

La bibliothèque `plsRglm`, modèles linéaires généralisés PLS sous R

Frédéric Bertrand¹, Myriam Maumy-Bertrand¹, Nicolas Meyer²

¹ Institut de Recherche Mathématique Avancée, Université de Strasbourg

² Laboratoire de Biostatistique - Faculté de Médecine - Université de Strasbourg

Chimiométrie 2009 — 30 Novembre 2009



Plan

- 1 Introduction
- 2 Méthode
- 3 Applications
- 4 Discussion

Contexte

La régression PLS est un outil important en chimiométrie

Contexte

La régression PLS est un outil important en chimiométrie


- en exploratoire
- en prédictif

Contexte

La régression PLS est un outil important en chimiométrie

- en exploratoire
- en prédictif
- en cas de colinéarité
- en raison des dimension de \mathbf{X} , $p \gg n$
- éventuellement des données manquantes

⇒ outils logiciels adéquats

- SIMCA, The Unscrambler, SAS, divers packages 

Contexte (2)

Pourquoi un nouveau package R consacré à la PLS ?

Contexte (2)

Pourquoi un nouveau package R consacré à la PLS ?

⇒ limites des bibliothèques actuelles

Contexte (2)

Pourquoi un nouveau package R consacré à la PLS ?

⇒ limites des bibliothèques actuelles

- pas de gestion des données manquantes (même si NIPALS)
- important en médecine (X-omics, allélotypage, data mining)
p.ê. moins en chimiométrie

Contexte (2)

Pourquoi un nouveau package R consacré à la PLS ?

⇒ limites des bibliothèques actuelles

- pas de gestion des données manquantes (même si NIPALS)
- important en médecine (X-omics, allélotypage, data mining)
p.ê. moins en chimiométrie
- rareté des critères de sélection autre que VC
par ex., LOO uniquement
- et VC uniquement sur données complètes
- pas de bootstrap ni de fonctions graphiques

Finalité de la bibliothèque `plsRglm`

- 1 extension de la régression PLS au cas des modèles linéaires généralisés
 - 2 notamment la régression logistique PLS (Bastien 2005)
 - 3 traitement des jeux de données incomplets par VC
 - 4 fonctions graphiques et bootstrap
- illustré sur données d'allélotypage (Meyer *et.al.*, 2009)

Plan

- 1 Introduction
- 2 Méthode**
- 3 Applications
- 4 Discussion

Notations en PLS

- Soit \mathbf{X} la matrice des prédicteurs $\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_p$ et \mathbf{y} .
- Reg. PLS : composantes orthogonales t_h , $\max(\text{cov}(\mathbf{y}, \mathbf{t}_h))$
- $\mathbf{y} = \mathbf{T}^t \mathbf{c} + \epsilon$, avec \mathbf{T} la matrice des composantes
- En posant $\mathbf{T} = \mathbf{XW}^*$, alors : $\mathbf{y} = \mathbf{XW}^{*t} \mathbf{c} + \epsilon$

Notations en PLS

- Soit \mathbf{X} la matrice des prédicteurs $\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_p$ et \mathbf{y} .
- Reg. PLS : composantes orthogonales t_h , $\max(\text{cov}(\mathbf{y}, \mathbf{t}_h))$
- $\mathbf{y} = \mathbf{T}^t \mathbf{c} + \epsilon$, avec \mathbf{T} la matrice des composantes
- En posant $\mathbf{T} = \mathbf{XW}^*$, alors : $\mathbf{y} = \mathbf{XW}^{*t} \mathbf{c} + \epsilon$

$$y_i = \sum_{h=1}^H (c_h w_{1h}^* x_{i1} + \dots + c_h w_{ph}^* x_{ip}) + \epsilon_i$$

où $H \leq \text{rg}(X)$

- Les coefficients $c_h w_{jh}^*$, où $1 \leq j \leq p$, (Wold *et al.*, 2001) : relation entre vecteur y et les x_j à travers les t_h .

Extension GLM de la PLS (1)

- réponse \mathbf{y} sur les $\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_p$ avec H composantes \mathbf{t}_h (Bastien *et al.* 2005)

$$g(\theta)_i = \sum_{h=1}^H c_h \mathbf{t}_h,$$

- θ : espérance ou vecteur des probabilités d'une loi discrète et

$$\mathbf{t}_h = w_{1h}^* x_{i1} + \dots + w_{ph}^* x_{ip}.$$

- fonction de lien $g(\theta)$ selon \mathbf{y} et qualité de l'ajustement du modèle aux données.
- les composantes PLS \mathbf{t}_h sont orthogonales.

Extension GLM de la PLS (2)

- pour obtenir w , on remplace
 - la régression linéaire simple (itérative)
 - la régression généralisée
- ⇒ calculer le coefficient a_{1j} de \mathbf{x}_j dans la régression linéaire généralisée de \mathbf{y} sur chaque prédicteur $\mathbf{x}_j, 1 \leq j \leq p$.
- idem pour c et les composantes :

Extension GLM de la PLS (2)

- pour obtenir w , on remplace
 - la régression linéaire simple (itérative)
 - la régression généralisée
- ⇒ calculer le coefficient a_{1j} de \mathbf{x}_j dans la régression linéaire généralisée de \mathbf{y} sur chaque prédicteur \mathbf{x}_j , $1 \leq j \leq p$.
- idem pour c et les composantes :
 - régression généralisée simple
 - régression généralisée multiple.

Extension GLM de la PLS (2)

première composante normer le vecteur a_1 : $w_1 = a_1 / \|a_1\|$,
puis calculer la composante

$$\mathbf{t}_1 = 1 / ({}^t w_1 w_1) \mathbf{X} w_1$$

h^{eme} composante calcul du coef a_{hj} de \mathbf{x}_j dans la reg. lin.
général. de \mathbf{y} sur $\mathbf{t}_1, \dots, \mathbf{t}_{h-1}$ et \mathbf{x}_j

- normer le vecteur colonne a_h : $w_h = a_h / \|a_h\|$
- calculer la matrice résiduelle \mathbf{X}_{h-1} de la régression linéaire de \mathbf{X} sur $\mathbf{t}_1, \dots, \mathbf{t}_{h-1}$, puis calculer la composante

$$\mathbf{t}_h = 1 / ({}^t w_h w_h) \mathbf{X}_{h-1} w_h.$$

- exprimer la composante \mathbf{t}_h en termes de prédicteurs \mathbf{X} :

$$\mathbf{t}_h = \mathbf{X} w * h.$$

- adaptation en cas de données incomplètes

Bootstrap dans les reg PLS

On suppose avoir retenu un nombre adéquat de composantes dans PLS1 de \mathbf{Y} sur $\mathbf{x}_1, \dots, \mathbf{x}_j, \dots, \mathbf{x}_p$.

- suivant Lazraq et al. (2003)
- pour faire des intervalles et tests bootstrap
- construction des IC avec plusieurs variantes
- normaux, percentiles ou BCa (Efron et Tibshirani 1993 ou Davison et Hinkley 1997).
- repose sur package boot

Bootstrap dans les reg PLS-GLM

On retient m composantes dans une reg PLS GLM.

- Soit $\widehat{F}_{(\mathbf{T}|\mathbf{y})}$ la fonction de répartition empirique étant donnée \mathbf{T} formée des m composantes PLS et la réponse \mathbf{y} .

Étape 1. Tirer B échantillons de $\widehat{F}_{(\mathbf{T}|\mathbf{y})}$.

Étape 2. Pour tout $b = 1, \dots, B$, calculer :

$$\mathbf{c}^{(b)} = (\mathbf{t}^{(b)\top} \mathbf{T}^{(b)} \mathbf{T}^{(b)})^{-1} \mathbf{t}^{(b)\top} \mathbf{y}^{(b)} \quad \text{et} \quad \mathbf{b}^{(b)} = \mathbf{W}^* \mathbf{c}^{(b)},$$

où $[\mathbf{T}^{(b)}, \mathbf{y}^{(b)}]$ est le b -ème échantillon bootstrap

- $\mathbf{c}^{(b)}$ coefficients des composantes
- $\mathbf{b}^{(b)}$ coefficients des p prédicteurs d'origine pour cet échantillon
- \mathbf{W}^* est la matrice fixe des poids des prédicteurs dans le modèle d'origine (m composantes)

Bootstrap dans les reg PLS-GLM

- Étape 3.** Pour chaque j , Φ_{b_j} approximation de Monte-Carlo de F de la statistique bootstrap de b_j .
- Pour chaque b_j , boîtes à moustaches et IC à partir des percentiles de Φ_{b_j} .
 - Un intervalle de confiance peut être défini par $I_j(\alpha) = \Phi_{b_j}^{-1}(\alpha), \Phi_{b_j}^{-1}(1 - \alpha)$
 - où $\Phi_{b_j}^{-1}(\alpha)$ et $\Phi_{b_j}^{-1}(1 - \alpha)$ sont les valeurs obtenues à partir de la fonction de répartition de la statistique bootstrap, niveau nominal de confiance de niveau $100(1 - 2\alpha)\%$

Le contenu des objets PLS du package

donne tous les résultats classiques nécessaires à la bonne interprétation d'un modèles PLS

Le contenu des objets PLS du package

donne tous les résultats classiques nécessaires à la bonne interprétation d'un modèles PLS pour réaliser de l'inférence et les cartes.

Le contenu des objets PLS du package

donne tous les résultats classiques nécessaires à la bonne interprétation d'un modèles PLS pour réaliser de l'inférence et les cartes.

- w , $||w||$, w^* , t_h , p , c ,
- \hat{Y} , \hat{Y}_{resid} ,
- YNA, residY, ExpliX, na.miss.X, XXNA, residXX, PredictY,
- press.ind, press.tot, ttPredictY, typeVC, computed nt,
- CoeffCFull, CoeffConstante, Std.Coeffs, press.ind2,
- RSSresidY, Yresidus, RSS, residusY, AIC.std, AIC,
- Nombre de mal classés, Proba d'affectation à une classe,
- standard/missingdata/adaptative
- R^2 residY, R^2 , PRESS, press.tot2, Q^2 , Q^2 lim, Q^2 cum,
- infos relatives à la VC, critères d'information

Plan

- 1 Introduction
- 2 Méthode
- 3 Applications**
- 4 Discussion

Données d'allélotypage

- données génétiques
- microsatellites : structures répétitives marqueurs de l'ADN
- \mathbf{X} , p variables, mesures binaires : altération présente / absente
- \mathbf{y} binaire (stade ou localisation métastase)
 - $\approx 30\%$ de données manquantes
 - $p \simeq n$ ou $p > n$, colinéarité

Données d'allélotypage

- données génétiques
- microsatellites : structures répétitives marqueurs de l'ADN
- \mathbf{X} , p variables, mesures binaires : altération présente / absente
- \mathbf{y} binaire (stade ou localisation métastase)
 - $\approx 30\%$ de données manquantes
 - $p \simeq n$ ou $p > n$, colinéarité

Exemp. cancer du colon, série de 33 microsatellites

- objectif descriptif et explicatif
- prédire stade de progression d'un cancer en 2 ou 4 classes
 - reg log binomiale/ordinale/multinomiale
- origine du package
- ici sur données complétées

Nbre composantes	0	1	2	3	4	5
AIC	145,83	118,14	109,96	105,16	103,84	104,73
BIC	148,47	123,43	117,89	115,74	117,06	120,60
Mal Classé	49	28	26	22	21	21
Préd. Significatifs	2	1	0	0	0	0
Mal Classé (10-CV)		64	62	57	61	62
$Q^2 \chi^2$ (10-CV)		-2,56	-23,90	$-1,00 \cdot 10^3$	$-1,17 \cdot 10^5$	$-4,12 \cdot 10^7$
χ^2 Pearson	104,00	100,54	99,18	123,38	114,78	98,88

Table: Résultats de la validation croisée, allélotypage, $k = 10$

BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS

Based on 250 bootstrap replicates

CALL :

```
boot.ci(boot.out = aze_compl.boot, conf = c(0.9, 0.95),  
type = c("norm", "basic", "perc", "bca"), index = 33)
```

Intervals :

Level	Normal	Basic
90%	(-1.0492, -0.0312)	(-1.0193, 0.0350)
95%	(-1.1467, 0.0663)	(-1.0933, 0.2430)

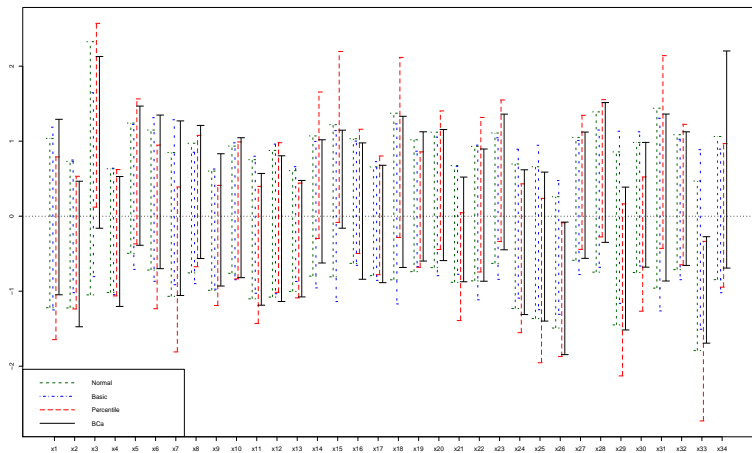
Level	Percentile	BCa
90%	(-1.2515, -0.1971)	(-1.1059, -0.1235)
95%	(-1.4594, -0.1231)	(-1.2855, -0.0875)

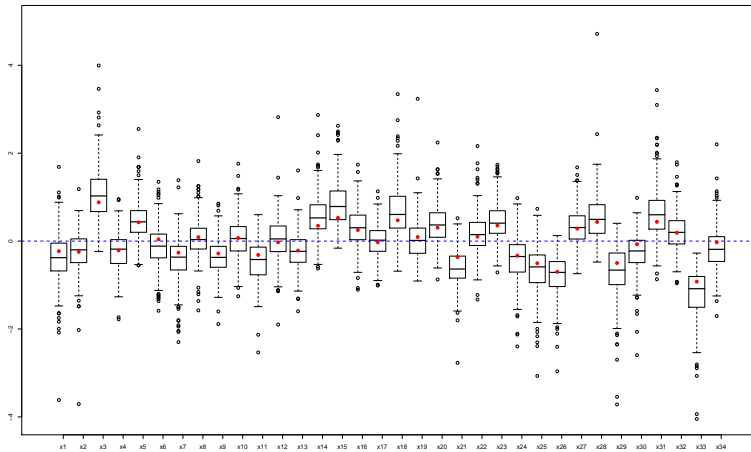
Calculations and Intervals on Original Scale

Some basic intervals may be unstable

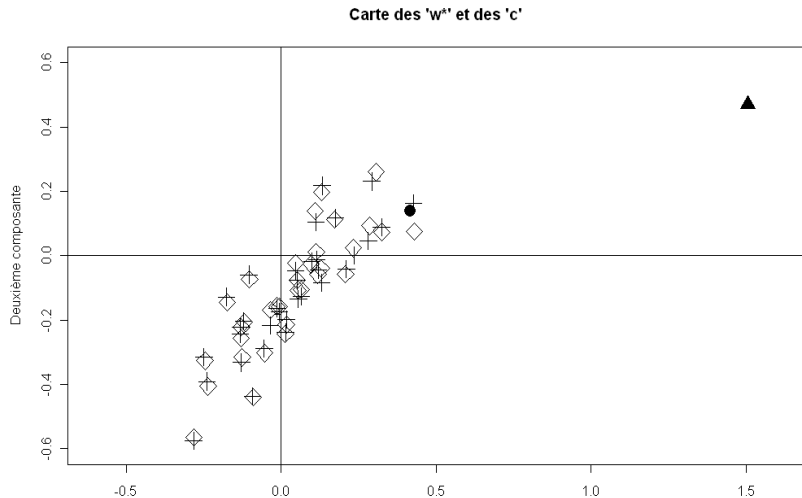
Some percentile intervals may be unstable

Some BCa intervals may be unstable





Carte des w^*,c (PLS et PLS logistique), données d'allélotypage.



Plan

- 1 Introduction
- 2 Méthode
- 3 Applications
- 4 Discussion**

Points forts de la bibliothèque

- modèles de régression PLS et PLS-GLM
- sur données complètes *et* incomplètes
- choix du nombre de composantes par différents critères
 - AIC / BIC
 - selon significativité des coefficients dans t_h
 - Q2 par validation croisée, même sur données incomplètes
- validation croisée « repeated k -folds cross-validation »
- bootstrap des coef des prédicteurs
 - PLS *et* PLS-GLM (reg logistique, survie, loi gamma, loi beta-binomiale etc).
 - données complètes *et* incomplètes

Disponibilité de plsRglm

- package fonctionnel
- page d'aide en cours de finalisation
- disponible sur le site

`http ://udsmed.u-strasbg.fr/labiostat/`

- en attendant de le mettre sur le CRAN

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.