# plsRglm, modèles linéaires généralisés PLS sous R

Frédéric Bertrand, Myriam Maumy, Nicolas Meyer

Chimiometrie 2009

Abstract/Résumé

# plsRglm, modèles linéaires généralisés PLS sous R

F. Bertrand[1] M. Maumy-Bertrand[2] N. Meyer[3]

[1] Institut de Recherche Mathématique Avancée, Université de Strasbourg, fbertran@math.u-strasbg.fr

[2] Institut de Recherche Mathématique Avancée, Université de Strasbourg, mmaumy@math.u-strasbg.fr

[3] Laboratoire de Biostatistique, Faculté de Médecine, Université de Strasbourg, nmeyer@unistra.fr

**Keywords:** Partial least squares, generalized linear models, high dimensional data, R software package.

## 1 Introduction

La finalité de la bibliothèque de fonctions plsRglm écrite par les auteurs et implémentée dans le logiciel R [1] est multiple et s'organise principalement autour de deux thématiques : l'extension de la régression PLS au cas des modèles linéaires généralisés, en particulier celui des régressions logistiques [2], et le traitement des jeux de données incomplets par validation croisée. Ces modèles ont été appliqués avec succès à des données de nature variée : par Bastien *et al.* [2] à des problèmes de régression multiple en liaison avec les données de Cornell [3], à des problèmes de régression linéaire généralisée et en particulier à une étude de la qualité de vins de Bordeaux à l'aide d'un modèle de régression logistique ordinale. Plus récemment, les auteurs se sont servis de modèles de régression logistique binaire PLS pour étudier des données d'allélotypage [4] qui interviennent dans la compréhension de mécanisme liés à l'évolution des cancers.

Nous commençons par rappeler comment étendre la régression PLS au cas des modèles linéaires généralisés, puis nous proposons trois exemples d'application de la régression PLS étendue aux modèles de régression logistique obtenus à l'aide de la bibliothèque de fonctions plsRglm disponible pour le logiciel R.

## 2 Théorie : Régression PLS étendue aux modèles de régression logistique

### 2.1 La regression PLS

Considérons les variables centrées y, $x_1$, …, $x_j$, …, $x_p$. Soit X la matrice des prédicteurs $x_1$, …, $x_j$, …, $x_p$.

La régression PLS est bien connue et décrite de manière exhaustive notamment par Höskuldsson [5] et Wold *et al.* [6]. La présentation classique de la régression PLS est sous forme algorithmique. Nous n'en rappellerons que les éléments utiles pour la suite.

La régression PLS est un modèle non-linéaire qui permet de construire des composantes orthogonales $t_h$ obtenues en maximisant les quantités cov(y,$t_h$). Soit T la matrice formée de ces composantes, nous avons :

$$\mathbf{y} = \mathbf{T}\,\mathbf{c}^{\mathbf{T}} + \boldsymbol{\varepsilon} \tag{1}$$

où ε est le vecteur des résidus et $c^T$ le vecteur des coefficients des composantes, $^T$ désignant la transposée.

En posant $T=XW^*$, où $W^*$ est la matrice des coefficients des variables $x_j$ dans chaque composante $t_h$, nous avons l'expression directe de la réponse y à l'aide des prédicteurs $x_j$ :

$$\mathbf{y} = \mathbf{X}\,\mathbf{W}^*\,\mathbf{c}^{\mathbf{T}} + \boldsymbol{\varepsilon} \tag{2}$$

En développant le membre de droite de (2), nous obtenons pour chaque composante $y_i$ de y :

$$y_i = \sum_{h=1}^{H}\left(c_h w^*_{1h} x_{i1} + \cdots + c_h w^*_{ph} x_{ip}\right) + \varepsilon_i \tag{3}$$

H étant le nombre de composantes retenues dans le modèle final avec $H \leq rg(X)$, H étant en général très inférieur au rang de X et p étant égal au nombre de variables contenues dans la matrice X. Les coefficients $c_h w^*_{jh}$, où $1 \leq j \leq p$, suivant la notation avec $^*$ de Wold *et al.* [6], traduisent la relation entre le vecteur y et les variables $x_j$ à travers les composantes $t_h$.

### 2.2 Extension aux modèles de régression logistique

La régression PLS étendue aux modèles de régression linéaire généralisée de la réponse y sur les variables $x_1, \ldots, x_j, \ldots, x_p$ avec H composantes $t_h = w^*_{1h}x_{i1} + \cdots + w^*_{ph}x_{ip}$ [2] s'écrit :

$$g(\theta)_i = \sum_{h=1}^{H} \left( c_h \sum_{j=1}^{p} w^*_{jh} x_{ij} \right) \tag{4}$$

où le paramètre θ peut être soit une espérance soit le vecteur des probabilités d'une loi discrète de support fini. La fonction de lien g est déterminée en fonction de la distribution de y et de la qualité de l'ajustement du modèle au données. Les composantes PLS $t_h$ sont orthogonales. L'algorithme permettant de déterminer les composantes PLS $t_h$ d'un modèle PLS-GLM est le suivant :

- Calcul de la première composante PLS $t_1$ :

1. Calculer le coefficient $a_{1j}$ de $x_j$ dans la régression linéaire généralisée de y sur $x_j$ pour chaque prédicteur $x_j$, $1 \leq j \leq p$.
2. Normer le vecteur colonne $a_1$ : $w_1 = a_1/||a_1||$, puis calculer la composante $t_1 = 1/(w_1^T w_1) X w_1$.

- Calcul de la seconde composante PLS $t_2$ :

1. Calculer le coefficient $a_{2j}$ de $x_j$ dans la régression linéaire généralisée de y sur $t_1$ et $x_j$ pour chaque prédicteur $x_j$, $1 \leq j \leq p$.
2. Normer le vecteur colonne $a_2$ : $w_2 = a_2/||a_2||$, calculer la matrice résiduelle $X_1$ de la régression linéaire de X sur $t_1$, puis calculer la composante $t_2 = 1/(w_2^T w_2) X_1 w_2$.
3. Exprimer la composante $t_2$ en termes de prédicteurs X : $t_2 = X w^*_2$.

- Calcul de la h-ème composante PLS $t_h$ :

1. Calculer le coefficient $a_{hj}$ de $x_j$ dans la régression linéaire généralisée de y sur $t_1, \ldots, t_{h-1}$ et $x_j$ pour chaque prédicteur $x_j$, $1 \leq j \leq p$.
2. Normer le vecteur colonne $a_h$ : $w_h = a_h/||a_h||$, calculer la matrice résiduelle $X_{h-1}$ de la régression linéaire de X sur $t_1, \ldots, t_{h-1}$, puis calculer la composante $t_h = 1/(w_h^T w_h) X_{h-1} w_h$.
3. Exprimer la composante $t_h$ en termes de prédicteurs X : $t_h = X w^*_h$.

Il est possible de modifier l'algorithme précédent pour pourvoir traiter les jeux de données incomplets [2].

## 3 Méthodes : points forts de l'implémentation

La bibliothèque de fonctions plsRglm possède plusieurs points forts.

- Régression PLS1 et PLS-GLM avec des données complètes ou incomplètes.
- Choix du nombre de composantes grâce à différents critères AIC, BIC, arrêt de significativité de la composante $t_{m+1}$ lorsqu'aucun des coefficients $a_{m+1}$ n'est plus significatif dans le modèle [2] ou en utilisant un critère $Q^2$ [2] ou le nombre de mal classés tous les deux estimés par validation croisée.
- Validation croisée « repeated k-fold cross validation » avec des données complètes ou incomplètes.

## 4 Résultats et discussion : application à trois exemples

Nous proposerons des exemples d'application à trois jeux de données classiques. Un seul est détaillé ici.

### 4.1 Microarray (Cancer du colon)

Alon et al. [7] ont analysé 62 échantillons (40 d'une tumeur, 22 d'une partie saine) prélevés dans le colon de 62 patients atteints du cancer du colon. 2000 parmi les 6500 gènes exprimés ont été sélectionnés par les auteurs [7]. Nous utilisons les fonctions de la bibliothèque plsRglm pour ajuster un modèle PLS-logistique qui permettra de modéliser la probabilité qu'un tissu soit sain ou cancéreux à l'aide de gènes bien choisis.

2 à 4 composantes devraient être retenues suivant les résultats du Tableau 1, les critères AIC et BIC étant connus pour être optimistes. Le $Q^2\chi^2$ a un comportement souvent surprenant en PLS logistique [4].

| Nbre composantes | 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|---|
| AIC | 82,65 | 60,58 | 36,01 | 17,53 | 10,00 | 12,00 |
| BIC | 84,78 | 64,83 | 42,39 | 26,04 | 20,64 | 24,76 |
| Préd. Significatifs | 352 | 1350 | 29 | 0 | 0 | 0 |
| Mal Classé (10-CV) | | 17 | 10 | 10 | 11 | 11 |
| $Q^2\chi^2$ (10-CV) | | -0,50 | -205,16 | $-8{,}57*10^{14}$ | $-3{,}09*10^{15}$ | $-6{,}47*10^{24}$ |
| $\chi^2$ Pearson | 62,00 | 49,73 | 36,77 | 11,66 | 0,00 | 0,00 |

Tableau 1 – Résultats de la validation croisée, k=10

### 4.2 Chimiométrie (phenyl) et chimiotaxonomie (hyptis)

Une régression PLS binaire et multinomiale sera appliquée à 2 jeux de données, phenyl et hyptis, de grande dimension étudiés dans [8] et disponibles dans la bibliothèque chemometrics du logiciel R.

## 5 Conclusion et perspectives

Notre objectif a été de mettre à la disposition des utilisateurs du logiciel libre R l'extension de la régression PLS au cas des modèles linéaires généralisés qui permet ainsi de faire bénéficier les régressions logistiques, aussi bien binaire, qu'ordinale ou multinomiale, des points forts de la régression PLS. En premier lieu, il s'agit de la possibilité de travailler avec des prédicteurs colinéaires, difficulté inévitable dans le cas de la modélisation des mélanges ou lors de l'analyse de spectres, de l'étude de données génétiques, protéomiques ou métabonomiques. En second lieu, la régression PLS, lorsqu'elle est réalisée par exemple avec l'algorithme NIPALS, peut être appliquée à des jeux de données incomplets.

La seconde thématique concerne le traitement des jeux de données incomplets. La bibliothèque de fonctions plsRglm vise à palier à certains manques des bibliothèques de fonctions existantes concernant le traitement des jeux de données incomplets à l'aide de la régression PLS classique. Dans ce cas, par exemple, et contrairement au logiciel SIMCA [9], aucune bibliothèque de fonctions dans le logiciel R ne propose pour le moment la sélection du nombre de composantes par validation croisée. Nous avons donc implémenté des fonctions permettant de choisir le nombre de composantes de régressions PLS classiques ou de régressions PLS-GLM par validation croisée « repeated k-fold cross validation » dans toutes les situations.

Nous souhaitons compléter la bibliothèque par un test de significativité des coefficients par bootstrap [2].

## 6 Bibliographie


[1] R Development Core Team : R : *A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2008. http://www.R-project.org.

[2] Bastien, Ph., Esposito Vinzi, V. & Tenenhaus, M. : PLS generalized linear regression, *Computational Statistics & Data Analysis*, 48(1), 17-46, 2005.

[3] Kettaneh-Wold, N. : Analysis of mixture data with partial least squares, *Chemometrics & Intelligent Laboratory Systems*, 14, 57-69, 1992.

[4] Meyer, N., Maumy-Bertrand, M. & Bertrand, F. : Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives : application aux données d'allélotypage, *Prépublication de l'IRMA*, 2009.

[5] Höskuldsson, A. : PLS regression methods, *Journal of Chemometrics*, 2, 211-228, 1988.

[6] Wold, S., Sjöström, M. & Eriksson, L. : PLS-regression: a basic tool of Chemometrics, *Chemometrics and Intelligent Laboratory Systems*, 58, 109-130, 2001.

[7] Alon, U., Barkai, N., Notterman, D. A., Gish, K., Ybarra, S., Mack, D. & Levine A. J. : Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissue probed by oligonucleotide arrays, *Proc. Natl. Acad. Sci. USA*, 96, 6745-6750, 1999.

[8] Varmuza, K. & Filzmoser, P. : *Introduction to Multivariate Statistical Analysis in Chemometrics*, CRC Press, Boca Raton, USA, 2009.

[9] Eriksson, L., Johansson, E., Kettaneh-Wold, N., Trygg, J., Wikström, C. & Wold, S. : *Multi- and Megavariate Data Analysis, Principles and Applications*. Umetrics Academy, Umeå, Sweden, 2001.