

Comparaison de variantes de régressions logistiques PLS et de régression PLS sur variables qualitatives : application aux données d'allélotypage

Nicolas Meyer¹, Myriam Maumy-Bertrand² et Frédéric Bertrand²

Title: Comparing the linear and the logistic PLS regression with qualitative predictors: application to allelotyping data

Résumé : Un microsatellite est une séquence non-codante de l'ADN. L'allélotypage consiste à rechercher le statut normal ou altéré d'un ensemble prédéfini de microsatellites, en général dans une cellule cancéreuse. Les données d'allélotypage rassemblent donc une série de variables binaires décrivant l'état global des chromosomes de la cellule. Ces données sont généralement utilisées pour expliquer une caractéristique, elle aussi qualitative binaire, du sujet ou de la tumeur. Les données d'allélotypage sont caractérisées par un nombre de variables (microsatellites) pouvant dépasser le nombre de sujets et par la présence éventuelle de colinéarité entre les microsatellites. La compréhension des mécanismes de cancérogenèse implique également une description multivariée des données. Le traitement statistique de ces données suggère donc l'utilisation de la régression PLS. Les variantes PLS des régressions linéaire et logistique ne font pas d'hypothèses sur le type de données pouvant être analysées. Nous trouvons dans la littérature l'utilisation sur des variables toutes qualitatives de modèles prévus *a priori* pour des données quantitatives. L'absence d'hypothèse sur les données impliquent par ailleurs une validation des modèles par des méthodes de type validation-croisée ou bootstrap. Nous comparons ici les performances des variantes PLS des régressions linéaire et logistique sur des données toutes qualitatives.

Abstract: A microsatellite is a non-coding DNA sequence. Allelotyping consists in establishing the normal or altered status of a set of predefined microsatellites, generally in a tumor cell. Allelotyping data thus gather a series of binary variables that describes the global state of the cell chromosomes. These binary data are generally used to explain a characteristic, binary also, of the subject or of the tumor. Allelotyping data are characterised by their number of variables (microsatellites) being sometimes larger than the number of subjects and by the possible collinearity of two microsatellites. The understanding of cancerogenesis mechanisms implies also a multivariate description of the data. The statistical processing of these data thus suggest using PLS regression. PLS variants of linear and logistic regression make no assumptions on the type of data on which the model can be run. In the literature, models theoretically devised for continuous data are sometimes used on binary data. The absence of assumption on data implies that the models be validated using either a bootstrap or a cross-validation method. We compare here the performances of linear and logistic regression on qualitative data.

Mots-clés : PLS, PLS-GLM, allélotypage

Keywords: PLS, PLS-GLM, allélotyping

Classification AMS 2000 : Primary 60K35, 60K35 ; secondary 60K35

¹ Laboratoire de Biostatistique - Faculté de Médecine, 4, rue Kirschleger 67085 Strasbourg Cedex.

E-mail : nmeyer@unistra.fr

² Université de Strasbourg - Institut de Recherche Mathématique Avancée - Laboratoire de Statistique, 7, rue René Descartes 67084 Strasbourg Cedex.

E-mail : mmaumy@math.u-strasbg.fr et E-mail : fbertran@math.u-strasbg.fr

1. Introduction

Dans cet article, nous allons nous intéresser aux données d'allélotypage dans lesquelles l'expérimentateur collige une information qualitative (présence ou absence d'une anomalie) sur une série d'une trentaine de segments d'ADN pour chaque sujet, en général dans des cellules cancéreuses [6]. Ces segments d'ADN sont nommés microsatellites et sont présents normalement en deux exemplaires dans chaque cellule. Lors de la cancérogenèse, ces microsatellites peuvent subir des altérations nommées « déséquilibre allélique » et notées AI pour *allelic imbalance*. Un AI indique une modification du nombre d'exemplaires du microsatellite et par extrapolation du nombre de segments chromosomiques portant le microsatellite. Le but de ces analyses d'allélotypage est de décrire et de trouver les éventuelles associations d'anomalies qui déclenchent ou entretiennent la cancérogenèse, ce qui revient à trouver la ou les voies d'altérations génétiques impliquées dans la cancérisation. Ce concept de voie d'altération implique qu'il y ait plusieurs anomalies agissant simultanément et/ou successivement pour déclencher un processus cancéreux. Sur le plan statistique, la mise en évidence de ces voies d'altérations nécessite généralement et assez naturellement une étude descriptive multivariée. Par ailleurs, nous souhaitons le plus souvent modéliser les relations entre les microsatellites et une variable réponse, tel que le stade du cancer ou un évènement particulier dans la maladie (présence ou absence d'une métastase, d'une rechute de la maladie, etc). Nous nous intéressons ici au lien entre la variable réponse et d'éventuelles voies d'altérations. Il s'agit alors d'un problème de modélisation qui de surcroît doit tenir compte de la structure multivariée des données.

Toutes les publications biomédicales traitant de données d'allélotypage ignorent la structure réelle des données. En effet, les analyses sont faites uniquement sous un angle univarié (voir par exemple [32]). Les quelques approches multivariées qui ont été tentées sont des analyses en cluster comme dans Weber [28], analyses qui ne permettent pas de répondre à toutes les questions posées par les biologistes. Il est à noter que les méthodes descriptives multivariées n'ont pas pour but de modéliser les données, ce qui limite leur utilisation pour étudier sur le plan biologique les relations possibles entre les voies d'altérations et une caractéristique clinique du patient ou de la tumeur cancéreuse.

Le point de vue que nous allons présenter dans cet article pour l'analyse des données d'allélotypage semble original dans la mesure où nous utilisons une approche globale du problème, à la fois inférentielle et descriptive. Les méthodes statistiques envisagées pour analyser ce type de données sont utilisées après avoir étudié et comparé leur propriétés par des simulations.

L'article est divisé en six parties. Après cette introduction, nous présentons dans la deuxième partie les problèmes statistiques posés par les données d'allélotypage. Dans la troisième partie, nous discutons des différents modèles envisageables pour traiter ces données avec leurs avantages et leurs inconvénients. L'objectif de cette étude découle de la comparaison des méthodes décrites. La quatrième section de l'article présente l'organisation des comparaisons entre les différentes méthodes retenues ainsi que les analyses réalisées. La cinquième partie présente les résultats des comparaisons. La sixième section contient la discussion et les conclusions suggérées par les résultats.

Notations Dans cet article, nous notons en capitale grasse une matrice de données (par ex. \mathbf{X}), en capitale non-grasse une variable aléatoire (par ex. Y), en minuscule grasse un vecteur (par ex.

t_h), en minuscule non-grasse une observation ou une réalisation d'une variable aléatoire.

2. Description des données

Sur le plan statistique, les données d'allélotypage sont composées d'une série de p variables explicatives qualitatives binaires regroupées dans une matrice \mathbf{X} à laquelle nous associons une variable réponse univariée \mathbf{y} également binaire. Les données ont été relevées sur un n -échantillon.

Les données étant binaires (AI ou non), les p variables correspondant à chaque microsatellite sont modélisées par des variables binomiales $B_j(n, \theta_j)$, $j \in \{1, \dots, p\}$, où n est l'effectif de l'échantillon et θ_j la probabilité qu'un sujet présente un AI pour le microsatellite j . La valeur de θ_j doit être estimée. Il faut noter que ces p variables binaires ne sont pas forcément indépendantes. Un des aspects du problème consiste justement à déterminer l'existence d'éventuelles associations entre les différents microsattellites, le nombre de microsattellites impliqués dans l'association étant inconnu.

Dans l'application présente, la variable réponse \mathbf{y} est une variable qualitative à deux modalités indiquant le stade du cancer selon la classification d'Astler-Coller [2]. Elle prend la valeur 0 quand la tumeur est dans un stade A ou B de la classification d'Astler-Coller et elle prend la valeur 1 quand le stade est C ou D.

À cette situation, se rajoutent les deux difficultés suivantes.

2.1. Les dimensions de la matrice \mathbf{X}

L'allélotypage est habituellement réalisé sur un nombre important de microsattellites, important en tout cas relativement par rapport au nombre de sujets. Classiquement le nombre de microsattellites se situe entre 20 et 50 quel que soit le cancer étudié. Ce nombre est considéré comme élevé car le nombre de sujets est souvent limité, allant de 50 à 100, cet effectif pouvant être parfois de l'ordre d'une dizaine de sujets. Nous obtenons dans ce dernier cas une matrice \mathbf{X} où le nombre de variables peut être supérieur au nombre de sujets. Les méthodes statistiques *classiques* ne permettent pas de travailler sur des matrices de ce type en raison de leur rang insuffisant. Des méthodes de sélection de variables ou de projection sur des sous-espaces doivent alors être envisagées. La situation est donc proche de celle rencontrée dans l'analyse des biopuces ou *microarrays* [21, 22, 4].

2.2. La colinéarité

Dans un certain nombre de situations biologiques, il est possible d'observer une colinéarité entre deux ou plus de deux microsattellites. Sur le plan statistique, la colinéarité est également une limite bien connue des méthodes statistiques classiques. En effet, en cas de colinéarité, ou lorsque les variables sont liées comme dans certains plans d'expérience, les analyses multivariées tendent à sélectionner des variables dites indépendantes, aux dépens de la cohérence des modèles [3, 27]. Ainsi, l'une de deux variables explicatives colinéaires peut être éliminée d'un modèle alors qu'elles sont toutes les deux pertinentes pour l'utilisateur.

Ces deux difficultés limitent donc l'usage des modèles tels que la régression logistique ou le modèle de Cox, d'utilisation classique dans le domaine biomédical. Il faudra donc tenter soit de changer de modèle statistique soit d'utiliser des modifications de ces modèles de manière à ce qu'ils soient adaptés au problème énoncé.

2.3. Les interactions

Si l'étude du rôle propre de chaque microsatellite est évidemment intéressante, il faut également envisager la possibilité que ce soit probablement les interactions entre microsatellites qui constituent les voies d'altérations et expliquent la cancérogenèse. Exceptés pour certains oncogènes déjà connus, la plupart des anomalies du génome ne sont pas suffisantes pour expliquer à elles seules la cancérisation d'une cellule. C'est donc dans l'association ou l'interaction entre microsatellites qu'il faut tenter de décrypter les voies de la cancérogenèse. Ceci suppose donc des modèles statistiques capables d'introduire des interactions en nombre suffisant compte tenu des deux difficultés énoncées, d'où notre intérêt pour les modèles PLS. Cependant, les données d'allélotypage sont caractérisées par un nombre important de données manquantes. Dans le présent article, cet aspect des données ne sera pas pris en compte et nous comblerons artificiellement les données manquantes par une méthode d'imputation simple [18]. Il ne s'agit donc pas des données originales mais de données ayant des caractéristiques proches aux données manquantes près. Par ailleurs, les interactions sont probablement plus sensibles au mécanisme d'imputation du fait du nombre important de valeurs manquantes dans les variables d'interaction. De ce fait, dans l'exemple d'application, nous n'étudierons pas les interactions même si dans le principe cela est réalisable.

3. Modèles envisagés

Il reste donc à définir le ou les modèle(s) pouvant être utilisé(s) dans le contexte présenté. Nous trouvons dans la littérature différentes références suggérant d'utiliser sur des variables toutes qualitatives des méthodes *a priori* destinées à des variables quantitatives.

Jolliffe [16] indique qu'une Analyse en Composantes Principales (ACP), *a priori* construite pour traiter des données continues, peut être utilisée sur des données qualitatives. Ceci est possible dans la mesure où l'ACP est une technique descriptive [13]. Il faut noter que l'ACP d'un tableau formé de variables binaires donne les mêmes résultats que l'Analyse Factorielle des Correspondances Multiples (AFCM). L'adaptation de l'ACP à notre problème supposerait d'utiliser les composantes principales obtenues comme variables explicatives dans un modèle de régression de la variable réponse y sur ces composantes. Cependant, les composantes ainsi obtenues ne maximisent pas la covariance entre le vecteur y et la matrice des prédictors X .

Dans notre contexte, une régression de type Partial Least Squares (PLS) semble être indiquée [30, 25]. Les données étant ici toutes binaires, cela revient donc à réaliser une analyse discriminante PLS. Comme l'ACP, ce modèle étant basé sur la construction de composantes, il permet d'obtenir des modèles cohérents même lorsque les variables x_j sont colinéaires, ou que la matrice X est horizontale ($p \gg n$). C'est cette propriété qui permettrait d'introduire un

grand nombre d'interactions dans la matrice \mathbf{X} . De plus, contrairement à la méthode indiquée ci-dessus, la régression PLS (RPLS) maximise la covariance entre la variable réponse \mathbf{y} et des composantes orthogonales $\mathbf{t}_h = \mathbf{X}\mathbf{w}_h^*$ avec $\|\mathbf{w}_h^*\| = 1$. Les aspects descriptifs multivariés, outre les aspects inférentiels de la régression de type PLS répondent donc bien aux objectifs des études d'allélotypage, présentés dans l'introduction.

Tenenhaus [25] mentionne dans l'introduction du chapitre 12 sur le traitement des données qualitatives que la RPLS ne repose sur aucune hypothèse probabiliste et que par conséquent elle peut être utilisée sur des variables qualitatives. Cette assertion n'est cependant valide que pour un usage purement descriptif de la méthode. En effet, la RPLS repose essentiellement sur un algorithme alors que ces propriétés statistiques ne sont pas encore suffisamment développées. En l'absence d'un modèle statistique idoine, la RPLS ne peut être utilisée de manière inférentielle sans recourir à des méthodes adaptées qui seront vues plus loin. C'est pourquoi Tenenhaus utilise dans une analyse uniquement descriptive une RPLS sur des variables toutes qualitatives, *via* les variables indicatrices des modalités de ces variables qualitatives.

Par ailleurs, dans ce même chapitre 12, Tenenhaus fait remarquer que Cazes [5] a proposé une méthode de régression après AFCM dans le cas particulier où toutes les variables sont qualitatives. Cette méthode comporte deux étapes successives. La première étape est une AFCM du tableau ${}^t\mathbf{YX}$, où ${}^t\mathbf{Y}$ désigne la transposée de la matrice \mathbf{Y} , suivi dans la seconde étape d'une régression du tableau des réponses \mathbf{Y} sur les profils lignes du tableau des prédicteurs \mathbf{X} rajoutées en supplémentaire sur ${}^t\mathbf{YX}$. Nous n'allons pas utiliser ici cette méthode du fait que l'AFCM et la régression ne se font pas simultanément, ce qui ne maximise pas forcément les relations entre \mathbf{Y} et \mathbf{X} . De plus, les méthodes comparées ici (qui sont donc des méthodes PLS) n'utilisent pas la matrice ${}^t\mathbf{YX}$ qui est à la base de la méthode de Cazes. Enfin, nous avons choisi dans notre contexte de nous focaliser sur les méthodes de régression PLS.

Une autre solution proposée par Tenenhaus *et al.* [27] consiste à faire l'Analyse Factorielle des Correspondances Multiples (AFCM) des réponses \mathbf{Y} , l'AFCM des prédicteurs \mathbf{X} et de faire la régression des composantes de \mathbf{Y} sur les composantes de \mathbf{X} . Nous pourrions adapter cette solution à notre problème mais elle aurait alors également l'inconvénient de ne pas maximiser la covariance entre les matrices \mathbf{Y} et \mathbf{X} , contrairement à une RPLS.

Tous les modèles présentés jusqu'ici font un usage uniquement descriptif de la RPLS et l'aspect inférentiel de la régression n'est pas exploité. Dans le cas de variables qualitatives, l'obtention des estimations ponctuelles des paramètres de la régression est possible, même si leur interprétation peut être délicate. En revanche, l'obtention de la variance des estimateurs des paramètres est problématique [14]. La difficulté est liée à l'absence de connaissance de la loi des erreurs sous-jacente à l'utilisation d'une régression linéaire (éventuellement dans sa version PLS) sur des données qualitatives. L'estimation de ces variances doit faire appel à des méthodes adaptées.

Xiong et Meullenet [31] utilisent explicitement des variables qualitatives dans une RPLS en mentionnant qu'une RPLS peut être utilisée sur des données qualitatives à condition de les remplacer par les variables indicatrices des modalités. Dans cette étude, la variable réponse est de nature ordinale (échelle de 0 à 9) et comme dans le cas de la régression linéaire, le modèle sous-jacent nécessaire à la prédiction n'est valide que si la loi des erreurs est gaussienne, hypothèse qui reste à vérifier sur le type de données analysées par Xiong et Meullenet.

Dans le même esprit, la documentation de SIMCA®[9], logiciel de référence pour la régression PLS, mentionne la possible utilisation de variables qualitatives, mais aucun commentaire

spécifique n'est fait sur les règles d'utilisation des variables qualitatives dans une RPLS. En contrepartie, comme dans l'étude de Xiong et Meullenet, la validation des modèles construits se fait par validation croisée. La validation croisée et le bootstrap sont des méthodes non-paramétriques qui permettent en effet de valider des modèles notamment lorsque la loi des erreurs est inconnue. Ce sont donc des outils qui apportent une solution au problème évoqué ici à savoir l'utilisation inférentielle d'une RPLS sur des données qualitatives. Si nous nous restreignons à un usage uniquement descriptif de la méthode, il n'y a pas de difficultés théoriques à l'utilisation de la RPLS sur des variables qualitatives, en revanche si nous souhaitons utiliser l'aspect inférentiel de la RPLS pour répondre à une partie de la question posée ici, il faut pouvoir valider les modèles et réaliser des tests statistiques de manière adéquate. Il faudra donc se tourner vers des méthodes non-paramétriques telles que la validation croisée [24], le bootstrap [8] ou les tests de permutation [12].

Si l'utilisation de méthodes non-paramétriques peut fournir une solution aux problèmes d'estimation des intervalles de confiance des paramètres lors de l'utilisation d'une RPLS sur des données uniquement qualitatives, une autre façon d'aborder le problème serait d'utiliser un modèle adapté spécifiquement au type de données que nous avons à étudier. Pour des données qualitatives, nous pouvons nous tourner vers des modèles de type régression logistique combinée aux principes de la RPLS, tels que ceux décrit par Bastien *et al.* [3], par Ding et Gentleman [7] ou encore par Nguyen et Rocke [21]. Nous retenons ici la méthode de Bastien *et al.*, facile à implémenter et à interpréter ainsi que celle de Ding et Gentleman. Dans le cas de la régression logistique PLS (RLogPLS) se pose également la question de la validation des modèles. La RPLS et la RLogPLS transformant les variables binaires initiales en composantes quantitatives, l'aléatoire sous-jacent à ce dernier modèle n'est pas clairement établi. Les méthodes non-paramétriques de bootstrap trouvent donc ici leur place pour permettre de réaliser des tests d'hypothèses. C'est d'ailleurs la méthode retenue notamment par Bastien *et al.* [3].

Objectif : La question qui découle alors de ces constats est de savoir quel modèle peut être utilisé pour traiter des données toutes qualitatives : la RPLS ou l'une des variantes actuelles de la RLogPLS. Dans la mesure où aucune de ces méthodes ne repose de manière explicite sur une théorie statistique sous-jacente et qu'il faut dans tous les cas recourir à des méthodes de type ré-échantillonnage, sans avantage décisif d'un modèle sur les autres, notre objectif est de comparer les performances de la RPLS et de plusieurs variantes de la RLogPLS sur des données qualitatives de type allélotypage.

4. Méthodes étudiées

La présente étude est réalisée dans l'idée de déterminer quelle est la méthode la plus adaptée pour l'analyse des données d'allélotypage. Nous rappelons que dans le cas des données d'allélotypage, l'objectif est à la fois descriptif et inférentiel et que l'analyse doit impérativement tenir compte de l'ensemble des données afin de faciliter la détection de voies d'altérations génétique.

4.1. Les régressions PLS

Il est inutile de donner les détails de la RPLS qui est bien connue et décrite de manière exhaustive notamment par Höskuldsson [15] et Wold *et al.* [30]. La présentation classique de la RPLS, telle

que nous pouvons la trouver dans l'article de Höskuldsson [15], est sous forme algorithmique. Nous n'en rappellerons que les éléments utiles pour notre propos.

4.1.1. La régression PLS

La RPLS est un modèle non-linéaire qui définit des composantes orthogonales qui forment une matrice \mathbf{T} contenant les composantes \mathbf{t}_h obtenues avec la contrainte suivante :

$$\max(\text{cov}(\mathbf{y}, \mathbf{t}_h)). \quad (1)$$

La RPLS peut alors s'écrire matriciellement de la façon suivante :

$$\mathbf{y} = \mathbf{T}^t \mathbf{c} + \varepsilon \quad (2)$$

avec ε le vecteur des résidus, \mathbf{c} le vecteur des coefficients des composantes.

En posant $\mathbf{T} = \mathbf{X}\mathbf{W}^*$, ceci peut encore s'écrire :

$$\mathbf{y} = \mathbf{X}\mathbf{W}^* \mathbf{c} + \varepsilon \quad (3)$$

où \mathbf{W}^* est une matrice de coefficients des variables \mathbf{x}_j dans chaque composante \mathbf{t}_h . D'autre part, cela peut également s'écrire

$$\mathbf{y} = \mathbf{X}\mathbf{B} + \varepsilon \quad (4)$$

en posant $\mathbf{B} = \mathbf{W}^* \mathbf{c}$.

En développant l'écriture, nous obtenons :

$$y_i = \sum_{h=1}^H (c_h w_{1h}^* x_{i1} + \dots + c_h w_{ph}^* x_{ip}) + \varepsilon_i, \quad (5)$$

où H est le nombre de composantes retenues dans le modèle final avec $H \leq \text{rg}(\mathbf{X})$, H étant en général très inférieur au rang de \mathbf{X} et p étant égal au nombre de variables contenues dans la matrice \mathbf{X} .

Pour être complet,

$$y_i = \sum_{j=1}^p b_j x_{ij} + \varepsilon_i$$

$$\text{où } b_j = \sum_{h=1}^H c_h w_{jh}^*.$$

Les coefficients $b_j = \sum_H c_h w_{jh}^*$ où $j \in \{1, \dots, p\}$, suivant la notation avec « * » de Wold *et al.* [30], traduisent la relation entre le vecteur \mathbf{y} et les variables \mathbf{x}_j à travers les composantes \mathbf{t}_h . Ce sont ces paramètres qui seront utilisés dans les analyses (§ 4.2) et les comparaisons pour évaluer les propriétés des différents modèles. Nous pourrions de prime abord penser utiliser les coefficients c_h mais l'utilisation du bootstrap ne le permet pas comme nous allons immédiatement le montrer.

La présentation algorithmique de la PLS permet de comprendre que la valeur de chaque élément c_h de \mathbf{c} est positive par construction.

Démonstration. En effet, nous avons :

$$\mathbf{w}_h = \frac{1}{{}^t\mathbf{y}_h\mathbf{y}_h} {}^t\mathbf{X}_{h-1}\mathbf{y}_{h-1}$$

où $\mathbf{X}_0 = \mathbf{X}$ et $\mathbf{y}_0 = \mathbf{y}$.

Donc, nous avons :

$$c_h = \frac{1}{{}^t\mathbf{t}_h\mathbf{t}_h} {}^t\mathbf{y}_{h-1}\mathbf{t}_h$$

où \mathbf{t}_h est non nul, et

$$\mathbf{t}_h = \frac{1}{{}^t\mathbf{w}_h\mathbf{w}_h} \mathbf{X}_{h-1}\mathbf{w}_h$$

en ré-exprimant c_h , nous notons que :

$$c_h = \frac{1}{{}^t\mathbf{t}_h\mathbf{t}_h {}^t\mathbf{w}_h\mathbf{w}_h {}^t\mathbf{y}_h\mathbf{y}_h} {}^t\mathbf{y}_{h-1}\mathbf{X}_{h-1} {}^t\mathbf{X}_{h-1}\mathbf{y}_{h-1}.$$

Nous reconnaissons le rapport de deux normes élevées au carré, par conséquent c_h est positif. \square

Les différentes répliques d'une méthode bootstrap fourniront donc toujours des coefficients c_h positifs. Par conséquent, la borne inférieure de l'intervalle de confiance du coefficient c_h basé sur les percentiles ne sera jamais inférieure à zéro. En pratique, les cas où cette borne inférieure est nulle ne s'observent jamais. Il n'est donc pas possible d'utiliser une méthode bootstrap pour tester les coefficients c_h puisque l'espérance mathématique de l'estimateur du coefficient c_h sous l'hypothèse nulle est égale à c_h qui est une valeur inconnue.

Par ailleurs, les c_h ne sont pas directement interprétables par l'utilisateur ce qui invite à utiliser les $b_j = \sum_H c_h w_{jh}^*$ pour exprimer le lien entre le vecteur \mathbf{y} et les variables \mathbf{x}_j . L'objectif de la RPLS est d'analyser les liens entre \mathbf{y} et \mathbf{X} en terme de covariance entre \mathbf{y} et \mathbf{t}_h : $\text{cov}(\mathbf{y}, \mathbf{t}_h)$, \mathbf{t}_h étant une combinaison linéaire des \mathbf{x}_j . L'interprétation des résultats se fera donc en utilisant les coefficients $b_j = \sum_H c_h w_{jh}^*$ qui quantifient la relation « physique [entre \mathbf{y} et \mathbf{X}] sous une forme directement compréhensible par l'utilisateur », selon l'expression de Tenenhaus [25] ; Wold *et al.* [30].

4.1.2. La régression logistique PLS

La RLogPLS modélise linéairement la relation entre la fonction $g(\pi)$ et la matrice des prédicteurs \mathbf{X} où π est la loi de probabilité de la variable réponse \mathbf{y} et g est une fonction *Logit*, $\text{Logit}(\pi) = \text{Ln}\left(\frac{\pi}{1-\pi}\right)$.

En RLogPLS, nous construisons à chaque étape la régression logistique de \mathbf{y} sur les composantes $\mathbf{t}_1, \dots, \mathbf{t}_h$. L'équation de régression logistique PLS est obtenue en exprimant cette équation en fonction des variables d'origine \mathbf{x}_j . Ainsi pour une réponse \mathbf{y} , en notant π la probabilité de l'évènement $Y = 1$, nous obtenons :

$$\begin{aligned}
 \widehat{\text{Logit}}(\boldsymbol{\pi}) &= c_1 \mathbf{t}_1 + \cdots + c_h \mathbf{t}_h \\
 &= c_1 \mathbf{X} \mathbf{w}_1^* + \cdots + c_h \mathbf{X} \mathbf{w}_h^* \\
 &= \mathbf{X} \mathbf{B}
 \end{aligned}$$

avec $h \in \{1, \dots, H\}$, H étant le nombre de composantes retenues dans le modèle final et $H \leq \text{rg}(\mathbf{X})$.

Dans la RLogPLS, les composantes \mathbf{t}_h sont construites de façon itérative à partir de régressions logistiques individuelles $\text{Logit}(\mathbb{P}(Y = 1 | \mathbf{x}_j)) = \beta_0 + \beta_j \mathbf{x}_j$.

Le paramètre d'intérêt exprimant la relation entre $\text{Logit}(\boldsymbol{\pi})$ et \mathbf{X} est alors $\boldsymbol{\beta}$.

Sur le plan pratique, les composantes tant dans la RPLS que dans la RLogPLS sont obtenues de manière itérative en utilisant l'algorithme NIPALS [30], [29].

4.1.3. La régression logistique sur composantes PLS

RLCPLS

La régression logistique sur composantes PLS est décrite par Tenenhaus dans le chapitre 12 de [26]. Elle se déroule en deux étapes. Tout d'abord, nous commençons par réaliser la régression PLS2 [25] de la matrice formée des indicatrices des modalités de la réponse \mathbf{y} sur les variables \mathbf{x}_j et choisir le nombre H adéquat de composantes. Puis nous ajustons un modèle de régression logistique à ces H composantes PLS.

4.1.4. Ding et Gentleman

gPLS

En 1996, Marx propose, dans [19], une extension de la régression PLS pour une variable réponse catégorielle. Sa méthode intègre les étapes classiques de la PLS dans l'algorithme itératif des moindres carrés repondérés donnant naissance à ce qu'il appelle l'algorithme itératif des moindres carrés partiels repondérés (IRWPLS). Malheureusement, il été observé que cet algorithme ne converge pas dans certains cas et en particulier ceux relevant de problèmes de classification. En 2005, Ding et Gentleman adaptent, dans [7], la méthode de Marx aux problèmes de classification en formulant ceux-ci dans le contexte des modèles linéaires généralisés et en utilisant la procédure de Firth .

4.1.5. Fort et Lambert-Lacroix

Ridge PLS

Dans leur article [10], Fort et Lambert-Lacroix proposent une nouvelle méthode de classification, la Ridge PLS, qui intègre à la fois la régression PLS et la régression logistique pénalisée avec contraintes. D'autre part, les auteurs présentent d'autres méthodes existantes basées sur la PLS ou des techniques de vraisemblance pénalisée.

Pour étudier et comparer les propriétés des cinq méthodes présentées sur des données toutes qualitatives, nous les avons confrontées sur des simulations puis nous les avons appliquées à des données d'allélotypage.

4.2. Simulations, critères de comparaisons et risque α

4.2.1. Simulations

Les données ont été construites suivant la méthode utilisée par Li *et al.* [17] dérivée de Næs et Martens [20]. Cette méthode permet de choisir le nombre de composantes PLS voulues dans le modèle final. Les variables générées sont toutes (\mathbf{y} et \mathbf{X}) issues de loi normales centrées réduites. Un premier modèle de RPLS est ajusté sur ces données continues de manière à vérifier que les critères de comparaison utilisés retrouvent bien le nombre de composantes PLS voulues. Puis, les données sont dichotomisées de part et d'autre de la valeur 0, de manière à générer des données toutes binaires sur lesquelles les différents modèles sont comparés. Le nombre de composantes PLS a été fixé à six de sorte que le nombre de composantes PLS retenues sur les modèles ajustés sur les données binaires retrouvent environ trois composantes selon les critères. Dans les simulations, nous avons utilisé des matrices de tailles différentes : la première de taille $(n = 100) \times (p = 20)$ et la seconde de taille $(n = 1000) \times (p = 20)$.

4.2.2. Critères de comparaisons

La sélection des composantes et des variables dans la construction des modèles PLS est encore sujet à débat. Malgré les résultats de Gauchi et Chagnon [11], il ne semble pas qu'une méthode domine particulièrement les autres, du moins pas de manière systématique. Pour cette raison, nous avons choisi de comparer les cinq régressions sur le nombre de composantes retenues par le critère d'information d'Akaike (AIC), [1], et par le Q2 obtenu par validation croisée [9], ainsi que sur le nombre de mal classés. Le détail des critères est donné dans les deux listes ci-dessous. Dans ces simulations, la gPLS n'a pas été retenue car le package `gp1s` du logiciel R ne propose pas de détermination du nombre optimal de composantes. Par ailleurs, ce package propose une détermination du nombre de mal classés par Leave-One-Out (LOO). Le LOO est une méthode de validation croisée qui est peu efficace, la validation croisée étant optimale lorsqu'elle est réalisée sur un faible nombre de sous-groupes, de l'ordre de 5 à 10 [23]. Par ailleurs, ce nombre de mal classés ne correspond pas au critère que nous utilisons qui est le nombre de mal classés pour le nombre de composantes retenues par l'un des deux critères. Dans nos résultats, le nombre minimal de mal classés n'est pas en soi estimé par validation croisée. En ce qui concerne la Ridge PLS, le package `pls.genomics` implémentant la fonction ne permet pas de calculer l'AIC sur ce modèle. Seule la validation croisée est disponible sur un critère qui n'est pas précisé dans la documentation.

4.2.3. Contrôle du risque α

Sur des données aléatoires et pour les cinq méthodes, la significativité des coefficients de chaque régression a été déterminée par l'utilisation d'un test de permutation [12]. En calculant la proportion de coefficients significatifs sur une grande série de simulations, nous pouvons déterminer le risque de première espèce α associé à l'utilisation de chacune des méthodes et vérifier l'adéquation entre le taux d'erreur de première espèce et un seuil nominal à atteindre de 5%.

Une variable aléatoire notée F est associée à chacun des coefficients b_j de chaque variable x_j sur la première composante du modèle, et ceci pour chacune des méthodes comparées. Sur l'ensemble des Q permutations du jeu de données (conservant la structure par ligne), nous avons tabulé $p = \mathbb{P}(F_{\text{coef}} \geq f_{\text{coef}})$ qui donne une estimation empirique de la p -valeur associée au paramètre d'intérêt. Les estimations ponctuelles des paramètres des modèles ont été calculées parallèlement sur les mêmes itérations afin d'annuler l'effet de l'erreur d'échantillonnage au cours des permutations. La valeur du risque de première espèce α est estimée sur 1000 tests de 999 permutations chacun. Les résultats donnés dans la Table 4 sont la proportion de tests significatifs au seuil 5% sur des données binaires aléatoires. De ce fait, la régression sur composantes PLS n'a pas été incluse puisque, dans ce modèle, les coefficients seraient ceux des composantes dans une régression linéaire multiple et non pas les coefficients des variables x_j dans une régression PLS. Le temps de calcul est de l'ordre de 20 minutes pour un test de permutation sur l'ensemble des méthodes.

Tous les calculs de cette étude ont été réalisés avec le logiciel R 2.9.1.

4.3. Données réelles : Les données d'allélotypage

Les données utilisées sont des données d'allélotypage recueillies sur 104 sujets porteurs d'un cancer du colon et pour lesquels la recherche d'AI a été faite sur 33 microsatellites. La variable réponse est le stade du cancer suivant la classification de Astler-Coller [2]. Cette classification comporte quatre stades A, B, C et D qui ont été ici regroupé en deux classes (A et B ; C et D) pour des raisons de simplicité, d'autant que dans la présente étude, la pertinence clinique de la classification subsiste malgré ce regroupement. Rappelons que ces données sont originellement incomplètes et que nous avons imputé les valeurs manquantes. L'imputation a été faite par une imputation simple en utilisant la méthode "sample" du package MICE du logiciel R. Il ne s'agit donc pas des données originales mais la structure en est très proche.

5. Les résultats

Les Tables 1 et 2 présentent le nombre de composantes retenues pour chacun des modèles comparés selon le critère retenu au cours des simulations. La liste ci-dessous énumère les critères qui ont été utilisés pour chacune des quatre méthodes comparées qui sont la RPLS, la RLogPLS, la RLCPLS et la Ridge PLS. Cette liste décrit aussi les onze lignes des Tables 1 et 2. Ces tables contiennent de plus dans les première et quatrième lignes les nombres de composantes obtenues par l'AIC et le Q2 sur les données simulées continues, avant dichotomisation, ce qui permet de vérifier la validité de ces deux critères. Ces indices permettent également de vérifier que le nombre attendu de six composantes est atteint. Nous rappelons que la gPLS n'est pas incluse dans cette liste car le package gPLS ne fournit pas de critère de validation.

1. l'AIC en RPLS sur données continues
2. l'AIC en RPLS sur données dichotomisées
3. l'AIC en RLogPLS sur données dichotomisées
4. le Q2 en RPLS sur données continues

TABLE 1
Nombre de composantes retenues selon les critères et les méthodes pour $n = 100$.

Critère	Nombre de composantes										
	0	1	2	3	4	5	6	7	8	9	10
AIC											
1	0	0	0	0	0	0	0	202	782	15	0
2	0	0	0	12	160	250	382	73	15	50	57
3	0	15	203	352	259	120	39	11	0	0	0
Q2											
4	0	0	0	0	0	3	989	7	0	0	0
5	0	17	742	160	56	10	6	2	1	1	4
6	0	874	125	0	0	0	0	0	0	0	0
7	0	72	404	283	121	61	26	21	11	0	0
Nombre de mal classés											
8	0	26	352	335	114	44	34	26	45	15	8
9	0	35	236	272	236	122	70	28	0	0	0
10	0	6	99	71	256	167	150	102	68	49	31
11	0	12	137	229	254	172	121	48	14	7	5

5. le Q2 en RPLS sur données dichotomisées
6. le Q2 en RLogPLS sur données dichotomisées
7. la validation croisée en Ridge PLS sur données dichotomisées
8. le nombre de mal classés en RPLS
9. le nombre de mal classés en RLogPLS
10. le nombre de mal classés en RLCPLS
11. le nombre de mal classés en Ridge PLS.

Il faut noter que les deuxième et cinquième lignes correspondent également aux résultats obtenus avec la RLCPLS. Par ailleurs, la validation croisée de la Ridge PLS se fait sur un critère qui n'est pas précisé dans la documentation du package `pls.genomics` ce qui explique la dénomination de « validation croisée » utilisée et non pas de « Q2 ».

Sur les données continues, permettant de vérifier la pertinence des deux critères de comparaison utilisés, l'AIC tend à sélectionner un nombre trop grand de composantes, à savoir huit au lieu de six. Le Q2 par validation croisée est nettement meilleur puisqu'il sélectionne le bon nombre de composantes dans 98,9% des cas (voir ligne 1 et 4 des Tables 1 et 2).

La RPLS appliquée à des données binaires obtenues par dichotomisation des données continues retrouve de quatre à six composantes (79,2% des cas) par l'AIC et deux ou trois selon le Q2, pour 90,2% des simulations (lignes 2 et 5). Si nous considérons que le Q2 est le meilleur des deux critères d'après les résultats pour les données continues, le nombre de composantes retenues par le modèle RPLS est donc diminué de trois ou de quatre composantes. L'AIC suggère une diminution de deux ou trois composantes. La RLogPLS retrouve de deux à cinq composantes dans 93,4% des simulations selon l'AIC (ligne 3) et soit une soit deux composantes dans la totalité des cas selon le Q2 (ligne 6). La Ridge PLS sélectionne entre deux et quatre composantes dans 808 des 1000 simulations. Si nous nous basons sur le nombre minimal de mal classés, la RPLS présente le

TABLE 2
Nombre de composantes retenues selon les critères et les méthodes pour $n = 1000$.

Critère	Nombre de composantes										
	0	1	2	3	4	5	6	7	8	9	10
AIC											
1	0	0	0	0	0	0	0	999	1		0
2	0	0	0	0	0	0	250	147	1	1	601
3	0	0	0	33	279	585	82	21	0	0	0
Q2											
4	0	0	0	0	0	0	1000	0	0	0	0
5	0	0	306	665	29	0	0	0	0	0	0
6	0	844	79	77	0	0	0	0	0	0	0
7	0	0	8	194	263	283	137	92	23	0	0
Nombre de mal classés											
8	0	0	112	252	447	23	18	0	0	11	137
9	0	0	91	158	145	188	216	201	1	0	0
10	0	0	1	7	160	138	97	73	71	270	183
11	0	0	8	88	151	145	108	102	115	144	139

nombre le plus faible de composantes retenues, de deux à quatre, puis la RLogPLS. Ce critère sélectionne de quatre à sept composantes pour la RLCPLS et de deux à six pour la Ridge PLS.

La Table 2, similaire à la Table 1 mais pour $n = 1000$, montre que le critère AIC sélectionne sept composantes dans 99,9% des cas et que le critère Q2 sélectionne six composantes dans 100% des simulations. Sur les données dichotomisées et en utilisant le critère AIC, la RPLS sélectionne 10 composantes dans 60,1% des cas, et six composantes dans 25% des cas. Le Q2 trouve deux et trois composantes dans 30,6% et 66,4% des simulations. Le nombre de composantes sélectionnées par AIC en RLogPLS est de quatre ou cinq dans 86,4% des simulations. Le nombre de composantes retenues par le Q2 est de un dans 84,4% des simulations.

La Table 3 ci-dessous donne les taux moyens de mal classés pour les quatre méthodes (RPLS, RLogPLS, RLCPLS, Ridge PLS) pour le nombre de composantes retenues par le critère AIC, par le critère Q2 puis par la validation croisée spécifique à la Ridge PLS. Cette table indique aussi les taux moyens de mal classés obtenus lorsque nous sélectionnons le nombre de composantes à partir du nombre minimum de mal classés. Les résultats sont donnés pour les deux séries de simulations ($n = 100$ et $n = 1000$).

Les proportions de mal classés sont très faibles, presque toujours inférieures à cinq pourcents. Les différences entre les méthodes sont peu importantes lorsque nous tenons compte des écart-types des taux de mal classés, les différentes méthodes ayant des performances finalement assez proches. Certains points peuvent cependant être soulignés. Les nombres de composantes sélectionnées par le Q2 étant plus petits que par l'AIC, les proportions de mal classés sont un peu plus importantes pour les trois modèles comparés. Sur $n = 100$, c'est la Ridge PLS qui donne la proportion la plus faible de mal classés (en moyenne 3,3%), puis la RLCPLS (4,3), la RPLS (4,6) et enfin la RLogPLS (5,8). Pour $n = 1000$, c'est la RLCPLS qui donne la proportion la plus basse, devant la Ridge PLS, la RPLS puis la RLogPLS.

Pour le critère du nombre minimum de mal classés, la RLogPLS et la RLCPLS ont des proportions minimum de mal classés identiques et les plus basses parmi les quatre méthodes. La RPLS a le nombre minimum de mal classés le plus élevé, la Ridge PLS ayant des résultats presque

TABLE 3
Taux moyens de mal classés selon les critères et les méthodes, pour $n = 100$ et $n = 1000$, exprimé en pourcents (écart-type).

Critère	$n = 100$	$n = 1000$
AIC en RPLS	4,2 (1,85)	5,1 (0,67)
AIC en RLogPLS	2,1 (2,00)	4,4 (0,66)
AIC en RLCPLS	2,5 (2,12)	4,3 (0,66)
Q2 en RPLS	4,6 (2,2)	5,2 (0,73)
Q2 en RLogPLS	5,8 (2,9)	7,2 (1,42)
Q2 en RLCPLS	4,3 (2,4)	5,4 (0,74)
Validation croisée en Ridge PLS	3,3 (2,1)	6,3 (1,45)
Nombre minimum de mal classés en RPLS	3,7 (1,7)	4,9 (0,60)
Nombre minimum de mal classés en RLogPLS	1,8 (1,6)	4,2 (0,64)
Nombre minimum de mal classés en RLCPLS	1,8 (1,7)	4,2 (0,62)
Nombre minimum de mal classés en Ridge PLS	1,9 (1,6)	4,2 (0,64)

TABLE 4
Adéquation entre le taux d'erreur de première espèce et un seuil nominal de 5%. La table indique le nombre de simulations parmi 1000 pour lesquels le coefficient b_j de la variable x_j est significatif au seuil $\alpha = 5\%$, par un test de permutation.

coef	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}
RPLS	85	66	85	87	88	63	84	85	72	91
RLogPLS	87	70	86	95	88	68	88	90	75	102
gPLS	78	69	84	85	79	60	81	86	66	98
Ridge PLS	105	91	102	107	106	91	100	113	96	112

coef	x_{11}	x_{12}	x_{13}	x_{14}	x_{15}	x_{16}	x_{17}	x_{18}	x_{19}	x_{20}
RPLS	84	83	67	75	82	74	100	83	62	80
RLogPLS	86	87	69	74	86	78	101	93	70	88
gPLS	83	83	64	67	77	67	96	86	68	81
Ridge PLS	93	104	91	92	101	96	116	105	81	98

aussi bons que la RLogPLS et que la RLCPLS, en raison du nombre important de composantes que cette méthode retient.

La Table 4 ci-dessous donne l'adéquation entre le taux d'erreur de première espèce et un seuil nominal à atteindre de 5% pour quatre méthodes : RPLS, RLogPLS, gPLS et Ridge PLS. La RLCPLS n'est pas incluse (voir la section *Méthodes étudiées*).

La vérification du seuil nominal des différentes méthodes montre que la RPLS et la RLogPLS ont des seuils empiriques proches mais supérieurs au niveau de référence de 5%. La gPLS respecte un peu mieux les seuils nominaux. Son risque α empirique se situe aux alentours de 7,8%. La Ridge PLS présente un taux de rejet de l'hypothèse nulle élevé, de l'ordre de 10%.

Application aux données d'allélotypage La RPLS appliquée aux données d'allélotypage sélectionne quatre composantes avec le critère AIC. Le critère Q2 ne retrouve aucune composante. La RLogPLS donne les mêmes résultats. Le nombre de mal classés avec quatre composantes est de 23 avec la RPLS et de 21 avec la RLogPLS. Les coefficients des microsatellites sur les premières et deuxièmes composantes sont numériquement très proches entre la RPLS et la RLogPLS.

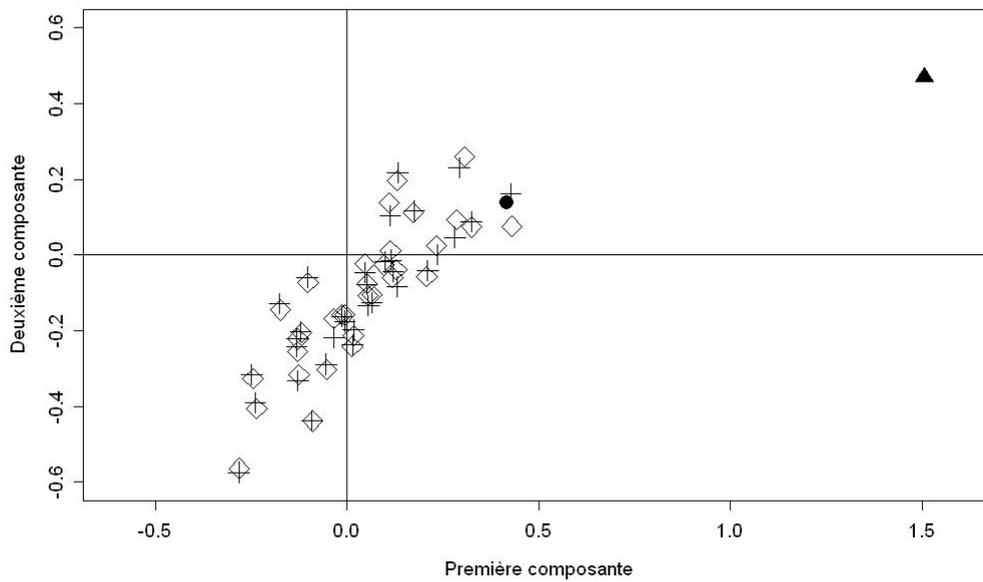


FIGURE 1. Carte des coefficients w^* et c des données d'allélotypage en RPLS et RLogPLS : deux premières composantes principales. Les coefficients w^* des microsatellites et c du stade obtenus par RPLS sont représentés par des + et un ●. Les coefficients obtenus par RLogPLS sont représentés par des ◇ et un ▲.

6. Discussion et conclusion

La RPLS présente des performances assez bonnes sur des données qualitatives binaires. Le nombre de composantes retenues, que ce soit par l'AIC ou par le Q2 est plus proche du nombre de composantes retenues pour les données continues servant de référence que le nombre de composantes retenues par la RLogPLS. En effet, en se basant sur l'AIC, la RPLS sélectionne deux composantes de plus que la RLogPLS. Par validation croisée sur Q2, elle retient une composante supplémentaire. Ce nombre plus grand de composantes retenues aboutit à un taux de mal classés plus faible pour la RPLS que pour la RLogPLS avec le critère Q2 mais plus grand avec le critère AIC. Ce critère du minimum de mal classés est cependant délicat à utiliser pour départager les méthodes en raison de son optimisme avéré.

Sur le plan calculatoire, il faut souligner un élément qui n'apparaît pas dans les résultats. Lors du déroulement de l'algorithme sur des données simulées, la RLogPLS aboutit relativement souvent à des matrices singulières dès que le nombre de composantes à calculer dépasse le nombre de composantes réellement présentes dans les données. Ainsi, lorsque l'utilisateur cherche à extraire $n + 2$ ou $n + 3$ composantes dans un modèle à n composantes, l'algorithme tend à s'interrompre. Il est donc difficile d'ajuster des modèles ayant un grand nombre de composantes. Cette situation a été observée dans quelques pourcents des simulations, obligeant à reprendre les simulations. Suite à l'observation de ce phénomène, nos routines incluent désormais un critère sur les matrices permettant de mener à bien les calculs. Ce critère de limitation s'avère nécessaire dans environ 10% des simulations. Ce cas est cependant peu susceptible de se présenter lors de l'analyse de données réelles.

Le graphique représentant les deux régressions RPLS et RLogPLS ajustées sur les données d'allélotypage montre des résultats similaires. En effet, les positions relatives des points sont très proches ce qui n'est pas surprenant puisqu'empiriquement, les résultats en terme de ratio paramètre/variance du paramètre sont très proches entre une régression linéaire et une régression logistique ajustées sur des données dichotomiques (en ne tenant pas compte des critères de validité dans le cas de la régression linéaire). Ceci montre que contrairement à ce que nous aurions pu penser initialement, une RPLS peut s'avérer très utile sur des données toutes binaires, en donnant des conclusions proches de celle d'une RLogPLS en terme de nombre de composantes retenues, de taux de mal classés pour des estimations de coefficients de régressions très proches. De plus, comme nous l'avons déjà souligné, sur un plan numérique, les calculs se déroulent plus facilement en RPLS qu'en RLogPLS. Évidemment, sur le plan de la quantification de la relation entre y et X , la RLog-PLS serait plus logique puisque la relation s'exprimerait en terme d'odds-ratio.

Dans le cas des données d'allélotypage, ce sont les mêmes microsattellites qui semblent liés au stade (y) pour les différentes méthodes. Les conclusions ne sont pas identiques mais très concordantes. Si nous nous en tenons à la recherche des microsattellites liés au stade, alors les différentes méthodes se valent.

Les deux méthodes principalement comparées ici, à savoir la RPLS et la RLogPLS, donnent des résultats très proches lorsqu'elles sont appliquées sur des données toutes binaires. L'utilisation de la RLogPLS sur des données toutes binaires semble plus logique en terme d'interprétation. Cependant, si l'intérêt se focalise sur la recherche des variables associées à la variable à prédire plutôt que sur la quantification des liens entre y et les x_j , les différentes méthodes semblent avoir des efficacités proches.

Pour l'utilisateur ayant à traiter des données binaires, il est cohérent d'utiliser une RLogPLS mais en pratique, les résultats d'une RPLS appliquée à ces données donnent des conclusions proches de celles de la RLogPLS. À l'heure actuelle, la RLogPLS n'est pas encore disponible en routine dans les logiciels statistiques courants. En attendant que cela soit le cas, l'utilisateur ne désirant pas programmer lui-même une RLogPLS pourra utiliser les programmes existants tel que SIMCA sur ces données binaires avec une certaine sécurité dans l'interprétation des résultats.

Différents packages de R proposent actuellement de la régression PLS. Leur niveau d'implémentation est assez variable, tant en ce qui concerne les paramètres d'entrée que les résultats fournis en sortie des fonctions. Par ailleurs, la PLS-GLM n'existait pas encore en R. Afin de répondre à l'ensemble des questions posées dans le présent travail, nous avons alors été amenés à rédiger un package R permettant de réaliser des RPLS et des PLS linéaires généralisées. Les fonctions proposées dans ce package permettent notamment de déterminer le nombre de composantes optimales d'un modèle PLS à partir de différents indices, tels que le Q2 par validation croisée, l'AIC, le BIC, par minimisation du taux de mal classés ou encore le nombre de variables significatives sur chaque composante. Par ailleurs, ces fonctions utilisent le principe d'un algorithme NIPALS et permettent donc de traiter des jeux de données incomplets et de réaliser la validation croisée sur ces données incomplètes. Le package, nommé `plsRglm`, est disponible sur le site suivant : <http://udsmcd.u-strasbg.fr/labioestat/>.

Remerciements

N. Meyer tient à remercier le Pr. P. Oudet, chef du Laboratoire de Biochimie et Biologie Moléculaire de l'Hôpital de Hautepierre, du CHU de Strasbourg, ainsi que toute son équipe pour l'avoir autorisé à utiliser les données d'allélotypage à l'origine de ce travail.

Par ailleurs, les auteurs remercient les deux relecteurs ayant évalué ce travail. Leurs remarques constructives nous ont permis d'améliorer grandement la qualité de l'article.

Documentation et sources

- [1] H. AKAIKE : A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974.
- [2] VB ASTLER et FA COLLIER : The prognostic significance of direct extension of carcinoma of the colon and rectum. *Ann Surg.*, 1954(6):846–851, 1954.
- [3] P BASTIEN, VE VINZI et M TENENHAUS : PLS generalised linear regression. *Computational Statistics & Data Analysis*, 48:17–46, 2005.
- [4] AL BOULESTEIX et K STRIMMER : Partial least squares : a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44, 2007.
- [5] P CAZES : Adaptation de la régression PLS au cas de la régression après analyse des correspondances multiples. *Revue de Statistique Appliquée*, 45(2):89–99, 1997.
- [6] GK CHAMBERS et ES MACAVOY : Microsatellites : consensus and controversy. *Comparative Biochemistry and Physiology Part B*, 126:455–476, 2000.
- [7] B DING et R GENTLEMAN : Classification using generalized partial least squares. *Journal of Computational & Graphical Statistics*, 14(2):280–298, 2005.
- [8] B EFRON : Bootstrap methods : another look at the jack-knife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [9] L ERIKSSON, E JOHANSSON, N KETTANEH-WOLD, J TRYGG, C WIKSTRÖM et S WOLD : *Multi- and Megavariate Data Analysis, Principles and Applications*. Umetrics Academy Umeå, 2001.

- [10] G FORT et S LAMBERT-LACROIX : Classification using partial least squares with penalized logistic regression. *Bioinformatics*, 21(8):1104–1111, 2005.
- [11] JP GAUCHI et P CHAGNON : Comparison of selection methods of explanatory variables in PLS regression with application to manufacturing process data. *Chemometrics and Intelligent Laboratory Systems*, 58:171–193, 2001.
- [12] P GOOD : *Permutation Tests : A Practical Guide to Resampling Methods for Testing Hypotheses*. Series in Statistics. Springer, 2000.
- [13] JC GOWER : Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, 53:325–338, 1966.
- [14] IS HELLAND : Some theoretical aspects of partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 58:97–107, 2001.
- [15] A HÖSKULDSSON : PLS regression methods. *Journal of Chemometrics*, 2:211–228, 1988.
- [16] IT JOLLIFFE : *Principal component analysis*. Springer, 2 édition, 2002.
- [17] B LI, J MORRIS et EB MARTIN : Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64:79–89, 2002.
- [18] RJA LITTLE et DB RUBIN : *Statistical Analysis with Missing Data*. John Wiley & Sons, 1987.
- [19] BD MARX : Iteratively reweighted partial least squares estimation for generalized linear regression. *Technometrics*, 38:374–381, 1996.
- [20] T NÆS et H MARTENS : Comparison of prediction methods for collinear data. *Commun. Stat., Simul.*, 14:545–576, 1985.
- [21] DV NGUYEN et DM ROCKE : Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18(1):39–50, 2008.
- [22] GP PAGE, SO ZAKHARKIN, K KIM, T MEHTA, L CHEN et K ZHANG : Microarray analysis. *Methods Mol Biol*, 404:409–30, 2007.
- [23] J SHAO : Linear model selection by cross-validation. *J. Am. Stat. Assoc*, 88:486–494, 1993.
- [24] M STONE : Cross-validated choice and assessment of statistical predictions. *Journal of the Royal Statistical Society, Series B.*, 36:111–133, 1974.
- [25] M TENENHAUS : *La régression PLS. Théorie et Pratique*. Technip, Paris, 1998.
- [26] M TENENHAUS : *Modèles statistiques pour données qualitatives*, chapitre La régression logistique PLS, pages 263–275. Technip, Paris, 2005.
- [27] M TENENHAUS, JP GAUCHI et C MÉNARDO : Régression PLS et applications. *Revue de Statistique Appliquée*, 43(1):7–63, 1995.
- [28] JC WEBER, N MEYER, E PENCREACH, A SCHNEIDER, E GUÉRIN, A NEUVILLE, C STEMMER, C BRIGAND, P BACHELLIER, S ROHR, M KEDINGER, C MEYER, D GUENOT, P OUDET, D JAECK et MP GAUB : Allelotyping analyses of synchronous primary and metastasis CIN colon cancers identified different subtypes. *Int J Cancer*, 120(3):524–32, 2007.
- [29] H WOLD : *Multivariate Analysis*, chapitre Estimation of principal component and related models by iterative least squares, pages 391–420. Academic Press, New York, 1966.
- [30] S WOLD, M SJÖSTRÖM et L ERIKSSON : PLS-regression : a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, 58:109–130, 2001.
- [31] P XIONG et JF MEULLENET : A PLS dummy variable approach to assess the impact of JAR attributes on liking. *Food Quality and Preference*, 17:188–198, 2006.
- [32] JJ ZHU, T SANTARIUS, X WU, J TSONG, A GUHA, JK WU, TJ HUDSON et P MCLBLACK : Screening for loss of heterozygosity and microsatellite instability in oligodendrogliomas. *Genes, Chromosomes & Cancer*, 21:207–216, 1998.