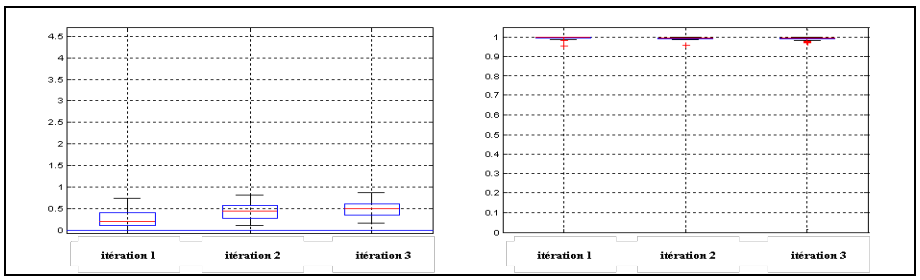


FIG. 4 – Distributions de l'erreur quadratique moyenne de prédiction EQMA (graphique de gauche) et du coefficient de validation R2 (graphique de droite) calculés sur les individus d'apprentissage pour les trois itérations.

On voit sur cet exemple (cf. remarque faite au paragraphe précédent) que l'erreur quadratique moyenne et le coefficient R2 calculés sur les données d'apprentissage ont tendance à surestimer (par sur-ajustement) les quantités qu'elles estiment, alors que l'erreur quadratique

moyenne et le coefficient Q2 calculés sur les données de test auraient plutôt tendance à les sous-estimer.

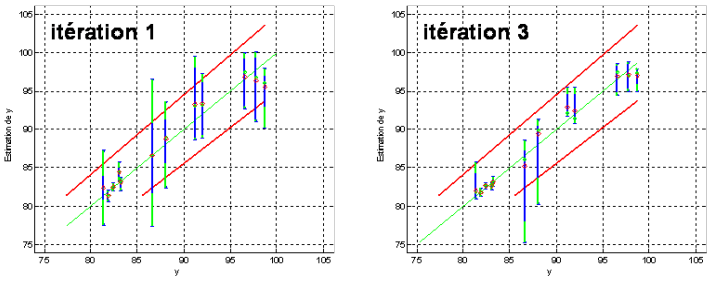


FIG. 5 – Distributions de l'erreur de prédiction de test pour les 12 individus lors des itérations 1 et 3 de l'algorithme. Les traits pleins verticaux (en bleu) représentent l'intervalle de confiance à 95%.

Sur la figure 6 ci-dessous sont représentées les projections des individus sur le plan factoriel ($t_1^{*\ell}, t_2^{*\ell}$) défini par les 2 premières composantes PLS pour 100 bootstrap pour la première itération (graphique de gauche) et la troisième itération (graphique de droite) de l'algorithme de sélection. On remarque que les individus sont moins bien séparés pour les 7 variables initiales (graphique de gauche) et mieux séparés pour les 3 variables sélectionnées (graphique de droite). Par ailleurs la valeur médiane du pourcentage de la variance restituée par les 2 premières composantes PLS croît de 85.6 % avec les 7 variables initiales à 93.6 % pour les 3 variables sélectionnées.

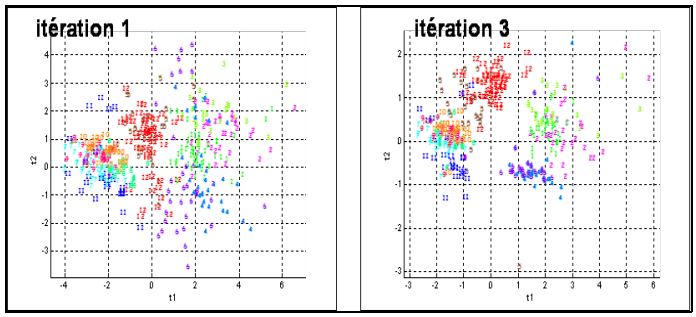


FIG. 6 – *Plans factoriels (t1,t2) des composantes PLS pour les 7 variables initiales à l'itération 1 (graphique de gauche) et pour les 3 variables sélectionnées par l'algorithme à l'itération 3 (graphique de droite). Les résultats sont obtenus pour 100 bootstraps.*

6 Conclusion

Nous avons présenté une version modifiée de la PLS-bootstrap pour la sélection de variables en régression PLS. Notre objectif était de définir une méthodologie itérative qui, tout

en répondant à cette problématique, se situe dans le contexte de l'analyse exploratoire des données :

- un modèle de régression PLS est construit à chaque étape (ré-échantillonnage par bootstrap) ;
- ce modèle est amélioré par élimination des variables non significatives (tests statistiques) ;
- des critères informant sur le pouvoir prédictif des modèles sont calculés ;
- les modèles sont utilisés comme outils de description (typologie) des individus.

Il ne faut pas perdre de vue que, lorsque le modèle inconnu n'est pas linéaire en paramètres, aucune méthode de sélection - si performante soit-elle - ne peut être efficace si elle consiste à construire un modèle linéaire (ce qui est le cas de la PLS). Une démarche exploratoire est dans ce cas nécessaire pour avoir quelques idées sur la nature du modèle. S'il est avéré que la régression linéaire n'est pas apte à modéliser les données, il faut chercher un modèle non linéaire (réseau de neurones par exemple) ou effectuer les transformations non linéaires adéquates sur les variables avant construction du modèle.

Références

Abrahamsson, C., J. Johansson, A. Sparén, D. Folkenberg, et F. lindgren (2003). Comparison of different selection methods conducted on NIR transmission measurements on intact tablets. *Chemometrics and Intelligent Laboratory Systems*, 69, 3-12.

Aji, S., S. Tavolaro, F. Lantz, et A. Faraj (2003). Apport du bootstrap à la régression PLS : application à la prédiction de la qualité des gazoles. *Oil and Gas Science and Technology - Revue de l'IFP*, Vol. 58, No 5, 599-608.

Aji, S., S. N. Schildknecht-Szydlowski, et A. Faraj (2004). Partial Least Square Modelling for the Control of Refining Processes on Mid-Distillates by Near Infrared Spectroscopy. *Oil & Gas Science and Technology – Rev. IFP*, Vol. 59, No. 3, 303-321.

Batten, G. D., S. Ciavarella, et A. B. Blakeney (2000). Modified jack-knifing in multivariate regression for variable selection in Near Infrared Spectroscopy. *Proceedings of the 9th International Conference.*

Breiman, L. et P. Spector (1992). Submodel selection and evaluation in regression: The X-random case. *International Statistical Review*, 60, 291-319.

Broadhurst, D., R. Goodacre, A. Jones, J. Rowland, et D. Kell (1997). Genetic algorithms as a method for variable selection in multiple linear regression and partial least squares regression, with application to pyrolysis mass spectrometry. *Analytica Chimica Acta*, 348, 71-86.

Cornell, J. A. (1990). *Experiments with mixture*, Wiley.

Efron, B. et R. Tibshirani (1993). *An introduction to the Bootstrap*, Chapman and Hall, London.

Faraj, A. et M. Constant (2004). Utilisation du bootstrap pour la sélection de variables et la typologie des individus en régression PLS. *39èmes Journées de Statistique de la SFdS, Montpellier, France.*

Forina, M., C. Casolino, et C. Pizzaro Millan (1999). Iterative predictor weighting (IPW) PLS: a technique for the elimination of useless predictors in regression problems. *Journal of Chemometrics*, 13, 165-184.

Gauchi, J.-P. et P. Chagnon (2001). Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58, 171-193.

Höskuldsson A. (2001). Variable and subset selection in PLS regression. *Chemometrics and Intelligent Laboratory Systems*, 55, 23-38.

Kubinyi, H., (1996). Evolutionary variable selection in regression and PLS analyses. *Journal of Chemometrics*, Vol. 10, 119-133.

Lazraq, A., R. Cléroux, et J.-P. Gauchi (2003). Selecting both latent and explanatory variables in the PLS1 regression model. *Chemometrics and Intelligent Laboratory Systems*, 66, 117-126.

Lindgren, F., P. Geladi, S. Rännar, et S. Wold (1994). Interactive variable selection (IVS) for PLS. Part I: Theory and algorithms. *Journal of Chemometrics*, Vol. 8, 349-363.

Lindgren, F., P. Geladi, A. Berglund, M. Sjöström, et S. Wold (1995). Interactive variable selection (IVS) for PLS. Part II: Chemical applications. *Journal of Chemometrics*, Vol. 9, 331-342.

Martens, H., M. Høy, F. Westad, D. Folkenberg, et M. Martens (2001). Analysis of designed experiments by stabilised PLS Regression and jack-knifing. *Chemometrics and Intelligent Laboratory Systems*, 58, 151-170.

Nocairi, H., et A. Faraj (2005). Optimisation de la sélection des variables pertinentes pour modèle de régression PLS par bootstrap. *Chimiométrie, 30 nov – 1 déc. 2005, Lille.*

Sarabia, L. A., M. C. Ortiz, M. S. Sánchez, et A. Herrero (2001). Dimension wise selection in partial least squares regression with a bootstrap estimated signal-noise relation to weight the loadings. *Proceedings of the PLS'01 International Symposium, CISIA-CERESTA Editeur, Paris*, 2001, pp. 327-339.

Tenenhaus, M., J.-P. Gauchy, et C. Ménardo (1995). Régression PLS et applications. *Rev. Stat. Appliquées*, XLIII (1), 7-63.

Tenenhaus, M., (1998). *La régression PLS - Théorie et pratique*, Ed. Technip, Paris.

Westad, F., et H. Martens (1999). Variable Selection in NIR based on significance testing in Partial Least Squares Regression. *Journal of Near Infrared Spectroscopy*, 8, 117-124.

Summary

The selection of model in PLS regression is crucial even if the number of variables, which can be superior to the one of individuals, is not imperative for the use of the method. A such selection consists in keeping, among models having good prediction capabilities, those with low number of input variables. The method that we present in this paper is based on the bootstrap. It allows to calculate the empirical distribution of the coefficients of the

model and to keep those which are most significant. It measures the prediction capacity of the regression model constructed as well for each individual as globally. This approach is applied to the analysis of a data set.