CONSERVATOIRE NATIONAL DES ARTS ET MÉTIERS

PARIS

THÈSE

Pour obtenir le grade de

DOCTEUR DU CONSERVATOIRE NATIONAL

DES ARTS ET MÉTIERS

Spécialité : INFORMATIQUE

Présentée et soutenue publiquement par

Philippe BASTIEN

18 Mars 2008

Sujet de la thèse :

RÉGRESSION PLS ET DONNÉES CENSURÉES

Composition du jury :

Gilbert SAPORTA	**Professeur, CNAM**	**Directeur de thèse**
Avner BAR HEN	**Professeur, Paris V**	**Rapporteur**
Douglas RUTLEDGE	**Professeur, INA P-G**	**Rapporteur**
Ali GANNOUN	**Professeur, CNAM**	**Examinateur**
El Mostapha QANNARI	**Professeur, ENITIAA**	**Examinateur**
Michel TENENHAUS	**Professeur, HEC**	**Examinateur**
Jean-Philippe VERT	**Docteur habilité,** **Ecole des Mines de Paris**	**Examinateur**

À mes parents

Remerciements

Je tiens tout d'abord à remercier Gilbert Saporta qui m'a permis de réaliser cette thèse au Conservatoire National des Arts et Métiers en acceptant de la diriger. Ses conseils m'ont été précieux dans la rédaction de la thèse. Je remercie aussi particulièrement les rapporteurs Avner Bar Hen et Douglas Rutlege pour avoir pris de leur temps pour examiner ma dissertation, ainsi qu'Ali Gannoun, El Mostapha Qannari, Michel Tenenhaus, et Jean-Philippe Vert qui m'ont fait l'honneur d'en constituer le jury.

Je remercie particulièrement L'Oréal qui m'a donné la possibilité d'effectuer cette thèse.

Je remercie Svante Wold pour m'avoir fait découvrir la régression PLS et conforté dans ma volonté d'étendre les principes de la régression PLS aux données censurées.

Je tiens à remercier Arthur Tenenhaus et François-Xavier Lejeune pour avoir partagé avec moi le temps de leur stage de DEA une passion commune.

Je remercie Vincenzo Esposito Vinzi pour notre collaboration à l'article de CSDA, son sens de l'hospitalité, et sa gentillesse.

J'ai une pensée toute particulière pour Michel Tenenhaus qui m'a beaucoup appris et à qui je dois certainement d'avoir tenté cette aventure. Ses compétences, son enthousiasme, ses qualités pédagogiques et humaines ont toujours été pour moi un exemple et une source de motivation.

Finalement je remercie Titouan et Timoté pour avoir épargné mes documents, et Jessica pour son intérêt désintéressé.

Table des matières

Introduction 5

1. Aspects algorithmiques de la régression PLS 10

1.1 NIPALS 11

1.1.1 Méthode de la puissance itérée 11

1.1.2 Algorithme NIPALS 12

1.2 PLS-NIPALS 13

1.2.1 PLS2-NIPALS 13

1.2.2 PLS1-NIPALS 15

1.3 Régression PLS et critère d'optimisation 16

1.3.1 Régression PLS et critère de Tucker 16

1.3.2 SIMPLS 17

1.4 Régression PLS1 et suite de Krylov 19

1.4.1 Orthonormalisation de Gram-Schmidt 19

1.4.2 Suite de Krylov 19

1.5 PLS et méthodes de Lanczos 20

1.5.1 Algorithme de Lanczos - Lanczos tridiagonalisation 21

1.5.2 Valeurs et vecteurs de Ritz 21

1.5.3 PLS1 et l'algorithme Bidiag2 22

1.5.4 PLS et décomposition bilinéaire 23

1.5.5 PLS et gradients conjugués 24

1.6 Résumé du chapitre 25

2. Régression linéaire généralisée PLS 26

2.1 Composantes PLS et moindres carrés ordinaires 26

2.2 Sélection des variables et détermination du nombre de composantes PLS 29

2.3 Régression linéaire généralisée PLS (PLS-GLR) 29

2.3.1 Algorithme de régression linéaire généralisée PLS 30

2.4 Article CSDA 2005 33

3. Données censurées et régression PLS généralisée 34

3.1 Les modèles de survie 34

3.1.1 Processus stochastiques 35

3.1.2 Le modèle de Cox 36

3.1.3 Modèle de Cox et vraisemblance monotone 40

3.2 Données de survie et processus de Poisson 41

3.2.1 Survie exponentielle et processus de Poisson pour la variable de censure 41

3.2.2 Survie exponentielle par morceaux et processus de Poisson pour la variable de censure: application au modèle de Cox 42

3.2.3 Données de survie paramétriques et processus de Poisson pour la variable de censure 43

3.3 Modèle de Cox et données discrètes 44

3.3.1 Données structurellement discrètes. 45

3.3.2 Données discrètes avec un sous-jacent continu. 46

3.4 Capacité prédictive d'un modèle 48

3.5 Données censurées et régression PLS généralisée 51

3.6 PLS-Cox 52

3.6.1 Algorithme PLS-Cox 52

3.7 Autres approches 53

3.8 Applications 54

4. Régression PLS et Imputation Multiple 55

4.1 Imputation simple 55

4.2 Algorithme EM 56

4.3 PLS EM 59

4.4 EM stochastique 62

4.5 PLS SEM 65

4.6 Imputation multiple 67
4.7 PLS-MI 69
4.8 Discussion et extension au modèle PLS-Cox. 70

5. Kernel PLS 72
5.1 Objectif 72
5.2 Historique 73
5.2.1 Kernel PLS (Lindgren, Geladi, et Wold S., 1993) 73
5.2.2 Modifications de Kernel PLS (de Jong et Ter Braak, 1994) 74
5.2.3 Modifications de Kernel PLS (Dayal et Mc Gregor, 1997) 74
5.2.4 Kernel PLS (Rännar, Geladi, et Wold S., 1994) 75
5.2.5 Canonical Partial least squares (de Jong, Wise et Ricker, 2001) 76
5.3 CANPLS 78
5.3.1 Algorithme CANPLS 78
5.3.2 Proposition 79
5.3.3 Kernel PCA 80
5.3.4 PLS par blocs pour des matrices très larges 80
5.3.5 Condition de signe 81
5.3.6 Continuum Power Regression 81
5.3.7 Relation entre PLS et PCR 82
5.3.8 Relation entre CANPLS et Kernel PLS (Rännar et al.) 82
5.3.9 CANPLS2 83
5.3.10 CANPLS dual 84
5.4 Généralisation au modèle de Cox 84
5.4.1 PC PLS-Cox 85
5.4.2 CANPLS-Cox 85
5.4.3 Application 87
5.5 Généralisation aux noyaux non linéaires 89
5.5.1 SAMPLS (Bush et Nachbar, 1993) 89
5.5.2 KPLS (Rosipal et Trejo, 2001) 90
5.5.2.1 Algorithme KPLS 91
5.5.3 DKPLS (Bennett et Embrecht, 2003) 92
5.5.4 KPLS Cox 92
5.5.5 Application 93

5.5.5.1 Simulation des descripteurs 93
5.5.5.2 Simulation d'une Gaussienne multivariée avec une structure de corrélation définie 93
5.5.5.3 Simulation des données de survie 94
5.5.5.4 Résultats 94

6 Deviance residuals based PLS regression 98

Conclusion 99

Bibliographie 103

Introduction

Le modèle à risques proportionnels proposé par Cox en 1972 [29] pour étudier la relation entre le temps d'apparition d'un évènement et un ensemble de descripteurs a eu un impact considérable dans l'analyse des données de survie, tant du point de vue théorique que pratique et est rapidement devenu le modèle le plus utilisé. Cependant, comme dans le cas de la régression linéaire multiple, il suppose qu'il y ait plus d'observations que de variables et des variables pas trop corrélées entre elles. Ces contraintes sont souvent impossibles à satisfaire dans la pratique. En oncologie par exemple, la recherche de descripteurs moléculaires liés à la durée de survie suppose de prendre en compte l'expression de milliers de gènes pour généralement seulement quelques dizaines d'individus.

Le développement de nouvelles méthodes analytiques en chimie dans les années 70-80, comme la spectroscopie NIR qui fut à l'origine de l'essor de la Chimiométrie, s'est accompagné d'une transformation de la structure des tableaux de données en générant des matrices larges avec de très nombreux descripteurs potentiellement corrélés entre eux pour seulement un nombre limité d'observations. La régression PLS [137], qui peut être vue comme une méthode de régularisation basée sur la réduction de la dimensionnalité, a été proposée à cette époque par Wold et al. [158], en particulier en calibration multivariée ou dans les analyses de type structure-activité (QSAR), comme une alternative à la régression linéaire multiple (OLS) en cas de forte multi-colinéarité des descripteurs ou lorsque le nombre de descripteurs dépasse le nombre d'individus. D'un point de vue algébrique on doit résoudre un problème dit mal posé associée à une matrice mal conditionnée.

Dans l'analyse des données censurées on peut aussi vouloir prendre en compte de nombreux descripteurs. Les problèmes sont alors les mêmes que ceux décrits précédemment. Par exemple, d'un point de vue transcriptomique, l'évolution des cancers est plus certainement

liée à l'association de nombreux gènes interagissant entre eux qu'à une origine monogénique. Les méthodes pouvant prendre en compte l'ensemble des gènes d'intérêt pour définir des profiles d'expression prédictifs de la survie vont permettre le développement de nouveaux outils diagnostiques. On peut en particulier faire un parallèle en analyse des données entre l'analyse spectrale apparue dans les années 70 et l'analyse transcriptomique 30 ans plus tard avec les mêmes problèmes en termes de régularisation. Il était par conséquent naturel de vouloir utiliser les principes de la régression PLS dans les modèles de régression avec données censurées et tout particulièrement dans le cadre du modèle de Cox. C'est l'objectif de cette thèse.

En utilisant une formulation alternative de la régression PLS, comme succession de régressions simples et multiples, Tenenhaus (1999) [138], et Esposito Vinzi et Tenenhaus (2001) [49] ont proposé une généralisation de la régression PLS à la régression logistique (PLS-LR) dans une approche beaucoup plus simple que celle développée par Marx en 1996 [98]. L'algorithme PLS-LR a ensuite été étendu par Bastien et Tenenhaus (2001) [11] à tout modèle de régression linéaire généralisé (PLS-GLR) et au modèle de Cox (PLS-Cox) dans le cas des données censurées. Des améliorations ont ensuite été apportées dans Bastien et al. (2005) [16] pour traiter le cas des descripteurs catégoriels et valider le modèle par ré-échantillonnage bootstrap. Dans le cadre des données censurées, des applications des modèles PLS-Poisson, PLS-Cloglog et PLS-Cox ont été proposées pour prédire la survenue du blanchiment précoce des cheveux [11,12,16] et pour prédire, à partir des données d'expressions des gènes, la probabilité de survie de patients atteints de lymphomes [15].

La régression PLS sous sa forme standard, avec l'utilisation de l'algorithme NIPALS [156], peut être utilisée avec des valeurs manquantes. Cette approche qui s'apparente d'une certaine façon à une imputation simple ne restitue pas l'incertitude associée à la prédiction des valeurs manquantes sous-estimant ainsi la variance de l'estimation des coefficients. L'imputation multiple introduite par Rubin dans les années 70 [118] remplace chaque valeur manquante par un ensemble de valeurs simulées qui tendent à reproduire l'incertitude associée à la prévision d'une valeur manquante à partir des données disponibles. Ce procédé conduit à une démarche confirmatoire valide reflétant l'incertitude associée aux valeurs manquantes. Nous proposerons différentes alternatives à NIPALS [14] en introduisant le concept d'imputation multiple en régression PLS simple et généralisée.

L'algorithme PLS-Cox est séquentiel dans les estimations des poids utilisés pour la détermination des composantes PLS. Quand le nombre de descripteurs devient très élevé, comme c'est le cas dans l'analyse des données d'expression des gènes ou le nombre de descripteurs peut atteindre plusieurs dizaines de milliers, les temps de calcul deviennent un facteur limitatif et des problèmes de capacité mémoire peuvent survenir. Des problèmes calculatoires similaires du fait de matrices très larges avec un grand nombre de descripteurs se sont déjà posés par le passé en Chimiométrie et des solutions appelées « Kernel PLS » ont été proposées dans les années 90 [38,40,91,113]. L'objectif de ces méthodes était de construire les composantes PLS en utilisant des matrices, kernels ou noyaux, de taille beaucoup plus petite que les matrices d'origine. Le terme noyau est utilisé ici plus avec pour objectif de réduire la complexité de l'espace des descripteurs que de définir une transformation non linéaire dans un espace de dimension plus élevée. Dans cet esprit nous proposerons des alternatives plus rapides à PLS-Cox en grande dimension. Une première alternative sera d'utiliser l'algorithme PLS-Cox sur les composantes principales (PC PLS-Cox). Une seconde alternative, CANPLS-Cox, consistera en la généralisation d'un algorithme original, CANPLS, inspiré de l'algorithme Canonical PLS de de Jong et al. (2001) [40]. Dans l'esprit de kernel logistic PLS proposée par Tenenhaus et al. [137], nous proposerons une généralisation de PLS-Cox à des noyaux non linéaires et gaussiens en particulier. Des applications à partir de données réelles [116] et simulées seront présentées.

Depuis 2001, afin de permettre la prise en compte d'information provenant de très nombreux descripteurs, de nombreux développements sont apparus autour du modèle de Cox dans des approches régularisées. La régularisation peut s'exprimer comme une réduction de la dimensionnalité, comme c'est le cas avec la régression PLS en recherchant la solution dans un sous-espace appelé espace de Krylov. Il en résulte des coefficients de régression biaisés mais à plus faible variance. La régularisation peut aussi s'exprimer par maximisation de la log vraisemblance partielle avec des contraintes de type L_1 sur les coefficients. Cette approche a été proposée par Tibshirani en 1997 [146] comme une extension de sa procédure de sélection de variables appelée Lasso [145] acronyme de « Least Absolute Shrinkage and Selection Operator ». Cependant la minimisation quadratique utilisée par la procédure Cox-Lasso ne fonctionne pas en grande dimension. En 2004, Efron et al. [45] ont proposé une procédure très efficace appelée LARS pour « Least Angle Regression for variable Selection » qui redonne le Lasso au prix d'une légère modification. En 2005, Gui et Li [60], en utilisant le lien étroit entre LARS et Lasso ont proposé une solution pour le Cox-lasso appelé LARS-Cox dans le

cas de données hautement multidimensionnelles. Cependant les moindres carrés pondérés itérés utilisés par l'algorithme LARS-Cox contrebalance l'efficacité de la procédure LARS-Lasso. Récemment, Segal [128] a décrit une stratégie qui restitue l'efficacité numérique de la procédure LARS-Lasso. Il a montré que l'expression que l'on cherche à minimiser dans la procédure Cox-lasso de Tibshirani peut être approchée par la somme des carrés des résidus de la déviance. Il a ainsi proposé d'utiliser l'algorithme original LARS-Lasso sur ces résidus spécifiques. En suivant l'approche de Segal on propose d'effectuer la régression PLS sur les résidus de la déviance comme une nouvelle alternative à PLS-Cox dans le cas de données hautement multidimensionnelles. Cette alternative est très intéressante car particulièrement simple à mettre en œuvre. Les performances prédictives des différentes approches ont été comparées sur des données d'expression de gènes en utilisant des courbes ROC généralisées [64,65].

La régression PLS sous sa forme standard prend en compte tous les descripteurs dans la construction des composantes PLS. Un modèle plus parcimonieux est proposé dans Bastien et al. [16] en ne retenant dans la construction des composantes PLS que les descripteurs associés à des coefficients de régression significatifs. On propose de généraliser cette approche dans une logique sélection de variables en considérant le seuil de significativité comme un méta-paramètre dans l'esprit de l'algorithme « Threshold Gradient Descent » proposé par Friedman et Popescu [54]. Cette approche, en définissant une famille de trajectoires dans l'espace des paramètres, propose une forme de continuum entre régression PLS et approche pénalisée de type L_1. En particulier, pour des valeurs faibles du seuil de significativité, on retrouve sur l'exemple des données de Rosenwald et al. [116] les descripteurs sélectionnés par le modèle LARS-Cox confirmant ainsi les résultats de Friedman et Popescu ou ceux de Gui et Li [61].

L'utilisation des principes de la régression PLS dans le contexte des données censurées à travers ces nouveaux outils apparaît particulièrement utile dans la détermination de bio-marqueurs, en oncologie en particulier, et dans le développement de nouveaux tests pharmacogénétiques. En opérant une réduction de la dimension de l'espace des gènes orientée vers l'explication de la fonction de risque, cette approche permet la détermination de signatures d'expression transcriptomique associées à la survie. De plus, la prise en compte dans le design et dans l'analyse des résultats des études cliniques des profiles d'expression estimés à partir de ces outils est de nature à améliorer la puissance des essais en mieux

contrôlant par ajustement au niveau génotypique les facteurs de variabilité associés à la réponse.

Chapitre 1

Aspects algorithmiques de la régression PLS

.

Dans ce chapitre nous présentons les notions principales associées à la régression PLS qui seront utiles par la suite. Nous rappelons aussi les relations intéressantes et moins connues entre la régression PLS et les méthodes de Lanczos [31, 46, 58, 59, 106, 111, 161].

La régression PLS [139,158] est une technique pour modéliser une relation linéaire entre un ensemble de variables réponses Y et un ensemble de variables explicatives X. Dans un premier temps PLS crée des composantes non corrélées qui sont des combinaisons linéaires des X avec des poids proportionnels à la covariance entre les variables explicatives et les variables de réponse. Un modèle de régression linéaire est ensuite construit sur ces composantes. Cela conduit à des coefficients biaisés mais cohérents et plus stables qu'avec les moindres carrés ordinaires. C'est une technique particulièrement utile en cas de forte multicolinéarité des descripteurs ou lorsque le nombre de descripteurs excède le nombre d'individus comme c'est le cas par exemple dans les applications en Chimiométrie (spectroscopie NIR) ou en Chemo ou Bioinformatique où on est confronté à des matrices très larges avec souvent des milliers de variables pour seulement quelques dizaines d'individus, et où les méthodes classiques deviennent inopérantes.

PLS a été d'abord introduit au début des années 70 par Herman Wold [140,157] dans le cadre de la modélisation de relations structurelles sur variables latentes pour l'analyse de plusieurs blocs de variables observées sur les mêmes individus. Cette méthode appelée par la suite « PLS Path Modelling » traduit en Français par « Approche PLS » est une alternative de type « Soft Modelling » à La méthode LISREL développée par Jöreskog [76], un élève de Wold, et

qui est basée sur le maximum de vraisemblance. Ce sont cependant, principalement en chimie, dans le cadre de deux blocs de variables qu'un développement particulier de cette méthode, appelée régression PLS, fut plus principalement utilisé. En effet à cette époque l'apparition de méthodes instrumentales multivariées (chromatographie, spectroscopie), rendit la régression PLS particulièrement utile.

1.1 NIPALS

L'algorithme de régression PLS classique, comme cas particulier de l'approche PLS, est basé sur des développements de l'algorithme NIPALS, acronyme de Non linear estimation by Iterative Partial Least Squares. L'algorithme NIPALS est une méthode robuste pour déterminer de façon itérative les valeurs et vecteurs propres d'une matrice. Elle permet en particulier de faire de l'analyse en composantes principales en présence de valeurs manquantes. Cet algorithme proposé par Herman Wold en 1966 [156] est inspiré de la méthode de la puissance itérée de Hotelling [68] pour le calcul du vecteur propre d'une matrice associé à la plus grande valeur propre, et son utilisation pour la détermination des composantes principales en ACP peut aussi être vue comme une réinvention d'une méthode de même nature suggérée par Fisher en 1923 [52,57].

1.1.1 Méthode de la puissance itérée

On initialise l'algorithme avec un vecteur w aléatoire, et on itère jusqu'à convergence :

$$\begin{cases} w = X'Xw \\ w = w/\|w\| \end{cases}$$

On obtient ainsi un vecteur w_1 vecteur propre de $X'X$ associé à la plus grande valeur propre λ_1. En effet, soit un vecteur b quelconque que l'on décompose dans la base canonique des vecteurs propres $(w_1,...,w_r)$ sous la forme $b = \sum_i \alpha_i w_i$, on a :

$$(X'X)^n b = \sum_i \alpha_i \lambda_i^n w_i \to \alpha_1 \lambda_1^n w_1 \propto w_1$$
$$avec\ \lambda_1 > \lambda_2 > ... > \lambda_r$$

La méthode converge après un nombre limité d'itérations, et on a la relation :

$$X'Xw_1 = \lambda_1 w_1$$

à partir de laquelle on déduit $\lambda_1 = \frac{\left\| w_1^{(iterration\ n+1)} \right\|}{\left\| w_1^{(iterration\ n)} \right\|}$

Notons que la méthode ne peut fonctionner dans le cas improbable ou le vecteur initial w serait choisi orthogonal à w_1.

On déflate ensuite la matrice $X'X$ en utilisant les relations ci-dessous pour la détermination de w_2 et des vecteurs propres de rang supérieur:

$$(X'X)^1 = X'X - \lambda_1 w_1 w_1^{'}$$

$$(X'X)^i = X'X - \sum_{j=1}^{i} \lambda_i w_i w_i^{'}$$

Cette méthode extrait ainsi les vecteurs propres l'un après l'autre dans l'ordre décroissant des valeurs propres.

1.1.2 Algorithme NIPALS

C'est un algorithme inspiré de la méthode de la puissance itérée appliquée à la matrice X, où l'on a remplacé le produit matrice-vecteur par des régressions simples.

1) $X_0 = X$ centré

2) pour h=1,…,a

2.1) t_h = une colonne de X_{h-1}, de préférence celle de plus forte variance.

2.2) Répéter jusqu'à convergence de p_h :

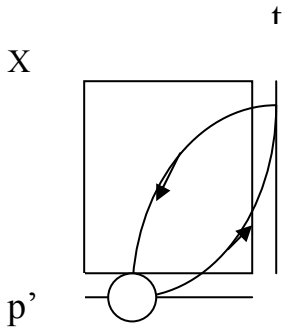
2.2.1) $p_h = X'_{h-1} t_h / t'_h t_h$

2.2.2) $p_h = p_h / \left\| p_h \right\|$

2.2.3) $t_h = X_{h-1} p_h / p'_h p_h$

2.3) $X_h = X_{h-1} - t_h p'_h$

L'algorithme NIPALS peut être schématisé comme décrit ci-dessous [57]:



L'algorithme peut fonctionner en présence de valeurs manquantes. Dans les étapes 2.2.1 et 2.2.3 p_h et t_h sont calculés comme les pentes des droites des moindres carrés passant par l'origine des nuages de points disponibles $(t_{hi}, x_{h-1,ji})$ et $(p_{hj}, x_{h-1,ji})$ respectivement.

1.2 PLS-NIPALS

L'algorithme PLS-NIPALS peut être vu comme un cas particulier de NIPALS qui utiliserait la ou les variables réponses pour la détermination des variables latentes.

1.2.1 Algorithme PLS2-NIPALS

1 *X, Y* centrés

2 Pour *h =1, a*

2.1 u_h colonne de *Y*

2.2 Répéter jusqu'à convergence de w_h

2.2.1 $w_h = X'_{h-1} u_h / u'_h u_h$

2.2.2 $\|w_h\| = 1$

2.2.3 $t_h = X_{h-1} w_h / w'_h w_h$

2.2.4 $c_h = Y' t_h / t'_h t_h$

2.2.5 $u_h = Y c_h / c'_h c_h$

2.3 $p_h = X'_{h-1} t_h / t'_h t_h$

2.4 $X_h = X_{h-1} - t_h p'_h$

3 $T = [t_1, ..., t_a]$, $\hat{y}^{PLS} = T(T'T)^{-1} T'y$

Les coordonnées des vecteurs w_h, t_h, u_h, p_h, *et* c_h représentent toutes des pentes de droites des moindres carrés passant par l'origine et peuvent donc être calculées avec les données manquantes.

On obtient en 2.2.2 un vecteur des coefficients normés dont les éléments sont proportionnels aux covariances associées, le terme $u_h' u_h$ disparaissant en normant en 2.2.2.

Calculer les covariances à partir de pentes permet ainsi de prendre en compte simplement les valeurs manquantes. Il est intéressant de noter que l'on réalise des régressions simples de $x_{h,j}$ *sur* u_h et non l'inverse, ce qui permet, u_h étant indépendant de $x_{h,j}$, d'éliminer u_h lorsque l'on norme les w_h.

Détermination du nombre de composantes

Le nombre de composantes PLS est généralement déterminé par validation croisée.

On définit pour la *h*-ème composante les critères du *PRESS* (PRedicted Residual Sum of Squares) et du *RSS* (Residual Sum of Squares) par les formules suivantes:

$$PRESS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{h(-i)})^2$$

Avec $\hat{y}_{h(-i)}$ la prévision de y_i à partir du modèle à *h* composantes estimé sans l'observation *i*.

$$RSS_h = \sum_{i=1}^{n} (y_i - \hat{y}_{hi})^2$$

On définit le pouvoir prédictif de la *h*-ème composante par :

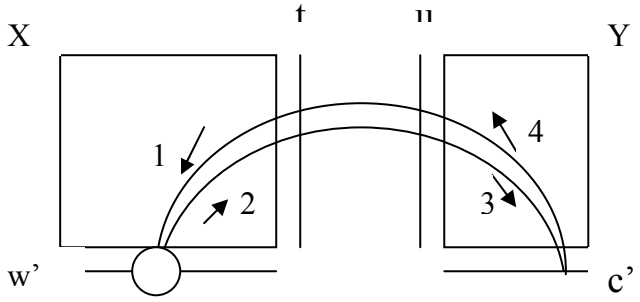
$$Q_h^2 = 1 - \frac{PRESS_h}{RSS_{h-1}} \text{ avec } RSS_0 = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

En particulier, une composante t_h est considérée comme significative par le logiciel SIMCA si : $\sqrt{PRESS_h} \le 0.95\sqrt{RSS_{h-1}} \Leftrightarrow Q_h^2 > 0.0975$

On peut finalement mesurer le pouvoir prédictif du modèle par le critère :

$$Q^2 = 1 - \prod_{h=1}^{H} \frac{PRESS_h}{RSS_{h-1}}$$

L'algorithme PLS-NIPALS peut être schématisé comme décrit ci-dessous [57]:



Dans le cas d'une seule réponse *y* l'algorithme PLS1-NIPALS n'est plus itératif.

1.2.2 Algorithme PLS1-NIPALS

1 X, y centrés

2 Pour $h = 1, a$

2.1 $w_h = X'_{h-1} y / y'y$

2.2 $\|w_h\| = 1$

2.3 $t_h = X_{h-1} w_h / w'_h w_h$

2.4 $p_h = X'_{h-1} t_h / t'_h t_h$

2.5 $X_h = X_{h-1} - t_h p'_h$

3 $T = [t_1, ..., t_a]$,

$$\hat{y}^{PLS} = T(T'T)^{-1} T'y$$

$$\hat{y}^{PLS} = \sum_{i=1}^{a} \frac{y'_i t_i}{t'_i t_i} t_i = \sum_{i=1}^{a} \frac{y'_i t_i}{t'_i t_i} X_{i-1} w_i = \sum_{i=1}^{a} \frac{y'_i t_i}{t'_i t_i} X w_i^*$$

Avec $W^* = W(P'W)^{-1}$

$W = [w_1, ..., w_a]$

$P = [p_1, ..., p_a]$

1.3 Régression PLS et critère d'optimisation

1.3.1 Régression PLS et critère de Tucker

PLS-NIPALS a tout d'abord été présentée comme une approche purement algorithmique et ce n'est qu'à la fin des années 80 qu'Höskuldsson [73] et Manne [94] montrèrent que cet algorithme était solution d'un problème d'optimisation. Ils démontrèrent que les premières composantes PLS $t_1 = Xw_1$ et $u_1 = Yc_1$ sont obtenues en maximisant le critère de Tucker de l'analyse factorielle inter-batteries [148] :

$$\mathrm{cov}^2(Xw_1, Yc_1) = \mathrm{var}(Xw_1)\,corr^2(Xw_1, Yc_1)\,\mathrm{var}(Yc_1)$$

Sous les contraintes: $\|w_1\| = 1,$ et $\|c_1\| = 1$

La solution à ce problème est obtenue en prenant le vecteur propre normalisé de la matrice $X'YY'X$ (respectivement $Y'XX'Y$) associé à la plus grande valeur propre commune.

On montre aussi que t_1 et u_1 sont vecteurs propres de respectivement $XX'YY'$ et $YY'XX'$ associés à la plus grande valeur propre commune.

Les autres composantes PLS $t_h = X_{h-1}w_h$ et $u_h = Y_{h-1}c_h$ sont obtenues en maximisant le critère :

$$\mathrm{cov}^2(X_{h-1}w_h, Y_{h-1}c_h) = \mathrm{var}(X_{h-1}w_h)\,\mathrm{cov}^2(X_{h-1}w_h, Y_{h-1}c_h)\,\mathrm{var}(Y_{h-1}c_h)$$

Sous les contraintes: $\|w_h\| = 1,$ et $\|c_h\| = 1$

Avec X_{h-1} le résidu de la régression de X sur $t_1,...,t_{h-1}$ et Y_{h-1} le résidu de la régression de Y sur $t_1,...,t_{h-1}$.

La solution w_h (respectivement c_h) est obtenue en prenant le vecteur propre normalisé de la matrice $X'_{h-1}Y_{h-1}Y'_{h-1}X_{h-1}$ (ou $Y'_{h-1}X_{h-1}X'_{h-1}Y_{h-1}$ respectivement) associé à la plus grande valeur propre commune.

On montre aussi [35] que t_h et u_h sont vecteurs propres de respectivement $X_{h-1}X'_{h-1}Y_{h-1}Y'_{h-1}$ et $Y_{h-1}Y'_{h-1}X_{h-1}X'_{h-1}$ associés à la plus grande valeur propre commune.

Quand $Y = y$, le rang de $X'YY'X$ est égal à 1, et le vecteur propre non normalisé est $X'y$ qui correspond à l'unique valeur propre non nulle $y'XX'y$.

En effet : $X'yy'XX'y = X'y\lambda = \lambda X'y$ avec λ un scalaire.

On a : $t_1 = Xw_1$ avec $w_{1j} = \dfrac{\mathrm{cov}(x_j, y)}{\sqrt{\sum_j \mathrm{cov}^2(x_j, y)}}$

Une autre façon, d'obtenir ce résultat est d'écrire l'expression du Lagrangien :

$$l = \mathrm{cov}(Xw_1, y) - \lambda(w_1'w_1 - 1),$$

puis, de le dériver suivant $w_{11}, ..., w_{1p}$, et λ.

Les autres composantes PLS $t_h = X_{h-1}w_h$ sont obtenues de façon similaire:

$$w_{hj} = \frac{\mathrm{cov}(x_{hj}, y)}{\sqrt{\sum_j \mathrm{cov}^2(x_{hj}, y)}}$$

avec x_{hj} le résidu de x_j dans la régression de x_j sur $t_1, ..., t_{h-1}$.

En 1997, Dayal et Mc Gregor [34] ont montré qu'il n'est pas nécessaire de déflater X et Y à la fois, mais que seulement l'un des deux suffisait.

1.3.2 SIMPLS

PLS-NIPALS n'offre pas de critère d'optimisation global. Il est nécessaire d'optimiser chaque composante PLS séparément. Une modification de l'algorithme PLS-NIPALS appelée SIMPLS par de Jong en 1993 [35] est elle solution d'un critère global.

SIMPLS cherche les composantes $t_h = Xa_h$ qui maximisent le critère suivant :

$$\max \sum_{k=1}^{q} \mathrm{cov}^2(y_k, Xa_h)$$

sous les contraintes a_h normés et $Xa_h \perp$

Une formulation équivalente est donnée par Abraham et Mérola en 2005 [1]

$$\max(d_j' Y' X a_j)^2$$

sous les contraintes : $a_j' a_j = 1,\ d_j' d_j = 1,\ a_j' X' X a_j = 0$ si $i > j$

La solution est donnée par les vecteurs propres correspondant aux plus grandes valeurs propres de :

$$(I_p - H_j) X' Y Y' X$$

avec : $H_j = X' T_{j-1} (T_{j-1}' X X' T_{j-1})^{-1} T_{j-1}')^{-1} T_{j-1}' X$ et $H_0 = 0$

Il est à noter que l'algorithme SIMPLS donne des solutions très proches de PLS-NIPALS dans le cas de PLS2, et des composantes PLS colinéaires avec PLS-NIPALS lorsqu'il n'y a qu'un seul vecteur réponse *y*, la première composante PLS t_1 étant identique pour les deux algorithmes. SIMPLS permet de plus une interprétation géométrique de PLS [110] plus simple que PLS-NIPALS, du fait que les composantes t_h sont exprimées à partir de la matrice initiale X. Il est aussi plus rapide en termes de temps de calculs que PLS-NIPALS.

Algorithme SIMPLS

$C_0 = X'Y$

Pour $h = 1 \rightarrow a$

r_h = vecteur singulier à gauche de C_{h-1} associé à la plus grande valeur propre.

$t_h = X r_h$

$p_h = X' t_h / t_h' t_h$

$C_h = (I - p_h (p_h' p_h)^{-1} p_h') C_{h-1}$ *pour que* r_h *soit* $\perp span(p_1, ..., p_{h-1})$

Fin

$$B^{PLS} = R(T'T)^{-1}T'Y$$

$$\hat{y}^{PLS} = XR(T'T)^{-1}T'Y = T(T'T)^{-1}T'Y$$

On peut noter que SIMPLS réalise la SVD de *X'Y*.

1.4 Régression PLS1 et suite de Krylov

La régression PLS est une méthode de régularisation par réduction de la dimensionnalité. La régression PLS1 recherche la solution $\hat{y}^{PLS}$ dans un sous espace de *X* appelé sous-espace de Krylov. $\hat{y}^{PLS}$ est la projection de *y* dans ce sous-espace de X engendré par les composantes PLS.

1.4.1 Orthonormalisation de Gram-Schmidt

L'orthonormalisation de Gram-Schmidt d'une suite de vecteurs $(u_1,...,u_n)$ en une base orthonormée $(v_1,...,v_n)$ utilisée par la suite est présentée ci-dessous :

$$v_1 = u_1,\ \|v_1\| = 1$$

$$v_2 = u_2 - v_1(v_1'u_2),\ \|v_2\| = 1$$

$$v_a = u_a - \sum_{k=1}^{a-1} v_k(v_k'u_a),\ \|v_a\| = 1$$

1.4.2 Suite de Krylov

En 1987-1988, Manne [94] et Helland [68] ont montré que la suite de vecteurs $(w_1,...,w_a)$ pouvait être obtenue par orthogonalisation de Gram-Schmidt de la suite de Krylov $K(s,S,a) = (s,Ss,S^2s,...,S^{a-1}s)$, avec $s = X'y$ et $S = X'X$. Ce résultat a ensuite été étendu aux autres suites en 1993 par de Jong [35]. On a en particulier :

$$(t_1,...,t_a) = GS\ K(XX'y, XX', a)$$

Phatak et de Jong [111] ont montré que le vecteur des coefficients de la régression PLS pouvait s'exprimer sous la forme :

$$\hat{\beta}_a^{PLS} = K_a (K_a' S K_a)^{-1} K_a' s$$

avec $K_a = K(s, S, a)$.

Le problème consiste alors à inverser $K_a' S K_a$, la matrice étant mal conditionnée. On montre cependant [111] que l'on peut remplacer K_a par toute base $V_a = (v_1, ..., v_a)$ qui engendre K_a.

$$\hat{\beta}_a^{PLS} = V_a (V_a' S V_a)^{-1} V_a' s$$

PLS-NIPALS consiste ainsi à proposer une base particulière: $V_a = GS(K_a) = W_a$

On a : $$\hat{\beta}_a^{PLS} = W_a (W_a' S W_a)^{-1} W_a' s$$

avec $W_a' S W_a$ matrice tridiagonale, facilement inversible

$$W_a' S W_a = \begin{pmatrix} \alpha_1 & \beta_1 & 0 \\ \beta_1 & \ddots & \beta_{a-1} \\ 0 & \beta_{a-1} & \alpha_a \end{pmatrix}$$

1.5 Régression PLS et méthodes de Lanczos

On peut voir ici le lien entre PLS1-NIPALS et la méthode de Lanczos. Cette méthode a été développée dans le domaine du calcul numérique par Lanczos en 1950 pour approximer rapidement les plus grandes valeurs propres de matrices symétriques *A.* Elle est particulièrement utile dans le cas de matrices large avec de nombreux zéros.

En partant d'un vecteur ***b***, la procédure construit une séquence de matrices tridiagonales $W_j' A W_j$ ou les vecteurs de Lanczos qui forment les colonnes de W_j sont construits par orthonormalisation de Gram-Schmidt de la série de Krylov $K(b, A, j) = (b, Ab, ...A^{j-1}b)$. Ils forment donc une base orthonormale de $K(b, A, j)$.

Remarque : si $V(n\ x\ n)$ est orthogonale ($V'V = I_n$), alors $B = V'AV$ est une transformation orthogonale de A qui préserve les valeurs et vecteurs propres de A.

On retrouve la régression PLS avec $b = X'y$, et $A = X'X$. Dans la base des vecteurs de Lanczos la matrice $X'X$ est tridiagonale.

L'algorithme original de Lanczos transforme une matrice symétrique A en une matrice $Q = W'AW$ tridiagonale à travers le schéma itératif suivant :

1.5.1 Algorithme de Lanczos

L'idée consiste à remplacer la matrice A par une matrice tridiagonale pour laquelle la détermination des valeurs et vecteurs propres est beaucoup plus simple.

$$w_i = N[Aw_{i-1} - w_{i-1}(w'_{i-1}Aw_{i-1}) - w_{i-2}(w'_{i-2}Aw_{i-1})],\ i \geq 2$$

avec N un facteur de normalisation.
On retrouve PLS1 en posant : $A = X'X$, $w_0 = 0$, et $w_1 = X'y / \|X'y\|$.

1.5.2 Valeurs et vecteurs de Ritz

Soient $\theta_i^{(m)}$ et $s_i^{(m)}$, $i = 1, m$ les valeurs propres et les vecteurs propres normés de $W_m'AW_m$

$\theta_i^{(m)}$ est appelé valeur de Ritz, et $\varphi_i^{(m)} = W_m s_i^{(m)}$ est appelé vecteurs de Ritz.

Le couple $(\theta_i^{(m)}, \varphi_i^{(m)})$ représente une approximation de (λ_i, γ_i) à partir des valeurs et vecteurs propres de $W_m'AW_m$. Il représente la meilleur approximation du $i^{\text{ème}}$ couple (λ_i, γ_i) de valeur et vecteur propre de A que l'on peut dériver à partir de $K(b, A, m)$. Dans le cas de PLS1 ($A=X'X$), ils peuvent s'interpréter comme la restriction des valeurs propres et des vecteurs propres de $X'X$ à l'espace de Krylov dans lequel PLS se projette pour déterminer les coefficients de régression.

On a : $$\hat{\beta}_m^{PLS} = \sum_{i=1}^{m} (\theta_i^{(m)})^{-1} \varphi_i^{(m)} \varphi_i^{(m)\prime} X'y$$

$$\hat{\beta}_m^{PCR} = \sum_{i=1}^{m} \lambda_i^{-1} \gamma_i \gamma_i' X' y$$

$$\hat{\beta}^{OLS} = \sum_{i=1}^{p} \lambda_i^{-1} \gamma_i \gamma_i' X' y$$

1.5.3 PLS1 et l'algorithme Bidiag2

On peut aussi voir PLS comme une façon particulière de décomposer (factoriser) la matrice X. La factorisation la plus simple consiste en une décomposition en valeurs singulières (SVD) de X : $X = U\Sigma V'$

a*vec* U et V matrice orthogonales, i.e. $U'U=V'V=I$ et Σ une matrice diagonale.

Pratiquement, la SVD consiste en une bidiagonalisation de la matrice X à l'aide d'une transformation de Householder, suivi d'une factorisation QR [58,59].

Cette décomposition est associée à la régression sur composantes principales, (PCR), qui revient à considérer un sous-ensemble des vecteurs propres formant U.

$$\hat{\beta}^{PCR} = X^{-} y = V\Sigma^{-1} U' y$$

Mais il existe d'autres façons de factoriser X. On peut aussi écrire :

$$X = URW'$$

avec U et W des matrices orthogonales et R une matrice non singulière.

Soit $\hat{y} = X\hat{\beta}$, on déduit :

$$\hat{\beta} = X^{-} y = WR^{-1} U' y$$

Le problème consiste à inverser $R = U'XW$. La solution de l'algorithme PLS-NIPALS revient à trouver une matrice R bidiagonale (à droite) facile à inverser. Les matrices W et U sont déterminées par un algorithme développé par Golub and Kahan en 1965 [58] et appelé bidiag2 par Page et Saunders en 1982 [106].

Algorithme Bidiag2

$u_0 = 0$

$w_1 = X'y / \|X'y\|$

$u_i = k_i[Xw_i - u_{i-1}(u'_{i-1}Xw_i)],\ i \geq 1$

$w_{i+1} = k'_i[X'u_i - w_i(w'_iX'u_i)],\ i \geq 1$

avec k_i *et* k'_i des constantes de normalisation.

On retrouve les matrices W et T de l'algorithme PLS-NIPALS si l'on considère des composantes PLS normées (T=U).

Important :

Il est intéressant de noter que X n'est pas explicitement déflatée, ce qui n'est pas le cas avec l'algorithme PLS-NIPALS. Cela peut être important si on veut conserver la structure de la matrice pour la détermination des composantes d'ordre supérieur particulièrement dans le cas de matrices larges avec de nombreux 0.

De plus, on a la relation $R'R = Q$, avec Q obtenue par tridiagonalisation de $X'X$ avec l'algorithme de Lanczos. Il existe donc une relation directe entre l'algorithme bidiag2 et la méthode de Lanczos.

1.5.4 PLS et décomposition bilinéaire

PLS revient à faire une décomposition bilinéaire de la matrice X des descripteurs de la forme $X = TP'$, avec $T = (t_1, ..., t_a)$ et $P = (p'_1, ..., p'_a)$. Pour avoir unicité de la décomposition on doit orthogonaliser une des deux matrices T ou P. On a vu que la régression PLS utilisait une décomposition de la forme $X = URW'$ avec R une matrice non singulière, alors que la régression sur composante principale (PCR) utilisait une décomposition en valeurs propres de X de la forme $X = U\Sigma V'$ avec Σ diagonale. L'algorithme PLS-NIPALS fait le choix de prendre les composantes t_i orthogonales, en utilisant une décomposition bilinéaire de la forme $X = U(RW')$ avec $T = U$ alors que la régression PLS non orthogonale de Martens a choisi de prendre les *vecteurs* p_i orthogonaux en utilisant la décomposition $X = (UR)W'$ avec $P = W$. Les deux algorithmes conduisent aux mêmes prédictions, les composantes PLS issues des deux méthodes engendrant les mêmes espaces [139].

1.5.5 PLS et gradient conjugué

La méthode du gradient conjugué consiste à résoudre le système d'équations linéaires $Ax = b$ en minimisant la fonction quadratique $\Phi(x) = \frac{1}{2} x'Ax - b'x$ dont le gradient est égal à $Ax - b$.

Etant donné une valeur initiale x_0, l'algorithme des GC converge vers la solution : $x = A^{-}b$ en p étapes. Il minimise successivement $\Phi(x)$ le long de p directions d_j, j=0,p-1 qui sont conjuguées par rapport à A (A orthogonale), i.e. $d_i'Ad_j =< d_i, d_j >_A = 0$.

Si on considère l'équation normale $X'X\beta = X'y$, alors les estimations β_j successives du GC, avec $A = X'X$ et $b = X'y$ en partant de $\beta_0 = 0$, sont identiques aux estimations $\hat{\beta}_j^{PLS}$ (modèle PLS à j composantes) [111,130]. Les minimisations successives de $\Phi(x)$ prennent place dans une suite de sous-espaces de Krylov de dimensions croissantes.

Algorithme du Gradient Conjugué (Hestenes et Stiefel, 1952)

$x_0 = 0,\ d_0 = r_0 = b - Ax_0 = b$

Pour *i=1* à *p-1*

$$a_i = \frac{d_i'r_i}{d_i'Ad_i}$$

$$x_{i+1} = x_i + a_i d_i$$

$$r_{i+1} = b - Ax_{i+1}$$

$$b_i = -\frac{r_{i+1}'Ad_i}{d_i'Ad_i}$$

$$d_{i+1} = r_{i+1} + b_i d_i$$

Fin i

On montre :

$$span(d_0, d_1, ..., d_{j-1}) = span(r_0, r_1, ..., r_{j-1}) = span(b, Ab, ..., A^{j-1}b) = K(b, A, j)$$

$$x_i = \sum_{j=0}^{i-1} d_j (d_j' A d_j)^{-1} d_j' b \;\Rightarrow\; \hat{\beta}_{PLS}^m = K_m (K_m' A K_m)^{-1} K_m' b = x_m$$

1.6 Résumé du chapitre

La régression PLS est une méthode qui possède de multiples facettes. Dans ce chapitre on s'est intéressé à en présenter quelques unes et en particulier les liens étroits entre la régression PLS et les méthodes de Lanczos développées dans le domaine du calcul numérique. Ce sont les espaces de Krylov qui forment le lien entre les différentes approches. L'objectif de ce chapitre n'est pas dans la comparaison des performances relatives des différentes approches mais réside plus dans la présentation d'aspects algorithmiques de la régression PLS qui à travers différant éclairages sur cette méthode permettent de mieux la comprendre avant d'aborder sa généralisation.

Chapitre 2

La régression linéaire généralisée PLS

Nous présentons ici les idées principales de l'article « PLS generalised linear regression » [16] reproduit à la fin du chapitre.

La régression PLS univariée est un modèle reliant une variable dépendante y à un ensemble de variables indépendantes quantitatives ou qualitatives x_1, …, x_p. Tenenhaus (1999) [138] a montré qu'elle pouvait être obtenue par une suite de régressions simples et multiples dans une approche similaire à celle de Garthwaite [56] mais qui permet aussi de prendre en compte les valeurs manquantes dans l'esprit de l'algorithme PLS-NIPALS. Une propriété remarquable de cette approche est, grâce à l'utilisation des tests statistiques associés à la régression linéaire, la possibilité de sélectionner les variables explicatives significatives à conserver dans la construction des composantes PLS et permettre ainsi de déterminer de façon naturelle le nombre de composantes PLS à retenir dans le modèle de prédiction.

2.1 Composantes PLS et moindres carrés ordinaires

Cette approche propose une interprétation des composantes PLS en fonction de simples régressions linéaires.

Calcul de la première composante PLS t_1

La première composante $t_1 = Xw_1$ est définie par :

$$t_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y, x_j)^2}} \sum_{j=1}^{p} \mathrm{cov}(y, x_j) x_j$$

La quantité cov(y,x_j) est aussi le coefficient de régression a_{1j} dans la régression linéaire simple reliant la variable y à la variable x_j/var(x_j) :

$$y = a_{1j}(x_j/\mathrm{var}(x_j)) + \varepsilon$$

En effet :

$$a_{1j} = \frac{\mathrm{cov}(y, \frac{x_j}{\mathrm{var}(x_j)})}{\mathrm{var}(\frac{x_j}{\mathrm{var}(x_j)})} = \mathrm{cov}(y, x_j)$$

Calcul de la deuxième composante PLS t_2

On construit les régressions simples de y et des x_j sur t_1 :

$$y = c_1 t_1 + y_1$$

et pour chaque $j = 1$ à p

$$x_j = p_{1j} t_1 + x_{1j}$$

La deuxième composante t_2 est définie par :

$$t_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j})^2}} \sum_{j=1}^{p} \mathrm{cov}(y_1, x_{1j}) x_{1j}$$

La quantité cov(y_1, x_{1j}) est aussi le coefficient de régression a_{2j} dans la régression multiple reliant la variable y aux variables t_1 et x_{1j}/var(x_{1j}) :

$$y = c_1 t_1 + a_{2j}(x_{1j}/\mathrm{var}(x_{1j})) + \varepsilon$$

Ce résultat provient de l'orthogonalité entre le résidu x_{1j} et la composante t_1.

Remarque : Il n'est pas nécessaire de déflater *y,* et on pourrait aussi utiliser le modèle :

$$y = a_{2j}(x_{1j}/\text{var}(x_{1j})) + \varepsilon$$

Il est cependant préférable s'il y a des valeurs manquantes de prendre en compte aussi t_1 car dans ce cas t_1 peut ne plus être orthogonal à x_{1j}.

La composante t_2 peut aussi s'exprimer en fonction des variables d'origine x_j puisque les résidus $x_{1j} = x_j - p_{1j}t_1$ sont fonctions des x_j. La composante t_2 exprimée en fonction des x_j s'écrit:

$$t_2 = Xw_2^* = X(I - w_1 p_1')w_2$$

Calcul des autres composantes

On procède de la même manière pour le calcul des autres composantes :

$$t_h = Xw_h^* = X\left(\prod_{i=1}^{h-1}(I - w_i p_i')\right)w_h$$

Formule de régression PLS :

$$\hat{y} = \sum_{h=1}^{m} c_h(\sum_{j=1}^{p} w_{hj}^* x_j)$$

Les coefficients c_h sont estimés par régression multiple de y sur les composantes PLS t_h. L'équation de régression estimée peut ensuite s'exprimer en fonction des variables d'origine x_j :

$$\hat{y} = \sum_{h=1}^{m} c_h (\sum_{j=1}^{p} w_{hj}^{*} x_j)$$
$$= \sum_{j=1}^{p} (\sum_{h=1}^{m} c_h w_{hj}^{*}) x_j$$
$$= \sum_{j=1}^{p} b_j x_j$$

2.2 Sélection des variables et détermination du nombre de composantes PLS.

Dans la construction de la composante PLS t_h, un test sur le coefficient de régression a_{hj} permet d'évaluer l'importance de la variable x_{hj} dans la construction de t_h. Dans la recherche d'un modèle parcimonieux, on pourrait décider de remplacer par 0 les coefficients non significatifs au seuil α. De plus, l'utilisation de α comme un méta paramètre permet de faire de la sélection de variables dans l'esprit de l'algorithme « Threshold Gradient Descent » proposé par Friedman et Popescu [54]. Cette approche que l'on peut intituler « Thresholding PLS regression » sera développée au paragraphe 6.

Une conséquence immédiate de cette approche concerne le nombre de composantes PLS à retenir dans le modèle de prédiction. On s'arrête naturellement à la composante t_h lorsque tous les coefficients $a_{h+1,j}$ dans la régression multiple reliant la variable y aux variables t_1 ,… , t_h et $x_{hj}/\mathrm{var}(x_{hj})$ pour la construction de la composante t_{h+1} sont non significatifs au seuil α. De façon pratique, les vecteurs $(t_1,...,t_h,x_{hj})$ et $(t_1,...,t_h,x_j)$ engendrant le même espace, il est équivalent de tester la significativité du coefficient de x_j dans la régression de y sur $(t_1,...,t_h,x_j)$.

2.3 Régression linéaire généralisée PLS (PLS-GLR)

La régression PLS peut être obtenue en utilisant des régressions simples et multiples. En remplaçant ces régressions par des régressions linéaires généralisées on a accès a une nouvelle gamme de modèles : la régression linéaire généralisée PLS.

La régression linéaire généralisée PLS de y sur $x_1,\ldots, x_p$ à m composantes s'écrit :

$$g(\theta) = \sum_{h=1}^{m} c_h (\sum_{j=1}^{p} w_{hj}^{*} x_j)$$

où le paramètre θ peut représenter la moyenne μ d'une variable continue y, ou le vecteur des probabilités des valeurs prises par une variable y discrète, ou encore la fonction de risque $h(t)$ en données de survie. La fonction de lien g est choisie par l'utilisateur en tenant compte de la loi de probabilité de y et de la qualité de l'ajustement du modèle aux données. On impose aux composantes PLS $t_h = \sum_{j=1}^{p} w_{hj}^{*} x_j$ d'être orthogonales. Les coefficients c_h sont estimés par régression linéaire généralisée de y sur les composantes PLS t_h.

Marx proposa dès 1996 [98] d'étendre la régression PLS aux modèles linéaires généralisés. Utilisant le fait que dans la famille exponentielle les estimateurs du maximum de vraisemblance peuvent s'obtenir via la méthode de Newton-Raphson dans sa variante « Fisher scoring » par une méthode de type moindre carrés pondérés itérés, il remplaça les moindres carrés par des régressions PLS (IRPLS). L'algorithme PLS-GLR est à la fois plus général dans son utilisation, car applicable à tout modèle de régression généralisé comme par exemple le modèle de Cox, mais il est aussi beaucoup plus simple à mettre en œuvre.

2.3.1 Algorithme de régression linéaire généralisée PLS

L'algorithme de régression linéaire généralisée PLS comprend trois étapes :

1) La recherche des composantes PLS t_h.
2) La régression linéaire généralisée de y sur les composantes PLS t_h.
3) L'expression de la régression linéaire généralisée en fonction des variables d'origine.
4) La validation bootstrap des coefficients de l'équation de régression linéaire généralisée finale.

Nous allons décrire en détail la première étape.

On note X la matrice dont les colonnes sont formées des valeurs des p variables explicatives x_j. On recherche successivement m composantes PLS orthogonales t_h combinaisons linéaires des x_j.

Recherche de la première composante PLS t_1

Etape 1 : Calculer le coefficient de régression a_{1j} de x_j dans la régression linéaire généralisée de y sur x_j pour chaque variable x_j, $j = 1$ à p.
Etape 2 : Normer le vecteur colonne a_1 formé des a_{1j} : $w_1 = a_1/||a_1||$
Etape 3 : Calculer la composante $t_1 = Xw_1/w_1'w_1$

Recherche de la deuxième composante PLS t_2

Etape 1 : Calculer le coefficient de régression a_{2j} de x_j dans la régression linéaire généralisée de y sur t_1 et x_j.
Etape 2 : Normer le vecteur colonne a_2 formé des a_{2j} : $w_2 = a_2/||a_2||$
Etape 3 : Calculer le résidu X_1 de la régression linéaire de X sur t_1 : $X_1 = X - t_1 p_1'$
Etape 4 : Calculer la composante $t_2 = X_1w_2/w_2'w_2$.
Etape 5 : Exprimer la composante t_2 en fonction de X : $t_2 = Xw_2^* = X(I - w_1 p_1')w_2$

Recherche de la h-ième composante PLS t_h

On a obtenu aux étapes précédentes les composantes PLS $t_1,\ldots, t_{h-1}$. On obtient la composante t_h en prolongeant la recherche de la deuxième composante.

Etape 1 : Calculer le coefficient de régression a_{hj} de x_j dans la régression linéaire généralisée de y sur $t_1, t_2, \ldots, t_{h-1}$ et x_j.
Etape 2 : Normer le vecteur colonne a_h formé des a_{hj} : $w_h = a_h/||a_h||$
Etape 3 : Calculer le résidu X_{h-1} de la régression linéaire de X sur $t_1,\ldots,t_{h-1}$: $X_{h-1} = X - \sum_{i=1}^{h-1} t_i p_i'$
Etape 4 : Calculer la composante $t_h = X_{h-1}w_h/w_h'w_h$.
Etape 5 : Exprimer la composante t_h en fonction de X :

$$t_h = Xw_h^* = X\left(\prod_{i=1}^{h-1}(I - w_i p_i')\right)w_h$$

Commentaires :

1) On peut simplifier le calcul des composantes PLS t_h en mettant à 0 les coefficients de régression a_{hj} non significatifs au seuil α. Seules les variables associées à des coefficients significatifs contribuent à la construction de la composante PLS. L'utilisation du seuil de signification comme un méta-paramètre sera utilisée au chapitre 6 pour faire de la sélection de variables.

2) Le nombre m de composantes PLS à retenir est fixé en constatant que la composante t_{m+1} n'est pas significative dans le sens où aucun $a_{m+1,j}$ ne l'est.

3) L'algorithme présenté peut fonctionner en présence de données manquantes. Notons $x_{h-1,i}$ la i-ième ligne de la matrice X_{h-1}. La valeur $t_{hi} = \dfrac{x_{h-1,i}w_h}{w_h' w_h}$ .de la composante t_h pour l'individu i représente la pente de la droite des moindres carrés sans constante du nuage de points (w_h, $x_{h-1,i}$). Cette pente est aussi calculable lorsqu'il y a des données manquantes. Ainsi dans les étapes de calcul des composantes PLS, le calcul du dénominateur n'est réalisé que sur les données disponibles au numérateur.

Cependant cette approche s'apparente d'une certaine façon à une imputation simple et ne restitue pas l'incertitude associée aux valeurs manquantes sous-estimant ainsi la variance de l'estimation des coefficients. Nous proposerons plus loin différentes alternatives à NIPALS [14,99] en introduisant le concept d'imputation multiple en régression PLS simple et généralisée.

4) Les intervalles de confiance des coefficients dans les applications ont été estimés par rééchantillonage bootstrap en utilisant la méthode des quantiles simples. La méthode d'échantillonnage utilisée est le balanced bootstrap qui a été introduit par Davison et al. en 1986 [33]. Elle permet de diminuer l'erreur de simulation inhérente au rééchantillonnage. Même si celle-ci n'améliore que légèrement le bootstrap uniforme pour ce qui concerne les estimateurs bootstrap de la distribution des quantiles, elle a été préférée dans les applications car sa mise en œuvre est très simple. Pour créer B échantillons bootstrap on empile B fois la matrice (y, x_1,…, x_p) des données, puis on permute aléatoirement suivant les lignes. En découpant la matrice résultante suivant les lignes en B parties égales, on obtient B

5) échantillons bootstrap équilibrés. En effet toutes les lignes ont la même fréquence d'apparition sur l'ensemble des *B* échantillons. Il existe d'autres types de rééchantillonnage bootstrap comme l'« *Importance Bootstrap Resampling* » qui améliore significativement les approximations des estimateurs bootstrap de la distribution des quantiles. Cependant leur mise en œuvre est plus complexe et n'a pas été implémentée ici.

La méthode des quantiles simples utilisée pour estimer les intervalles de confiance bootstrap des coefficients peut être améliorée en particulier en corrigeant pour le biais. La méthode des quantiles d'Efron et Tibshirani (1993) [44] avec correction pour le biais et accélération est décrite et implémentée dans l'article « *PLS generalised linear regression* » reproduit ci-dessous.

2.4 Article CSDA 2005

Available online at www.sciencedirect.com
SCIENCE DIRECT®
ELSEVIER
Computational Statistics & Data Analysis 48 (2005) 17–46
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

PLS generalised linear regression

Philippe Bastien[a], Vincenzo Esposito Vinzi[b,c,*], Michel Tenenhaus[c]

[a] *L'Oréal Recherche, Aulnay, France*
[b] *Department of Mathematics and Statistics University "Federico II", Naples, Italy*
[c] *HEC, School of Management Jouy-en-Josas, France*

Accepted 6 February 2004

Abstract

PLS univariate regression is a model linking a dependent variable $\mathbf{y}$ to a set $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_p\}$ of (numerical or categorical) explanatory variables. It can be obtained as a series of simple and multiple regressions. By taking advantage from the statistical tests associated with linear regression, it is feasible to select the significant explanatory variables to include in PLS regression and to choose the number of PLS components to retain. The principle of the presented algorithm may be similarly used in order to yield an extension of PLS regression to PLS generalised linear regression. The modifications to classical PLS regression, the case of PLS logistic regression and the application of PLS generalised linear regression to survival data are studied in detail. Some examples show the use of the proposed methods in real practice. As a matter of fact, classical PLS univariate regression is the result of an iterated use of ordinary least squares (OLS) where PLS stands for *partial least squares*. PLS generalised linear regression retains the rationale of PLS while the criterion optimised at each step is based on maximum likelihood. Nevertheless, the acronym PLS is kept as a reference to a general methodology for relating a response variable to a set of predictors. The approach proposed for PLS generalised linear regression is simple and easy to implement. Moreover, it can be easily generalised to any model that is linear at the level of the explanatory variables.
© 2004 Elsevier B.V. All rights reserved.

Keywords: Partial least squares; Stepwise regression; Variable selection; Modified PLS regression

* Corresponding author. Dipartimento di Matematica e Statistica, Università degli Sutdi di Napoli "Federico II", Via Cintia—Complesso Monte Sant'Angelo, 80126 Napoli, Italy. Tel.: +39-081-675112; fax: +39-081-675113.
E-mail address: vincenzo.espositovinzi@unina.it (V.E. Vinzi).

0167-9473/$ - see front matter © 2004 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2004.02.005

1. PLS regression background

All variables $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_j, \ldots, \mathbf{x}_p$ are assumed to be centred.

The PLS regression model with m components is written as

$$\mathbf{y} = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right) + \text{residual}, \tag{1}$$

with the constraint that the PLS components $\mathbf{t}_h = \sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j$ are orthogonal. We can consider that the parameters c_h and w_{hj}^{*} in model (1) are to be estimated. This is the nonlinear aspect of the model.

In the following, the same notation is used for the model parameters and their estimates. The context will clarify the actual meaning of the notation.

PLS regression (Wold et al., 1983; Tenenhaus, 1998; Garthwaite, 1994) is an algorithm for estimating the parameters of model (1). In the following, this algorithm is presented in a new version by linking each step to a simple or multiple OLS regression.

Computation of the first PLS component $\mathbf{t}_1$. The first component $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1^{*}$ is defined as

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j)^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j) \mathbf{x}_j. \tag{2}$$

The weight $\operatorname{cov}(\mathbf{y}, \mathbf{x}_j)$ for the variable $\mathbf{x}_j$ may be also written as $\operatorname{cor}(\mathbf{y}, \mathbf{x}_j)^{*} s(y)^{*} s(\mathbf{x}_j)$ where $s(y)$ and $s(\mathbf{x}_j)$ are, respectively, the standard deviation of y and $\mathbf{x}_j$. As a consequence, in order for a variable $\mathbf{x}_j$ to be important in building up $\mathbf{t}_1$, it needs to be strongly correlated with $\mathbf{y}$ and to bear enough variability in terms of standard deviation.

The quantity $\operatorname{cov}(\mathbf{y}, \mathbf{x}_j)$ is also the regression coefficient a_{1j} in OLS simple regression between $\mathbf{y}$ and the modified explanatory variable $\mathbf{x}_j / \operatorname{var}(\mathbf{x}_j)$:

$$\mathbf{y} = a_{1j} (\mathbf{x}_j / \operatorname{var}(\mathbf{x}_j)) + \text{residual}. \tag{3}$$

Actually,

$$a_{1j} = \frac{\operatorname{cov}\left(\mathbf{y}, \dfrac{\mathbf{x}_j}{\operatorname{var}(\mathbf{x}_j)} \right)}{\operatorname{var}\left(\dfrac{\mathbf{x}_j}{\operatorname{var}(\mathbf{x}_j)} \right)} = \operatorname{cov}(\mathbf{y}, \mathbf{x}_j).$$

Thus, a test on the regression coefficient a_{1j} allows to assess the importance of the variable $\mathbf{x}_j$ in building $\mathbf{t}_1$. On this basis, the simple regression of $\mathbf{y}$ on $\mathbf{x}_j$ may be studied:

$$\mathbf{y} = a'_{1j} \mathbf{x}_j + \text{residual}. \tag{4}$$

As a matter of fact, there is an equivalence between testing whether a_{1j} or a'_{1j} are different from 0. In (2), each nonsignificant covariance may be replaced by a 0 so as to disregard the related explanatory variable.

Computation of the second PLS component $\mathbf{t}_2$. At first, the $p+1$ simple regressions of, respectively, $\mathbf{y}$ and each $\mathbf{x}_j$ on $\mathbf{t}_1$ are run:

$$\mathbf{y} = c_1\mathbf{t}_1 + \mathbf{y}_1, \tag{5}$$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}. \tag{6}$$

The second PLS component $\mathbf{t}_2$ is defined as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})\mathbf{x}_{1j}. \tag{7}$$

The quantity $\operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j})$ is also the regression coefficient a_{2j} in the OLS multiple regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_{1j}/\operatorname{var}(\mathbf{x}_{1j})$

$$\mathbf{y} = c_{1j}\mathbf{t}_1 + a_{2j}(\mathbf{x}_{1j}/\operatorname{var}(\mathbf{x}_{1j})) + \text{residual}. \tag{8}$$

This result is a consequence of the orthogonality between the residual $\mathbf{x}_{1j}$ and the component $\mathbf{t}_1$.

The partial correlation between $\mathbf{y}$ and $\mathbf{x}_j$ given $\mathbf{t}_1$ is defined as the correlation between the residuals $\mathbf{y}_1$ and $\mathbf{x}_{1j}$. Similarly, the partial covariance between $\mathbf{y}$ and $\mathbf{x}_j$ given $\mathbf{t}_1$ is defined as the covariance between the residuals $\mathbf{y}_1$ and $\mathbf{x}_{1j}$:

$$\operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1) = \operatorname{cov}(\mathbf{y}_1, \mathbf{x}_{1j}). \tag{9}$$

Therefore, the second PLS component may be alternatively written as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1)^2}} \sum_{j=1}^{p} \operatorname{cov}(\mathbf{y}, \mathbf{x}_j \mid \mathbf{t}_1)\mathbf{x}_{1j}. \tag{10}$$

A test on the regression coefficient a_{2j} allows to assess the importance of $\mathbf{x}_{1j}$ in building $\mathbf{t}_2$. It is also possible to test the contribution of $\mathbf{x}_j$ to the building of the second PLS component by studying directly the regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_j$:

$$\mathbf{y} = c'_{1j}\mathbf{t}_1 + a'_{2j}\mathbf{x}_j + \text{residual}. \tag{11}$$

As a matter of fact, there is an equivalence between testing a_{2j} or a'_{2j} being different from 0 because the pairs of vectors $(\mathbf{t}_1, \mathbf{x}_{1j})$ and $(\mathbf{t}_1, \mathbf{x}_j)$ span the same space.

In (7), each nonsignificant covariance may be replaced by a 0 so as to disregard the related explanatory variable from contributing to the model.

For interpretation purposes, the component $\mathbf{t}_2$ is better expressed as a function of the raw variables $\mathbf{x}_j$. This is possible because the residuals $\mathbf{x}_{1j} = \mathbf{x}_j - p_{1j}\mathbf{t}_1$ are a function of $\mathbf{x}_j$.

When expressed in terms of $\mathbf{x}_j$, the component $\mathbf{t}_2$ is written as $\mathbf{t}_2 = \mathbf{X}\mathbf{w}_2^*$. From the operational point of view, this is achieved by replacing the residual $\mathbf{x}_{1j}$ in (10) by its expression $\mathbf{x}_j - p_{1j}\mathbf{t}_1$.

Computation of the next components and stopping rule. We follow the same procedure for computing the next components $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^*$. The search of new components is stopped either in accordance with a cross-validation procedure (as implemented in SIMCA-P 10.5 Umetri, 2004) or when all partial covariances are not significant.

PLS regression formula. In (1) the coefficients c_h are estimated by multiple regression of $\mathbf{y}$ on the PLS components $\mathbf{t}_h$. The estimated regression equation may be then expressed in terms of the original variables $\mathbf{x}_j$'s:

$$\begin{aligned}\hat{\mathbf{y}} &= \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right), \\ &= \sum_{j=1}^{p} \left(\sum_{h=1}^{m} c_h w_{hj}^{*} \right) \mathbf{x}_j, \\ &= \sum_{j=1}^{p} b_j \mathbf{x}_j. \end{aligned} \tag{12}$$

2. PLS generalised linear regression (PLS-GLR)

It has been shown that PLS regression can be obtained by using simple and multiple regression. If these regressions are replaced by GLR, a new series of models is accessed: PLS-GLR.

PLS-GLR of $\mathbf{y}$ on $\mathbf{x}_1, \ldots, \mathbf{x}_p$ with m components is written as:

$$g(\theta) = \sum_{h=1}^{m} c_h \left(\sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j \right), \tag{13}$$

where the parameter θ may be either the mean of a continuous variable $\mathbf{y}$, or the probability vector of the values taken by a discrete variable $\mathbf{y}$, or the hazard ratio in survival data for proportional hazard ratio models. The link function g is chosen by the user according to the probability distribution of $\mathbf{y}$ and the model goodness of fit to the data. The new PLS components $\mathbf{t}_h = \sum_{j=1}^{p} w_{hj}^{*} \mathbf{x}_j$ are constrained to be orthogonal. In the following, for the sake of simplicity, the same notation as before is kept for all elements in the model related to PLS-GLR. This new approach is firstly described and then applied to different examples.

2.1. PLS-GLR algorithm

The algorithm consists of four steps:

(1) computation of the m PLS components $\mathbf{t}_h$ $(h = 1, \ldots, m)$;
(2) generalised linear regression of $\mathbf{y}$ on the m retained PLS components;
(3) expression of PLS-GLR in terms of the original explanatory variables;
(4) Bootstrap validation of coefficients in the final model of PLS-GLR.

Hereinafter, the first and the last steps are described in details while the middle two steps are shown directly in the examples as they are trivial.

Computation of the PLS components. Let $\mathbf{X}=\{\mathbf{x}_1,\ldots,\mathbf{x}_p\}$ be the matrix comprising the previously centred p explanatory variables $\mathbf{x}_j$'s. The objective is to search for m PLS orthogonal components $\mathbf{t}_h$'s defined as linear combinations of $\mathbf{x}_j$.

Computation of the first PLS component $\mathbf{t}_1$

Step 1: Compute the regression coefficient a_{1j} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_1$ made by a_{1j}'s: $\mathbf{w}_1=\mathbf{a}_1/\|\mathbf{a}_1\|$,

Step 3: Compute the component $\mathbf{t}_1=\mathbf{X}\mathbf{w}_1/\mathbf{w}_1'\mathbf{w}_1$.

Computation of the second PLS component $\mathbf{t}_2$

Step 1: Compute the regression coefficient a_{2j} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_2$ made by a_{2j}'s: $\mathbf{w}_2=\mathbf{a}_2/\|\mathbf{a}_2\|$,

Step 3: Compute the residual matrix $\mathbf{X}_1$ of the linear regression of $\mathbf{X}$ on $\mathbf{t}_1$,

Step 4: Compute the component $\mathbf{t}_2=\mathbf{X}_1\mathbf{w}_2/\mathbf{w}_2'\mathbf{w}_2$,

Step 5: Express the component $\mathbf{t}_2$ in terms of $\mathbf{X}$: $\mathbf{t}_2=\mathbf{X}\mathbf{w}_2^*$.

Computation of the hth PLS component $\mathbf{t}_h$. In the previous steps, the PLS components $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$ have been yielded. The component $\mathbf{t}_h$ is obtained by iterating the search for the second component.

Step 1: Compute the regression coefficient a_{hj} of $\mathbf{x}_j$ in the generalised linear regression of $\mathbf{y}$ on $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$ and $\mathbf{x}_j$ for each variable $\mathbf{x}_j$, $j=1$ to p,

Step 2: Normalise the column vector $\mathbf{a}_h$ made by a_{hj}'s: $\mathbf{w}_h=\mathbf{a}_h/\|\mathbf{a}_h\|$,

Step 3: Compute the residual matrix $\mathbf{X}_{h-1}$ of the linear regression of $\mathbf{X}$ on $\mathbf{t}_1,\ldots,\mathbf{t}_{h-1}$,

Step 4: Compute the component $\mathbf{t}_h=\mathbf{X}_{h-1}\mathbf{w}_h/\mathbf{w}_h'\mathbf{w}_h$,

Step 5: Express the component $\mathbf{t}_h$ in terms of $\mathbf{X}$: $\mathbf{t}_h=\mathbf{X}\mathbf{w}_h^*$.

Remarks

(1) Computation of the PLS component $\mathbf{t}_h$ can be simplified by setting to 0 those regression coefficients a_{hj} that show to be not significant. Only significant variables will then contribute to the computation of the PLS component.

(2) The number m of PLS components to be retained may be chosen by cross-validation on the predictive power of the model or by observing that the component $\mathbf{t}_{m+1}$ is not significant because none of the coefficients $a_{m+1,j}$ is significantly different from 0.

(3) The proposed algorithm may cope with missing data. Let $\mathbf{x}_{h-1,i}$ be the column vector obtained by transposing the ith row of $\mathbf{X}_{h-1}$. The value $\mathbf{t}_{hi}=\mathbf{x}_{h-1,i}'\mathbf{w}_h/\mathbf{w}_h'\mathbf{w}_h$ of the ith case on the component $\mathbf{t}_h$ represents the slope of the OLS line without constant term related to the cloud of points $(\mathbf{w}_h,\ \mathbf{x}_{h-1,i})$. This slope may be computed even when there are missing data. In such a case, in computing the hth PLS component, the denominator of Step 4 is computed only on the data available also for the numerator.

Bootstrap validation procedure. Once PLS model has been estimated and the proper number of components to retain chosen, it is important to ascertain statistical significance of explanatory variables. Hereinafter, this is done in a nonparametric

framework by means of a bootstrap procedure under the assumption that PLS components $\mathbf{t}_h$'s ($h=1,\ldots,m$) are given as they represent the most stable part in the model. Most commonly used software for PLS regression (SIMCA-P (Umetri, 2004) and The Unscrambler® (Camo, 2003)) provide with jackknife estimates of PLS regression coefficients. However, the algorithms behind the construction of such estimates are not clearly given. For this reason, the development of an ad hoc nonparametric procedure is preferred in the following with related steps clearly stated.

Let $\hat{F}_{(\mathbf{T}|\mathbf{y})}$ be the empirical distribution function given by the PLS components matrix $\mathbf{T}=\{\mathbf{t}_1,\ldots,\mathbf{t}_h,\ldots,\mathbf{t}_m\}$ and the response variable $\mathbf{y}$.

Step 1: Draw B random samples from $\hat{F}_{(\mathbf{T}|\mathbf{y})}$;

Step 2: For $b=1,\ldots,B$, compute:

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

where $[\mathbf{T}^{(b)},\mathbf{y}^{(b)}]$ is the bth bootstrap sample, $\mathbf{c}^{(b)}$ is the vector of the corresponding estimates of PLS coefficients linking components to $\mathbf{y}$, $\mathbf{b}^{(b)}$ is the vector of p regression coefficients linking original explanatory variables to $\mathbf{y}$ (see Eq. (12)) for the bth sample, and $\mathbf{W}^{*}$ is the unchanged matrix of the original explanatory variables weights in constructing the m PLS components,

Step 3: For each $j=1,\ldots,p$ Φ_{b_j} is the Monte Carlo approximation of the bootstrap cumulative distribution function (*cdf*) of b_j.

For each b_j, box-plots and confidence intervals can be derived by means of the percentiles of Φ_{b_j}. The confidence interval is defined as $I_j(\alpha)=[\Phi_{b_j}^{-1}(\alpha),\Phi_{b_j}^{-1}(1-\alpha)]$ where $\Phi_{b_j}^{-1}(\alpha)$ and $\Phi_{b_j}^{-1}(1-\alpha)$ are the values from the bootstrap *cdf* such that a nominal $100(1-2\alpha)\%$ confidence interval is yielded. In order to improve the coverage accuracy (i.e. the capability of $I_j(\alpha)$ to give accurate coverage probabilities) of the above intervals, the so-called BC_a (*Bias-Correction and accelerated*) intervals, proposed in Efron and Tibshirani (1993), are considered. A BC_a interval is defined as

$$I_j^{\mathrm{BC}_a}(\alpha)=[\Phi_{b_j}^{-1}(\alpha_{j,1}),\Phi_{b_j}^{-1}(1-\alpha_{j,2})],$$

where

$$\alpha_{j,1}=\Theta\left(\hat{z}_0+\frac{\hat{z}_0+z(\alpha)}{1-\hat{a}[\hat{z}_0+z(\alpha)]}\right) \quad \text{and}$$

$$\alpha_{j,2}=\Theta\left(\hat{z}_0+\frac{\hat{z}_0+z(1-\alpha)}{1-\hat{a}[\hat{z}_0+z(1-\alpha)]}\right).$$

Here, $\Theta(\cdot)$ is the standard normal *cdf* and $z(\alpha)$ is the 100αth percentile of a standard normal distribution. The bias correction factor $\hat{z}_0$ and the acceleration factor $\hat{a}$ can be obtained, respectively, as

$$\hat{z}_0=\Theta^{-1}(\Phi_{b_j}(b_j)),$$

where $\Theta^{-1}(\cdot)$ is the inverse function of the standard normal *cdf*, and

$$\hat{a}=\frac{\sum_{i=1}^{n}(b_{j(\cdot)}-b_{j(i)})^3}{6\left[\sum_{i=1}^{n}(b_{j(\cdot)}-b_{j(i)})^3\right]^{3/2}},$$

where $b_{j(\cdot)}$ is the jack-knife estimate of b_j and $b_{j(i)}$ is the estimate of b_j when the ith observation is dropped out of the sample. If both $\hat{z}_0$ and $\hat{a}$ are equal to zero, then $I_j(\alpha)$ and $I_j^{\mathrm{BC}_a}(\alpha)$ coincide.

One of the advantages of BC_a intervals concerns their accuracy, i.e. a $100(1-2\alpha)\%$ confidence interval is supposed to have an overall probability 2α of not covering. Moreover, it can be shown that intervals $I_j^{\mathrm{BC}_a}(\alpha)$'s converge toward unbiasedness with a higher speed (relative to the number B of bootstrap samples) than intervals $I_j(\alpha)$'s. As a matter of fact, they are second order accurate intervals while simple bootstrap intervals are only first order accurate. Roughly speaking, the performance improvement of BC_a intervals is due to the fact that simple bootstrap confidence intervals implicitly assume symmetry of the bootstrap distribution of the estimator while BC_a intervals automatically take into account the shape (in terms of skewness and kurtosis) and the bias of the distribution.

It is worth noticing that the above-mentioned confidence intervals are not simultaneous (Shao and Tu, 1995). Therefore, they are not meant for pairwise or multiple comparisons but can only be interpreted one at a time.

2.2. The case of multiple regression

The application of the general algorithm proposed above to the specific case of multiple regression yields a natural modification of PLS regression that was earlier proposed in (Shenk and Westerhaus, 1991) with the name of "modified PLS regression". In this case, response variable $\mathbf{y}$ is standardised while explanatory variables $\mathbf{x}_j$'s are simply centred.

Computation of the first PLS component $\mathbf{t}_1$. The first component $\mathbf{t}_1 = \mathbf{X}\mathbf{w}_1$ is defined as

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{1j}^2}} \sum_{j=1}^{p} a_{1j}\mathbf{x}_j, \tag{14}$$

where a_{1j} is the regression coefficient of the centred variable $\mathbf{x}_j$ in the regression of $\mathbf{y}$ on each $\mathbf{x}_j$. By replacing a_{1j} by its value $\mathrm{Cov}(\mathbf{x}_j,\mathbf{y})/\mathrm{Var}(\mathbf{x}_j)$ expression (14) becomes

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \left[\frac{\mathrm{Cov}(\mathbf{x}_j,\mathbf{y})}{\mathrm{Var}(\mathbf{x}_j)}\right]^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})\mathbf{x}_j^*, \tag{15}$$

where $\mathbf{x}_j^*$ shows that $\mathbf{x}_j$ is standardised. It is worth noticing that the explanatory variables are automatically reduced when changing from Eq. (14) to Eq. (15). It is certainly preferable to change the normalisation so as to obtain

$$\mathbf{t}_1 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_j,\mathbf{y})\mathbf{x}_j^*. \tag{16}$$

Only those explanatory variables that are significantly correlated with $\mathbf{y}$ contribute to the computation of $\mathbf{t}_1$. There is no difference at all between this new component and the usual PLS component computed on standardised data.

Computation of the second PLS component $\mathbf{t}_2$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the explanation of $\mathbf{y}$ in addition to $\mathbf{t}_1$ are looked for. Therefore, the multiple regressions of $\mathbf{y}$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$ are built

$$\mathbf{y} = c_1\mathbf{t}_1 + a_{2j}\mathbf{x}_j + \text{residual}. \tag{17}$$

In order to obtain a second component $\mathbf{t}_2$ that is orthogonal to $\mathbf{t}_1$, it is necessary to build it starting from the residuals $\mathbf{x}_{1j}$ of the regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}. \tag{18}$$

Thus, the second component $\mathbf{t}_2$ is defined as

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{2j}^2}} \sum_{j=1}^{p} a_{2j}\mathbf{x}_{1j}, \tag{19}$$

where a_{2j} is also the regression coefficient of $\mathbf{x}_{1j}$ in the regression of $\mathbf{y}$ on $\mathbf{t}_1$ and $\mathbf{x}_{1j}$

$$\begin{aligned}\mathbf{y} &= c_1\mathbf{t}_1 + a_{2j}(p_{1j}\mathbf{t}_1 + \mathbf{x}_{1j}) + \text{residual}\\ &= (c_1 + a_{2j}p_{1j})\mathbf{t}_1 + a_{2j}\mathbf{x}_{1j} + \text{residual}.\end{aligned} \tag{20}$$

The orthogonality between $\mathbf{x}_{1j}$ and $\mathbf{t}_1$, by modifying the normalisation, leads to

$$\mathbf{t}_2 = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{1j}, \mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{1j}, \mathbf{y})\mathbf{x}_{1j}^*, \tag{21}$$

where $\mathbf{x}_{1j}^*$ represents the standardised version of the residual vector $\mathbf{x}_{1j}$.

Computation of the hth PLS component $\mathbf{t}_h$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the explanation of $\mathbf{y}$ in addition to $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ are looked for. Therefore, the multiple regressions of $\mathbf{y}$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ and each $\mathbf{x}_j$ are built

$$\mathbf{y} = c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \cdots + c_{h-1}\mathbf{t}_{h-1} + a_{hj}\mathbf{x}_j + \text{residual}. \tag{22}$$

In order to obtain a hth component $\mathbf{t}_h$ that is orthogonal to $\mathbf{t}_{h-1}$, it is necessary to build it starting from the residuals $\mathbf{x}_{(h-1)j}$ of the regressions of each $\mathbf{x}_j$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$

$$\mathbf{x}_j = p_{1j}\mathbf{t}_1 + p_{2j}\mathbf{t}_2 + \cdots + p_{(h-1)j}\mathbf{t}_{h-1} + \mathbf{x}_{(h-1)j}. \tag{23}$$

The component $\mathbf{t}_h$ is then defined as

$$\mathbf{t}_h = \frac{1}{\sqrt{\sum_{j=1}^{p} a_{hj}^2}} \sum_{j=1}^{p} a_{hj}\mathbf{x}_{(h-1)j}, \tag{24}$$

where a_{hj} is also the regression coefficient of $\mathbf{x}_{(h-1)j}$ in the regression of $\mathbf{y}$ on $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$ and $\mathbf{x}_{(h-1)j}$

$$\begin{aligned}\mathbf{y} &= c_1\mathbf{t}_1 + c_2\mathbf{t}_2 + \cdots + c_{h-1}\mathbf{t}_{h-1} + a_{hj}(p_{1j}\mathbf{t}_1 + p_{2j}\mathbf{t}_2 + \cdots\\ &\quad + p_{(h-1)j}\mathbf{t}_{(h-1)} + \mathbf{x}_{(h-1)j}) + \text{residual}\\ &= (c_1 + a_{hj}p_{1j})\mathbf{t}_1 + (c_2 + a_{hj}p_{2j})\mathbf{t}_2 + \cdots + (c_{h-1} + a_{hj}p_{(h-1)j})\mathbf{t}_{h-1}\\ &\quad + a_{hj}\mathbf{x}_{(h-1)j} + \text{residual}.\end{aligned} \tag{25}$$

The orthogonality between $\mathbf{x}_{(h-1)j}$ and $\mathbf{t}_1, \mathbf{t}_2, \ldots, \mathbf{t}_{h-1}$, by modifying the normalisation, leads to

$$\mathbf{t}_h = \frac{1}{\sqrt{\sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{(h-1)j}, \mathbf{y})^2}} \sum_{j=1}^{p} \mathrm{Cor}(\mathbf{x}_{(h-1)j}, \mathbf{y}) \mathbf{x}^*_{(h-1)j}, \tag{26}$$

where $\mathbf{x}^*_{(h-1)j}$ represents the standardised version of the residual vector $\mathbf{x}_{(h-1)j}$.

Stopping rule. The search for new components stops when all regression coefficients are not significant.

A case study: Cornell data on gasoline mixture. The proposed algorithm is now applied to a multiple regression problem related to the Cornell data (Kettaneh-Wold, 1992; Tenenhaus, 1998). Only the significant variables are retained for computing the PLS components. The dataset consists of 12 gasoline samples whose mixture is known in terms of proportions (Table 1). he influence of the mixture on the octane ratio $\mathbf{y}$ needs to be determined. Correlations between variables are given in Table 2.

Variables:
$\mathbf{y}$ = motor octane rating
$\mathbf{x}_1$ = straightrun (between 0 and 0.21)
$\mathbf{x}_2$ = reformate (between 0 and 0.62)
$\mathbf{x}_3$ = thermally cracked naphta (between 0 and 0.12)
$\mathbf{x}_4$ = catalytically cracked naphta (between 0 and 0.62)
$\mathbf{x}_5$ = polymer (between 0 and 0.12)
$\mathbf{x}_6$ = alkylate (between 0 and 0.74)
$\mathbf{x}_7$ = natural gasoline (between 0 and 0.08)

Table 1
Cornell data

$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$	$\mathbf{y}$
0.00	0.23	0.00	0.00	0.00	0.74	0.03	98.70
0.00	0.10	0.00	0.00	0.12	0.74	0.04	97.80
0.00	0.00	0.00	0.10	0.12	0.74	0.04	96.60
0.00	0.49	0.00	0.00	0.12	0.37	0.02	92.00
0.00	0.00	0.00	0.62	0.12	0.18	0.08	86.60
0.00	0.62	0.00	0.00	0.00	0.37	0.01	91.20
0.17	0.27	0.10	0.38	0.00	0.00	0.08	81.90
0.17	0.19	0.10	0.38	0.02	0.06	0.08	83.10
0.17	0.21	0.10	0.38	0.00	0.06	0.08	82.40
0.17	0.15	0.10	0.38	0.02	0.10	0.08	83.20
0.21	0.36	0.12	0.25	0.00	0.00	0.06	81.40
0.00	0.00	0.00	0.55	0.00	0.37	0.08	88.10

The sum of all $\mathbf{x}_j$'s for a given sample is equal to 1.

Multiple regression. This dataset causes some problems when analysed with multiple regression. As a matter of fact, the output from Proc REG of SAS on a model without the constant term (Table 3) shows results that are uncoherent with the observed

Table 2
Correlation matrix

Pearson correlation coefficients, $N = 12$							
	Y	$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	$\mathbf{X}_5$	$\mathbf{X}_6$
Y	**1.000**						
X1	**−0.837**	1.000					
X2	−0.071	0.104	1.000				
X3	**−0.838**	**0.999**	0.101	1.000			
X4	**−0.707**	0.371	−0.537	0.374	1.000		
X5	0.494	−0.548	−0.293	−0.548	−0.211	1.000	
X6	**0.985**	−0.805	−0.191	−0.805	−0.646	0.463	1.000
X7	**−0.741**	0.603	−0.590	0.607	**0.916**	−0.274	−0.656

In bold, significant correlations with $p < 0.0001$.

correlations for variables $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \mathbf{x}_5$ and $\mathbf{x}_7$. For variables $\mathbf{x}_1$, $\mathbf{x}_3$ and $\mathbf{x}_7$ this is due to a multicollinearity problem. For variables $\mathbf{x}_2$ and $\mathbf{x}_5$ an explication may be found by using PLS regression or, as it is reported in the following, by running a modified PLS regression.

Table 3
Multiple regression: SAS output

Dependent variable: Y
NOTE: No intercept in model. *R*-Square is redefined.
Root MSE 0.83619 R-Square 1.0000
Dependent Mean 88.58333 Adj R-Sq 0.9999
Coeff Var 0.94396

Variable	DF	Parameter estimate	Standard error	t Value	$\Pr > \lvert t \rvert$
$\mathbf{X}_1$	1	34.32023	209.12698	0.16	0.8761
$\mathbf{X}_2$	1	85.92283	1.24820	68.84	<0.0001
$\mathbf{X}_3$	1	141.25193	375.32840	0.38	0.7221
$\mathbf{X}_4$	1	77.18010	9.21351	8.38	0.0004
$\mathbf{X}_5$	1	87.75022	5.81572	15.09	<0.0001
$\mathbf{X}_6$	1	100.30083	3.47382	28.87	<0.0001
$\mathbf{X}_7$	1	116.92128	81.09563	1.44	0.2089

Forward stepwise modified PLS regression. The response variable $\mathbf{y}$ is taken to be standardised ($\mathbf{y}^*$).

Computation of $\mathbf{t}_1$. Component $\mathbf{t}_1$ is computed only with those variables $\mathbf{x}_j$'s that are significantly correlated to $\mathbf{y}$.

Then, formula (16) yields

$$\mathbf{t}_1 = \frac{-0.8373\mathbf{x}_1^* - 0.83796\mathbf{x}_3^* - 0.70671\mathbf{x}_4^* + 0.98507\mathbf{x}_6^* - 0.74112\mathbf{x}_7^*}{\sqrt{0.8373^2 + 0.83796^2 + 0.70671^2 + 0.98507^2 + 0.74112^2}}$$

$$= 0.4526\mathbf{x}_1^* - 0.4530\mathbf{x}_3^* - 0.3820\mathbf{x}_4^* + 0.5325\mathbf{x}_6^* - 0.4006\mathbf{x}_7^*.$$

Computation of $\mathbf{t}_2$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_2$ are looked for. Therefore, 7 multiple regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$ are run. Significance levels (p-values) are reported in Table 4.

Table 4
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$ and each $\mathbf{x}_j$

Variable	p-values
$\mathbf{x}_1$	0.6225
$\mathbf{x}_2$	**0.0101**
$\mathbf{x}_3$	0.6016
$\mathbf{x}_4$	0.9055
$\mathbf{x}_5$	0.7221
$\mathbf{x}_6$	$<$**0.0001**
$\mathbf{x}_7$	0.0532

Only variables $\mathbf{x}_2$ and $\mathbf{x}_6$ are significant at a risk level $\alpha = 0.05$. The residual vectors $\mathbf{x}_{12}$, $\mathbf{x}_{16}$ from the regressions of $\mathbf{x}_2$, $\mathbf{x}_6$ on $\mathbf{t}_1$ are computed. They are orthogonal to $\mathbf{t}_1$ and linear combinations of $\mathbf{x}_j$. As a consequence, any linear combination of $\mathbf{x}_{12}$ and $\mathbf{x}_{16}$ is also a linear combination of $\mathbf{x}_j$ and orthogonal to $\mathbf{t}_1$. The application of (21) leads to

$$\mathbf{t}_2 = \frac{1}{\sqrt{0.19565^2 + 0.25762^2}}(-0.19565\mathbf{x}_{12}^* + 0.25762\mathbf{x}_{16}^*),$$

$$= 0.4211\mathbf{x}_1^* - 0.6098\mathbf{x}_2^* + 0.4214\mathbf{x}_3^* + 0.3554\mathbf{x}_4^* + 1.5363\mathbf{x}_6^* + 0.3727\mathbf{x}_7^*.$$

It is worth noticing that variable $\mathbf{x}_2$ is not at all significant in terms of correlation with $\mathbf{y}$. Nevertheless, its partial correlation with $\mathbf{y}$ conditionally to $\mathbf{t}_1$ is significant.

Computation of $\mathbf{t}_3$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_3$ are looked for. Therefore, 7 multiple regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$ and each $\mathbf{x}_j$ are run. Table 5 reports the p-values.

Table 5
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$ and each $\mathbf{x}_j$

Variables	p-values
$\mathbf{x}_1$	**0.0289**
$\mathbf{x}_2$	**0.0294**
$\mathbf{x}_3$	**0.0258**
$\mathbf{x}_4$	**0.0177**
$\mathbf{x}_5$	0.6356
$\mathbf{x}_6$	**0.0294**
$\mathbf{x}_7$	0.0922

Conclusion: Variables $\mathbf{x}_1$, $\mathbf{x}_2$, $\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_6$ are significant at a risk level $\alpha = 0.05$; a third PLS component $\mathbf{t}_3$ may be then computed.

The third PLS component $\mathbf{t}_3$ is then computed on the standardised residual vectors $\mathbf{x}_{21}^*$, $\mathbf{x}_{22}^*$, $\mathbf{x}_{23}^*$, $\mathbf{x}_{24}^*$, $\mathbf{x}_{26}^*$ from the multiple regressions of the significant explanatory variables on the components $\mathbf{t}_1$ and $\mathbf{t}_2$

$$\mathbf{t}_3 = \frac{0.0859\mathbf{x}_{21}^* + 0.0857\mathbf{x}_{22}^* + 0.0872\mathbf{x}_{23}^* - 0.0912\mathbf{x}_{24}^* + 0.0857\mathbf{x}_{26}^*}{\sqrt{0.0859^2 + 0.0857^2 + 0.0872^2 + 0.0912^2 + 0.0857^2}}$$

$$= 1.600\mathbf{x}_1^* + 1.248\mathbf{x}_2^* + 1.622\mathbf{x}_3^* - 0.243\mathbf{x}_4^* + 3.260\mathbf{x}_6^* + 0.541\mathbf{x}_7^*.$$

Computation of $\mathbf{t}_4$. At first, the variables $\mathbf{x}_j$'s with a significant contribution to the construction of $\mathbf{t}_4$ are looked for by studying the regressions of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ and each $\mathbf{x}_j$. Table 6 reports the p-values.

Table 6
Results from multiple regression of $\mathbf{y}^*$ on $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ and each $\mathbf{x}_j$

Variable	p-values
$\mathbf{x}_1$	0.7096
$\mathbf{x}_2$	0.9378
$\mathbf{x}_3$	0.8517
$\mathbf{x}_4$	0.5711
$\mathbf{x}_5$	0.6867
$\mathbf{x}_6$	0.9378
$\mathbf{x}_7$	0.3351

Conclusion: There is no significant explanatory variable at the risk level $\alpha = 0.05$; only the first three PLS components are retained.

PLS regression equation with three components. Regression of the original response variable $\mathbf{y}$ on the three components $\mathbf{t}_1$, $\mathbf{t}_2$, $\mathbf{t}_3$ leads to the equation

$$\hat{\mathbf{y}} = 88.58 + 3.25\mathbf{t}_1 + 2.53\mathbf{t}_2 + 1.20\mathbf{t}_3,$$

that can be also written in terms of the original explanatory variables so as to show the related regression coefficients:

$$\hat{\mathbf{y}} = 87.682 - 5.920\mathbf{x}_1 - 2.034\mathbf{x}_2 - 10.060\mathbf{x}_3 - 3.892\mathbf{x}_4 + 15.133\mathbf{x}_6 - 26.429\mathbf{x}_7$$

$$= 87.682\left(\sum_{j=1}^{7}\mathbf{x}_j\right) - 5.920\mathbf{x}_1 - 2.034\mathbf{x}_2 - 10.060\mathbf{x}_3$$

$$-3.892\mathbf{x}_4 + 15.133\mathbf{x}_6 - 26.429\mathbf{x}_7$$

$$= 81.762\mathbf{x}_1 + 85.648\mathbf{x}_2 + 77.622\mathbf{x}_3 + 83.790\mathbf{x}_4 + 87.682\mathbf{x}_5$$

$$+ 102.815\mathbf{x}_6 + 61.253\mathbf{x}_7.$$

The maximum octane ratio is obtained by setting $\mathbf{x}_2 = 0.14$, $\mathbf{x}_5 = 0.12$, $\mathbf{x}_6 = 0.74$ such that

$$\hat{\mathbf{y}} = 85.648 \times 0.14 + 87.682 \times 0.12 + 102.815 \times 0.74 = 98.59.$$

Use of classical PLS regression and bootstrap validation. PLS regression has been also run on Cornell data. Cross-validation leads to 3 PLS components. Fig. 1 shows a multiple box-plot representation of bootstrap distributions (with $B = 1000$) for the standardised regression coefficients related to the seven explanatory variables. All coefficients, except the one related to $\mathbf{x}_5$, have their bootstrap distributions well below or above zero and thus can be considered statistically significant. However, the graphical representation clearly shows asymmetry for almost all bootstrap distributions. Namely, the medians of the distributions for standardised b_1, b_3 and b_7 are closer to the respective third quartiles while the medians of the distributions for standardised b_2 and b_4 are close to the first quartiles; the distributions of standardised b_2 and b_7 show right heavy tails.

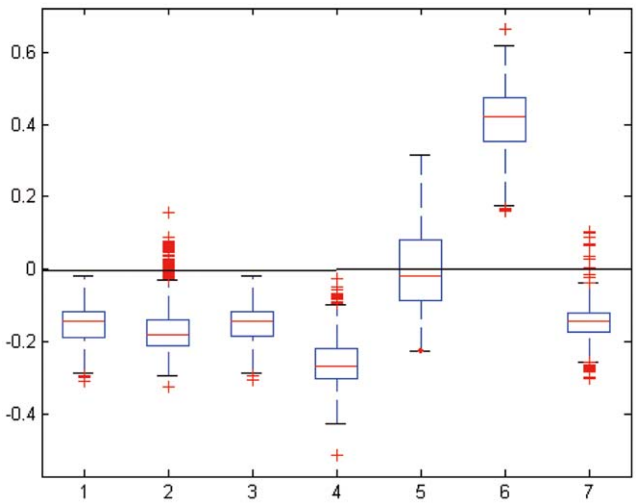


Fig. 1. Bootstrap distributions of standardised b_j ($j = 1, \ldots, 7$).

Fig. 2 shows 95% confidence intervals in the case of both classical Bootstrap (continuous lines) and BC_a (long dotted lines) intervals. By comparing (Table 7) the values of $\alpha_{j,1}$ and $\alpha_{j,2}$ with 0.025 and 0.975 (nominal confidence bounds), it can be seen that the bias-correction and the acceleration factor shift the intervals and may even change the proportion of empirical (bootstrap) distribution comprised in between. Actually, the shift takes into account shape and bias of the empirical distributions and ensures a 95% confidence level.

BC_a confidence interval related to variable $\mathbf{x}_5$, polymer, comprises zero and thus the related coefficient is considered to be not statistically significant.

BC_a confidence intervals after removing variable $\mathbf{x}_5$ (short dotted lines in Fig. 2) show how all remaining variables are significant. Also in this case, both $\alpha_{j,1}$ and $\alpha_{j,2}$ are shifted (Table 7) and the related confidence intervals are by far more accurate than the previous ones.

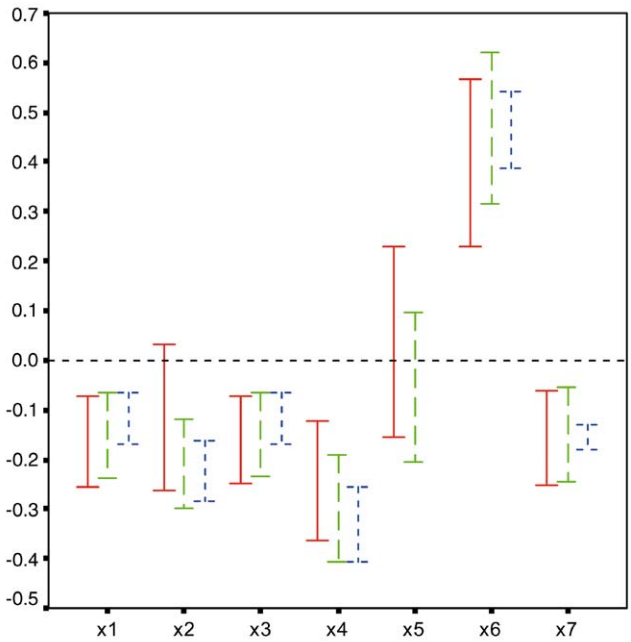


Fig. 2. 95% bootstrap confidence intervals: continuous lines indicate $I_j(\alpha)$'s, long dotted lines indicate $I_j^{\mathrm{BC}_a}(\alpha)$'s, short dotted lines indicate $I_j^{\mathrm{BC}_a}(\alpha)$'s after removing variable $\mathbf{x}_5$.

Table 7
Order of percentiles for BC_a confidence intervals (with and without $\mathbf{x}_5$)

All variables included				After removing $\mathbf{x}_5$		
	$\alpha_{j,1}$,	$\alpha_{j,2}$	Covered proportion of bootstrap distributions	$\alpha_{j,1}$,	$\alpha_{j,2}$	Covered proportion of bootstrap distributions
$\mathbf{x}_1$	0.0397	0.9840	0.9443	0.037	0.983	0.946
$\mathbf{x}_2$	0.0012	0.8077	0.8065	0.011	0.951	0.940
$\mathbf{x}_3$	0.0402	0.9842	0.9440	0.037	0.983	0.946
$\mathbf{x}_4$	0.0027	0.8794	0.8767	0.014	0.959	0.944
$\mathbf{x}_5$	0.0034	0.7832	0.7798		Removed	
$\mathbf{x}_6$	0.1539	0.9982	0.8443	0.057	0.992	0.935
$\mathbf{x}_7$	0.0280	0.9778	0.9498	0.028	0.977	0.949

2.3. The case of logistic regression: Bordeaux wine data

The proposed algorithm for generalised linear regression is now applied to logistic regression by referring to a dataset on Bordeaux wines.

Table 8
Bordeaux wine data

OBS	Year	Temperature	Sunshine	Heat	Rain	Quality
1	1924	3064	1201	10	361	2
2	1925	3000	1053	11	338	3
3	1926	3155	1133	19	393	2
4	1927	3085	970	4	467	3
5	1928	3245	1258	36	294	1
6	1929	3267	1386	35	225	1
7	1930	3080	966	13	417	3
8	1931	2974	1189	12	488	3
9	1932	3038	1103	14	677	3
10	1933	3318	1310	29	427	2
11	1934	3317	1362	25	326	1
12	1935	3182	1171	28	326	3
13	1936	2998	1102	9	349	3
14	1937	3221	1424	21	382	1
15	1938	3019	1230	16	275	2
16	1939	3022	1285	9	303	2
17	1940	3094	1329	11	339	2
18	1941	3009	1210	15	536	3
19	1942	3227	1331	21	414	2
20	1943	3308	1366	24	282	1
21	1944	3212	1289	17	302	2
22	1945	3361	1444	25	253	1
23	1946	3061	1175	12	261	2
24	1947	3478	1317	42	259	1
25	1948	3126	1248	11	315	2
26	1949	3458	1508	43	286	1
27	1950	3252	1361	26	346	2
28	1951	3052	1186	14	443	3
29	1952	3270	1399	24	306	1
30	1953	3198	1259	20	367	1
31	1954	2904	1164	6	311	3
32	1955	3247	1277	19	375	1
33	1956	3083	1195	5	441	3
34	1957	3043	1208	14	371	3

The following variables (Table 8) were measured in 34 years (1924–1957):

Temperature:	Sum of average day temperatures (°C)
Sunshine:	Duration of sunshine (h)
Heat:	Number of very warm days
Rain:	Rain height (mm)
Wine quality:	$1 =$ good, $2 =$ average, $3 =$ poor

Ordinal logistic regression. Ordinal logistic regression of quality on the four *standardised* predictors corresponds to the following model

$$\text{Prob}(\mathbf{y} \leqslant k) = \frac{e^{\alpha_k+\beta_1\text{Temperature}+\beta_2\text{Sunshine}+\beta_3\text{Heat}+\beta_4\text{Rain}}}{1+e^{\alpha_k+\beta_1\text{Temperature}+\beta_2\text{Sunshine}+\beta_3\text{Heat}+\beta_4\text{Rain}}}, \tag{27}$$

where $k = 1, 2$, relates to the first two levels of quality.

It is a proportional odds ratio model that is here accepted by means of a Score test given in Table 9. This table shows also the results from a Proc Logistic run in SAS on the standardised variables. The p-values yielded by the Wald test on the four logistic regression coefficients are, respectively, $0.0573, 0.1046, 0.4568, 0.0361$. Only Temperature and Rain are significant variables at the 10% risk level. The estimation of model (27) allows to compute the probability for a year to be of good, average or poor quality. After assigning a year to the most probable quality, Table 10 is obtained by crossing observed quality with the predicted one. There are seven misclassified years.

Table 9
Logistic regression of quality on the standardised meteorological variables

Score test for the proportional odds assumption Chi-square = 2.9159 with 4 DF ($p = 0.5720$)					
Analysis of maximum likelihood estimates					
Variable	DF	Parameter estimate	Standard error	Wald chi-square	Pr > chi-square
INTERCP1	1	−2.6638	0.9266	8.2641	0.0040
INTERCP2	1	2.2941	0.9782	5.4998	0.0190
TEMPERA	1	3.4268	1.8029	3.6125	0.0573
SUNSHINE	1	1.7462	1.0760	2.6335	0.1046
HEAT	1	−0.8891	1.1949	0.5536	0.4568
RAIN	1	−2.3668	1.1292	4.3931	0.0361

Table 10
Prediction quality of model (27) by using classical logistic regression

Quality observed counts	Predicted 1	2	3	Total
1	8	3	0	11
2	2	8	1	11
3	0	1	11	12
Total	10	12	12	34

PLS ordinal logistic regression. In the Bordeaux wine data, the multicollinearity between predictors leads to two difficulties: first, influent variables such as Sunshine and Heat become nonsignificant in model (27) even though they are significant when taken one at a time; second, the variable Heat shows a negative coefficient in the model while it has a positive influence on quality in reality.

Generally, PLS logistic regression allows to obtain a coherent model at the level of coefficients while retaining all predictors. Moreover, it is equally feasible when predictors have missing data. Separate logistic regressions of quality on each standardised predictor yields the coefficients a_{1j} of Temperature, Sunshine, Heat and Rain

equal to, respectively, 3.0117 (0.0002), 3.3401 (0.0002), 2.1445 (0.0004) and −1.7906 (0.0016), with the p-values given in parentheses. These coefficients are all significant with coherent signs.

After normalizing the coefficients, the first PLS component is defined as

$$\mathbf{t}_1 = \frac{3.0117 \times \text{Temperature} + 3.3401 \times \text{Sunshine} + 2.1445 \times \text{Heat} - 1.7906 \times \text{Rain}}{\sqrt{(3.0117)^2 + (3.3401)^2 + (2.1445)^2 + (-1.7906)^2}},$$

$$= 0.5688 \times \text{Temperature} + 0.6309 \times \text{Sunshine} + 0.4050 \times \text{Heat} - 0.3382 \times \text{Rain}.$$

The results from logistic regression of quality on component $\mathbf{t}_1$ are given in Table 11. It is satisfactory to notice that there are only 6 misclassified years, i.e. one misclassified year less than in classical logistic regression.

Table 11
Results from logistic regression of quality on component $\mathbf{t}_1$

Analysis of parameter	Maximum DF	Likelihood estimate	Estimates std.error	Chi-square	Pr > Chi sq
Intercept	1	−2.2650	0.8644	6.8662	0.0088
Intercept2	1	2.2991	0.8480	7.3497	0.0067
$t1$	1	2.6900	0.7155	14.1336	0.0002

Cross-table of observed and predicted quality

Quality observed Counts	Predicted 1	2	3	Total
1	9	2	0	11
2	2	8	1	11
3	0	1	11	12
Total	11	11	12	34

In order to search for the variables whose contribution to the second component $\mathbf{t}_2$ is significant, logistic regressions of quality on $\mathbf{t}_1$ and each standardised predictor $\mathbf{x}_j^*$ are built

$$\text{Prob}(\mathbf{y} \leqslant k) = \frac{e^{\alpha_k + \beta_1 \mathbf{t}_1 + \beta_{2j} \mathbf{x}_j^*}}{1 + e^{\alpha_k + \beta_1 \mathbf{t}_1 + \beta_{2j} \mathbf{x}_j^*}}. \tag{28}$$

The following p-values are obtained for the predictors coefficients: 0.6765, 0.6027, 0.0983, 0.2544. Therefore, the second PLS component is not significant and, as a consequence, the model with only one component is retained.

By expressing the component $\mathbf{t}_1$ in terms of the standardized variables Temperature, Sunshine, Heat and Rain, the final estimates of the parameters in model (27) are more coherent than the ones earlier obtained by classical logistic regression

$$\text{Prob}(\mathbf{y} = 1) = \frac{e^{-2.265 + 1.53 \times \text{Temperature} + 1.70 \times \text{Sunshine} + 1.09 \times \text{Heat} - 0.91 \times \text{Rain}}}{1 + e^{-2.5265 + 1.53 \times \text{Temperature} + 1.70 \times \text{Sunshine} + 1.09 \times \text{Heat} - 0.91 \times \text{Rain}}},$$

and

$$\text{Prob}(\mathbf{y} \leqslant 2) = \frac{e^{2.2991+1.53\times\text{Temperature}+1.70\times\text{Sunshine}+1.09\times\text{Heat}-0.91\times\text{Rain}}}{1+e^{2.2991+1.53\times\text{Temperature}+1.70\times\text{Sunshine}+1.09\times\text{Rain}-0.91\times\text{Rain}}}.$$

Validation issues for PLS logistic regression: *randomised and balanced bootstrap*. In order to derive an empirical distribution for the parameters of PLS Logistic Regression, a simpler variant of bootstrap is adopted with the name of randomised bootstrap. It was earlier proposed in (Efron, 1983) for the case of a binary ($K = 2$) categorical response variable $\mathbf{y}$.

Let $\hat{F}$ be the empirical distribution of $(y_i, \mathbf{x}_i)$; $\hat{F}$ puts equal mass $1/n$ on each observed unit. Efron (1983) suggests two possible modifications of $\hat{F}$:

case 1: $\pi_i(\bar{y}_i, \mathbf{x}_i) = 0.1$ (empirical assessment of the probability) where $\pi_i(\bar{y}_i, \mathbf{x}_i)$ represents the probability of the complementary units to $(y_i, \mathbf{x}_i)$;

case 2: $\pi_i(y_i, \mathbf{x}_i) = \frac{e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i\hat{\boldsymbol{\beta}}}}$ (probability based on predictions from logistic regression).

Let $\hat{F}^{(\text{RAND})}$ be the modified empirical distribution. The bootstrap estimates are then based on the Monte Carlo approximation of the parameters bootstrap distributions.

An extension of "randomised" bootstrap to PLS Logistic Regression (generalised to the case where $K > 2$) is the following.

Let $\pi_i(y_i = k \mid \mathbf{x}_i)$ be the probability to observe $y_i = k$ $(k = 1, 2, \ldots, K)$ given a row vector $\mathbf{x}_i$ for the explanatory variables. These probabilities can be then derived in two different alternative ways: take the conditional probability estimates from classical or PLS logistic regression model or refer to the empirical distribution of $\mathbf{y}$.

Let B be the number of resamples, "randomised" boostrap performs the following steps:

Step 1: For each $b = 1, 2, \ldots, B$, let $X^{(b)}$ be the bth resample made by n i.i.d. random variables drawn from the explanatory variables part of $\hat{F}$. In order to properly assign a value $y_i^{(b)} = k$ $(k = 1, 2, \ldots, K)$ to each vector $\mathbf{x}_i$ $(i = 1, 2, \ldots, n)$, a pseudo random number is drawn from $\pi_i(y_i = k \mid \mathbf{x}_i)$,

Step 2: For each resample $(y_i^{(b)}, x_i^{(b)})$, the estimate $\hat{\beta}^{(b)}$ of PLS logistic regression coefficients is yielded,

Step 3: For each explanatory variable $\mathbf{x}_j$ $(j = 1, \ldots, 4)$ the Monte Carlo approximation $\Phi^{(B)}_{\hat{F}^{(\text{RAND})}, \hat{\beta}_j}$ of the bootstrap distribution of $\hat{\beta}_j$ is obtained.

Finally, the percentiles of $\Phi^{(\text{B})}_{\hat{F}^{(\text{RAND})}, \hat{\beta}_j}$ may be used for building up confidence intervals for β_j.

It is worth noticing that efficiency of bootstrap computation can be easily improved by means of balanced bootstrap (or permutation bootstrap (Efron and Tibshirani, 1993)), a simple modification of the ordinary bootstrap sampling schema. In the procedure of randomised bootstrap for PLS logistic regression, the balanced bootstrap (Tenenhaus, 2002) assures that each row $\mathbf{x}_i$ appears B times in the set of nB resampled rows. This is easily accomplished by running the following procedure:

(1) Stack B copies of $\mathbf{X}$ into a matrix, say $\boldsymbol{\Upsilon}$ with nB rows,
(2) Take a random permutation of $\boldsymbol{\Upsilon}$,

(3) The first bootstrap sample consists in the first n rows of Υ; the second bootstrap sample in the second set of n rows of Υ and so on until the Bth set coinciding with the last n rows of Υ.

The nonparametric validation procedure suggested above for PLS Logistic Regression has been applied to Bordeaux wine data (Fig. 3) with $B = 1000$. Actually, 12 resamples out of 1000 were unuseful due to divergence toward infinite of the coefficient estimates. This is a very well-known problem in literature and usually appears when a complete (or almost complete) separation of data is achieved in the bootstrap resample. In Moulton and Zager (1991) a modified estimation algorithm for logistic regression is proposed in order to overcome the problem. However, this algorithm was not worth using here because of the very small number of problematic samples. Finally, due to the strong significance enhanced by confidence intervals in Fig. 3, it was not even worth computing the BC_a intervals discussed earlier.

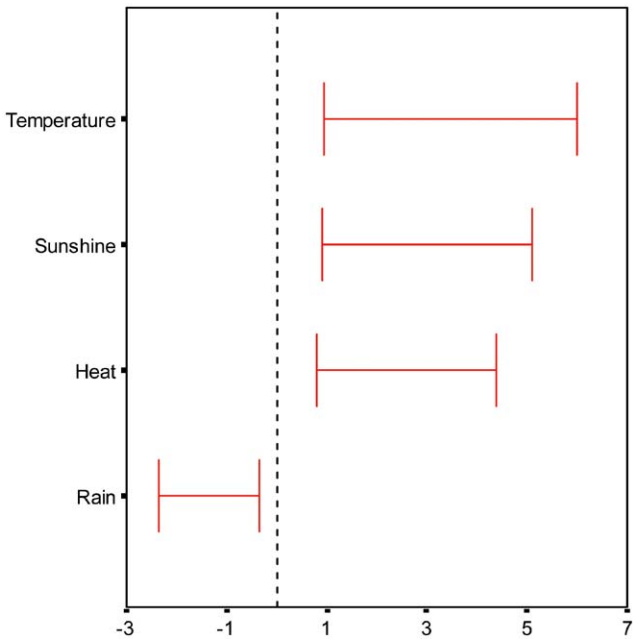


Fig. 3. 95% randomized/balanced bootstrap confidence intervals for Bordeaux wine data.

2.4. The case of censored data: determining risk factors associated with the age when first white hair appears in males

2.4.1. Data presentation

Started in 1994 by Prof. Serge Hercberg, the epidemiological study SU.VI.MAX (*SUpplémentation en VItamines et Minéraux Antioxidants*) aimed to evaluate the

Table 12
Data pre-treatment and construction of the response function

I	t	$\mathbf{y}_{it}$
Case 1	$\leqslant 30$	0
Case 1	31–35	0
Case 1	36–40	1
Case 2	$\leqslant 30$	0
Case 2	31–35	0
Case 2	36–40	0
Case 2	41–45	0

nutritional state of the French population and study the influence of an oral supplementation of vitamins and anti-oxidant minerals on a set of health indicators: heart attack, cardiovascular illnesses, cancers, where free radicals are strongly involved (Hercberg et al., 1998). This study was meant to finish after 8 years (in 2003) and comprises more than 12,000 volunteers between 25 and 65 years old, constituting a representative sample of the French population. Half of them were given an anti-oxidant treatment, the others a placebo.

Starting from this cohort, L'Oréal (a French cosmetics company) run a study on the health state of hair and fingernails on 10,323 cases (4057 males and 6266 females). On the basis of responses to a questionnaire covering more than 150 items, an attempt was made to demonstrate the risk factors associated with the premature onset of greying hair in men.

2.4.2. Variables description

The age of appearance of the first white hair in males is the response variable. It is a discrete variable with 6 categories: up to 30 years old, 31–35 years, 36–40 years, 41–45 years, 45–50 years, more than 50 years old.

The needed background for analysing this type of data is given in Allison (1995). Each case, say i, is duplicated m_i times where m_i is the number of his/her age class at the time of appearance of the first white hair or his/her age class at present if no white hair has appeared yet. The response variable $\mathbf{y}_{it}$ is then defined. It is equal to 0 if the ith case has no white hair at age t and 1 otherwise. For the last record in the list ($t = m_i$), $\mathbf{y}_{it}$ is equal to 1 if the event occurred, and 0 otherwise.

For instance, Table 12 shows two cases: case 1 is a 45-year-old male whose first white hair appeared at the age of 38 (3 records for this case); case 2 is a 45-year-old male with no white hair (4 records for this case).

The risk factors that have been selected from the SUVIMAX database are the following:

Previous appearances in the family (*dummy variables*: 1 = *yes*, 0 = *no*):

- The mother had white hair before being 30 years old ($\mathbf{x}_1$)
- The sister had white hair before being 30 years old ($\mathbf{x}_2$)
- The mother had no white hair after being 60 years old ($\mathbf{x}_3$)

- The father had white hair before being 30 years old ($\mathbf{x}_4$)
- The father had no white hair after being 60 years old ($\mathbf{x}_5$)

Hair features

- Thickness (very thin, thin, medium, thick)
- Natural color (red, blond, light brown, brown, dark brown, black)

Only significant factors have been retained for the application. Moreover, for a comparison between classical generalised linear model (McCullagh and Nelder, 1989) and PLS-GLR, only the cases with no missing data on the selected risk factors are considered.

2.4.3. The model

Let P_{it} be the probability that the first white hair for case i appeared at time t knowing that it has not appeared at previous times $1, 2, \ldots, t-1$. Then

$$P_{it} = \text{Prob}(\mathbf{y}_{it} = 1/\mathbf{y}_{i1} = 0, \ldots, \mathbf{y}_{i,t-1} = 0) \tag{29}$$

is the conditional probability of observing "$\mathbf{y}_{it} = 1$" knowing that "$\mathbf{y}_{i1} = 0, \ldots, \mathbf{y}_{i,t-1} = 0$".

Let us suppose that the underlying continuous model is a Cox model (Cox, 1972) with proportional risks

$$\begin{aligned}\text{Log}[h_i(t)/h_0(t)] &= \beta_1\mathbf{x}_1 + \beta_2\mathbf{x}_2 + \beta_3\mathbf{x}_3 + \beta_4\mathbf{x}_4 + \beta_5\mathbf{x}_5 \\ &\quad + \beta_6\text{Red} + \beta_7\text{Blond} + \beta_8\text{Light Brown} + \beta_9\text{Brown} \\ &\quad + \beta_{10}\text{Dark Brown} + \beta_{11}\text{Very Thin} + \beta_{12}\text{Thin} + \beta_{13}\text{Medium} \\ &= \sum_{j=1}^{13} \beta_j\mathbf{v}_j,\end{aligned} \tag{30}$$

where $\mathbf{v}_j$ relates to any predictor indiscriminately.

A generalised linear model with the "Complementary Log Log" (Prentice and Gloeckler, 1978) link function may be used for modeling the response variable in function of the predictors

$$\text{Log}[-\text{Log}(1 - P_{it})] = \alpha_t + \sum_j \beta_j\mathbf{v}_j. \tag{31}$$

The interpretation of the regression coefficients in (31) in terms of hazard ratios coincides with the one related to the underlying model (30). For instance, appearance of the first white hair in the mother before being 30 years old increases the risk by $100(\exp(\beta_1) - 1)\%$, once the other variables are fixed.

The particular form of the response vector associated to the same case (all values are zero if the case is censored, all values are zero but the last one that is equal to 1 if the case is not censored) allows a factorisation of the likelihood by means of P_{it}

conditional probabilities:

$$
\begin{aligned}
L &= \prod_{i=1}^{N} \text{prob}(\mathbf{y}_{i1}=0,\mathbf{y}_{i2}=0,\ldots,\mathbf{y}_{it_i-1}=0,\mathbf{y}_{it_i}=1)\\
&= \prod_{i=1}^{N} \text{prob}(\mathbf{y}_{it_i}=1/\mathbf{y}_{i1}=0,\ldots,\mathbf{y}_{it_i-1}=0)\times \text{prob}(\mathbf{y}_{i1}=0,\ldots,\mathbf{y}_{it_i-1}=0)\\
&= \prod_{i=1}^{N} P_{it_i}(1-P_{it_i-1})\times\cdots\times(1-P_{i1})\\
&= \prod_{i=1}^{N}\prod_{l\leqslant t_i} P_{il}^{\mathbf{y}_{il}}(1-P_{il})^{1-\mathbf{y}_{il}}. \qquad (32)
\end{aligned}
$$

It is all set as if the responses $\mathbf{y}_{il}$ were independently distributed as Bernoulli probability distributions with parameter P_{il}. Model (31) is a specific generalised linear model that is going to be studied by means of Proc GENMOD in SAS and PLS-GLR.

2.4.4. Generalised linear model

Table 13 reports the results from Proc GENMOD in SAS version 8.2 applied to model (31).

Remarks

(1) The ratio between deviance and degrees of freedom is less than 1 and tends to validate the model goodness of fit to data.
(2) Conditional probability of first white hair appearance increases with age. This is the statistical translation of what is a natural phenomenon in the studied population.
(3) Results concerning the family history were as expected.
(4) The risk of appearance of white hair increases with hair thickness.
(5) The risk of appearance of white hair increases with the darkness of the natural hair colour.
(6) The red colour is located between light brown and brown. This result confirms the peculiarity of this colour from the bio-chemical point of view and underlines the need to treat this factor as categorical.

2.4.5. PLS generalised linear model

In Marx (1996) the parameters of a generalised linear regression are estimated by introducing PLS regression in the iterated weighted least-squares algorithm used for likelihood maximisation. The approach proposed in the following is much simpler, its implementation needs a little and easy programming and it can be easily generalised to all models that are linear at the level of explanatory variables.

2.4.5.1. Computation of the first PLS component. A generalised linear model is built with age and each predictor separately taken. Categorical variables are considered

Table 13
GENMOD output

Criteria for assessing goodness of fit

Criterion	DF	Value	Value/DF
Deviance	6147	5161.5374	0.8397
Scaled Deviance	6147	5161.5374	0.8397
Pearson Chi-Square	6147	5761.4509	0.9373
Scaled Pearson X2	6147	5761.4509	0.9373
Log Likelihood		−2580.7687	

LR statistics for type 3 analysis

Source	DF	Chi-square	Pr > Chi sq
Age	5	1448.91	<.0001
X1	1	14.61	0.0001
X2	1	19.97	<.0001
X3	1	30.20	<.0001
X4	1	45.62	<.0001
X5	1	23.95	<.0001
Thickness	3	16.75	0.0008
Color	5	31.12	<.0001

Analysis of parameter estimates

Parameter	DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chi Sq
Intercept	1	1.3745	0.1434	1.0935	1.6556	91.89	<.0001
<= 30	1	−3.6202	0.1345	−3.8837	−3.3566	724.63	<.0001
31–35	1	−3.1961	0.1233	−3.4378	−2.9544	671.75	<.0001
36–40	1	−2.2025	0.1039	−2.4061	−1.9988	449.09	<.0001
41–45	1	−1.3685	0.0974	−1.5595	−1.1776	197.30	<.0001
46–50	1	−0.9598	0.1014	−1.1586	−0.7610	89.54	<.0001
> 50	1	0.0000	0.0000	0.0000	0.0000	—	—
X1	1	0.4977	0.1244	0.2539	0.7415	16.01	<.0001
X2	1	0.6281	0.1330	0.3675	0.8887	22.31	<.0001
X3	1	−0.4532	0.0865	−0.6227	−0.2837	27.46	<.0001
X4	1	0.8153	0.1102	0.5994	1.0312	54.77	<.0001
X5	1	−0.4323	0.0928	−0.6141	−0.2505	21.72	<.0001
Very thin	1	−0.4598	0.1157	−0.6865	−0.2330	15.80	<.0001
Thin	1	−0.3518	0.0982	−0.5442	−0.1594	12.84	0.0003
Medium	1	−0.2791	0.0975	−0.4701	−0.0881	8.20	0.0042
Thick	0	0.0000	0.0000	0.0000	0.0000	—	—
Red	1	−0.2751	0.3141	−0.8908	0.3405	0.77	0.3811
Blond	1	−0.6806	0.1452	−0.9652	−0.3959	21.96	<.0001
Light brown	1	−0.3959	0.1055	−0.6027	−0.1891	14.07	0.0002
Brown	1	−0.3141	0.1032	−0.5165	−0.1118	9.26	0.0023
Dark brown	1	−0.1310	0.1120	−0.3506	0.0886	1.37	0.2422
Black	0	0.0000	0.0000	0.0000	0.0000	—	—

as a whole in the model. All variables are significant (Table 14). This result is not surprising as only significant risk factors have been retained for this application. Therefore, all variables contribute to the construction of the first component $\mathbf{t}_1$ (Table 15).

Table 14
Generalised linear regression of **y** on age and each predictor

Variables	Wald		
	DoF	Chi-square	p-value
$\mathbf{x}_1$	1	42.53	<0.0001
$\mathbf{x}_2$	1	59.76	<0.0001
$\mathbf{x}_3$	1	35.16	<0.0001
$\mathbf{x}_4$	1	74.05	<0.0001
$\mathbf{x}_5$	1	31.93	<0.0001
Thickness	3	27.87	<0.0001
Colour	5	46.93	<0.0001

Table 15
Generalised linear regression of **y** on age and each dummy or categorical predictor

Variables ($\mathbf{v}_j$)	Regression coefficients (a_{1j})
$\mathbf{x}_1$	0.73
$\mathbf{x}_2$	0.92
$\mathbf{x}_3$	−0.50
$\mathbf{x}_4$	0.93
$\mathbf{x}_5$	−0.51
Very thin	−0.56
Thin	−0.38
Medium	−0.28
Thick	0.00
Red	−0.41
Blond	−0.80
Light brown	−0.49
Brown	−0.40
Dark brown	−0.19
Black	0.00

2.4.5.2. Computation of the second PLS component. A generalised linear model is built with age, $\mathbf{t}_1$, and each predictor separately taken. Table 16 shows how all variables are not significant. Only the first PLS component $\mathbf{t}_1$ is retained.

2.4.5.3. Generalised linear model on $\mathbf{t}_1$. Table 17 reports the results from generalised linear regression McCullagh and Nelder (1989) of **y** on age and $\mathbf{t}_1$.

Remark

Table 18 shows how the goodness of fit criteria for the cloglog and the PLS cloglog models are really similar. They lead to accept these models as the values of deviance and Pearson's Chi-square are close to the degrees of freedom.

Table 16
Generalised linear regression of $\mathbf{y}$ on age, $\mathbf{t}_1$ and each predictor

Variables	Wald		
	DoF	Chi-square	p-value
$\mathbf{x}_1$	1	1.23	0.27
$\mathbf{x}_2$	1	1.63	0.20
$\mathbf{x}_3$	1	0.41	0.52
$\mathbf{x}_4$	1	0.48	0.49
$\mathbf{x}_5$	1	0.26	0.61
Thickness	3	0.41	0.94
Colour	5	0.26	0.99

Table 17
Generalised linear regression of $\mathbf{y}$ on age and $\mathbf{t}_1$

Criteria for assessing goodness of fit			
Criterion	DF	Value	Value/DF
Deviance	6159	5164.8371	0.8386
Scaled deviance	6159	5164.8371	0.8386
Pearson chi-square	6159	5779.5776	0.9384
Scaled pearson X2	6159	5779.5776	0.9384
Log likelihood		−2582.4185	

Analysis of parameter estimates							
Parameter	DF	Estimate	Standard error	Wald 95% confidence limits		Chi-square	Pr > Chi Sq
< = 30	1	−2.9125	0.1041	−3.1165	−2.7085	783.28	<0.0001
31–35	1	−2.4855	0.0897	−2.6614	−2.3096	767.06	<0.0001
36–40	1	−1.4883	0.0610	−1.6078	−1.3688	595.66	<0.0001
41–45	1	−0.6544	0.0498	−0.7521	−0.5567	172.38	<0.0001
46–50	1	−0.2500	0.0580	−0.3637	−0.1363	18.56	<0.0001
> 50	1	0.7060	0.0829	0.5436	0.8684	72.59	<0.0001
$\mathbf{t}_1$	1	1.6864	0.1036	1.4833	1.8894	264.98	<0.0001

Table 18
Cloglog and PLS cloglog models goodness of fit

	Cloglog model			PLS cloglog model		
	Value	DoF	Value/DoF	Value	DoF	Value/DoF
Deviance	5162	6147	0.84	5165	6159	0.84
Pearson's Chi-square	5761	6147	0.94	5780	9159	0.94

2.4.5.4. PLS model in terms of the original variables. PLS generalised linear model may be expressed in function of the original variables by using the expression $\mathbf{t}_1$. Estimates of coefficients in cloglog and PLS cloglog models, as well as their 95% confidence intervals (Wald for classical model and bootstrap for PLS model) are very close to each other (Table 19).

Table 19
Coefficients of cloglog model and PLS cloglog model with 1 component in terms of the original variables

Variables	Cloglog				Cloglog PLS			
	Parameter estimates	Standard error	Hazard ratio	Wald 95% CI	Parameter estimates	Standard error	Hazard ratio	Bootstrap 95% CI
$\leqslant 30$	−3.62	0.13	0.03	0.02–0.03	−3.65	0.15	0.03	0.02–0.03
31–35	−3.20	0.12	0.04	0.03–0.05	−3.22	0.13	0.04	0.03–0.05
36–40	−2.20	0.10	0.11	0.09–0.14	−2.21	0.11	0.11	0.09–0.13
41–45	−1.37	0.10	0.25	0.21–0.31	−1.38	0.09	0.25	0.21–0.30
46–50	−0.96	0.10	0.38	0.31–0.47	−0.97	0.10	0.38	0.30–0.46
> 50	0.00	—	—	—	0.00	0.00	—	—
$\mathbf{x}_1$	0.50	0.12	1.65	1.28–2.10	0.59	0.12	1.80	1.45–2.22
$\mathbf{x}_2$	0.63	0.13	1.88	1.45–2.44	0.74	0.12	2.09	1.72–2.59
$\mathbf{x}_3$	−0.45	0.09	0.64	0.54–0.76	−0.39	0.06	0.67	0.58–0.75
$\mathbf{x}_4$	0.82	0.11	2.27	1.82–2.80	0.74	0.12	2.11	1.79–2.97
$\mathbf{x}_5$	−0.43	0.09	0.65	0.54–0.78	−0.40	0.07	0.67	0.58–0.77
Very thin	−0.46	0.12	0.63	0.50–0.79	−0.44	0.09	0.64	0.54–0.75
Thin	−0.35	0.10	0.70	0.58–0.85	−0.30	0.08	0.74	0.63–0.87
Medium	−0.28	0.10	0.76	0.63–0.91	−0.22	0.08	0.80	0.70–0.93
Thick	0.00	—	—	—	0.00	0.00	—	—
Red	−0.28	0.31	0.76	0.41–1.40	−0.30	0.24	0.74	0.49–1.19
Blond	−0.68	0.15	0.51	0.38–0.67	−0.64	0.13	0.53	0.42–0.64
Light brown	−0.40	0.11	0.67	0.55–0.83	−0.39	0.09	0.68	0.56–0.79
Brown	−0.31	0.10	0.73	0.59–0.90	−0.32	0.10	0.73	0.62–0.84
Dark brown	−0.13	0.11	0.88	0.70–1.09	−0.15	0.11	0.86	0.72–1.03
Black	0.00	—	—	—	0.00	—	—	—

Remarks

(1) Confidence intervals are obtained by means of the "Balanced bootstrap" resampling method (Davison et al., 1986) as it performs better than classical "Uniform bootstrap" in reducing the simulation error.
(2) Confidence intervals are estimated by referring to the percentiles of the bootstrap empirical distribution, however, the use of the bootstrap estimate for the coefficients standard deviation would yield very comparable results.
(3) A PLS model with two components yields the same estimates as the classical model with a precision up to the second decimal digit.

2.4.6. PLS Cox model

The Cox model (30) could also be used with discrete data (here, the age when white hair appears takes values $t = 1$–6). Efron approximation is used for taking into

account ties. As for model (31) on complementary log log, only one PLS component is retained. Table 20 reports results from classical Cox model and PLS Cox.

Table 20
Coefficients of classical Cox model and PLS Cox model with 1 component in terms of the original variables

Variables	Cox				PLS Cox			
	Parameter estimates	Standard error	Hazard ratio	Wald 95% CI	Parameter estimates	Standard error	Hazard ratio	Bootstrap 95% CI
$\mathbf{x}_1$	0.47	0.12	1.60	1.27–2.02	0.55	0.11	1.72	1.44–2.16
$\mathbf{x}_2$	0.58	0.13	1.78	1.39–2.29	0.69	0.11	1.99	1.62–2.36
$\mathbf{x}_3$	−0.43	0.08	0.65	0.55–0.77	−0.37	0.06	0.69	0.61–0.77
$\mathbf{x}_4$	0.76	0.11	2.15	1.75–2.64	0.69	0.10	2.01	1.64–2.46
$\mathbf{x}_5$	−0.41	0.09	0.67	0.56–0.79	−0.39	0.07	0.68	0.61–0.77
Very thin	−0.43	0.11	0.65	0.52–0.81	−0.42	0.08	0.66	0.57–0.79
Thin	−0.33	0.10	0.72	0.60–0.87	−0.29	0.07	0.75	0.66–0.85
Medium	−0.26	0.09	0.77	0.64–0.93	−0.21	0.07	0.81	0.70–0.95
Thick	0.00	—	—	—	0.00	—	—	—
Red	−0.26	0.30	0.77	0.42–1.39	−0.30	0.21	0.74	0.46–1.11
Blond	−0.65	0.14	0.52	0.40–0.69	−0.61	0.12	0.54	0.43–0.69
Light brown	−0.38	0.10	0.69	0.56–0.84	−0.37	0.09	0.69	0.56–0.81
Brown	−0.30	0.10	0.74	0.61–0.90	−0.30	0.09	0.74	0.62–0.89
Dark brown	−0.13	0.11	0.88	0.71–1.09	−0.15	0.10	0.86	0.71–1.06
Black	0.00	—	—	—	0.00	—	—	—

Remarks

(1) As it could be expected, results from Cox model and from cloglog model are very similar.
(2) Here, the same remarks previously done in comparing cloglog model with PLS cloglog apply.
(3) Coefficients are estimated by maximising partial likelihood (Cox, 1972).
(4) The Cox model is not strictly speaking a generalised linear model McCullagh and Nelder (1989). PLS generalised linear regression can also be extended beyond the exponential family.
(5) The Cox model does not need the duplication of patient data as for the cloglog model.

In Nguyen and Rocke (2002), the so-called Partial Least Squares Proportional Hazard Regression is proposed for the application to gene expression data from DNA microarrays. Their proposal actually consists of a two-stage strategy of analysis: PLS regression at the first stage in order to reduce data dimensionality and extract PLS components; Proportional Hazard regression model at the second stage in order to estimate survival distribution. This two-stage strategy does not take into account the censoring information in the estimation of PLS components, thus inducing bias in their

estimates. The approach proposed above goes further than that as it is meant to use PLS criterion inside generalised linear regression model, so as to enrich its interpretation and enlarge its feasibility to peculiar and difficult situations often encountered in real practice.

3. Conclusion

The PLS regression algorithm that has been re-formulated above shows several advantages:

(1) Classical PLS regression is directly linked to the usual procedures for simple and multiple regressions and is therefore enriched by the classical testing procedures of such methods. In this framework, the statistical tests aim at identifying those explanatory variables that do not significantly contribute to the construction of PLS components and, consequently, with a low explanatory power on the response variable. A PLS components is judged as not significant when there is no explanatory variable with a significant weight. In its construction. In the example of Section 2, the proposed approach led to the same variable selection as backward stepwise classical PLS regression. The approach might be further validated on a wider variety of examples also in comparison with other approaches to variable selection in PLS recently proposed in Forina et al. (1999), Gauchi and Chagnon (2001), Höskuldsson (2001), Lazraq et al. (2003), Lingren et al. (1994) and Sarabia et al. (2001).
(2) In practice, when a strong degree of multicollinearity shows up, stepwise multiple regression is commonly used. The inconvenience of this method is the elimination of explanatory variables strongly correlated to the response variable and thus important for the user. On the contrary, PLS regression allows to retain in the model all variables with a stronger explanatory power.
(3) In case of missing data, PLS components are computed in accordance with the NIPALS algorithm. However, PLS components are correlated in this case. The original PLS regression algorithm does not consider this feature. On the contrary, the new formulation let the correlation between PLS components play a role as multiple regression is used.
(4) There is an immediate extension to generalised linear regression. Some preliminary results were already obtained in PLS logistic regression Esposito Vinzi and Tenenhaus (2001) and in survival data with the Cox PLS model Bastien and Tenenhaus (2001).

Acknowledgements

We are very grateful to Silvano Amato for his very valid help in setting up the bootstrap-based validation procedures presented in the paper.

References

Allison, P.D., 1995. Survival Analysis Using the SAS System: A Practical Guide. SAS Inc, Cary, NC, USA.

Bastien, P., Tenenhaus, M., 2001. PLS generalised linear regression. Application to the analysis of life time data. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 131–140.

Camo, ASA., 2003. The Unscrambler® 8.0, CAMO ASA, Nedre Vollgate 8, 0158 Oslo, Norway.

Cox, D.R., 1972. Regression models and life-tables (with discussion). J. Roy. Statist. Soc., Series B 34, 187–220.

Davison, A.C., Hinkley, D.V., Schechtman, E., 1986. Efficient bootstrap simulations. Biometrika 73, 555–566.

Efron, B., 1983. Estimating the error rate of a prediction rule: some improvements on cross-validation. J. Amer. Statist. Assoc. 78, 316–331.

Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman & Hall, New York.

Esposito Vinzi, V., Tenenhaus, M., 2001. PLS Logistic Regression. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 117–130.

Forina, M., Casolino, C., Pizarro Millan, C., 1999. Iterative predictor weighting PLS (IPW): a technique for the elimination of useless predictors in regression problems. J. Chemometrics 13, 165–184.

Garthwaite, P.H., 1994. An interpretation of partial least squares. J. Amer. Statist. Assoc. 89 (425), 122–127.

Gauchi, J.-P., Chagnon, P., 2001. Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. Chemometrics & Intelligent Laboratory Systems 58 (2), 171–193.

Hercberg, S., Preziosi, P., Briancon, S., Galan, P., Triol, I., Malvy, D., Roussel, A.M., Favier, A., 1998. A primary prevention trial of nutritional doses of antioxydant vitamins and minerals on cardiovascular diseases and cancers in general populations: the SU.VI.MAX Study—design, methods and participant characteristics. Supplementation en Vitamines et Mineraux AntioXydants. Control Clin. Trials 19 (4), 336–351.

Höskuldsson, A., 2001. Variable and subset selection in PLS regression. Chemometrics & Intelligent Laboratory Systems 55, 23–38.

Kettaneh-Wold, N., 1992. Analysis of mixture data with partial least squares. Chemometrics & Intelligent Laboratory Systems 14, 57–69.

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

Lingren, F., Geladi, P., Rannar, S., Wold, S., 1994. Interactive variable selection (IVS) for PLS. Part 1: theory and algorithms. J. Chemometrics 8, 349–363.

Marx, B.D., 1996. Iteratively reweighted partial least squares estimation for generalized linear regression. Technometrics 38 (4), 374–381.

McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, 2nd Edition. Chapman & Hall, London.

Moulton, L., Zager, S., 1991. Bootstrapping generalized linear models. Computat. Statist. Data Anal. 11, 53–63.

Nguyen, D.V., Rocke, D.M., 2002. Partial least squares proportional hazard regression for application to DNA microarray survival data. Bioinformatics 18 (12), 1625–1632.

Prentice, R.L., Gloeckler, L.A., 1978. Regression analysis of grouped survival data with application to breast cancer data. Biometrics 34, 57–67.

Sarabia, L.A., Ortiz, M.C., Sánchez, M.S., Herrero, A., 2001. Dimension-wise selection in partial least squares regression with a bootstrap estimated signal-noise relation to weight the loadings. In: Esposito Vinzi, V., Lauro, C., Morineau, A., Tenenhaus, M. (Eds.), PLS and Related Methods, Proceedings of the PLS'01 International Symposium, CISIA-CERESTA, Paris, pp. 327–339.

Shao, J., Tu, D., 1995. The Jackknife and the Bootstrap. Springer, New York.

Shenk, J.S., Westerhaus, M.O., 1991. Population structuring of near infrared spectra and modified partial least squares regression. Crop Science 31, 1548–1555.

Tenenhaus, M., 1998. La régression. PLS Technip, Paris.

Tenenhaus, A., 2002. Régression Logistique PLS validée par bootstrap. Mémoire de DEA de Statistique, Université Pierre et Marie Curie, Paris and L'Oréal, Aulnay-sous-Bois.

Umetri, AB., 2004. SIMCA-P 10.5, User Guide and Tutorial. Umetri AB, Box 7960, S-90719 Umeå, Sweden.

Wold, S., Martens, H., Wold, H., 1983. The multivariate calibration problem in chemistry solved by the PLS method. In: Ruhe, A., Kåstrøm, B. (Eds.), Proceedings of the Conference Matrix Pencils, March 1982, Lecture Notes in Mathematics, Springer, Heidelberg, pp. 286–293.

Chapitre 3

Données censurées et régression PLS généralisée

Au début des années 80, Aitkin [2,4], Clayton et Cuzick [27], ou Whitehead [154], ont montré qu'il était possible d'utiliser le logiciel GLIM [3] dédié aux modèles linéaires généralisés dans le cadre des données de survie. Ils ont utilisé le fait que les vraisemblances pour les distributions de survie exponentielles et les lois de Poisson étaient équivalentes conduisant (principe de vraisemblance) à des estimations identiques pour les paramètres. La reformulation des modèles de survie sous la forme de modèles linéaires généralisés va nous permettre dans un premier temps d'utiliser l'algorithme PLS-GLR avec des données censurées. En particulier Bastien et Tenenhaus [12] ont proposé en 2002 dans le cadre de modèles PLS-GLR un modèle de régression PLS-Poisson comme alternative au modèle de Cox. On verra ensuite dans le cadre du modèle de Cox que l'algorithme PLS-GLR peut être utilisé directement sans transformation préalable des données.

3.1 Les modèles de survie

On peut faire remonter l'analyse des données de survie à 1693 avec l'astronome Halley [6] qui après une étude des relevés d'état civil de Londres donna les premières tables de mortalité et enseigna le moyen d'y lire la probabilité de survie d'un individu. Si dans les années qui ont suivi la seconde guerre mondiale ont s'est intéressé à l'analyse des données de survie plus pour des applications industrielles en utilisant des modèles paramétriques avec des lois exponentielles ou de Weibull, ce n'est que plus récemment, motivées par des applications

médicales, que sont apparues les méthodes non-paramétriques (Kaplan-Meier 1958, pour l'estimation non-paramétrique d'une fonction de survie) et semi-paramétriques (Cox 1972).

Les modèles de survie forment une classe de méthodes statistiques qui ont pour but d'étudier le nombre et la répartition des temps d'apparition des évènements. On peut s'intéresser à des modèles où l'on ne considère que le temps d'apparition des évènements, mais on s'intéresse plus généralement à des modèles où le risque d'apparition d'un évènement dépend de covariables. On retrouve ainsi l'expression d'un modèle de régression. Une des principales caractéristiques des données de survie est de posséder des données censurées. Une observation sur une variable T est censurée à droite si ce que l'on sait sur T est $T \geq c$, i.e. l'observation prend fin avant que l'évènement ne survienne. Les données peuvent aussi être censurées par intervalle si on ne connaît pas l'instant précis d'apparition de l'évènement mais seulement qu'il a eu lieu dans un intervalle donné. Les approches classiques en analyse des données de survie sont de type stochastique, le temps d'apparition d'un évènement est supposé être la réalisation d'un processus aléatoire associé à une distribution particulière.

3.1.1 Processus stochastiques

Il y a différentes façons de caractériser ces processus aléatoires. On peut les caractériser par la probabilité de survie $S(t)$, la densité $f(t)$, ou la fonction de risque $h(t)$.

$$S(t) = \mathrm{P}(\mathrm{T} > t)$$

$$F(t) = 1 - \mathrm{S}(t)$$

$$f(t) = \frac{dF(t)}{dt} = - \frac{dS(t)}{dt}$$

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq \mathrm{T} < t + \Delta t \ / \ \mathrm{T} \geq t)}{\Delta t} = \frac{f(t)}{S(t)}$$

La dernière caractérisation est très utile. C'est une densité conditionnelle, la probabilité qu'un évènement apparaisse dans l'intervalle (t, $t + \Delta t$) sachant qu'il n'était pas encore apparu à l'instant t. On peut aussi l'interpréter comme un taux : le nombre d'évènements par intervalle de temps. Ou bien si on prend $1/h(t)$, on peut estimer le temps moyen entre deux évènements. Si on fait varier la fonction de risque on obtient différents modèles de survie que l'on peut exprimer en fonction de covariables $x_1, ..., x_k$ et du temps t. Par exemple :

Exponentiel : $\ln h(t)=\mu+\beta_1 x_1+...+\beta_k x_k$

Gompertz : $\ln h(t)=\mu+\alpha\, t+\beta_1 x_1+...+\beta_k x_k$

Weibull : $\ln h(t)=\mu+\alpha \ln t+\beta_1 x_1+...+\beta_k x_k$

Ces trois modèles font partie d'une classe de modèles dits à risques proportionnels, les fonctions de risque pour deux individus étant parallèles au cours du temps.

3.1.2 Le modèle de Cox

C'est le modèle le plus utilisé dans l'analyse des données de survie. Il s'écrit sous la forme:

$$\ln h(t) = \ln h_0(t) + \beta_1 x_1 + ... + \beta_k x_k$$

Il permet d'estimer les coefficients des covariables pour tout modèle à risques proportionnels sans avoir à spécifier le type exact du modèle, i.e. $h_0(t)$, grâce à l'estimation par maximum de vraisemblance partielle. C'est pourquoi il est appelé semi-paramétrique. Il est pour cette raison très robuste par rapport aux modèles paramétriques. Le modèle de Cox peut être stratifié pour prendre en compte des facteurs de nuisance. Il permet la prise en compte de covariables dépendant du temps, et s'adapte aux données aussi bien continues que discrètes. Il ne s'intéresse qu'à l'ordre d'apparition des évènements et non à leur valeur.

Risques proportionnels et vraisemblance partielle

Dans l'article princeps de 1972, qui est l'article le plus référencé en statistique, Cox fait deux propositions. Le terme "Modèle de Cox" fait référence à ces deux points.

1) Il propose un modèle auquel on fait généralement référence sous le nom de "Proportional Hazards Model" même si ce modèle peut se généraliser facilement aux risques non proportionnels.

$$\frac{h_i(t)}{h_j(t)} = \exp\left[\beta_1\left(x_{i1}-x_{j1}\right)+...+\beta_k\left(x_{ik}-x_{jk}\right)\right]$$

$h_0(t)$ a disparu du ratio, les fonctions de risque pour deux individus sont parallèles au cours du temps.

2) Il propose une nouvelle méthode d'estimation qui fut appelée ensuite "Partial Likelihood" ou "Maximum Partial Likelihood".
La fonction de vraisemblance pour le modèle à risques proportionnels peut se factoriser en deux parties :

- une partie qui dépend à la fois de $h_0(t)$ et de β
- une autre partie qui ne dépend que de β

La vraisemblance partielle élimine la première partie et traite la deuxième comme si on avait une fonction de vraisemblance ordinaire. Comme on perd une partie de l'information sur β les estimations ne sont pas à variance minimale. Les écarts types sont un peu supérieurs à ceux que nous aurions calculés en prenant en compte la fonction de vraisemblance complète. La perte est cependant faible dans la plupart des cas comme cela a été montré par Efron en 1977 [43]. En contrepartie on gagne en robustesse. La vraisemblance partielle s'exprime sous la forme :

$$\mathrm{PL}(\beta) = \prod_{i=1}^{n} \left[\frac{\exp(\beta' x_i)}{\sum Y_{ij} \exp(\beta' x_j)} \right]^{\delta_i}$$

avec $Y_{ij} = 1$ si $t_j \geq t_i$

$Y_{ij} = 0$ si $t_j < t_i$

$\delta_i = 0$ si l'évènement est censuré

ce qui peut aussi s'écrire sous la forme :

$$\prod_{i=1}^{D} \frac{\exp(\beta' x_i)}{\sum_{j \in R(t_i)} Y_{ij} \exp(\beta' x_j)}$$

avec D le nombre d'instants ou l'évènement est apparu et $R(t_i)$ le sous-ensemble des individus à risque à l'instant t_i.

Estimation :

Une fois la fonction de vraisemblance partielle établie il faut la maximiser en β.

Soit : $L(\beta) = \ln(PL(\beta))$

$$U(\beta) = \frac{\partial L}{\partial \beta} = \left(\frac{\partial L}{\partial \beta_1}, \ldots, \frac{\partial L}{\partial \beta_p} \right)$$ le vecteur score des dérivées premières de $L(\beta)$

$$I(\beta)_{i,j} = - \frac{\partial^2 L(\beta)}{\partial \beta_i \partial \beta_j}$$ le $(i,j)^{\text{ème}}$ élément de la matrice d'information de Fisher $I(\beta)$

$I^{-1}(\beta)$ est un estimateur asymptotiquement consistant de la matrice de variance-covariance de β. Le $\hat{\beta}$ qui maximise la vraisemblance partielle est solution de $U(\beta)=0$.

Il n'y a pas de solution explicite à cette équation, on utilise généralement l'algorithme de Newton-Raphson [28]. A partir d'une valeur initiale l'algorithme calcule itérativement jusqu'à convergence :

$$\hat{\beta}_{(k+1)} = \hat{\beta}_{(k)} + I^{-1}(\hat{\beta}_{(k)}) U(\hat{\beta}_{(k)})$$

A partir de $\hat{\beta}$ on peut estimer $\hat{H}_0(t) = \int_0^t h_0(u)du$ en utilisant l'estimateur de Breslow:

$$\hat{H}_0(t,x) = \sum_{i/t_i<t} \left(\sum_{l \in R(t_l)} \exp x_l' \hat{\beta} \right)^{-1}$$

La fonction de survie pour un individu de covariables x est alors estimée par :

$$S(t,x) = \exp\left(-\hat{H}_0(t,x) \exp(x' \hat{\beta}) \right)$$

Inférence sur les paramètres :

Soit H_0 : $\beta = \beta_0$, les statistiques présentées ci-dessous pour tester cette hypothèse suivent une loi du χ^2 à p degrés de liberté.

- Statistique du rapport de vraisemblance : $2\left(L(\hat{\beta}) - L(\beta_0)\right)$
- Statistique de Wald : $(\hat{\beta} - \beta_0)' I(\hat{\beta})(\hat{\beta} - \beta_0)$
- Statistique du score : $U'(\beta_0) I^{-1}(\beta_0) U(\beta_0)$

Evènements simultanés

Dans le cas d'évènements simultanés, la formulation de la vraisemblance est légèrement modifiée. En particulier Breslow et Efron ont proposés des méthodes approximatives pour prendre en compte les *ex aequo* [6]. Il existe aussi des formulations exactes de la vraisemblance partielle suivant que le processus a un sous-jacent continu ou discret.

Méthodes approximatives :

Soit d_i le nombre d'évènements survenus au temps t_i.

Méthode approximative de Breslow :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{\left(\sum_{j \in R(ti)} \exp(\beta' x_j) \right)^{d_i}}$$

avec D le nombre d'évènements distincts et $S_i = \sum x_j$ sur les unités décédées à t_i.

Méthode approximative d'Efron :

$$L(\beta) = \prod_{i=1}^{D} \frac{\exp(\beta' S_i)}{\prod_{k=1}^{di} \left(\sum_{j \in R(t_i)} \exp(\beta' x_j) - \frac{k-1}{d_i} \sum_{j \in D(t_i)} \exp(\beta' x_j) \right)}$$

Méthodes exactes :

Sous-jacent continu: Si à un temps t on a k ex aequo on aura $k!$ possibilités pour arranger les ex aequo. La contribution à la vraisemblance sera $\sum_{j=1}^{k!} P(A_j)$ avec P(A_j) la probabilité associée au $j^{\text{ième}}$ arrangement.

Sous-jacent discret: Si à un temps t on a k ex aequo, et soit N_t la population à risque au temps t. Il y a $Q = \binom{N_t}{k}$ combinaisons de k individus parmi N_t. La contribution au temps t de la vraisemblance sera $L_1 = \frac{\Psi_1}{\sum_{j=1}^{Q} \Psi_j}$

avec : Ψ_j = produit des odds pour les individus dans la combinaison j $= \prod_{l=1}^{k} \frac{P_{lj}}{1-P_{lj}}$ avec P_{lj} la probabilité pour l'individu l de la combinaison j qu'un évènement survienne à l'instant t sachant qu'il n'est pas apparu précédemment.

Ψ_1 = produit des odds pour la combinaison observée à l'instant t.

S'il y a J instants ou les évènements ont lieu, il y aura J termes dans l'expression de la vraisemblance partielle : $PL = \prod_{j=1}^{J} L_j$

3.1.3 Modèle de Cox et vraisemblance monotone

Un problème de vraisemblance monotone peut apparaître avec le modèle de Cox quand une covariable ou une combinaison linéaire des covariables prend à chaque instant de décès les valeurs les plus élevées (ou les plus faibles) pour les individus qui décèdent que pour ceux qui survivent. On observe un phénomène similaire en régression logistique en cas de séparation ou quasi-séparation. Dans ce cas la vraisemblance converge vers une valeur finie alors qu'au moins un des paramètres tend vers plus ou moins l'infini. Heinze et Schemper (2001) [66] ont proposé une solution à ce problème en utilisant les résultats de Firth (1992) [51] pour réduire le biais des estimateurs du maximum de vraisemblance dans le cadre des modèles linéaires généralisés. La solution consiste à pénaliser la vraisemblance partielle par un terme correspondant à un *a priori* non informatif de Jeffreys.

On a : $LnL(\beta)^* = LnL(\beta) + 0.5Ln\left|I(\beta)\right|$

ou encore: $L(\beta)^* = L(\beta)\left|I(\beta)\right|^{0.5}$

avec: $\left|I(\beta)\right|$ le déterminant de la matrice d'information évalué à β.

La pénalisation de la vraisemblance partielle par l'*a priori* de Jeffreys tend à modifier l'estimation des paramètres vers une valeur où le déterminant de l'inverse de la matrice de variance-covariance (Hessien) est maximisé, ou encore où la variance des estimations des paramètres est minimisée. Sous SAS on peut simplement remplacer l'appel de la procédure PHREG par l'appel de la macro développée par Heinze et Ploner (2002) [67] et disponible sur le net.

3.2 Données de survie et processus de Poisson

Dans ce chapitre nous allons montrer comment on peut utiliser les modèles linéaires généralisés et en particulier la régression de Poisson dans le cadre des données de survie. On pourra ainsi utiliser l'algorithme PLS-GLR directement dans le cadre des données censurées. En particulier, utilisant les résultats de Whitehead (1980) [154] qui montra l'équivalence entre le modèle de Poisson pour la variable de censure et le modèle de Cox pour la fonction de risque, les vraisemblances étant proportionnelles, Bastien et Tenenhaus [12] ont proposé en 2002 dans le cadre de la régression PLS généralisée un modèle de régression PLS-Poisson comme alternative au modèle de Cox. Le modèle PLS-Cloglog a aussi été proposé [12] comme alternative au modèle de Cox dans le cas de données discrètes avec un sous-jacent continu.

3.2.1 Survie exponentielle et processus de Poisson pour la variable de censure

Nous allons montrer ci-dessous la proposition suivante :

Proposition :

Il y a équivalence entre les vraisemblances pour les distributions de survie exponentielles et les lois de Poisson.

Cette proposition implique (principe de vraisemblance) des estimations identiques pour les paramètres.

Démonstration :

Soit n temps de survie t_i et w_i l'indicateur de censure qui vaut 1 si l'évènement s'est produit et 0 sinon. La vraisemblance s'écrit:

$$L = \prod_{i=1}^{n} f(t_i)^{w_i} [1 - F(t_i)]^{1-w_i}$$

$$\ln(L) = \sum_i w_i \ln[f(t_i)] + \sum_i (1 - w_i) \ln[1 - F(t_i)] \qquad (1)$$

Soit x_i le vecteur des covariables pour le sujet i, $\lambda \exp(\beta' x_i)$ sa fonction de risque pour une loi exponentielle de paramètre $\lambda \exp(\beta' x_i)$. L'espérance du nombre d'évènements pour l'individu i est estimée par :

$$m_i = t_i \lambda \exp(\beta' x_i) \qquad (2)$$

En substituant (2) dans l'expression (1), $f(t_i)$ étant égale à :

$$\lambda \exp(\beta' x_i) \exp[-\lambda \exp(\beta' x_i) t_i]$$

On obtient :

$$\ln(L) = \sum_i w_i [ln(\lambda \exp \beta' x_i) - \lambda \exp \beta' x_i t_i] + \sum_i -(1-w_i)\lambda \exp \beta' x_i t_i$$

$$= \sum_i w_i [\ln\left(\frac{m_i}{t_i}\right) - m_i] + \sum_i -(1-w_i) m_i$$

$$= \sum_i w_i \ln(m_i)_i - \sum_i m_i - \sum_i w_i \ln(t_i)$$

Seuls les deux premiers termes font intervenir β. Ils sont identiques à la log-vraisemblance pour des variables aléatoires indépendantes w_i distribuées suivant une Poisson de paramètre m_i. Par conséquent il aurait été équivalent de maximiser la vraisemblance pour le modèle de Poisson suivant :

$$\ln(m_i) - \ln(t_i) = \ln(\lambda) + \beta' x_i$$

3.2.2 Survie exponentielle par morceaux et processus de Poisson pour la variable de censure: application au modèle de Cox

L'hypothèse de risque constant est souvent trop forte. On peut supposer maintenant le risque constant par intervalle. La fonction de risque devient :

$$h_k(t, x_i) = \lambda_k \exp(\beta' x_i)$$

pour t dans l'intervalle k.

Soit la table de contingence décrivant le nombre d'évènements avec une dimension pour le temps discrétisé et les autres dimensions associées aux covariables, on montre qu'un modèle log-linéaire de Poisson pour cette table est équivalent à un modèle log-linéaire pour la fonction de risque d'une distribution exponentielle par morceaux [132]. Si le nombre d'intervalles devient arbitrairement élevé, le modèle exponentiel par morceaux s'apparente essentiellement à une approche non-paramétrique. Cela revient à approcher la fonction de risque par une fonction en escalier. On converge par généralisation vers le modèle de Cox : $h(t,x_i) = h_0(t)\exp(\beta' x_i)$

3.2.3 Données de survie paramétriques et processus de Poisson pour la variable de censure

Au-delà de la distribution de survie de type exponentielle, Aitkin et Clayton [2], au début des années 80, ont montré comment ajuster des modèles de régression paramétriques aux données de survie. En supposant la fonction de risque de la forme $h(t,x) = h_0(t)\exp(\beta' x)$, avec $h_0(t)$ une fonction de risque paramétrique, ils ont montré que la vraisemblance était proportionnelle à celle d'un modèle de Poisson pour la variable de censure avec comme paramètre $\ln \mu_i = \ln H_0(t) + \beta' x$ correspondant à un modèle log-linéaire pour la fonction de risque.

Soit :

$$h(t,x) = h_0(t)\exp\left(\beta' x\right)$$

$$H(t,x) = H_0(t)\exp(\beta' x)$$

$$S(t,x) \;=\; \exp[-H(t,x)] = \exp[-H_0(t)\exp(\beta' x)]$$

$$f(t,x) \;=\; \text{h(t,x)S(t,x)}=h_0(t)\;\exp(\;\beta' x - H_0(t)\exp(\beta' x))$$

Alors la vraisemblance peut s'écrire sous la forme:

$$L = \prod_{i=1}^{n} f(t_i, x_i)^{w_i} S(t_i, x_i)^{1-w_i}$$

$$= \prod_{i=1}^{n} \left[h_0(t_i) \exp[\beta' x_i - H_0(t_i) \exp(\beta' x_i)] \right]^{w_i} \exp(-H_0(t_i) \exp(\beta' x_i))^{1-w_i}$$

$$= \prod_{i=1}^{n} \left[h_0(t_i) \exp(\beta' x_i) \right]^{w_i} \exp(-H_0(t_i) \exp(\beta' x_i))$$

$$= \prod_{i=1}^{n} \left[H_0(t_i) \exp(\beta' x_i) \frac{h_0(t_i)}{H_0(t_i)} \right]^{w_i} \exp(-H_0(t_i) \exp(\beta' x_i))$$

$$= \prod_{i=1}^{n} \left[\mu_i^{w_i} e^{-\mu_i} \right] \left[h_0(t_i) / H_0(t_i) \right]^{w_i} \qquad (3)$$

avec $\mu_i = H_0(t_i) \exp(\beta' x_i)$

Le premier terme de (3) correspond à la vraisemblance pour la variable aléatoire w_i qui suit une Poisson de paramètre μ_i :

$$\ln(\mu_i) = \ln H_0(t_i) + \beta' x_i.$$

Le modèle log-linéaire sur $H(t)$ est donc associé à un modèle log-linéaire sur μ_i. Il est intéressant de noter que la fonction $H_0(t)$ à la particularité de transformer l'échelle de temps en une échelle où la distribution des temps d'apparition des évènements est exponentielle [27].

3.3 Modèle de Cox et données discrètes

Jusqu'à présent nous avons considéré uniquement les évènements apparaissant de façon continue au cours du temps. Ce n'est pas toujours le cas et dans certaines situations les évènements ne pourront être observés qu'à des instants donnés. C'est par exemple le cas avec une étude clinique où les évènements sont recueillis à des visites qui ont lieu tous les deux mois. Il existe deux types de données discrètes. Le premier est associé à des données structurellement discrètes avec des évènements qui ne peuvent se produire qu'à des instants précis dans le temps. Le second type correspond à des données qui apparaissent discrètes du fait d'un manque de précision dans le recueil de l'information. A chacun de ces types sera

associé un modèle de régression binaire modélisant la probabilité qu'un évènement apparaisse dans chaque unité de temps. Cette modélisation sera utilisée ensuite dans le cadre de l'utilisation de la régression PLS généralisée avec des données de survie.

3.3.1 Données structurellement discrètes.

Lorsque les évènements ne peuvent survenir qu'à des instants réguliers dans le temps, Cox a proposé dans son article de référence de 1972 une généralisation de son modèle à risques proportionnels :

$$\frac{\lambda(t,x)dt}{1-\lambda(t,x)dt} = \exp(x\beta)\frac{\lambda_0(t)dt}{1-\lambda_0(t)dt} \qquad (4)$$

Dans le cas continu cette expression se résume à l'expression classique :

$$\lambda(t,x) = \lambda_0(t)\exp(x\beta)$$

L'expression (4) ne traduit plus un modèle à risques proportionnels, c'est un modèle à *odds* proportionnels. On utilisera un modèle de régression logistique pour modéliser la probabilité d'apparition d'un évènement dans chaque intervalle de temps.

Les données pour chaque individu sont dupliquées en autant d'intervalles où l'individu est à risques. Sur chacun de ces intervalles la variable réponse vaut 1 ou 0 suivant que l'évènement est apparu dans cet intervalle ou non.

Soit P_{it} la probabilité pour l'individu i qu'un évènement survienne dans l'intervalle t sachant qu'il n'est pas apparu précédemment: P_{it} = Prob(y_{it} = 1 / y_{i1} = 0, …, $y_{i,t-1}$ = 0). C'est la probabilité conditionnelle d'observer « y_{it} = 1 » sachant que « y_{i1} = 0, …, $y_{i,t-1}$ = 0 ». C'est une version discrète de la fonction de risque.

$$\ln\left[\frac{P_{it}}{1-P_{it}}\right] = \alpha_t + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

3.3.2 Données discrètes avec un sous-jacent continu.

Si nous supposons que les données continues sous-jacentes sont générées par un modèle de Cox, Prentice et Gloeckler ont montré en 1978 [112] que le modèle est un modèle linéaire généralisé avec une fonction de lien de type « *Complementary Log Log* »

$$\ln\left[-\ln(1-P_{it})\right] = \alpha_t + \beta_1 x_{1i} + \ldots + \beta_p x_{pi}$$

Supposons une distribution continue avec une densité $f(t)$, une fonction de survie $S(t)$ et une fonction de risque $h(t)$ regroupée en s intervalles $A_j =]t_{j-1}, t_j]$ $j = 1,\ldots,s$, avec $t_0 = 0$, $t_s = \infty$.

Soit f_j, s_j, et h_j les probabilités, fonctions de survie, et fonctions de risque respectivement pour les distributions discrètes associées.

Alors:

$$f_j = \Pr(T \in A_j) = S(t_{j-1}) - S(t_j) = s_j - s_{j+1}$$

$$avec \quad s_j = \Pr(T > t_{j-1})$$

$$\begin{aligned} h_j &= \Pr\left(T \in A_j / T > t_{j-1}\right) \\ &= f_j / \left(f_j + f_{j+1} + \ldots + f_s\right) \\ &= f_j / s_j \end{aligned}$$

Remarques :

$$h_j = 1 - \frac{s_{j+1}}{s_j}$$

$$\frac{s_{j+1}}{s_j} = 1 - h_j \qquad (5)$$

$$s_{j+1} = \prod_{k=1}^{j} \frac{s_{k+1}}{s_k}$$

$$Ainsi: \quad s_j = \prod_{k=1}^{j-1} \left(1 - h_k\right) \qquad (6)$$

Soit w_{ij} la variable de censure et h_{ij} la fonction de risque pour le i^{ème} individu dans le j^{ème} intervalle, la contribution à la vraisemblance est :

$$f_{ij(i)}^{w_{ij(i)}} S_{ij(i)+1}^{1-w_{ij(i)}} \tag{7}$$

$$= (h_{ij(i)} S_{ij(i)})^{w_{ij(i)}} (1-h_{ij(i)})^{1-w_{ij(i)}} \qquad \textit{d'après (5)} \text{ et (7)}$$

$$= S_{ij(i)} h_{ij(i)}^{w_{ij(i)}} (1-h_{ij(i)})^{1-w_{ij(i)}}$$

$$= h_{ij(i)}^{w_{ij(i)}} (1-h_{ij(i)})^{1-w_{ij(i)}} \prod_{j=1}^{j(i)-1} (1-h_{ij(i)}) \qquad \textit{d'après (6)}$$

$$= \left(\frac{h_{ij(i)}}{1-h_{ij(i)}}\right)^{w_{ij(i)}} (1-h_{ij(i)}) \prod_{j=1}^{j(i)-1} (1-h_{ij})$$

$$= \left(\frac{h_{ij(i)}}{1-h_{ij(i)}}\right)^{w_{ij(i)}} \prod_{j=1}^{j(i)} (1-h_{ij})$$

$$= \prod_{j=1}^{j(i)} \left(\frac{h_{ij}}{1-h_{ij}}\right)^{w_{ij}} (1-h_{ij}) \qquad w_{ij(i)} = 0 \text{ si } j < j(i)$$

$$= \prod_{j=1}^{j(i)} h_{ij}^{w_{ij}} (1-h_{ij})^{1-w_{ij}}$$

Pour le modèle à risque proportionnel :

$$h(t,x) = \lambda_0(t) e^{\beta' x}$$
$$S(t,x) = \exp(-H_0(t) e^{\beta' x})$$

ce qui donne sous forme discrète :

$$s_j = \exp(-H_0(t_{j-1}) e^{\beta' x})$$
$$h_j = 1 - \frac{s_{j+1}}{s_j}$$
$$= 1 - \exp\left[-e^{\beta' x}\left\{H_0(t_j) - H_0(t_{j-1})\right\}\right]$$

ainsi :

$$h_{ij} = 1 - \exp\left(-e^{\beta' x_i + \psi_j}\right)$$

avec

$$\psi_j = \ln\left(H_0(t_j) - H_0(t_{j-1})\right)$$

La vraisemblance sur l'ensemble des individus est donc :

$$L = \prod_{i=1}^{n} \prod_{j=1}^{j(i)} h_{ij}^{w_{ij}} \left(1 - h_{ij}\right)^{1-w_{ij}} = \prod_{j=1}^{s} \prod_{i \in R_j} h_{ij}^{w_{ij}} \left(1 - h_{ij}\right)^{1-w_{ij}}$$

avec R_j l'ensemble des individus à risque dans le k-ème intervalle. C'est un produit de Bernoulli, une pour chaque individu par intervalle. La probabilité conditionnelle de survenue de l'événement pour le $i^{\text{ème}}$ individu dans le $j^{ème}$ intervalle est h_{ij}.

On a : $$\ln\left\{-\ln\left(1 - h_{ij}\right)\right\} = \beta' x_i + \psi_j$$

Ainsi le modèle à temps discrets associé à un modèle sous-jacent continu à risques proportionnels peut être ajusté en traitant les observations comme étant indépendantes d'un intervalle à l'autre. La censure est la variable réponse avec une distribution de Bernoulli. La fonction de lien est de type « *Complementary log-log* ». Les paramètres ψ_j associés aux intervalles doivent être estimés en incluant un facteur bloc identifiant l'intervalle dans le modèle. L'interprétation des coefficients de régression en termes de risques relatifs est identique à celle du modèle de Cox sous-jacent.

3.4 Capacité prédictive d'un modèle de survie

Les méthodes de validation croisée basées sur la somme des carrés des résidus utilisées pour évaluer les qualités prédictives d'un modèle ne sont plus valides dans le cas de données censurées. Les données censurées, qui sont caractérisées à la fois par la durée de suivi du sujet et son statut à la fin de la période de suivi, partagent à la fois les caractéristiques des données continues et binaires. Par conséquent il était naturel d'essayer d'étendre aux données censurées les critères utilisés avec les données binaires (courbes ROC). En particulier, Heagerty et Zheng (2003)[64] ont proposés un critère de validation du modèle basé sur la notion de courbes ROC dépendant du temps en utilisant une représentation de type processus pour les données de survie. Cette approche actuellement très largement utilisée en particulier en bioinformatique [83,84,85,66,128] sera utilisée par la suite pour comparer les qualités prédictives des différents algorithmes. Nous allons la décrire dans le paragraphe ci-dessous en rappelant dans un premier temps la notion de courbe ROC.

Courbes ROC

Quand la variable réponse est binaire, la qualité d'un modèle en prédiction est généralement évaluée par le taux de bien classés. Elle est caractérisée par 2 critères complémentaires que sont la sensibilité $p(\hat{p}_i > c / y_i = 1)$ et la spécificité $p(\hat{p}_i \leq c / y_i = 0)$ où $\hat{p}_i$ est la prédiction du modèle et *c* un seuil tel que : $\hat{y}_i = 1 \; si \; \hat{p}_i > c$ et $\hat{y}_i = 0 \; si \; \hat{p}_i \leq c$

En faisant varier *c* on peut tracer une courbe avec en abscisse la sensibilité (vrais positifs) et en ordonné 1-spécificité (faux positifs). La surface sous la courbe mesure la probabilité que la valeur prédictive $\hat{p}_i$ pour un cas $(y_i = 1)$ tiré au hasard soit supérieure à celle d'un témoin $(y_i = 0)$ tiré au hasard. On montre que l'AUC est directement lié à la statistique *U* du test de Mann-Whitney (Hanley et Mc Neil, 1982) [62]. Comme le précisent Heagerty et Zheng, les courbes ROC sont particulièrement utiles pour discriminer la capacité de différents biomarqueurs potentiels. En particulier un marqueur sera dit uniformément de meilleur diagnostic s'il possède une meilleur sensibilité quelque soit la spécificité.

Extension aux données de survie

Si l'on considère les données de survie dans une approche processus, elles peuvent être vues sous la forme d'une variable binaire dépendant du temps :

$$N_i(t) = 1(T_i \leq t)$$

avec T_i la durée de survie de l'individu *i* et 1 une fonction indicatrice.

En utilisant cette approche processus, différentes extensions des concepts de sensibilité et spécificité ont été proposées [64]. Une version plus spécifiquement adaptée aux modèles de régression avec données de survie a été proposée par Heagerty et Zheng (2003) et est décrite ci-dessous.

Soit M_i la valeur d'un marqueur prédictif de la survie, qui peut être égale à la partie explicative $x_i'\beta$ d'un modèle de régression. M_i peut par exemple être issu de la phase de calibration d'un modèle PLS-Cox.

Soit : $dN_i(t) = N_i(t) - N_i(t-)$

Heagerty et Zheng définissent la sensibilité et la spécificité comme :

$$sensibilité(c,t) = P(M_i > c / T_i = t) = P(M_i > c / dN_i(t) = 1)$$
$$spécificité(c,t) = P(M_i \leq c / T_i > t) = P(M_i \leq c / N_i(t) = 0)$$

La sensibilité mesure l'espérance de la proportion d'individus avec $M_i > c$ parmi les sujets qui décèdent à l'instant *t*.
La spécificité mesure l'espérance de la proportion de sujets avec $M_i \leq c$ parmi les individus qui sont à risque à l'instant *t* et survivent au delà de *t*.

La binarisation de l'ensemble des sujets à risque à chaque instant *t* entre ceux qui décèdent et ceux qui survivent, implicite dans les deux définitions précédentes, apparaît étroitement liée à la vraisemblance partielle du modèle de Cox et permettra d'estimer sensibilité et spécificité à partir des estimations du modèle.

Estimation de la sensibilité et de la spécificité

Soit $h(t / M_i) = h_0(t)\exp(M_i'\gamma)$

Soit $\pi_i(\gamma,t) = \dfrac{R(t_i)\exp(M_i'\gamma)}{\sum_j R(t_j)\exp(M_j'\gamma)}$

avec $R(t_j)$ défini en 3.1.2 comme le sous-ensemble des individus à risque à t_j, Xu et O'Quigley (2000) ont montré que l'on pouvait estimer la sensibilité $\widehat{TP}_t(c)$ à partir de $\pi_i(\gamma,t)$.

$$\widehat{TP}_t(c) = P(M_i > c / T_i = t) = \sum_k 1(M_k > c)\pi_k(\gamma,t)$$

Pour estimer $\widehat{FP}_t(c)$*=1-sensibilité*, ils utilisent un estimateur empirique:

$$\widehat{FP}_t(c) = P(M_i > c / T_i > t) = \sum_k 1(M_k > c)R_k(t+)/W^r(t+)$$

avec:

$R_k(t)$=1 si l'individu *k* est à risque à l'instant *t*, et 0 sinon.

$R_k(t+) = \lim_{\delta \to 0^+} R_k(t+\delta)$, et $W^r(t+) = \sum_k R_k(t+)$ qui correspond au nombre de sujets à risque à l'instant *t* et qui survivent au-delà de cet instant.

Estimation des aires sous la courbe en fonction de t

Etant donné les estimations de $\hat{TP}_t(c)$ et de $\hat{FP}_t(c)$, la courbe ROC à l'instant *t* peut être estimée par :

$$\hat{ROC}_t(p) = \hat{TP}_t\left(c_p\right)$$

où $$c_p = \inf_c \left\{ c : \hat{FP}_t(c) \le p \right\}$$

L'aire sous la courbe ROC à l'instant *t*, AUC(*t*), est donnée par l'intégration suivant *p* :

$$\hat{AUC}(t) = \int \hat{ROC}_t(p)dp$$

Pratiquement, cette intégrale peut être approchée en utilisant une méthode standard d'intégration numérique comme la méthode des trapèzes.

3.5 Données censurées et régression PLS généralisée.

Le tableau ci-dessous résume les différents cas de figure observés dans l'analyse des données censurées avec les modèles de régression PLS associés.

Table 1: Régression PLS et données censurées

	Modèles de survies	**Modèles linéaires généralisés**	**PLS-GLR**
continues	$h_0(t)$ paramétré	POISSON (Aitkin & Clayton 80)	*POISSON PLS*
	$h_0(t)$ non paramétré (Cox)	POISSON (Whitehead 80) *(approx : exponentielle par morceaux)*	*POISSON PLS*
discrètes	sous-jacent continu	CLOGLOG (Prentice & Glockler 78)	*CLOGLOG PLS*
	sous-jacent discret	LOGIT (Cox 72) *(discrete Cox model : prop odds model)*	*LOGIT PLS*

Cependant la reformulation en prenant la variable de censure comme variable de réponse accroît la dimension du problème. En effet des observations multiples doivent être crées pour chaque individu, autant que d'intervalles entre évènements successifs (décès) pendant lesquels l'individu est à risque. Nous allons voir dans le paragraphe suivant que nous pouvons éviter cette complication en étendant directement l'algorithme PLS-GLR au modèle de Cox.

3.6 PLS-Cox

Nous avons vu précédemment que les modèles de régression pour données censurées pouvaient être traités dans le cadre des modèles linéaires généralisés. L'algorithme PLS-GLR décrit précédemment ne nous contraint cependant pas à rester dans le cadre de la famille exponentielle et peut s'étendre directement à tout modèle de régression linéaire généralisé et au modèle de Cox en particulier. Cette approche se révèle être beaucoup plus simple d'un point de vue calculatoire, les unités expérimentales devenant les individus et non les individus par intervalle comme dans les approches précédentes.

3.6.1 Algorithme PLS-Cox

L'algorithme PLS-Cox comprend trois étapes :

1) La recherche des composantes PLS t_h
2) Le modèle de Cox sur les composantes PLS t_h
3) L'expression du modèle de Cox en fonction des variables d'origine
4) La validation bootstrap des coefficients du modèle de Cox final

Nous allons décrire en détail la première étape.

Recherche des composantes PLS t_h

On note X la matrice dont les colonnes sont formées des valeurs des p variables explicatives x_j. On recherche successivement m composantes PLS orthogonales t_h combinaisons linéaires des x_j.

Recherche de la première composante PLS t_1

Etape 1 : Calculer le coefficient a_{1j} de x_j dans le modèle de Cox sur x_j pour chaque variable x_j, $j = 1$ à p.

Etape 2 : Normer le vecteur colonne a_1 formé des a_{1j} : $w_1 = a_1/||a_1||$

Etape 3 : Calculer la composante $t_1 = Xw_1/w_1'w_1$

Recherche de la deuxième composante PLS t_2

Etape 1 : Calculer le coefficient a_{2j} de x_j dans le modèle de Cox sur t_1 et x_j.

Etape 2 : Normer le vecteur colonne a_2 formé des a_{2j} : $w_2 = a_2/||a_2||$

Etape 3 : Calculer le résidu X_1 de la régression linéaire de X sur t_1 : $X_1 = X - t_1 p_1'$

Etape 4 : Calculer la composante $t_2 = X_1w_2/w_2'w_2$.

Etape 5 : Exprimer la composante t_2 en fonction de X : $t_2 = Xw_2^* = X(I - w_1 p_1')w_2$

Recherche de la h-ième composante PLS t_h

On a obtenu aux étapes précédentes les composantes PLS $t_1,\ldots, t_{h-1}$. On obtient la composante t_h en prolongeant la recherche de la deuxième composante.

Etape 1 : Calculer le coefficient a_{hj} de x_j dans le modèle de Cox sur $t_1, t_2, \ldots, t_{h-1}$ et x_j.

Etape 2 : Normer le vecteur colonne a_h formé des a_{hj} : $w_h = a_h/||a_h||$

Etape 3 : Calculer le résidu X_{h-1} de la régression linéaire de X sur $t_1,\ldots,t_{h-1}$: $X_{h-1} = X - \sum_{i=1}^{h-1} t_i p_i'$

Etape 4 : Calculer la composante $t_h = X_{h-1}w_h/w_h'w_h$.

Etape 5 : Exprimer la composante t_h en fonction de X :

$$t_h = Xw_h^* = X\left(\prod_{i=1}^{h-1}(I - w_i p_i')\right)w_h$$

3.7 Autres approches

En utilisant aussi les résultats de Whitehead, Park et al. [108], ont aussi proposé d'utiliser la régression PLS généralisée de Marx dans le cas d'une régression de Poisson pour traiter des données de survie. Cependant des problèmes de convergence ont été observés par Li et Gui (2004) [84] avec l'algorithme de Park et al. pour un nombre élevé de descripteurs.

Une autre approche a aussi été proposée par Nguyen and Rocke en 2002 [103] dans le cadre de l'analyse de données d'expression issues de bio-puces. Cependant leur approche en deux phases ne prend pas en compte l'information sur la censure dans l'estimation des composantes PLS, induisant ainsi un biais dans l'estimation des coefficients du modèle [19].

En 2004, Li et Gui [84] ont proposé un algorithme appelé « Partial Cox Regression » (PCR) très similaire à PLS-Cox, les deux algorithmes différant légèrement dans la construction des composantes PLS. Résultant d'une généralisation littérale de l'algorithme de Garthwaite, PCR construit les composantes PLS en pondérant les descripteurs proportionnellement à des pseudo-covariances généralisées alors que PLS-Cox, plus dans l'esprit de l'algorithme PLS-NIPALS, utilise des pentes généralisées. Il est à noter que PCR contrairement à PLS-Cox ne permet pas la prise en compte des valeurs manquantes.

3.8 Applications

Deux applications du modèle PLS-GLR dans le cadre des données censurées sont présentées. La première application [12] à trait à la détermination de facteurs de risque associés à l'apparition précoce des premiers cheveux blancs. Cette application correspond à la première utilisation des principes de la régression PLS dans le cadre des données de survie et du modèle de Cox en particulier. Cette communication a été présentée en octobre 2001 au congrès PLS'01 à Capri. Elle a été complétée ensuite dans l'article de CSDA 2005 (cf 2.4) pour traiter plus spécifiquement le cas de descripteurs catégoriels et valider le modèle par ré-échantillonnage bootstrap.

La seconde application [15] utilise des données issues de bio-puces pour prédire à partir d'un profile d'expression génomique la probabilité de survie de patients atteints de lymphomes à grande cellules *B* diffuses.

PLS generalized linear regression
Application to the analysis of life time data

Philippe BASTIEN[1] , Michel TENENHAUS[2]
(1) L'Oréal Research, Clichy, France (pbastien@recherche.loreal.com)
(2) HEC School of management, Jouy-en-Josas, France (tenenhaus@hec.fr)

SUMMARY

Problems encountered in multiple regression due to multicolinearity or missing data can be overcome by using PLS regression. Several versions of the PLS regression algorithm exist. In this paper, we present a new version of this algorithm which can be extended to generalized linear regression models such as ordinal or multinomial logistic regression, generalized linear models, and Cox regression models. An application with discrete censored data concerning the occurrence of prematurely graying hair in more than 4000 adult males is presented. The Cox regression model has been used. For discrete data summarizing a continuous time process, this model is equivalent to a generalized linear model using a complementary log-log link function.

key words : PLS regression; Generalized linear regression; Life time data; Survival analysis; Cox model.

I PLS generalized linear regression

The PLS regression method in its basic form (PLS 1) applies for one single response variable Y and is non iterative (Tenenhaus 1998). It is particularly useful when the X variables are closely correlated with each other.

Marx (1996) proposed a generalization of the PLS algorithm to generalized linear models (Dobson 1990). He used the fact that, in the

context of the exponential family, maximum likelihood estimates are obtained by an iterative weighted least squares procedure. His approach consisted of replacing the iterative weighted least squares step by a sequence of PLS regressions.

Esposito Vinci and Tenenhaus (2001) propose a much simpler approach consisting of adapting the PLS regression algorithm described by Wold, Martens and Wold, (1983) to the case of binary or ordinal logistic regression. In this paper, we present a generalization of this algorithm to generalized linear regression.

The generalized linear regression model is described below :

Let Y be a response function and τ some parameter related to the distribution of Y. For example τ can be the mean μ, the probability π of occurrence of an event, or a hazard function h(t). Let $x_1,..,x_p$ be the explanatory variables. Let g be the link function.

The generalized linear regression is the following :

(1) $$g(\tau) = \sum_j \beta_j x_j$$

The model parameters β_j are estimated by maximum likelihood.

The PLS generalized linear regression algorithm is described below:

Let X_0 be the matrix of the standardized input variables $x_{01},..,x_{0p}$.

Determination of the first PLS component t_1

1: For each $j = 1$ to p, compute the regression coefficient w_{1j} of x_{0j} in the generalized linear regression model (1) of y on x_{0j}.
2: The vector $w_1 = (w_{11},\ldots,w_{1p})'$ is normalized.
3: $t_1 = X_0 w_1 / w_1' w_1$

Determination of the second PLS component t_2

1: Compute the residual X_1 of the regression of X_0 on t_1.
2: Compute the regression coefficient w_{2j} of x_{1j} in the generalized linear regression of y on t_1 and x_{1j}, for *j = 1* to *p*.
3: The vector $w_2 = (w_{21},\ldots,w_{2p})'$ is normalized.
4: $t_2 = X_1w_2/w_2'w_2$

This procedure is iterated for the other PLS components t_h.

When some data are missing, the calculation of the PLS components is modified by applying the NIPALS algorithm principle. The calculation of the numerator and denominator of $t_h = X_{h-1}w_h/w_h'w_h$ is carried out on the basis of the available data. The denominator is in fact calculated only on data available to the numerator.

At each step the generalized linear regression of y on components $t_1,\ldots, t_h$ is carried out. We stop the procedure and the component t_h is not included in the model if it is not significant. The number *m* of PLS components t_h can also be determined by cross-validation.. The final regression equation is obtained by expressing the generalized linear regression of y on $t_1,\ldots, t_m$ as a function of the original variables.
In the case of ordinary multiple regression, this algorithm gives usual PLS regression when there is no missing data. When some missing data are present this algorithm takes into account the correlation between the PLS components. This is not the case for the usual PLS regression.

II Application

II.1 Material and objectives

Begun in 1994 by Professor Serge Hercberg, the cohort study SUVIMAX *(«SUpplémentation en VItamines et Minéraux Antioxydants»)* assumed the task of evaluating the nutritional state of

the population in France and assessing the influence of antioxidant minerals and vitamins on various indications of the state of health such as cancer or cardiovascular disease (Hercberg 1997). This study, scheduled for completion after 8 years in 2002, has enrolled more than 12000 volunteers aged between 35 and 65 years and representative of the population of France, half of whom received antioxidants and the other half a placebo.

We conducted a study of the state of health of hair and nails in 10323 subjects (4057 men and 6266 women) from this cohort. On the basis of responses to a questionnaire covering more than 150 items, an attempt was made to demonstrate the risk factors associated with the premature onset of graying hair in men.

The response variable is the age at which the first gray hairs appear in men. This is a discrete variable divided into 6 stages: less than 30 years, 31 to 35 years, 36 to 40 years, 46 to 50 years and over 50 years. The Cox regression model was used to fit the data. As the process of the advent of the first gray hairs is intrinsically continuous, the Cox regression model is equivalent to a generalized linear model using a complementary log-log link function (Allison, 1995, page 216).

Family factors, natural hair color, and hair thickness are predictor variables significantly related to the time at which the first gray hairs appear. A Cox model was used to relate the response variable to the predictors.

II 2 Discrete life time data

The basic ideas described in Allison (1995) are simple. Each individual's survival history is broken down into a set of discrete time units that are treated as distinct observations. After pooling these observations, the next step is to estimate a binary regression model to predict whether an event did or did not occur in each time unit. Even if multiple observations are created for a single individual, there is no concern about dependence here. The creation of multiple observations

follows directly from factoring the likelihood function for the data. It follows immediately from the definition of conditional probability.

Two models for processing these data exist: one that assumed that events really occur at the same discrete time and another that assumed that ties result from imprecise measurement.

When events can only occur at regular discrete points in time, the appropriate model is a logit model.

Let P_{it} be the conditional probability that individual i has an event at time t, given that the event has not already occurred in the case of that individual. The model states that P_{it} is related to the predictors by the logit regression equation:

(2) $$\log \frac{P_{it}}{1-P_{it}} = \alpha_t + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

For most applications, however, ties occur because event times are measured coarsely even though events can actually occur at any point in time. If we now assume that events are generated by Cox's proportional hazards model it follows that:

(3) $$\log\left[-\log(1-P_{it})\right] = \alpha_t + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$

i.e. the probability of an event at some well-defined interval of time is given by the complementary log-log model. Furthermore the β coefficients in the model have the same relative risk interpretation as in the underlying Cox proportional hazard model. The expression $-\log(1-P_{it})$ in formula 3 is the discrete version of the hazard function $h_i(t)$. Hence, formula 3 is the discrete version of the Cox proportional hazard model :

(4) $$\log\left[h_i(t)\right] = \alpha_t + \beta_1 x_{1i} + ... + \beta_p x_{pi}$$